ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ

ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS

ΣΧΟΛΗ
ΔΙΟΙΚΗΣΗΣ
ΕΠΙΧΕΙΡΗΣΕΩΝ
SCHOOL OF
BUSINESS

MSc IN
BUSINESS ANALYTICS

Department of Management Science & Technology
MSc in Business Analytics

«A Business Application of a Regression Model for Apartment Rental Ads Pricing»

By
Antonis Kouklinos

Student ID Number: BAPT1713
Name of Supervisor: Dimitrios Karlis

April 2020
Athens, Greece

Table of Contents

SCOPE

The scope of this study is to examine the apartment rentals section of the real estate market in Greece focusing on developing a model that will explain and predict the fair price of a rental.

Combining the knowledge provided in the Masters Courses, there would be a holistic approach to the subject and not a strictly scientific statistical method of developing a forecasting model.

In Section (I) the Literature Review will present the recent studies that were used as a reference during the process of variable selection and transformation in order to build the optimal model.

Sections (II) and (III) utilize the theory and practices of Data Management and Big Data courses in order to create the dataset that will be used for analysing the apartment rental market and for developing the model.

Section (II) explains how the data are extracted from the websites that ads are uploaded and Section (III) describes the process through which the raw data that were extracted from the internet are transformed and loaded with the desired form in a data warehouse.

Sections (IV) and (V) use the statistical methods taught in the relevant courses in order to examine the dataset, build the forecasting model and test its credibility.

Sections (VI) and (VII) use Business Intelligence techniques and tools in order to present the insights of the data through visualizations and a user interface for the forecasting model.

Finally, Section (VIII) presents how this study can fit in a business, explaining the real time processes that will embed this work in a website that will act as an e-real estate agent.

I. LITERATURE REVIEW

The main subject of this study is to develop a model with a high explanatory power that will be used in order to calculate the fair price of an apartment that is listed in an ads website.

There is no regulation in Greece for keeping the prices within certain limits according to the apartment's characteristics. The owner is free to charge the apartments at whatever price wishes. Moreover, there is no calculation for an objective or fair price of the rental that could also be used as imputed income in taxation.

There is, though, a clear connection that is implied with the objective pricing for houses which is regulated through a government law (Article 41, Law. 1249/1982).

This law states that the determinants for the price of a house are:

- The location zone,
- The building blocks within the area
- The typology of the building

More specifically, regarding the apartments, the price determinants are:

- The age
- The location within the block
- The floor

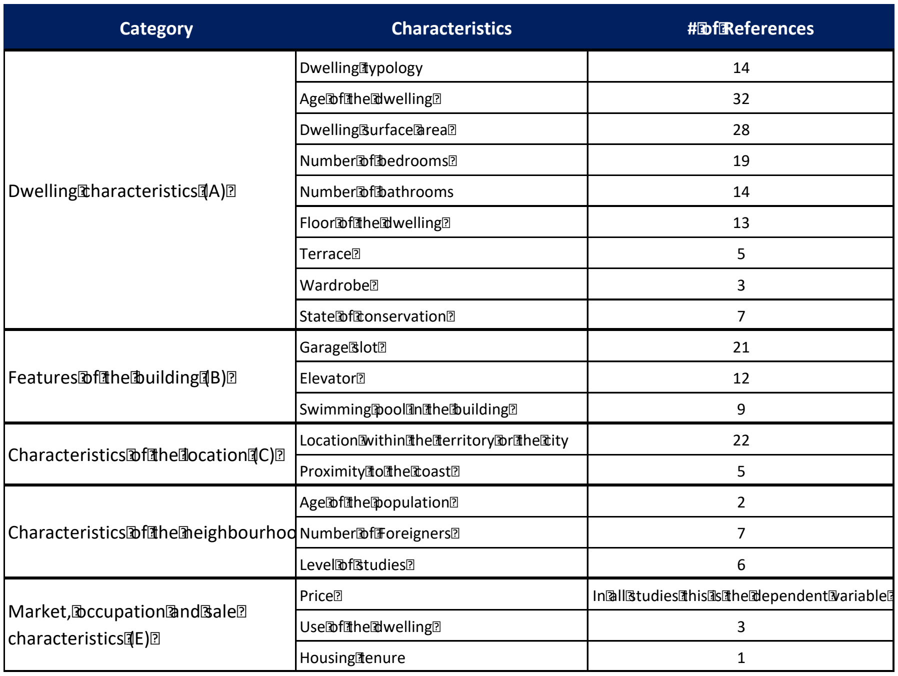
This seems like a poor valuation model, especially when it comes to rental, where the apartments may have renovated or equipped with several features that would add value.

A few studies that have been published for the subject from Greek researchers indicate that except for these characteristics there may also be some other but still concentrating in the location category, such as the view or the surrounding area (Papakyriakou, 2017)

Extending to the real estate market of other countries, there is a number studies showing the many characteristics that can be used as determinants of the price of housing, and how these can be grouped into categories. An interesting research by Smith (1988) groups these characteristics into four categories:

durability, heterogeneity, spatial fixation and government involvement. A more elaborate study by Sirmans et al. (2005) which collected data from different past studies group them into seven categories: internal, external, structural characteristics, public services, environmental, market & occupation & sales factors, funding.

A further step on this categorization is made by a recent study by Raul-Tomas Mora-Garcia et al. (2019) on the determinants of the price of housing in Spain, in which the authors gathered several articles (57) from the past decade with models that try to capture the influence of different determinants in the price of dwelling. They summarised the characteristics used that have proven to be statistically significant (at >95% level) in the corresponding models in Table 1.



Category	Characteristics	# of References
Dwelling characteristics (A)	Dwelling typology	14
	Age of the dwelling	32
	Dwelling surface area	28
	Number of bedrooms	19
	Number of bathrooms	14
	Floor of the dwelling	13
	Terrace	5
	Wardrobe	3
	State of conservation	7
Features of the building (B)	Garage slot	21
	Elevator	12
	Swimming pool in the building	9
Characteristics of the location (C)	Location within the territory or the city	22
	Proximity to the coast	5
Characteristics of the neighbourhoo	Age of the population	2
	Number of Foreigners	7
	Level of studies	6
Market, occupation and sale characteristics (E)	Price	In all studies this is the dependent variable
	Use of the dwelling	3
	Housing tenure	1

Table 1. References by Characteristic

Other researchers emphasize on the quality of life that the surface area of the house combined with the proximity to the city center offers (Bohl, 2012).

A closer approach to the Greek market can be met in studies that were conducted for Istanbul, the capital of the neighbour country. All studies agree that the location plays a crucial role for the price determination but it was also introduced the central heating as a significant factor. (Kayar & Atan, 2017; Yayar & Demir, 2014; Keskin & Watkins, 2017)

The location seems to be the number one determinant according to most of the researchers. McGreal &Taltavull (2013) have used a very large dataset of over 2.3 million houses (of the Spain market) to come up with another interesting conclusion, that there are market characteristics that are particular to different regions. The value of the attributes changes over time, which is evidence of the economic cycle of the real estate market. Their study also highlights the importance of income, population, accessibility and structural characteristics, to explain the price of housing and spatial differences.

Moving to the Asian continent, the researchers there embed other variables as well that are probably irrelevant to a study for the Greek market, but there are noticeable in order to understand that each market can be affected by its own characteristics and environment. Thus, in Beijing a very important determinant is the size of the apartment, the number of rooms as concluded by Zhang & Yi (2017) which are pretty common. The new variables were the air pollution levels and noise levels which were highly significant with a negative effect in the price. (Chasco Y. & Sanchez R., 2012)

All these studies focus on the price of sale for an apartment. As already mentioned, the rental price has to do with the price of the underlying asset, the house. So, the studies are relevant to the topic under discussion. Unfortunately, there was lack of more dedicated papers that would address to the exact subject.

However, with the rise of the Airbnb platform there are several researchers that developed models which try to predict the rental prices for Airbnb apartments using the characteristics that are available through the platform. There is a clear connection between the pricing for Airbnb and for the classic rental and most of the characteristics that are used in these studies as explanatory variable appear dataset that will be analyzed, in contrast with the studies for the housing prices were a lot of the characteristics were not easily accessible.

The location is still the number one factor that affects the price. Unfortunately, for this study there was no easy access to data for proximity to points of interest such as hospitals, schools, universities, landfill sites, subway stations, restaurants. These factors would be assumed to be covered from the detailed location that will be examined through the neighbourhood variable. (Yuanhang L. & Xuanyu Z.& Yulian Z., 2019),

The characteristics regarding the typology of the house and its sizing are also present in most of the models and finally a number of features are used as determinants such as the autonomous heating or the air-condition. The drawback of the Airbnb related studies is that a crucial factor in some models has to do with the comments of the previous visitors or with the profile of the owner which are irrelevant in an apartment rental situation.

II. DATA EXTRACTION

Data Source

The necessary data to build the model that will explain and predict the fair rental price of an apartment had to be pulled from an online platform that hosts advertisements with detailed descriptions for the apartments.

In Greek real estate market, there are basically two websites where almost all available apartments for rent are uploaded. The potential tenants visit these two sites to explore the available apartments and contact the owner or the real estate agent for more information. These two website are spitogatos.gr and xe.gr.

For the purpose of this study, the website of spitogatos.gr will be used for gathering the data that will be analyzed to explain the rental price and the xe.gr will be used for testing the power of the prediction model that will be the outcome of this study.

It should be clear that the credibility of the data depends in the user’s willingness to be honest. This mean that it is not totally certain whether the input of an ad reflects the reality. However, this is not a stopper for this study, since its scope is to develop a model that will predict the “fair price” as it should be depicted in an ad search based on its input and not the real fair price of a rental agreement.

Data Crawler versus Data Scraper

In order to pull the data from the websites and create the dataset for the further analysis, the techniques of data (or web) crawling and scraping were used.

In general, the term crawler means the ability of a program to navigate web pages on its own, possibly even without a clearly defined end goal or goal, endlessly exploring what a site or network can offer. Web crawlers are actively used by search engines to extract content for a URL, check this page for other links, get URLs for these links and so on.

On the other hand, web scraping is a process of extracting specific data. Unlike web crawling, a web scraper searches for specific information on specific websites or pages.

The process of data extraction that follows splits in two parts. The first one will extract the urls of the ads and the second will pull the information of these ads. For the shake of this study, even though it will not be totally accurate according to the technical terminology, the first step will be referred as the data crawler and the second as the data scraper. (Ahuja; Bal, 2014 & Zhao, 2018)

Data Crawler

The data needed for a single apartment can be found in the particular ad page. The first challenge of the data extraction is to reach each url that hosts the ad page.

An internet user can conduct his search by visiting first the home page of the website and then selecting the type of the real estate, the type of transaction (either rental or sale), the area and optionally the price and the size of his interest (Image 1).

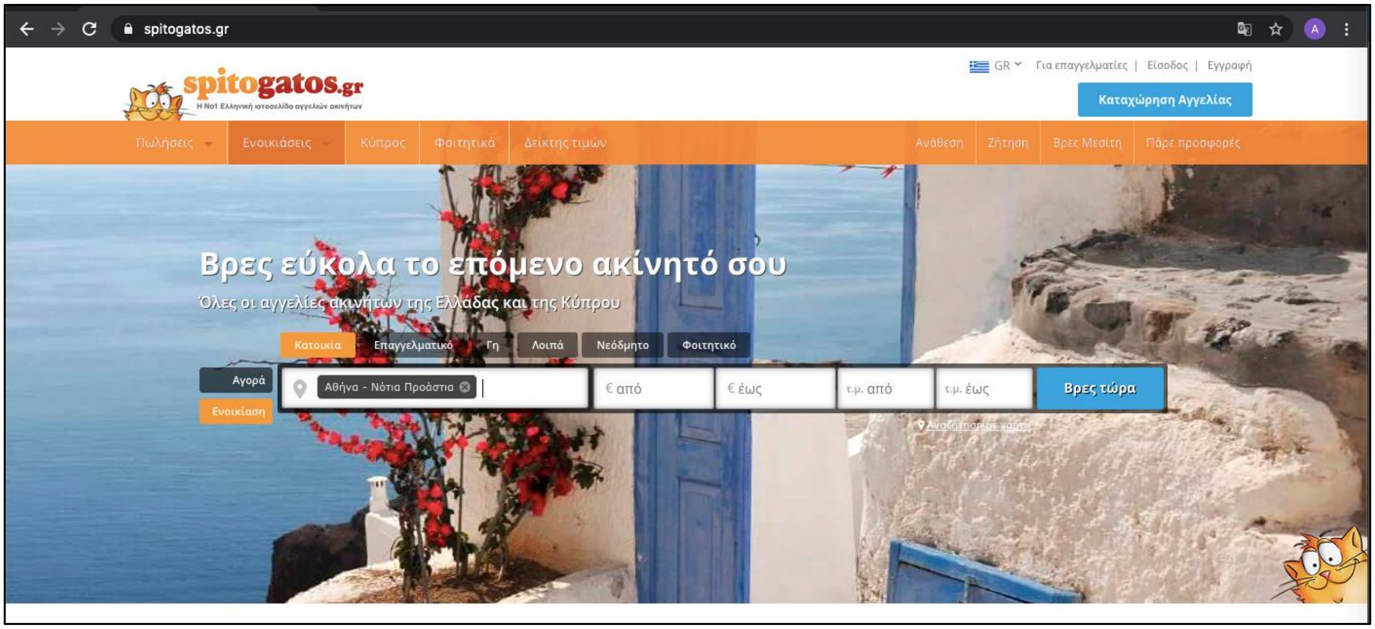
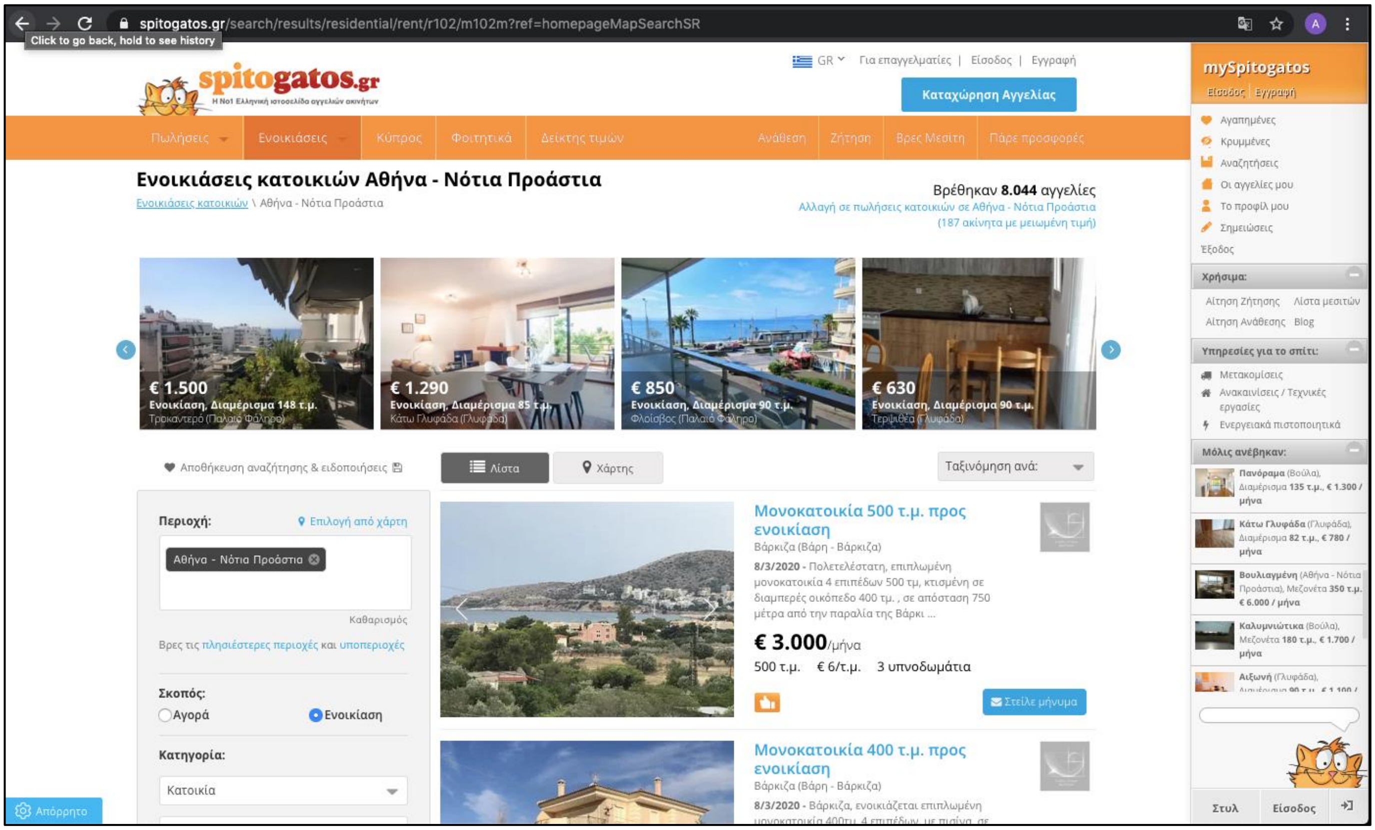


Image 1. Spitogatos.gr Search Page

The selection of the homepage will lead to the listing pages with every available ad under the specified criteria. In order to collect the data for this research, the starting point of the data crawling is the first listing page for the areas under consideration. The areas from which the data were collected are: South Attica, North Attica, West Attica, East Attica, Center of Athens, and Piraeus. The process that will be presented was repeated for each area.

The listings page has the structure that appears in Image 2. Every listing page has an area of the available filters for the user to select the specific characteristics in which he is interested. The results in each listing page are 10 ads with a short description and the basic characteristics of the apartment. However, in order to retrieve the full characteristics of each apartment, the ad page should be visited.



spitogatos.gr/search/results/residential/rent/r102/m102m?ref=homepageMapSearchSR

Ενοικιάσεις κατοικιών Αθήνα - Νότια Προάστια

Ενοικιάσεις κατοικιών \ Αθήνα - Νότια Προάστια

Βρέθηκαν **8.044** αγγελίες

Αλλαγή σε πωλήσεις κατοικιών σε Αθήνα - Νότια Προάστια (187 ακίνητα με μειωμένη τιμή)

€ 1.500 Ενοικίαση, Διαμέρισμα 148 τ.μ. Τροκαντερό (Παλαιό Φάληρο)

€ 1.290 Ενοικίαση, Διαμέρισμα 85 τ.μ. Κάτω Γλυφάδα (Γλυφάδα)

€ 850 Ενοικίαση, Διαμέρισμα 90 τ.μ. Φλοίσβος (Παλαιό Φάληρο)

€ 630 Ενοικίαση, Διαμέρισμα 90 τ.μ. Τερψιθέα (Γλυφάδα)

Περιοχή: Αθήνα - Νότια Προάστια

Βρες τις πλησιέστερες περιοχές και υποπεριοχές

Σκοπός: ○ Αγορά ● Ενοικίαση

Κατηγορία: Κατοικία

Μονοκατοικία 500 τ.μ. προς ενοικίαση

Βάρκιζα (Βάρη - Βάρκιζα)

8/3/2020 - Πολετελέστατη, επιπλωμένη μονοκατοικία 4 επιπέδων 500 τμ, κτισμένη σε διαμπερές οικόπεδο 400 τμ. , σε απόσταση 750 μέτρα από την παραλία της Βάρκι ...

€ 3.000/μήνα

500 τ.μ. € 6/τ.μ. 3 υπνοδωμάτια

Μονοκατοικία 400 τ.μ. προς ενοικίαση

Βάρκιζα (Βάρη - Βάρκιζα)

8/3/2020 - Βάρκιζα, ενοικιάζεται επιπλωμένη

Πανόραμα (Βούλα), Διαμέρισμα **135 τ.μ., € 1.300 / μήνα**

Κάτω Γλυφάδα (Γλυφάδα), Διαμέρισμα **82 τ.μ., € 780 / μήνα**

Βουλιαγμένη (Αθήνα - Νότια Προάστια), Μεζονέτα **350 τ.μ. € 6.000 / μήνα**

Καλυμνιώτικα (Βούλα), Μεζονέτα **180 τ.μ., € 1.700 / μήνα**

Αιξωνή (Γλυφάδα),

Image 2. Spitogatos.gr Listing Page

In order to collect every available ad, the data crawler filters only the type to 'Apartments' in the options of the listing pages. Starting from the first page listing page, the data crawler selects the hyperlink that includes the url of every ad in the listing page. One row of data is extracted for each listing page, including ten fields each one holding the url of an ad appearing in the visited listing page. This action is iterated through every listing page by visiting the listing page that corresponds to the hyperlink of the next button after the completion of the data extraction. This iteration process is called pagination. The pagination stops only when there is no next button, meaning that the data crawler reached the final page of the listings.

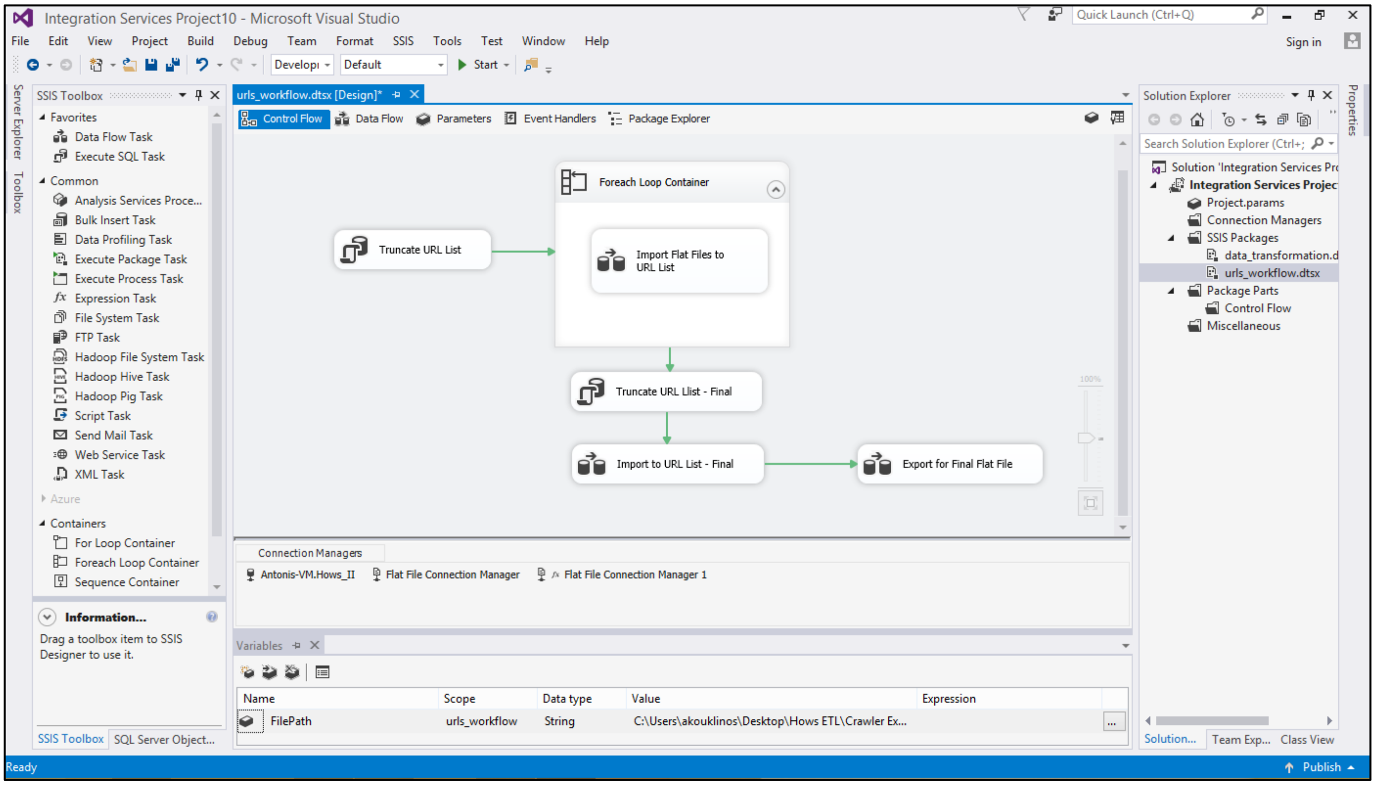
The final output of the data crawler is extracted to a flat file in which the number of rows equals to the listing pages that were visited and the ten fields equal to the respective ad urls of each page.

As mentioned before, this process is executed for each of the areas under consideration. Thus, the final outcome will be 6 flat files that contain the url of every available ad for apartment rental in these areas.

The next step was to merge the different flat files into one file in the form of a list of urls that will be used for visiting each particular ad the collecting its data.

In order to merge the flat files, an integration package was built in SSIS tool of Visual Studio. The package, as show in Image 3 consisted of five steps

1. Truncate the sql table where the data will be put with the same format that are stored in the flat files generated from the data crawler.
2. For each file found in the specified folder that the data crawler stores its output, the data of the file are transferred to the sql table and then the file is moved to a backup folder. At the end of this step, all data have been gathered in an sql table with the total rows of all listing pages and ten fields with the urls of the ads from each listing page. Also, the folder that contains the flat files is empty since after processed the flat files moved to a backup folder.
3. Truncate the sql table where the final list of urls will be stored.
4. Transform the data of the initial sql table in order to create a list of urls and transfer them to the final data table which contains a single field with the urls a number of rows equal to the total number of urls of the ads.
5. Export the sql table with the final list to a flat file in a specific location, from where it will be reached from the data scraper and used as the input file with the urls that need to be visited.

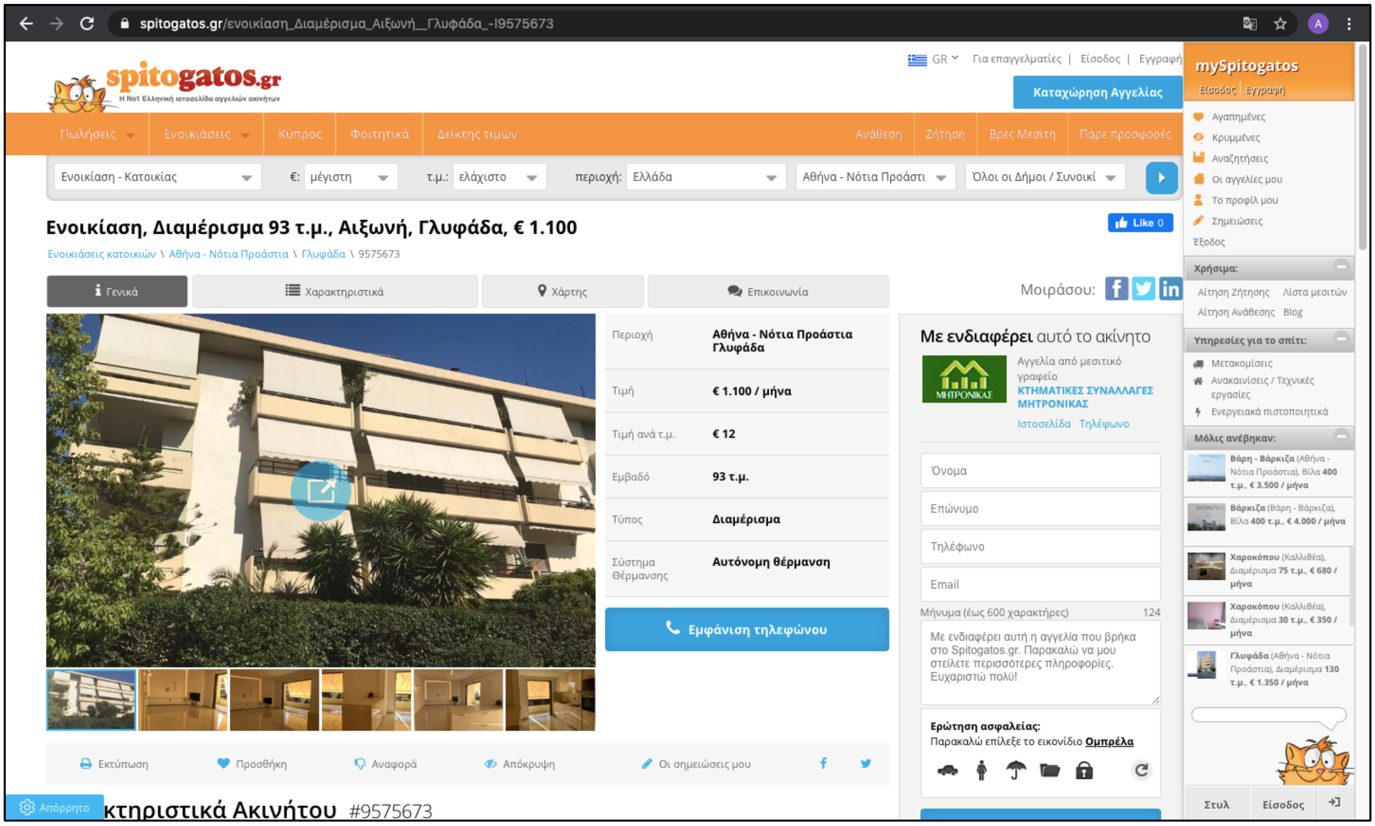


Name	Scope	Data type	Value	Expression
FilePath	urls_workflow	String	C:\Users\akouklinos\Desktop\Hows ETL\Crawler Ex...	

Image 3. SSIS – URLs List Extraction

Data Scraper

The output of the data crawler is a list of urls that will be visited from the data scraper in order to collect the characteristics of each ad. Every url corresponds to the homepage of an ad (Image 4).



Ενοικίαση, Διαμέρισμα 93 τ.μ., Αιξωνή, Γλυφάδα, € 1.100

Ενοικιάσεις κατοικιών \ Αθήνα - Νότια Προάστια \ Γλυφάδα \ 9575673

Περιοχή	**Αθήνα - Νότια Προάστια Γλυφάδα**
Τιμή	**€ 1.100 / μήνα**
Τιμή ανά τ.μ.	**€ 12**
Εμβαδό	**93 τ.μ.**
Τύπος	**Διαμέρισμα**
Σύστημα Θέρμανσης	**Αυτόνομη θέρμανση**

Με ενδιαφέρει αυτό το ακίνητο

Αγγελία από μεσιτικό γραφείο
ΚΤΗΜΑΤΙΚΕΣ ΣΥΝΑΛΛΑΓΕΣ ΜΗΤΡΟΝΙΚΑΣ
Ιστοσελίδα Τηλέφωνο

Όνομα
Επώνυμο
Τηλέφωνο
Email
Μήνυμα (έως 600 χαρακτήρες) 124

Με ενδιαφέρει αυτή η αγγελία που βρήκα στο Spitogatos.gr. Παρακαλώ να μου στείλετε περισσότερες πληροφορίες. Ευχαριστώ πολύ!

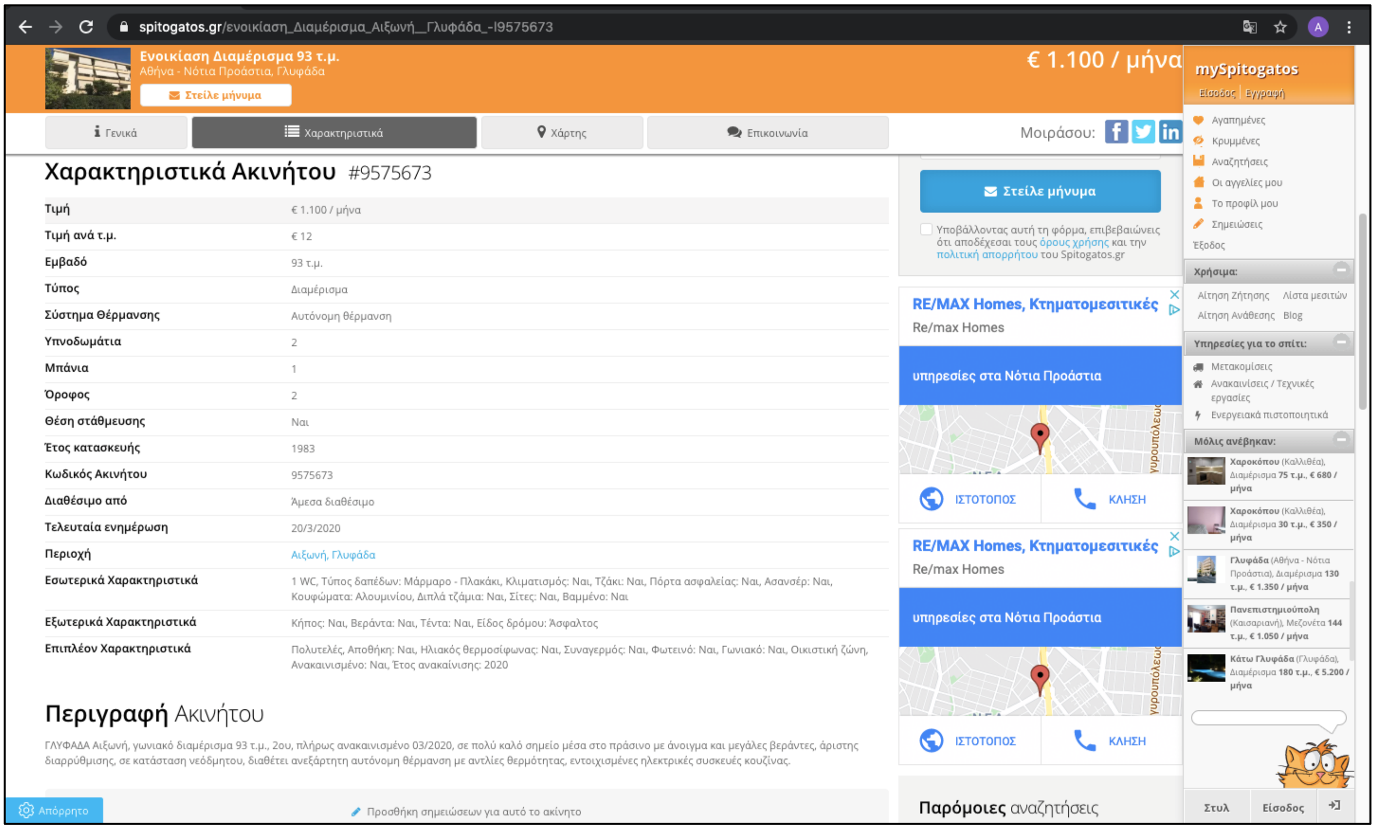
Ερώτηση ασφαλείας:
Παρακαλώ επίλεξε το εικονίδιο **Ομπρέλα**

Βάρη - Βάρκιζα (Αθήνα - Νότια Προάστια), Βίλα **400 τ.μ., € 3.500 / μήνα**
Βάρκιζα (Βάρη - Βάρκιζα), Βίλα **400 τ.μ., € 4.000 / μήνα**
Χαροκόπου (Καλλιθέα), Διαμέρισμα **75 τ.μ., € 680 / μήνα**
Χαροκόπου (Καλλιθέα), Διαμέρισμα **30 τ.μ., € 350 / μήνα**
Γλυφάδα (Αθήνα - Νότια Προάστια), Διαμέρισμα **130 τ.μ., € 1.350 / μήνα**

Χαρακτηριστικά Ακινήτου #9575673

Image 4. Spitogatos.gr Ad Page

After landing to the homepage, the data scraper collects every available information about the apartments from the list of characteristics section (Image 5).


spitogatos.gr/ενοικίαση_Διαμέρισμα_Αιξωνή__Γλυφάδα_-l9575673
Ενοικίαση Διαμέρισμα 93 τ.μ.
Αθήνα - Νότια Προάστια, Γλυφάδα
Στείλε μήνυμα
€ 1.100 / μήνα
Γενικά
Χαρακτηριστικά
Χάρτης
Επικοινωνία
Μοιράσου:
mySpitogatos
RE/MAX Homes, Κτηματομεσιτικές
Re/max Homes
υπηρεσίες στα Νότια Προάστια
ΙΣΤΟΤΟΠΟΣ
ΚΛΗΣΗ
Απόρρητο
Προσθήκη σημειώσεων για αυτό το ακίνητο
Παρόμοιες αναζητήσεις

Χαρακτηριστικά Ακινήτου #9575673

Τιμή	€ 1.100 / μήνα
Τιμή ανά τ.μ.	€ 12
Εμβαδό	93 τ.μ.
Τύπος	Διαμέρισμα
Σύστημα Θέρμανσης	Αυτόνομη θέρμανση
Υπνοδωμάτια	2
Μπάνια	1
Όροφος	2
Θέση στάθμευσης	Ναι
Έτος κατασκευής	1983
Κωδικός Ακινήτου	9575673
Διαθέσιμο από	Άμεσα διαθέσιμο
Τελευταία ενημέρωση	20/3/2020
Περιοχή	Αιξωνή, Γλυφάδα
Εσωτερικά Χαρακτηριστικά	1 WC, Τύπος δαπέδων: Μάρμαρο - Πλακάκι, Κλιματισμός: Ναι, Τζάκι: Ναι, Πόρτα ασφαλείας: Ναι, Ασανσέρ: Ναι, Κουφώματα: Αλουμινίου, Διπλά τζάμια: Ναι, Σίτες: Ναι, Βαμμένο: Ναι
Εξωτερικά Χαρακτηριστικά	Κήπος: Ναι, Βεράντα: Ναι, Τέντα: Ναι, Είδος δρόμου: Άσφαλτος
Επιπλέον Χαρακτηριστικά	Πολυτελές, Αποθήκη: Ναι, Ηλιακός θερμοσίφωνας: Ναι, Συναγερμός: Ναι, Φωτεινό: Ναι, Γωνιακό: Ναι, Οικιστική ζώνη, Ανακαινισμένο: Ναι, Έτος ανακαίνισης: 2020

Περιγραφή Ακινήτου

ΓΛΥΦΑΔΑ Αιξωνή, γωνιακό διαμέρισμα 93 τ.μ., 2ου, πλήρως ανακαινισμένο 03/2020, σε πολύ καλό σημείο μέσα στο πράσινο με άνοιγμα και μεγάλες βεράντες, άριστης διαρρύθμισης, σε κατάσταση νεόδμητου, διαθέτει ανεξάρτητη αυτόνομη θέρμανση με αντλίες θερμότητας, εντοιχισμένες ηλεκτρικές συσκευές κουζίνας.

Image 5. Spitogatos.gr Apartment Characteristics

There are three types of fields in the section. The first contains a single value (such as the ‘Price’), the second may contain a series of characteristics (such as the ‘Outdoor Characteristics’) where in most cases each feature is followed by the value ‘YES’ or ‘NO’ and finally the ‘Description’ which is a free text box that the owner of the ad can write a short description of the apartment highlighting its key features.

A demonstration ad was created in order to explore all available fields and option that an owner has when uploading an apartment rental ad. The result showed that the information of the ‘Code’, ‘Area’, ‘SubArea’ and ‘Price’ of the apartment reside in certain positions –since they are mandatory fields, in the header of the ad page. The same occurs for the ‘Description’ field in the bottom of the characteristics list. However, there were 15 more positions where the other characteristics could be found.

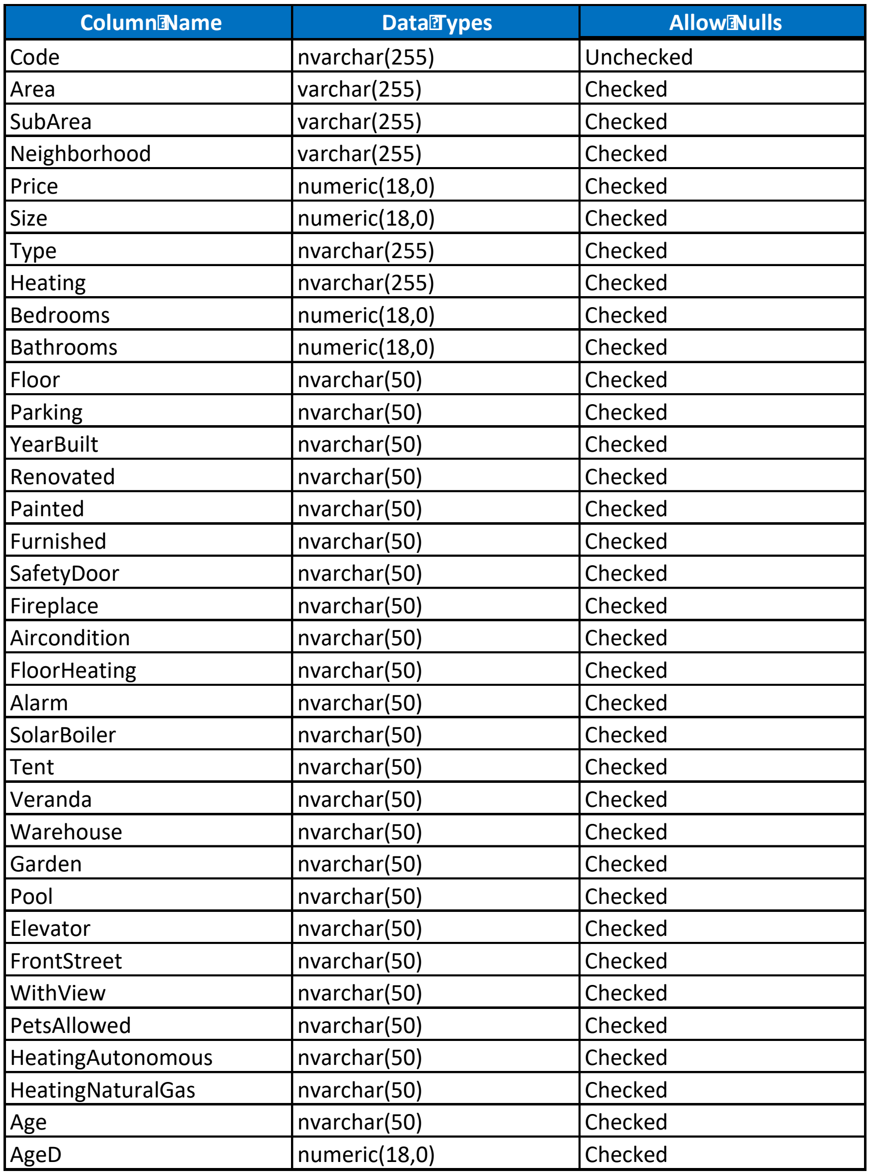
The fact that the position of all other characteristics is not stable in the ad page makes the extraction of the data more complicated. For every possible position of the characteristics list of the ad page, the field name and the field value was being extracted and saved. The url of each ad was also added to the list of the extracted values. The final list of the extraction for an ad is shown in the Table 2.

FIELD POSITION	FIELD VALUE
Code	7058919
Area	Αθήνα Βόρεια Προάστια
SubArea	Κηφισιά
Price	€ 5.000 / μήνα
F0_N	Εμβαδό
F0_V	600 τ.μ.
F1_N	Τύπος
F1_V	Βίλα
F2_N	Σύστημα Θέρμανσης
F2_V	Αυτόνομη Θέρμανση (Πετρέλαιο)
F3_N	Υπνοδωμάτια
F3_V	5
F4_N	Μπάνια
F4_V	5
F5_N	Όροφος
F5_V	Ισόγειο
F6_N	Επίπεδα
F6_V	4
F7_N	Θέση στάθμευσης
F7_V	Ναι
F8_N	Έτος κατασκευής
F8_V	1998
F9_N	Κωδικός Ακινήτου
F9_V	PA-117
F10_N	Διαθέσιμο από
F10_V	
F11_N	Τελευταία ενημέρωση
F11_V	08/08/2019
F12_N	Περιοχή
F12_V	Κεφαλάρι, Κηφισιά
F13_N	Εσωτερικά Χαρακτηριστικά
F13_V	3 Σαλόνια, 1 Κουζίνα, 1 WC, Τύπος δαπέδων: Ξύλο, Κλιματισμός: Ναι, Σοφίτα: Ναι, Τζάκι: Ναι, Playroom: Ναι, Πόρτα ασφαλείας: Ναι, Εσωτερική σκάλα: Ναι, Κουφώματα: Ξύλινα, Διπλά τζάμια: Ναι, Σίτες: Ναι
F14_N	Εξωτερικά Χαρακτηριστικά
F14_V	Κήπος: Ναι, Επιφάνεια Οικοπέδου: 800 τ.μ., Βεράντα: Ναι, Είδος δρόμου: Ασφαλτος, Προσανατολισμός: Ανατολικομεσημβρινός
F15_N	Επιπλέον Χαρακτηριστικά
F15_V	Αποθήκη: Ναι, Ηλιακός θερμοσίφωνας: Ναι, Συναγερμός: Ναι, Δορυφορική κεραία: Ναι, Διαμπερές: Ναι, Φωτεινό: Ναι, Θέα: Ναι, Πισίνα: Ναι, Κατοικίδια ευπρόσδεκτα: Ναι, Πρόσοψης: Ναι, Γωνιακό: Ναι, Οικιστική ζώνη, Νυχτερινό ρεύμα: Ναι, Ενεργειακή κλάση: Β+
Description	ΚΗΦΙΣΙΑ ΚΕΦΑΛΑΡΙ. Αρχοντική μονοκατοικία 600τμ, τεσσάρων επιπέδων (playroom, ισόγειο, 1ος όροφος, σοφίτα), εντός οικοπέδου 800τμ, με μεγάλη πισίνα και BBQ. Η κατοικία αναπτύσσεται ως εξής:Ισόγειο: Μεγάλο σαλόνι με τζάκι, τραπεζαρία, μεγάλο γραφείο με τζάκι, WC, αποθηκάκι, κλειστή κουζίνα με εντοιχισμένες ηλεκτρικές συσκευές που επικοινωνεί με δική της στεγασμένη βεράντα με BBQ. Α' όροφος: Μια μεγάλη νεοκλασικού τύπου σκάλα οδηγεί από το σαλόνι στο επίπεδο αυτό, το οποίο διαθέτει μεγάλο εξώστη, ελεύθερο χώρο κατάλληλο για γραφείο ή καθημερινό tv room, τρία μεγάλα υπνοδωμάτια με πολλές ντουλάπες (εκ των οποίων το ένα μάστερ) και δύο πλήρη μπάνια. Β' όροφος (σοφίτα): Το επίπεδο αυτό διαθέτει ένα μεγάλο υπνοδωμάτιο με μπάνιο, τύπου σουίτας.Ημι-ισόγειο (250τμ): Διαθέτει μεγάλο playroom με τζάκι, ξενώνα, μεγάλο laundry, δύο πλήρη μπάνια, γκαράζ για 3 αυτοκίνητα. Το επίπεδο αυτό έχει ανεξάρτητη είσοδο και μπορεί να λειτουργήσει και ως ανεξάρτητο διαμέρισμα.Με ανεξάρτητη θέρμανση πετρελαίου, κλιματιστικά, ξύλινα κουφώματα με διπλά τζάμια, δρύινα δάπεδα σε όλους τους χώρους, δυνατότητα προσθήκης καμπίνας ανελκυστήρα, συναγερμό, σε καταπράσινο περιβάλλον και με πολύ ωραία θέα από τα υπνοδωμάτια, αποτελεί μοναδική επιλογή για ενοικίαση κατοικίας στο καλύτερο σημείο της Κηφισιάς. Με πολύ εύκολη πρόσβαση στην πλατεία και τα ξενόγλωσσα σχολεία της περιοχής, εύκολη πρόσβαση στις κεντρικές οδικές αρτηρίες και την εθνική οδό, συστήνεται ανεπιφύλακτα σε υψηλόβαθμα στελέχη πολυεθνικών εταιρειών ή πρεσβειών. Προσφέρεται στην τιμή των 5.000€.Πάρα πολλά ακόμα ακίνητα,μικρότερα-μεγαλύτερα,στην Κηφισιά κ στους όμορους Δήμους www.panteleonrealestate.gr Μεσιτικό Κώστας Παντελαίων-Κηφισιά-6945-392060 info@panteleon.gr Σε περίπτωση ενδιαφέροντος καλέστε και στο 6974-793435 από 11:00 21:00.
Page_URL	https://www.spitogatos.gr/%CE%B5%CE%BD%CE%BF%CE%B9%CE%BA%CE%AF%CE%B1%CF%83%CE%B7_%CE%92%CE%AF%CE%BB%CE%B1_%CE%9A%CE%B5%CF%86%CE%B1%CE%BB%CE%AC%CF%81%CE%B9__%CE%9A%CE%B7%CF%86%CE%B9%CF%83%CE%B9%CE%AC_-l7058919

Table 2. Ad Elements Extracted

III. ETL PROCESS

The next step is to create an sql table with the available characteristics that can be extracted from an ad. The available characteristics were written down during the creation of the demonstration ad. In Table 3, the design of the resulting table is shown. This will be the target table where each value of the raw extraction should be placed in the appropriate field and format (data type).



Column Name	Data Types	Allow Nulls
Code	nvarchar(255)	Unchecked
Area	varchar(255)	Checked
SubArea	varchar(255)	Checked
Neighborhood	varchar(255)	Checked
Price	numeric(18,0)	Checked
Size	numeric(18,0)	Checked
Type	nvarchar(255)	Checked
Heating	nvarchar(255)	Checked
Bedrooms	numeric(18,0)	Checked
Bathrooms	numeric(18,0)	Checked
Floor	nvarchar(50)	Checked
Parking	nvarchar(50)	Checked
YearBuilt	nvarchar(50)	Checked
Renovated	nvarchar(50)	Checked
Painted	nvarchar(50)	Checked
Furnished	nvarchar(50)	Checked
SafetyDoor	nvarchar(50)	Checked
Fireplace	nvarchar(50)	Checked
Aircondition	nvarchar(50)	Checked
FloorHeating	nvarchar(50)	Checked
Alarm	nvarchar(50)	Checked
SolarBoiler	nvarchar(50)	Checked
Tent	nvarchar(50)	Checked
Veranda	nvarchar(50)	Checked
Warehouse	nvarchar(50)	Checked
Garden	nvarchar(50)	Checked
Pool	nvarchar(50)	Checked
Elevator	nvarchar(50)	Checked
FrontStreet	nvarchar(50)	Checked
WithView	nvarchar(50)	Checked
PetsAllowed	nvarchar(50)	Checked
HeatingAutonomous	nvarchar(50)	Checked
HeatingNaturalGas	nvarchar(50)	Checked
Age	nvarchar(50)	Checked
AgeD	numeric(18,0)	Checked

Table 3. SQL Target Table

The process that will transform the extraction list to the resulting sql table is built through an SSIS package in three steps as shown in Image 6. Before starting the ETL process, since the data that will be extracted from the web are in Greek language, the SQL Server and the SSIS tools of the Visual Studio should be set accordingly. This means that the Locale ID should be set to ‘Greek’ and the Default CodePage to ‘1253’ enabling also the option AlwaysUseDefaultCodePage. Otherwise, the data could be transferred to the SQL but they could not be readable using the Greek language in the query conditions.

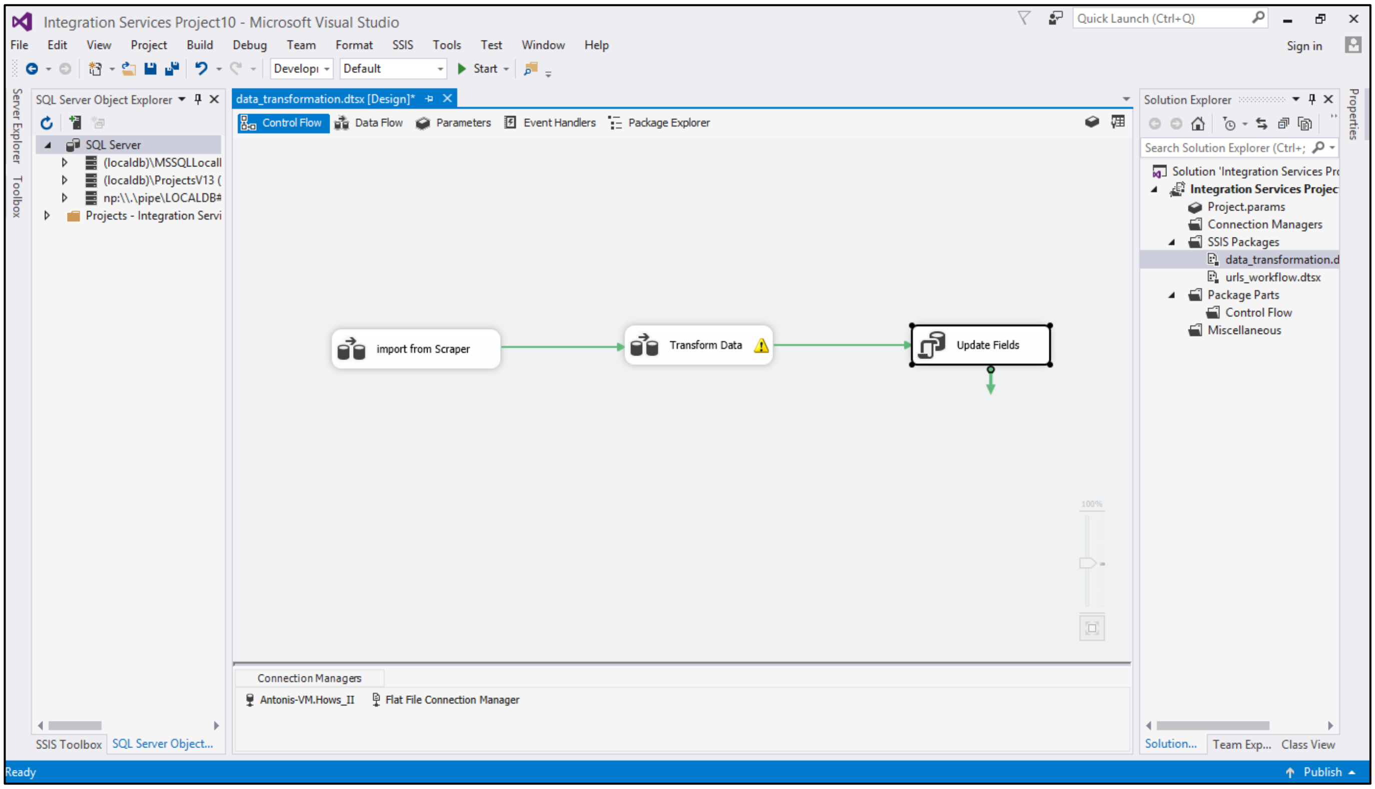


Image 6. SSIS – Data Transformation

The first step moves the data from the flat file that the data scraper exports in the form of the table 1 to a stage sql table. The second step inserts the records of the staging table to the fields of the resulting table (Table 3) by reading each line and searching for the values that should be placed in each field. This step also performs a transformation of the data that are extracted in a different form than the desired. For example, the value of the Price is extracted from the website as a text ‘€ 5.000 / μήνα’ and this should be cleared to keep and transform to numeric the value 5000. The script that can be found in Appendix (SQL Query – Data Transformations) executes these transformations.

The third step performs an update to a number of fields of the resulting table in order the values of these fields to be in the proper form to apply the statistical analysis. For example, the floor should be defined as a numeric variable in the statistical analysis, but the raw data contain also text description. Thus, an update is applied to transform the values 'Ισόγειο', 'Ημιόροφος', 'Υπόγειο' to the respective 0, 0.5, -1 numeric values.

At this step, there is also a creation of the Age Variables (Age and AgeD) which based on the field YearBuilt calculate the age of the apartment both in absolute values and in decades.

The update script can be found in the Appendix (SQL Query – Data Transformations Updates).

The final step of the ETL process aims to cleanse the data in order to assure that every line of the sql table refers in an apartment rental ad. It has been observed that there are fault uploads in the website where the data are retrieved from. Ads that refer to Sales are uploaded to the Rental section. This can cause a lot of outliers in the statistical analysis that will follow. These ads will be deleted from the dataset using a rule of thumb which states that the max Price per Square Meter (PpM) for a rental should not be over 50 €/m^2. All ads with a PpM value over 50, will be considered as fault sales ads and will be removed from the dataset with the script that can be found in the Appendix (SQL Query – Data Cleansing).

IV. DESCRIPTIVE STATISTICS

After completing the ETL process and gathered the available data from the web to the resulting table in the desired form, a connection is made between SQL Server and R Studio in order to import the data table into an R Dataframe for further analysis.

The Dataset

The main purpose of the analysis in R is to build a regression model that will capture the variables that describe the rental prices. The y value (dependent variable) of the regression that the model will try to explain will be the Price per Square Meter (PpM). It was decided to use this measure instead of the Price because of its smoother distribution.

Before starting to building the model, all the potential explanatory variable of the dataset should be set the appropriate datatypes of the variables.

First, they will be separated into two basic groups, the numerical variables and non-numerical variables (or factors). As numerical variables will be considered the ones that can have an integer value.

These are the Price, Size, Floor, Bathrooms, Bedrooms, YearBuilt and AgeD.

The factor explanatory variables are the dummy variables that consist of all the available features of an apartment that can be found in an ad.

These are the Heating, Parking, Renovated, Painted, Furnished, SafetyDoor, Fireplace, Aircondition, FloorHeating, Alarm, SolarBoiler, Tent, Veranda, Warehouse, Garden, Pool, Elevator, FrontStreet, WithView, PetsAllowed, HeatingAutonomous and HeatingNaturalGas. All these are variables that can have a value of ‘YES’ or ‘NO’ and will appear in the model as the dummy variable ***_YES.

There are, also, the location factors which are the Area, SubArea and Neighborhood which have multiple values (levels of the factor).

The resulting dataset is as follows.

> str(hows)

```
> str(hows)
data.frame': 6329 obs. of 34 variables:
 $ Area              : Factor w/ 7 levels "Αθήνα","Βόρεια Προάστια",..: 3 3 3 3 3 3 3 3 3 3 ..
 $ SubArea           : Factor w/ 69 levels "Αγία Παρασκευή",..: 13 13 13 13 13 13 13 13 13 13 ..
 $ Neighborhood      : Factor w/ 224 levels "Hilton","NULL",..: 2 2 2 2 2 2 2 2 2 2 ..
 $ Price             : num 550 580 1400 1200 850 900 900 1450 1450 600 ..
 $ Size              : num 50 46 150 98 55 102 115 132 150 75 ..
 $ Type              : Factor w/ 1 level "Διαμέρισμα": 1 1 1 1 1 1 1 1 1 1 ..
 $ Heating           : Factor w/ 31 levels "Fan coil","NULL",..: 10 10 2 3 21 10 3 2 3 2 ..
 $ Bedrooms          : num 1 1 3 2 1 3 2 2 3 2 ..
 $ Bathrooms         : num 1 1 1 2 1 1 2 1 2 1 ..
 $ Floor             : num 0 1 1 1 3 6 2 NA 1 1 ..
 $ Parking           : Factor w/ 2 levels "No","Yes": 1 1 1 2 1 2 1 1 1 1 ..
 $ YearBuilt         : num 2002 1980 1980 2000 1978 ..
 $ Renovated         : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ..
 $ Painted           : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ..
 $ Furnished         : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ..
 $ SafetyDoor        : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ..
 $ Fireplace         : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ..
 $ Aircondition      : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ..
 $ Alarm             : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ..
 $ SolarBoiler       : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ..
 $ Tent              : Factor w/ 2 levels "No","Yes": 1 1 1 2 1 2 1 2 1 2 ..
 $ Veranda           : Factor w/ 2 levels "No","Yes": 2 1 1 2 2 2 1 2 1 2 ..
 $ Warehouse         : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ..
 $ Garden            : Factor w/ 2 levels "No","Yes": 2 1 1 2 1 1 1 2 1 2 ..
 $ Pool              : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ..
 $ Elevator          : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ..
 $ FrontStreet       : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ..
 $ WithView          : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ..
 $ PetsAllowed       : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ..
 $ HeatingAutonomous: Factor w/ 2 levels "No","Yes": 2 2 1 2 1 2 2 1 2 1 ..
 $ HeatingNaturalGas: Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ..
 $ Age               : Factor w/ 4 levels "<20",">40","20-40",..: 1 3 3 1 2 2 1 4 1 4 ..
 $ AgeD              : num 2 4 4 2 5 5 2 10 2 10 ..
 $ ppm               : num 11 12.61 9.33 12.24 15.45 ..
```

Table 4. R Dataframe

Numerical Variables

In Figure 1 it is obvious that neither the dependent variable nor any of the explanatory variables seem to follow the normal distribution.

The PpM, which is a combination of the price and size distributions follows also the log normal distribution, as expected from the literature review.

This leads to a model where the dependent variable should be the log(PpM) and also the size variable should be used as log(size) as well.

The values Bedrooms, Bathrooms seems also to follow the log normal distribution and if used in the model, the most appropriate form should be the log. However, it is highly probable there is collinearity among the variables Size and Bedrooms or Bathrooms since it is common sense that the larger the apartment, the more the extra rooms, so the assumption will be that the Size capture already the information that Bathrooms or Bedrooms give. The first hint of a possible collinearity is between these variables is shown in the Correlation matrix in the next paragraph.

The discussion of collinearity is more obvious between the variables YearBuilt and AgeD (Age in decades) where the second is create from the first. For simplicity reasons in the user interaction with the final model, the variable AgeD will be used (if significant) in the model, instead of YearBuilt.

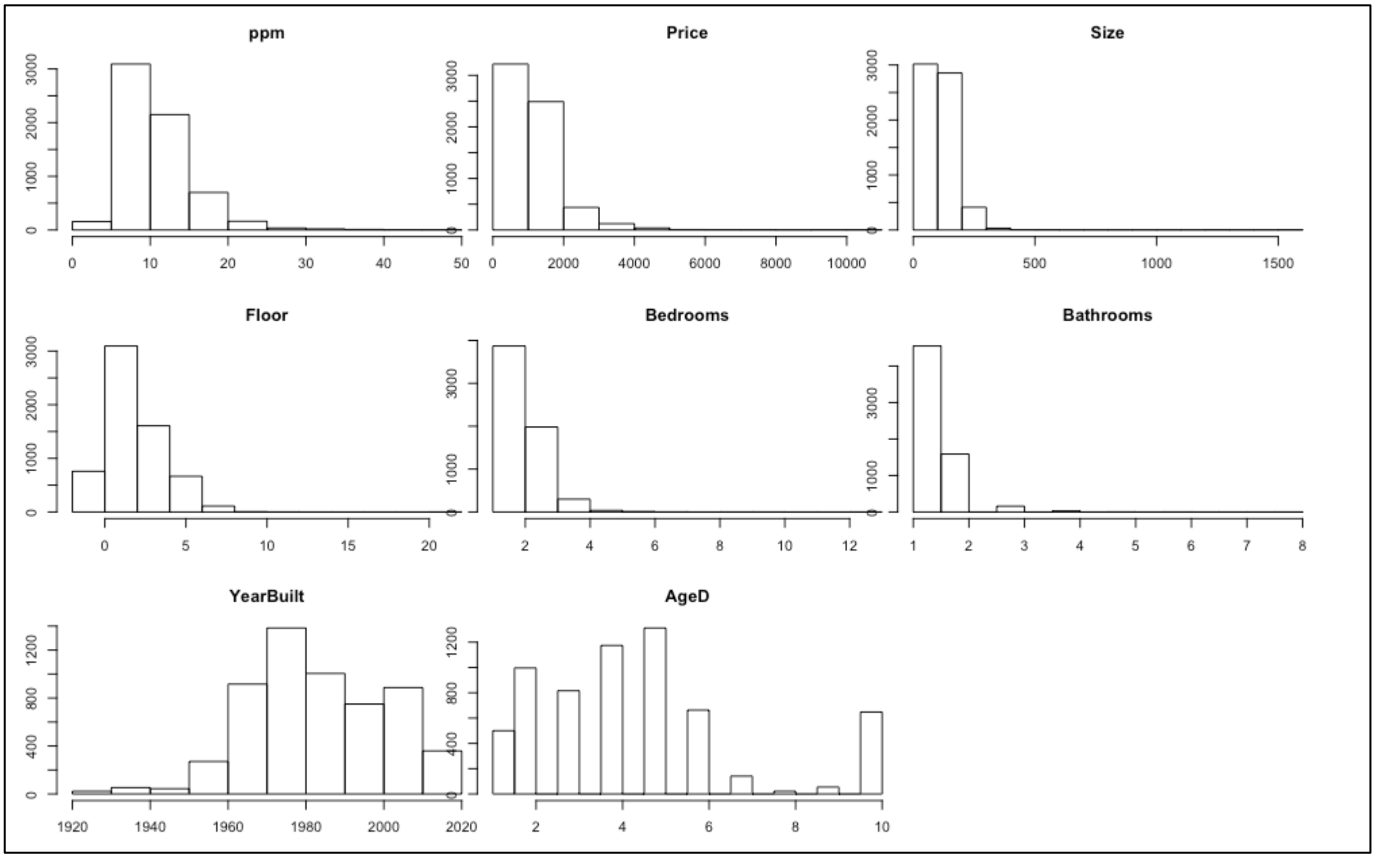


Figure 1. Numerical Variables Histograms

Pairwise Comparisons

Before proceeding to the creation of the model, it is necessary to check the relations among the variables of the dataset. In numerical variables, the best way to do that is through the correlation matrix (Figure 2).

Numerical Variables

As mentioned previously, the Age and AgeD have a perfect negative correlation.

Size is highly correlated with Price, Bedrooms and Bathrooms.

The correlation matrix shows also the signs the direction of the influence of the explanatory numeric variables to the dependent variable. Thus, this matrix will be taken into consideration when building the final model.

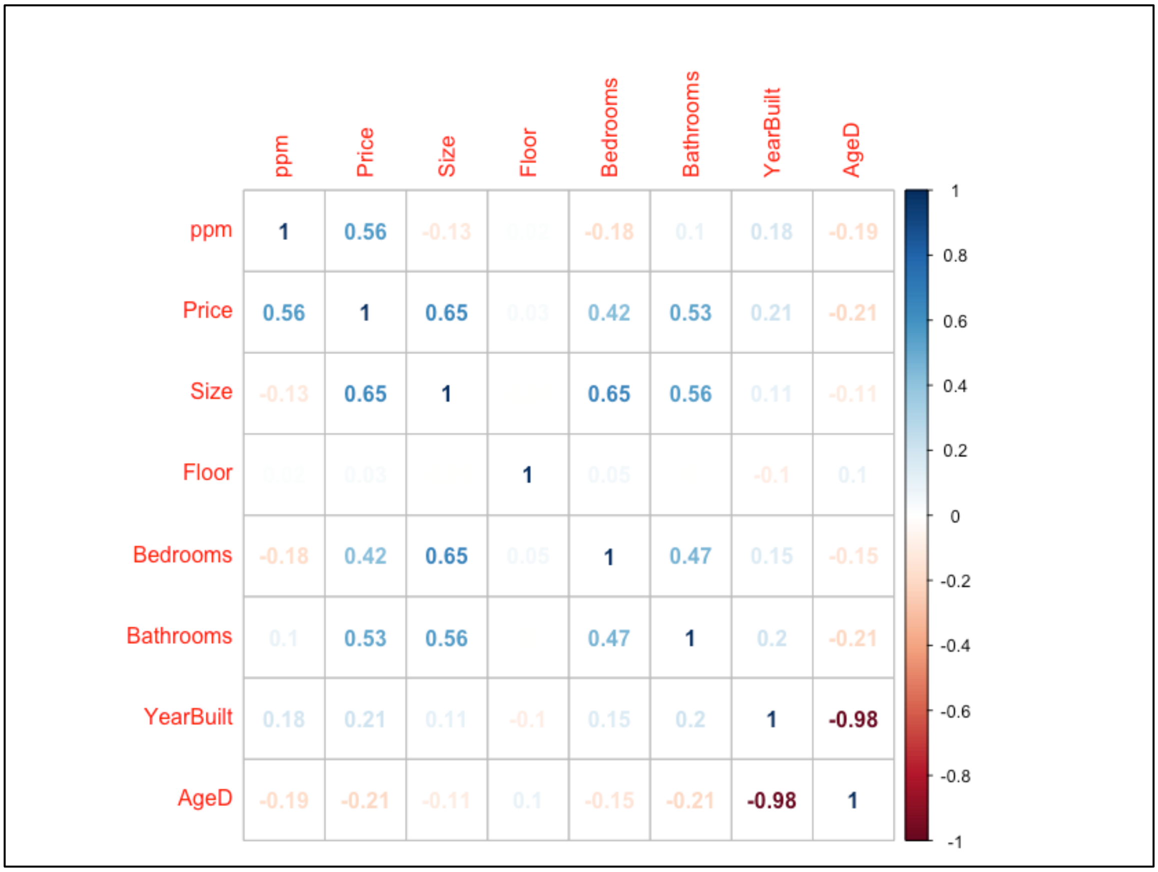


Figure 2. Numerical Variables Correlations

Non-Numerical Variables

In Figure 3-6, the relation of the non-numerical variables to the PpM are depicted, giving insights that will be used as expectations for the model.

Area, SubArea, Neighborhood and Heating are variables with multiple levels and the box plot representation does not offer much information about certain levels and their relation with the PpM.

As regards all other variables that present the features of an apartment and take values either 'YES' or 'NO', there are cases that the avg PpM seem to be independent of the value of the for the most of them (SafetyDoor, Fireplace, FloorHeating, Alarm, SolaBoiler, Tent, Warehouse, Garden, Pool, Elevator, FloorHeating, WihView, PetsAllowed) by showing the same avg for the two levels of the factor.

On the other hand, the features Parking, Renovated, Painted, Furnished, Aircondition, Veranda, HeatingAutonomous and HeatingNaturalGas are more probable to be determinants of the PpM.

In any case, the fact that there are a lot of values outside the interquartile range makes every interpretation of the charts uncertain.

At this point, it should be taken into consideration that the values of the most features as will later be presented in the visualization section, are far from balanced. This means that even if there is a significant difference in the average PpM between the two levels of the variable, if the values of variable are in a proportional 95%-5% or even less, their use in a forecasting model should probably be avoided because this proportion could be indirectly considered as the missing values of the variables (even if added, the model will throw it off as non-significant through the t-test).

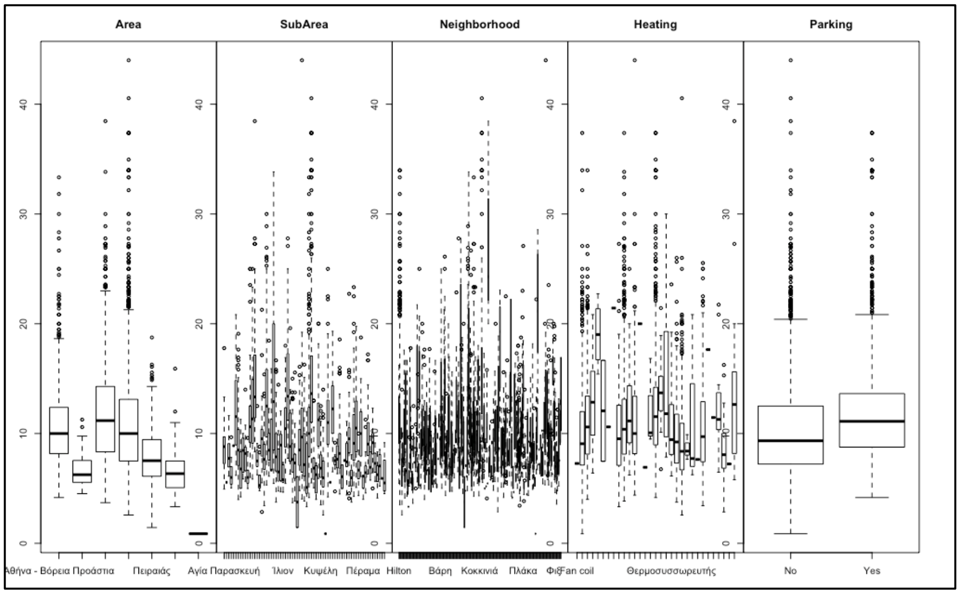


Figure 3. Non-Numerical Variables Boxplots - I

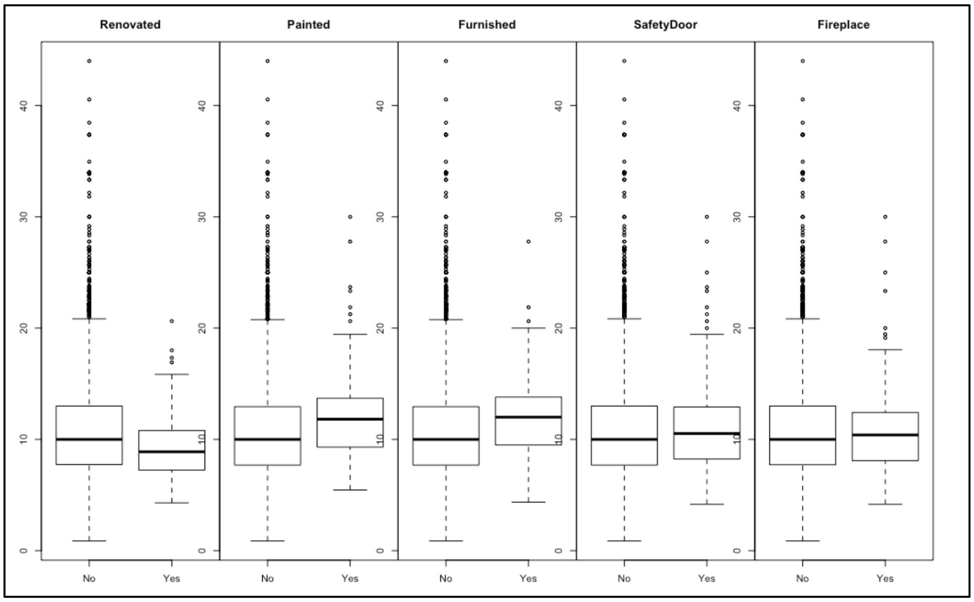


Figure 4. Non-Numerical Variables Boxplots - II

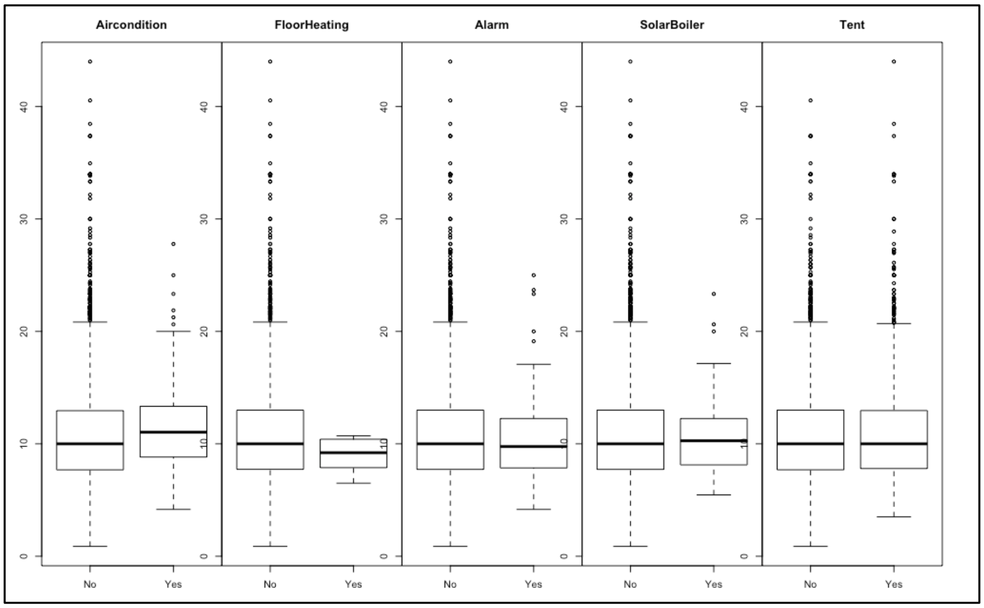


Figure 5. Non-Numerical Variables Boxplots - III

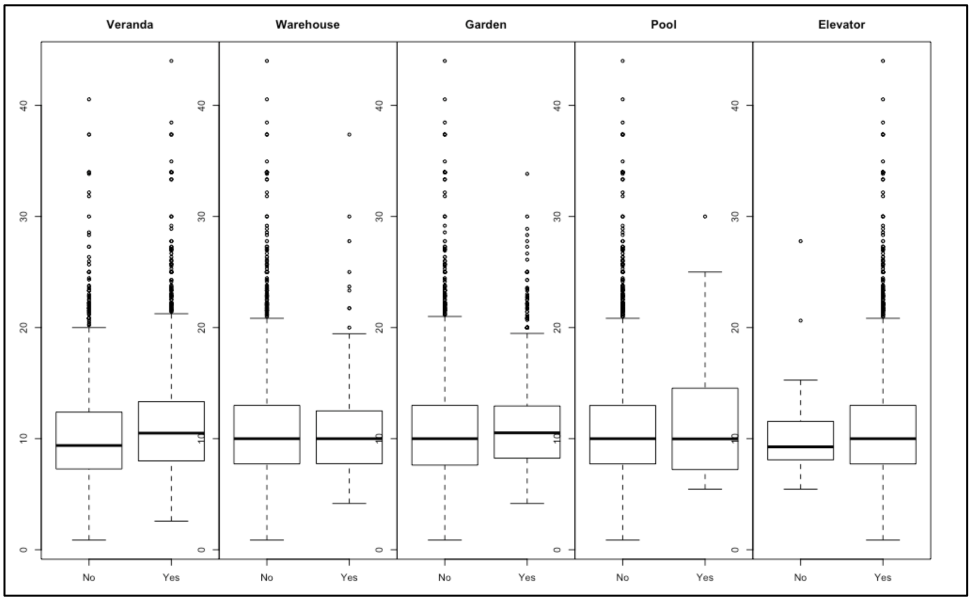


Figure 6. Non-Numerical Variables Boxplots - IV

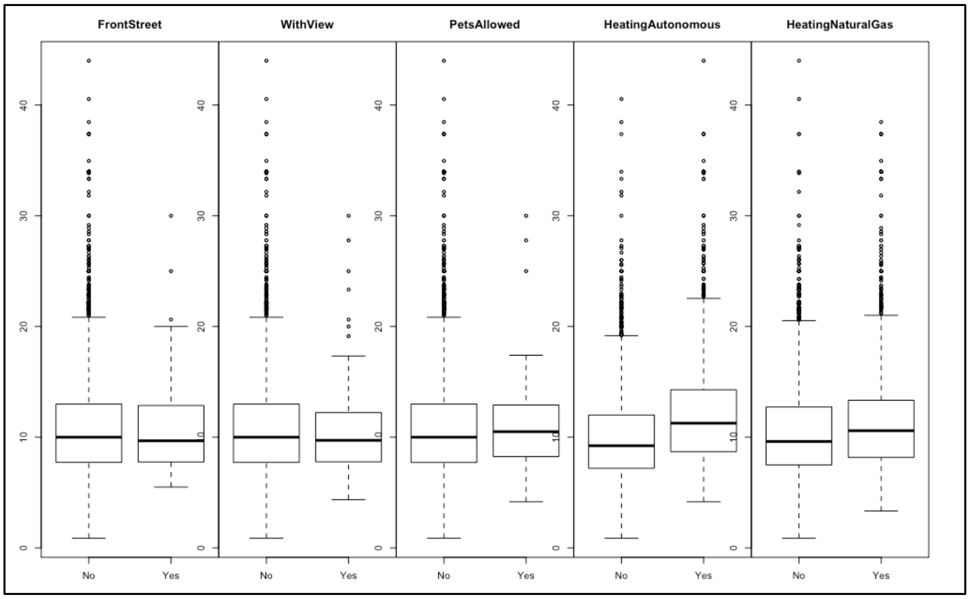


Figure 7. Non-Numerical Variables Boxplots - V

V. FAIR PRICE MODEL

Several models were tested before reaching to the final model. The indicator that was used to compare the models was the R square adjusted (adjR^2) which measures the explanatory power of the model.

Starting from the full model, where the PpM was regressed against all the explanatory variables, there should be a reduction in the variables in order to make the model simpler for the end user. This variable reduction also targeted to improve the adjR^2 and the model's performance in the residuals tests. The rational that used to exempt variables from the model took under consideration the collinearity among some variables as described earlier, the suggestions of similar studies as described in literature review and of course the t-test for variable significance.

The final model includes the explanatory variables:

SubArea, Neighborhood due to the strong recommendations of the literature for geo-location variables.

Size is also suggested by the previous studies, but as the location variables are of great significance for the user search using the Fair Price tool that is the main purpose of the model.

AgeD and Floor was found to be significant variables that contribute to adjR^2 and also would be very useful for the user's search.

As regards the features, the variables Furnished, Parking, Tent, Aircondition, HeatingAutonomous and HeatingNaturalGas were selected.

The signs of the coefficients of the explanatory variables follow the common sense and the analysis that were presented in the previous section. More specifically, the size has a negative sign -which means as the size increases, the PpM falls. The same happens for the age in decades, which indicates that the older the house, the lower the price per square meter. The features have all except for the aircondition positive signs, which means that their presence increase the PpM. The negative direction of the aircondition relation is explained due to the fact that in most cases the aircondition is used instead of other means of heating. Thus, the aircondition means absence of central heating which gives meaning to the negative sign.

After selecting the variables, the Size and AgeD needed to be transformed in order to improve the performance of the model in the residuals' tests that are presented in the next section. The distributions of the two variables, but mostly the visualizations against the avg PpM indicate that the Size should be transformed to log(Size) and AgeD to poly(AgeD,3). The result of these transformation was to increase the adjR^2 and improve the residuals tests.

The Final Model is:

$$\mathbf{Log(PpM) = 3.964 - 0.320log(Size) - 0.032Aircondition + 0.040Floor - 0.003Floor^{2} + 0.000Floor^{3}}$$
$$\mathbf{+ 0.059Furnished + 0.059HeatingAutonomous + 0.048HeatingNaturalGas + 0.089Parking}$$
$$\mathbf{+ 0.015Tent - 0.297AgeD + 0.052AgeD^{2} - 0.003AgeD^{3}}$$
$$\mathbf{+ \varepsilon\ ,\ \varepsilon \sim N(0,0.252^{2})}$$

Final Model

Coefficients	**Estimate**	**Std. Error**	**t Value**	**p Value**	**Pr(>|t|)**
(Intercept)	3.964	0.074	53.290	0.000	***
log(Size)	-0.320	0.009	-36.367	0.000	***
AirconditionYes	-0.032	0.019	-1.647	0.100	.
poly(Floor, 3, raw = TRUE)1	0.040	0.007	5.994	0.000	***
poly(Floor, 3, raw = TRUE)2	-0.003	0.001	-2.029	0.042	*
poly(Floor, 3, raw = TRUE)3	0.000	0.000	1.738	0.082	.
FurnishedYes	0.059	0.027	2.175	0.030	*
HeatingAutonomousYes	0.059	0.008	7.312	0.000	***
HeatingNaturalGasYes	0.048	0.008	6.108	0.000	***
ParkingYes	0.089	0.010	9.136	0.000	***
TentYes	0.015	0.008	1.942	0.052	.
poly(AgeD, 3, raw = TRUE)1	-0.297	0.020	-15.094	0.000	***
poly(AgeD, 3, raw = TRUE)2	0.052	0.004	11.729	0.000	***
poly(AgeD, 3, raw = TRUE)3	-0.003	0.000	-9.837	0.000	***
SubArea ###					
Neighborhood ###					

Multiple R-squared: 0.5807, Adjusted R-squared: 0.5594

Residual standard error: 0.252 on 5933 degrees of freedom, (88 observations deleted due to missingness)

F-statistic: 27.33 on 299 and 593 p-value: < 2.2e-16

Table 5. Final Model

Testing the Model Assumptions

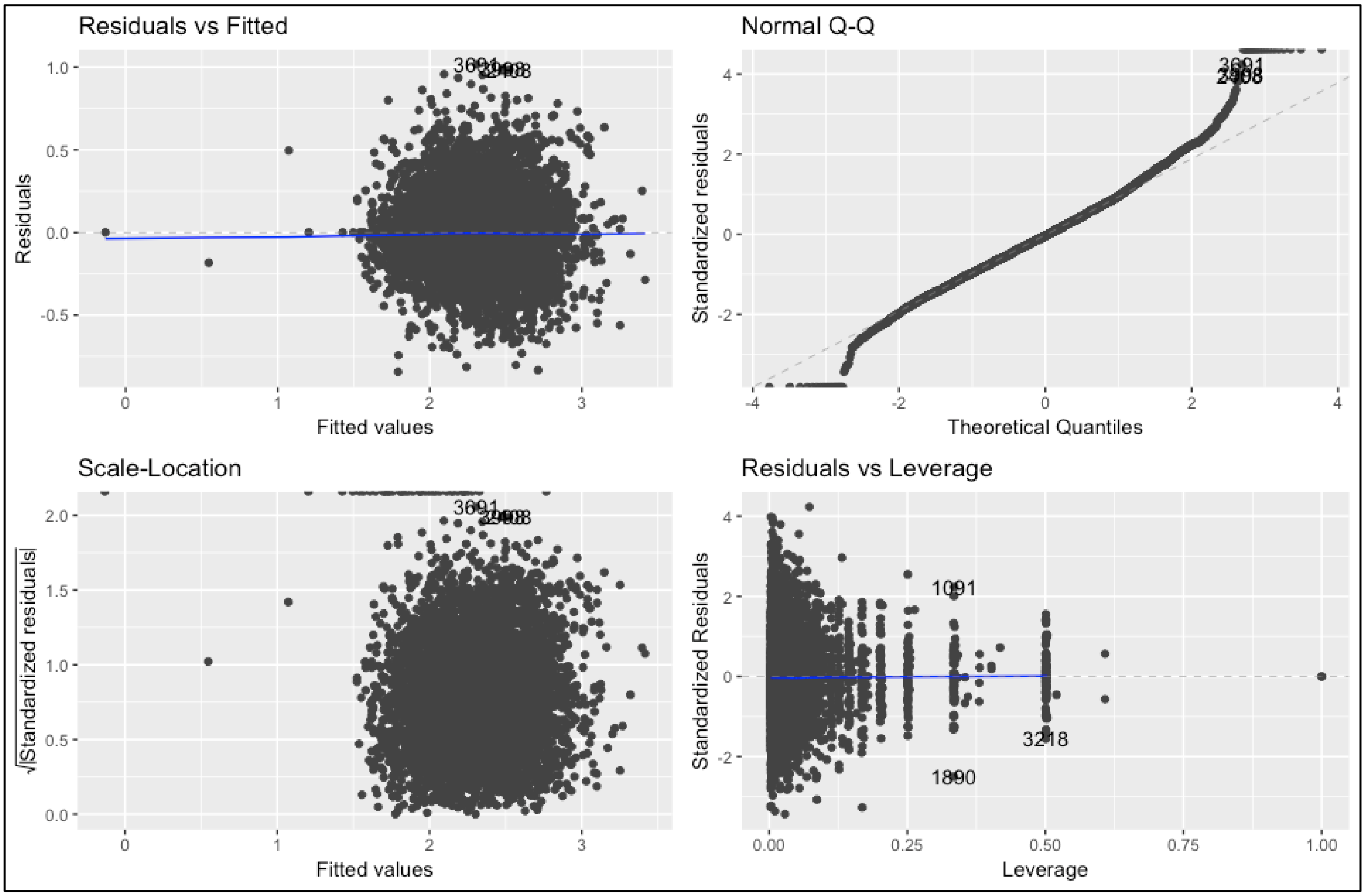


Figure 8. Residuals Tests

Residuals vs Fitted

The plot of residuals versus predicted values is useful for checking the assumption of linearity and homoscedasticity. If the model does not meet the linear model assumption, then the residuals would draw a defined shape or a distinctive pattern. For example, if the plot looked like a parabola, that would be bad. Instead, the scatterplot of residuals looks more likely to the night sky – without any distinctive patterns. The blue line through the scatterplot is also straight and horizontal, not curved, which means that the linearity assumption is satisfied. Finally, in order the homoscedasticity assumption to be met, the residuals should be equally spread around the y = 0 line, which is also observed in the chart.

Normal Q-Q

The normality assumption can be also evaluated using a QQ-plot by comparing the residuals to an ideal, normal observations along the 45-degree line. In general, the normality assumption is met, with a problem in the extreme values of the range.

Another way to check this assumption is the distribution of Studentized Residuals as shown in Figure 9, which confirms that the (studentized) residuals follow the normal distribution.

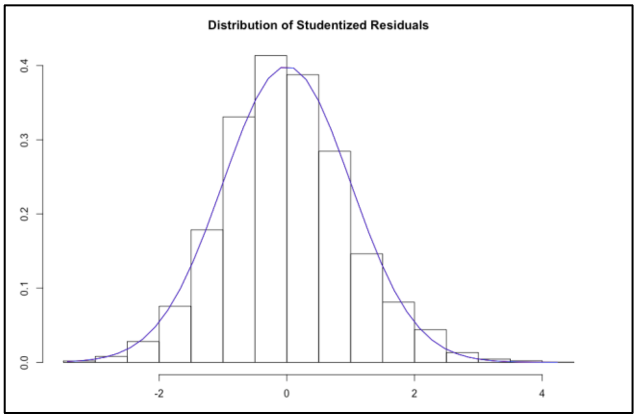


Figure 9. Studentized Residuals Distribution

Scale – Location

The third plot is a scale-location plot (square rooted standardized residual vs. predicted value). This is useful for checking the assumption of homoscedasticity. In this particular plot it should be checked whether an obvious pattern in the residuals exists. The absence of a red line, means that it is flat and horizontal with equally and the data points are randomly in both sides. This is a sign that the assumption is not violated.

Residuals vs Leverage

The fourth plot helps find influential cases, if any are present in the data. It should be noticed that the outliers may or may not be influential points. The influential outliers are of the greatest concern. They could alter the results, depending on whether they are included or excluded from the analysis. If there was a dash red curve in the chart, it would mean that influential cases exist in the dataset. Since there is no red (Cook's distance) curved line on the plot, there is no problem, even though in Figure 10 appear some data points that can be considered as outliers. These outliers will not be removed from the dataset because they are influential.

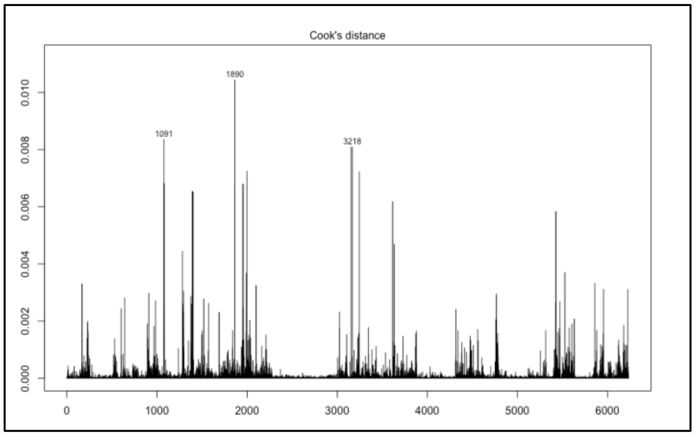


Figure 10. Outliers according to Cook's Distance

Finally, the assumption of Independence of errors is also met as shown in Figure 11 since there is no clear trend.

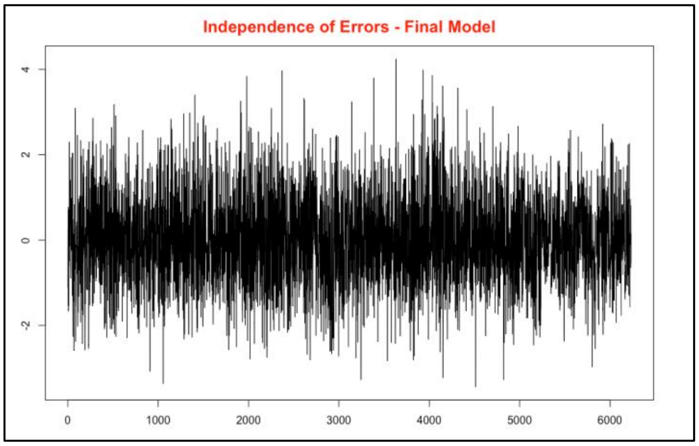


Figure 11. Independence of Errors

VI. DATA VISUALIZATIONS

After building the model in RStudio, the dataset and the model are exported in a flat file in a specific location where from QlikSense which is used as the Business Intelligence tool imports the data to produce the necessary visualizations that will offer more insights about the PpM, which is the key metric, and how it interacts with all other data.

As already mentioned, the location variables should not be missing from the model, because according to the literature are maybe the most crucial factor of the rental price fluctuations.

Through a Map Chart (Image 7), an image of the apartments per SubArea (the name of the variable that stands closely to the Municipality) shows the number of apartments in each SubArea represented by the size of the pentagon figure and the average PpM represented by the colour of the figure.

The higher number of ads that were used for this study located at Kifissia and Marousi from the Northern suburbs of Attica, at Glyfada from the Southern suburbs and at the city center of Athens.

The city center holds also the first place in terms of the higher average PpM as it can be seen from the deep red colour of the figure. Along with the city center, Vouliagmeni a prestigious northern suburb seems to be very expensive as expected. On the other hand, the lowest average PpM is found in few ads that were located at Markopoulo, an Eastern suburb of Attica or in apartments from neighbours near the city center but with very bad reputation such as Platia Vathis or Attiki.

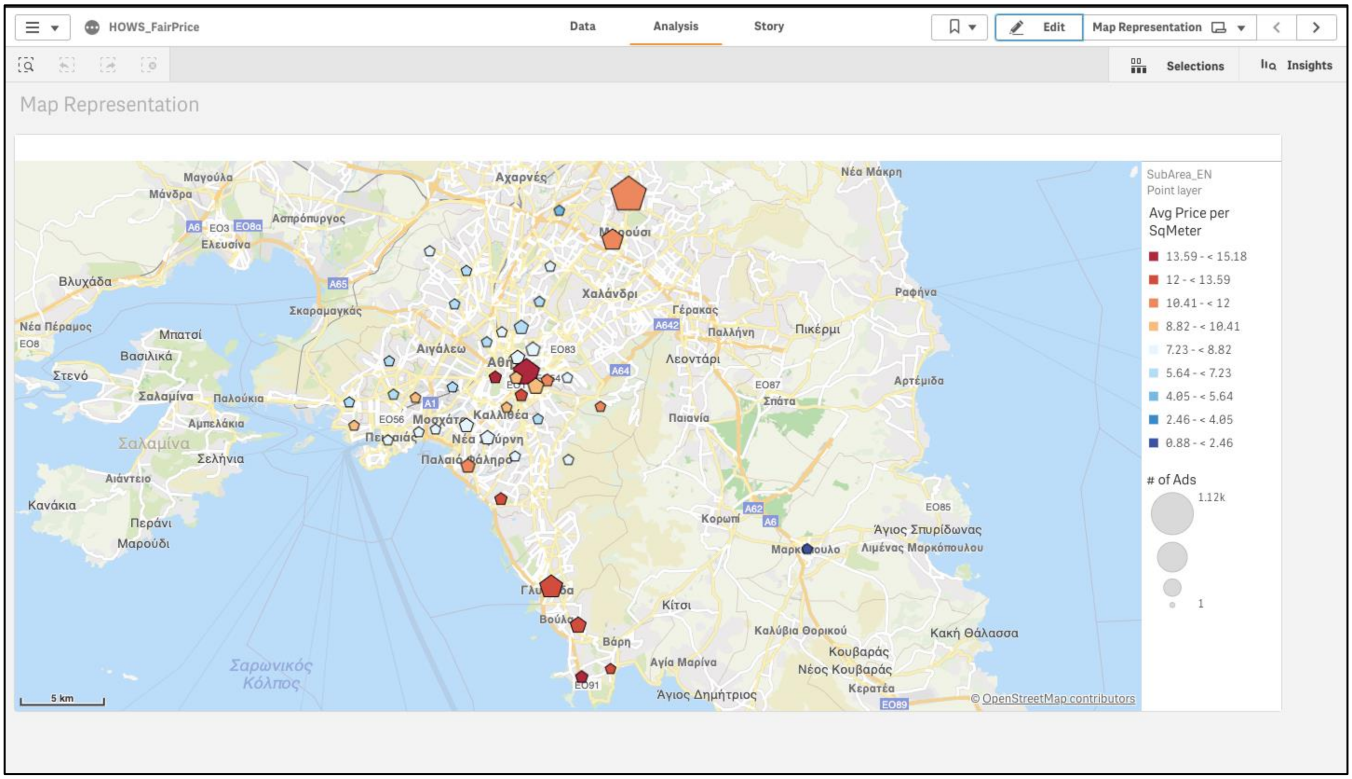
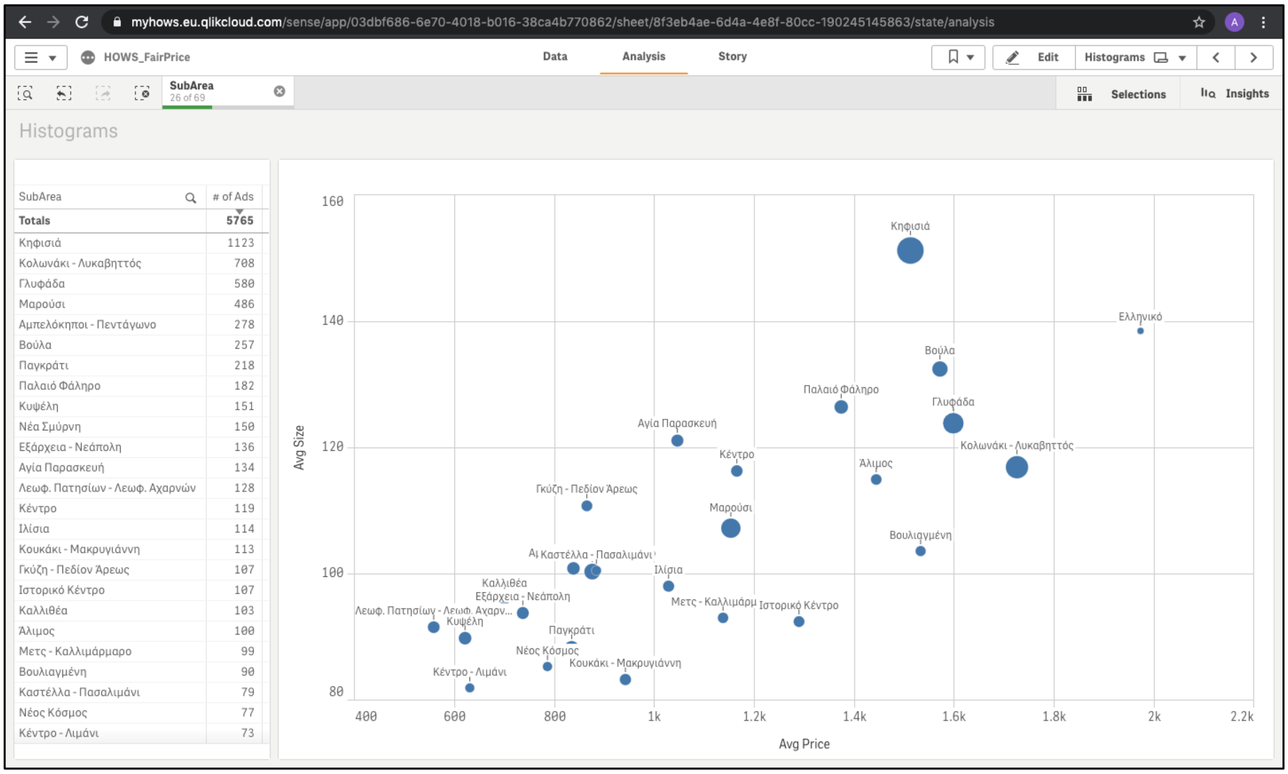


Image 7. Map Chart

A more detailed view of the average PpM per SubArea is shown in the Image 8, where an extra dimension is added, the average Size. The bubbles found in the right side of the graph have the higher average PpM while the bubbles found on the upper side have the higher average Size. The diameter of the bubble represents the number of ads in the SubArea and gets bigger as the number of ads is higher.

The filter pane on the left can be used to narrow down the SubAreas shown in the graph in order to have a more clear view, but also to find where a specific SubArea is located in the graph. Finally, an extra column shows the exact number of ads in each SubArea.

An interesting fact that was missed from the Map Chart is that the SubArea of Elliniko has the higher average PpM and this is probably due to some luxurious apartments of large size since the average size of apartments is relatively high compared to other northern suburbs such as Glyfada or Vouliagmeni.

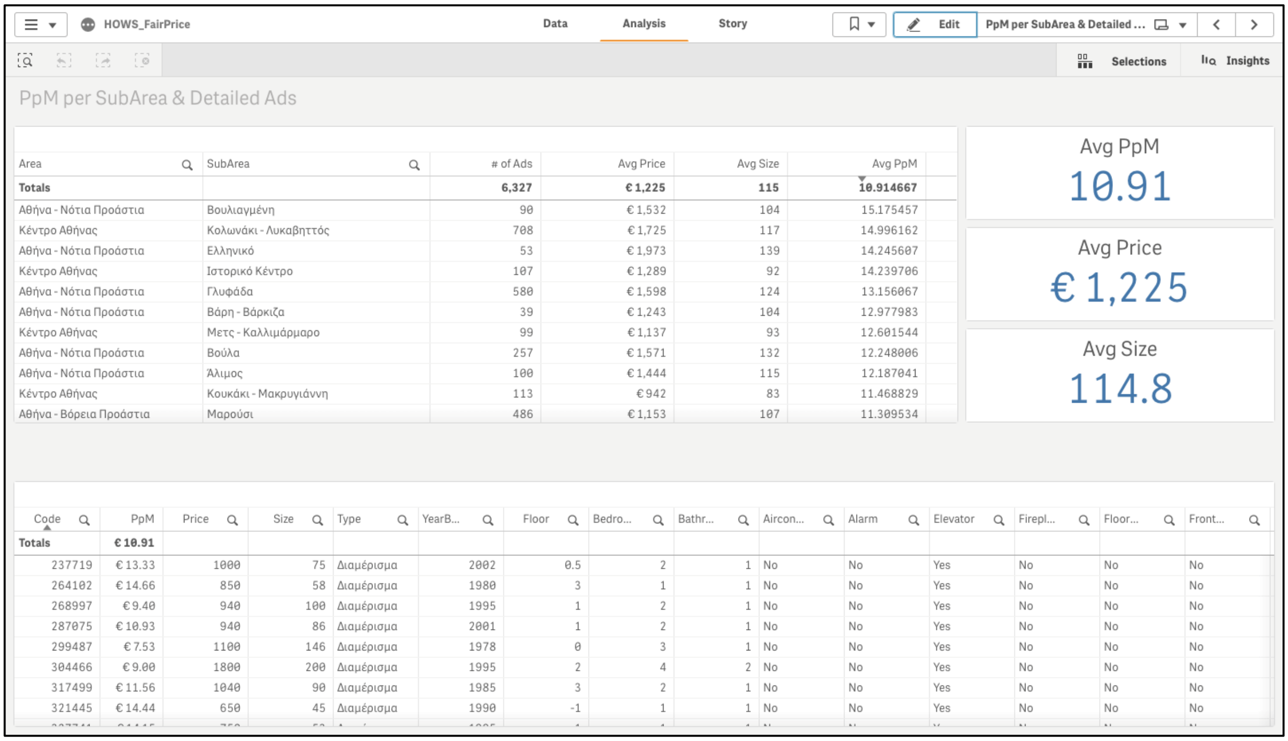


SubArea	# of Ads
Totals	**5765**
Κηφισιά	1123
Κολωνάκι - Λυκαβηττός	708
Γλυφάδα	580
Μαρούσι	486
Αμπελόκηποι - Πεντάγωνο	278
Βούλα	257
Παγκράτι	218
Παλαιό Φάληρο	182
Κυψέλη	151
Νέα Σμύρνη	150
Εξάρχεια - Νεάπολη	136
Αγία Παρασκευή	134
Λεωφ. Πατησίων - Λεωφ. Αχαρνών	128
Κέντρο	119
Ιλίσια	114
Κουκάκι - Μακρυγιάννη	113
Γκύζη - Πεδίον Άρεως	107
Ιστορικό Κέντρο	107
Καλλιθέα	103
Άλιμος	100
Μετς - Καλλιμάρμαρο	99
Βουλιαγμένη	90
Καστέλλα - Πασαλιμάνι	79
Νέος Κόσμος	77
Κέντρο - Λιμάνι	73

Image 8. Bubble Chart for Avg Price and Avg Size per SubArea

Finally, in Image 9 a straight table shows the number of Ads, average Price, average Size and average PpM in numbers, where the user can easily sort the data for any of the columns or easily search for a specific SubArea.

Additionally, there is a fully detailed table below, where all the ads along with all their characteristics are shown. The two tables are also connected, so that when filtering a specific SubArea from the upper table with the average results, the ads of this SubArea(s) only are shown in the detailed table at the bottom of the page.



PpM per SubArea & Detailed Ads

Area	SubArea	# of Ads	Avg Price	Avg Size	Avg PpM
Totals		**6,327**	**€ 1,225**	**115**	**10.914667**
Αθήνα - Νότια Προάστια	Βουλιαγμένη	90	€ 1,532	104	15.175457
Κέντρο Αθήνας	Κολωνάκι - Λυκαβηττός	708	€ 1,725	117	14.996162
Αθήνα - Νότια Προάστια	Ελληνικό	53	€ 1,973	139	14.245607
Κέντρο Αθήνας	Ιστορικό Κέντρο	107	€ 1,289	92	14.239706
Αθήνα - Νότια Προάστια	Γλυφάδα	580	€ 1,598	124	13.156067
Αθήνα - Νότια Προάστια	Βάρη - Βάρκιζα	39	€ 1,243	104	12.977983
Κέντρο Αθήνας	Μετς - Καλλιμάρμαρο	99	€ 1,137	93	12.601544
Αθήνα - Νότια Προάστια	Βούλα	257	€ 1,571	132	12.248006
Αθήνα - Νότια Προάστια	Άλιμος	100	€ 1,444	115	12.187041
Κέντρο Αθήνας	Κουκάκι - Μακρυγιάννη	113	€ 942	83	11.468829
Αθήνα - Βόρεια Προάστια	Μαρούσι	486	€ 1,153	107	11.309534

Avg PpM
10.91

Avg Price
€ 1,225

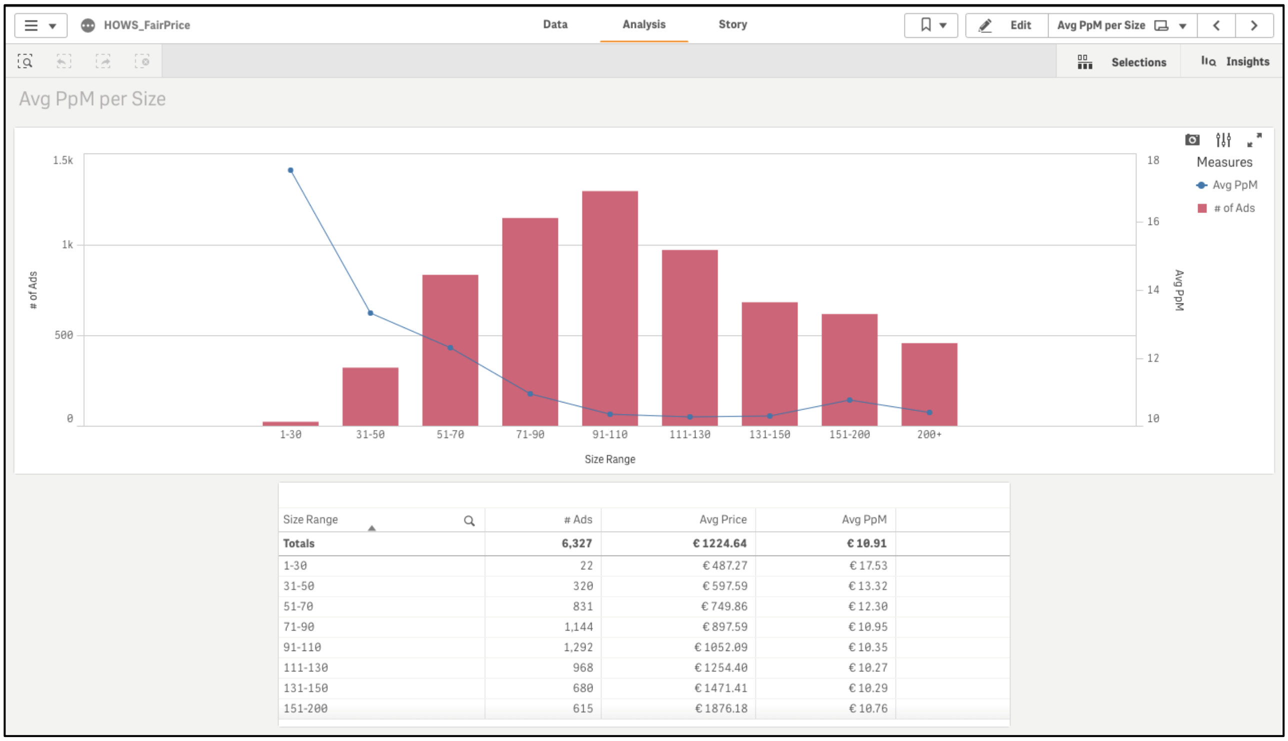
Avg Size
114.8

Code	PpM	Price	Size	Type	YearB...	Floor	Bedro...	Bathr...	Aircon...	Alarm	Elevator	Firepl...	Floor...	Front...
Totals	**€ 10.91**													
237719	€ 13.33	1000	75	Διαμέρισμα	2002	0.5	2	1	No	No	Yes	No	No	No
264102	€ 14.66	850	58	Διαμέρισμα	1980	3	1	1	No	No	Yes	No	No	No
268997	€ 9.40	940	100	Διαμέρισμα	1995	1	2	1	No	No	Yes	No	No	No
287075	€ 10.93	940	86	Διαμέρισμα	2001	1	2	1	No	No	Yes	No	No	No
299487	€ 7.53	1100	146	Διαμέρισμα	1978	0	3	1	No	No	Yes	No	No	No
304466	€ 9.00	1800	200	Διαμέρισμα	1995	2	4	2	No	No	Yes	No	No	No
317499	€ 11.56	1040	90	Διαμέρισμα	1985	3	2	1	No	No	Yes	No	No	No
321445	€ 14.44	650	45	Διαμέρισμα	1990	-1	1	1	No	No	Yes	No	No	No

Image 9. Aggregate Table per SubArea and Detailed Listing

The next three Images 10-12 offer a valuable insight about the numerical variables that are used in the model in respect to the PpM trend. The charts follow the same logic. They are combo charts that represent the Number of ads in bars which refers to the left y-axis and the average PpM in a line which refers to the right y-axis. The x-axis of the chart shows the respective numeric variable. Also, a table below the chart shows the same information in numbers.

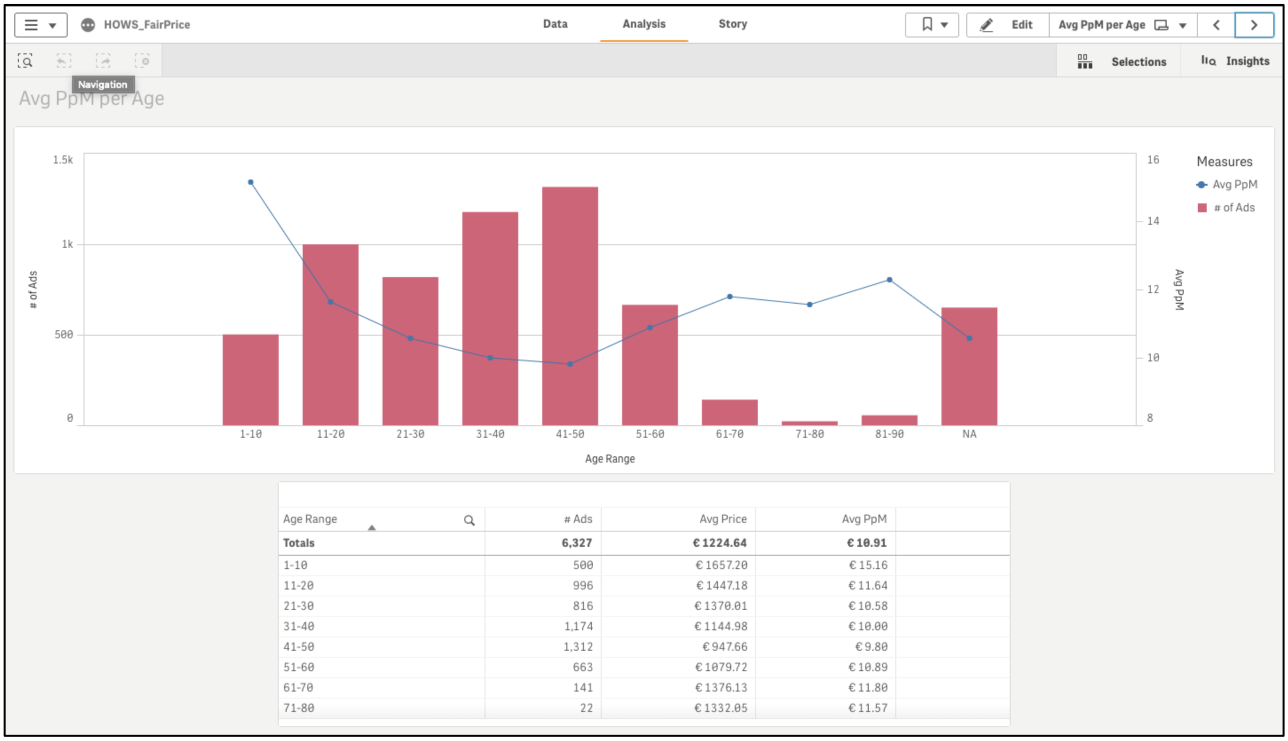
The most important numerical explanatory variable is the Size. The results of the model show that the coefficient of Size is negative and this is clearly confirmed in the trend of the average PpM as shown in Image 10. While moving to the right in x-axis, to apartments with higher Size, the average PpM drops. A noticeable difference in the number of ads can also be seen, which might consider uncertain the very high peak in the average PpM for very small apartments due to the possible effect from few outliers which could have raised this average.



Size Range	# Ads	Avg Price	Avg PpM
Totals	**6,327**	**€ 1224.64**	**€ 10.91**
1-30	22	€ 487.27	€ 17.53
31-50	320	€ 597.59	€ 13.32
51-70	831	€ 749.86	€ 12.30
71-90	1,144	€ 897.59	€ 10.95
91-110	1,292	€ 1052.09	€ 10.35
111-130	968	€ 1254.40	€ 10.27
131-150	680	€ 1471.41	€ 10.29
151-200	615	€ 1876.18	€ 10.76

Image 10. Number of Ads & Avg PpM per Size Range

A similar with the Size trend was expected to be seen in Image 11 which presents the Age Range (AgeD). The rational should be that as the apartment grows older, the average PpM should drop. This claim seems to stand for ages up to 50 years but then the average PpM starts to increase. A possible explanation for the phenomenon is that apartments over 50 years are mansions, expensive apartments of another era in other words that still have a high value or are renovated. Another possible explanation is that many old buildings can be found in the city center which, as shown previously, has the higher average PpM. It should also be mentioned that the N/A bar should not confuse because it is in the right of the diagram. It represents the apartments without the Year of Built information and seem to have an average value for the metric. This chart led to the polynomic use in the variable AgeD of the model since there is no clear linear relationship between the explanatory and the dependent variable



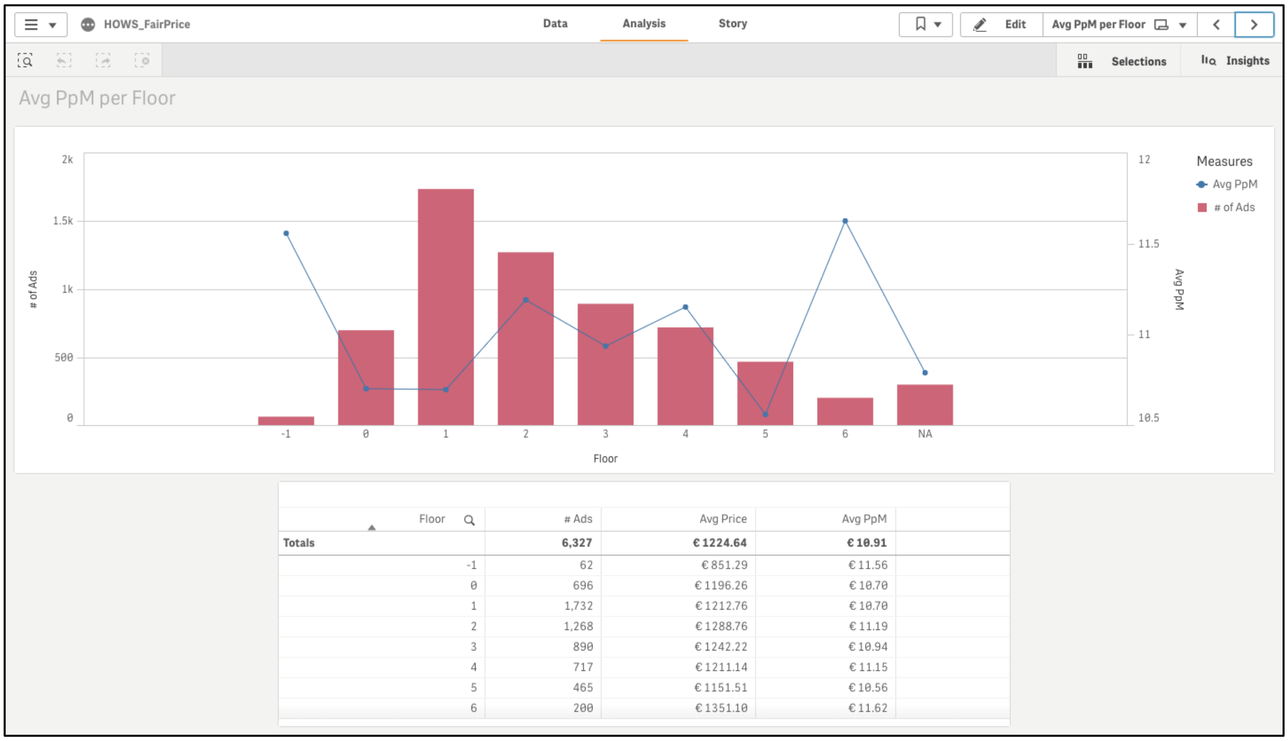
Age Range	# Ads	Avg Price	Avg PpM
Totals	**6,327**	**€ 1224.64**	**€ 10.91**
1-10	500	€ 1657.20	€ 15.16
11-20	996	€ 1447.18	€ 11.64
21-30	816	€ 1370.01	€ 10.58
31-40	1,174	€ 1144.98	€ 10.00
41-50	1,312	€ 947.66	€ 9.80
51-60	663	€ 1079.72	€ 10.89
61-70	141	€ 1376.13	€ 11.80
71-80	22	€ 1332.05	€ 11.57

Image 11. Number of Ads & Avg PpM per Age Range

Finally, Image 12 depicts the variations in the Floor variable, where there is no specific trend between the average PpM and the number of the Floor. An interesting fact is that the average PpM for apartments in the basement are surprisingly high. At this point, the table below the chart helps to notice that the average

Price of these apartments is, as expected, lower than all the others and due to the fact that these apartments are relatively smaller, the average PpM appears to be higher.

The curve of the line may also support the polynomic factors that were used to measure the contribution of the Floor variable in the model.



Floor	# Ads	Avg Price	Avg PpM
Totals	**6,327**	**€ 1224.64**	**€ 10.91**
-1	62	€ 851.29	€ 11.56
0	696	€ 1196.26	€ 10.70
1	1,732	€ 1212.76	€ 10.70
2	1,268	€ 1288.76	€ 11.19
3	890	€ 1242.22	€ 10.94
4	717	€ 1211.14	€ 11.15
5	465	€ 1151.51	€ 10.56
6	200	€ 1351.10	€ 11.62

Image 12. Number of Ads & Avg PpM per Floor

The next block of charts (Image 13) focuses on the presence of the extra features in the add that have a binary option. At this point, it should be reminded that through the ETL process, all null values were considered as missing characteristics and replaced with the value ‘NO’ (except in the case of the Elevator where the default value was ‘YES’). Most of the variables have a very small presence in the data and thus it should be inconvenient to use them in the model. Even if used, the t-test would mark them as statistically non-significant. Studying these charts and having also in mind the common sense for the key drivers of an apartment rental price, after a lot of trials where the adjR^2 value was the key indicator, the features that were selected are shown in Image 14.

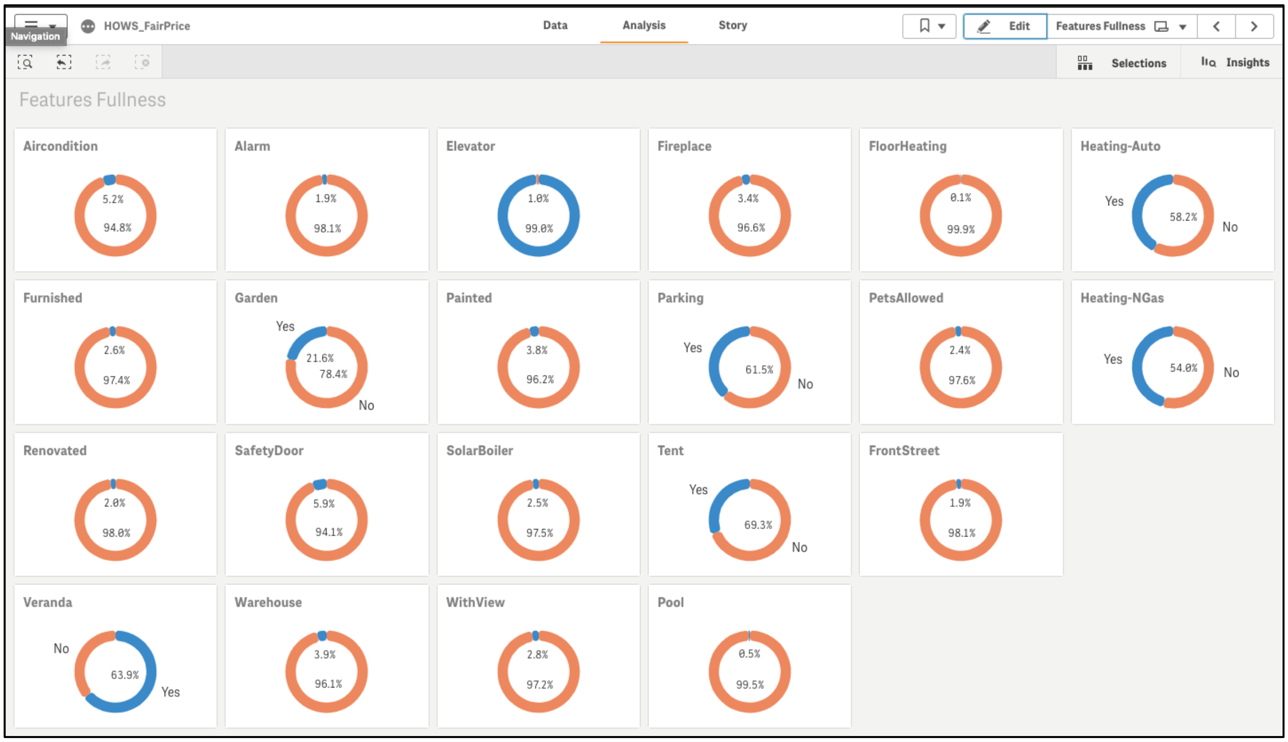
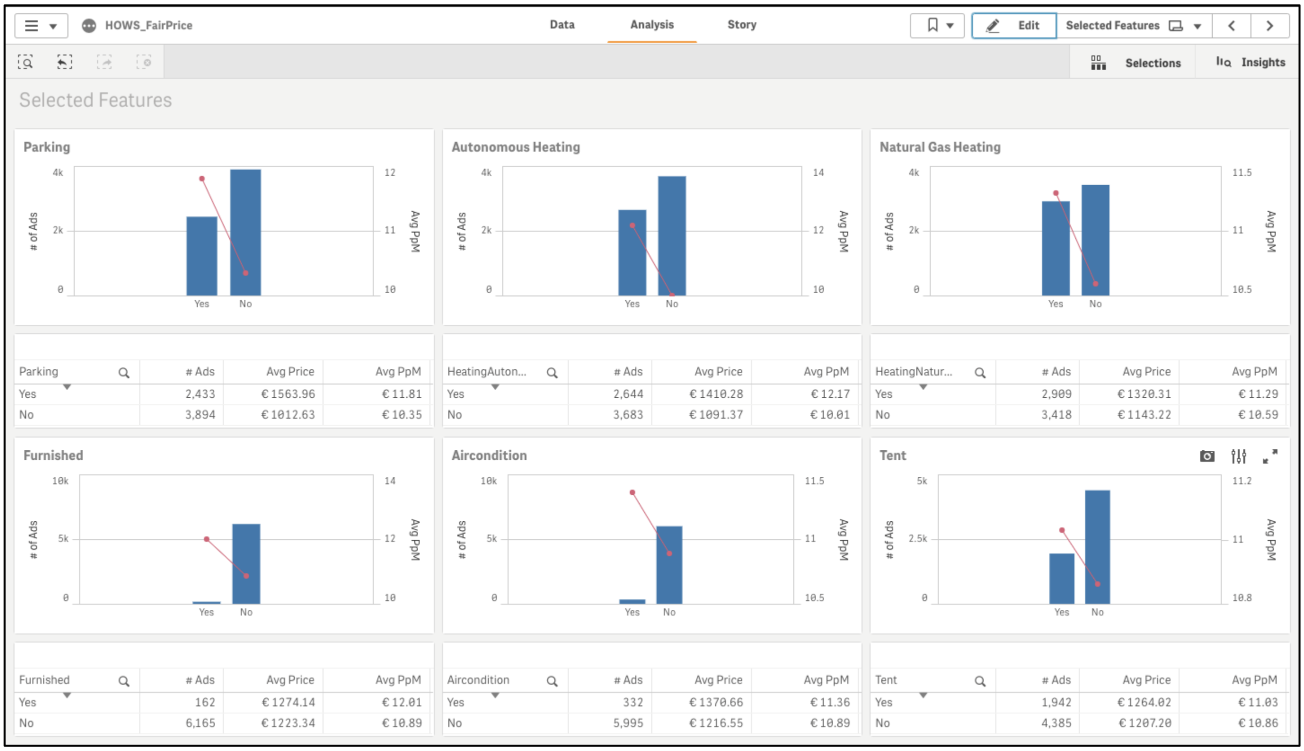


Image 13. Apartment Features Presence Rate

As in numerical variables, Image 14 presents the average PpM in respect to the presence or absence of each characteristic. The responses of the Parking, Autonomous Heating and Natural Gas Heating features were nearly balanced, but the average PpM seemed to have a clear increase where the value was 'Yes' especially in the Autonomous Heating and Parking where the difference approached the 2 €/m^2.

On the other hand, the responses of Tent, Aircondition and Furnished were not as balanced, but the contribution in the adjR^2 and their significant coefficient in the model indicated that the difference that is seen in the average PpM is influenced in a degree from these characteristics.



Parking	# Ads	Avg Price	Avg PpM
Yes	2,433	€ 1563.96	€ 11.81
No	3,894	€ 1012.63	€ 10.35

HeatingAuton...	# Ads	Avg Price	Avg PpM
Yes	2,644	€ 1410.28	€ 12.17
No	3,683	€ 1091.37	€ 10.01

HeatingNatur...	# Ads	Avg Price	Avg PpM
Yes	2,909	€ 1320.31	€ 11.29
No	3,418	€ 1143.22	€ 10.59

Furnished	# Ads	Avg Price	Avg PpM
Yes	162	€ 1274.14	€ 12.01
No	6,165	€ 1223.34	€ 10.89

Aircondition	# Ads	Avg Price	Avg PpM
Yes	332	€ 1370.66	€ 11.36
No	5,995	€ 1216.55	€ 10.89

Tent	# Ads	Avg Price	Avg PpM
Yes	1,942	€ 1264.02	€ 11.03
No	4,385	€ 1207.20	€ 10.86

Image 14. Avg PpM and Apartment Feature Presence Rate

VII. FAIR PRICE RESULTS

The scope of this study was not only to build a model that can explain the variability of the apartment rental prices in the region of Attica, but to create a tool were this model could be used to show the fair price of an ideal apartment using the explanatory variables as the filtering characteristics in an online search.

Thus, a sheet in QlikSense, which can easily be embedded in a web application (the most common way is through an i-Frame), can be used as a calculator for the fair price of a search.

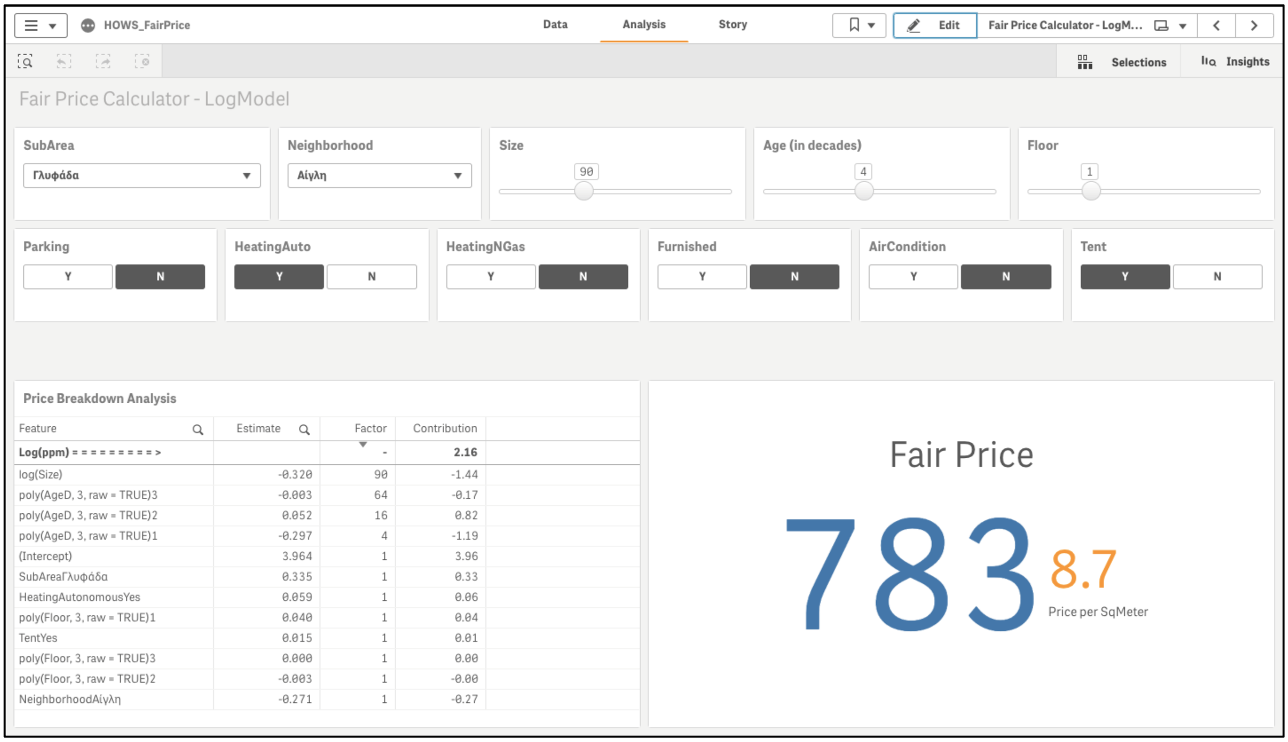
On the upper side of the screen, there are the parameters that the user should pass into the model.

There are three types of input boxes used to declare the value of the variables.

- Drop down lists are used for the SubAreas and their relative Neighborhoods.
- Sliders are used for the numerical variables Size, AgeD and Floor
- Binary buttons with the values ‘Y’ for ‘YES’ and ‘N’ for ‘No’ are used to declare the presence of the features that are used in the model

While adding the values in the variables, the table in the bottom left shows the contribution of each variable added in the log(PpM) value which is regressed in our model and the KPI box in the right shows the Fair Price which equals to the exponential of the sum of each coefficient as shown in the ‘Price Breakdown Analysis’ table multiplied by the Size value. The 2nd number (in orange) show also the Fair PpM value resulted for the specific search according to the model.

In Image 15 is displayed a user’s search where the desired apartment is located in a neighbourhood of **Glyfada** named **Egli**, has a size of **90** square meters, have been built in the **1980s** and is in the **1st** floor. The user’s preferences also include **Autonomous Heating** and **Tent**. The model’s output for the Fair Price for an apartment with these characteristics is **783€ per month**.



Fair Price Calculator - LogModel

SubArea: Γλυφάδα

Neighborhood: Αίγλη

Size: 90

Age (in decades): 4

Floor: 1

Parking: Y [✓ N]

HeatingAuto: [✓ Y] N

HeatingNGas: Y [✓ N]

Furnished: Y [✓ N]

AirCondition: Y [✓ N]

Tent: [✓ Y] N

Price Breakdown Analysis

Feature	Estimate	Factor	Contribution
Log(ppm) = = = = = = = = = >		**-**	**2.16**
log(Size)	-0.320	90	-1.44
poly(AgeD, 3, raw = TRUE)3	-0.003	64	-0.17
poly(AgeD, 3, raw = TRUE)2	0.052	16	0.82
poly(AgeD, 3, raw = TRUE)1	-0.297	4	-1.19
(Intercept)	3.964	1	3.96
SubAreaΓλυφάδα	0.335	1	0.33
HeatingAutonomousYes	0.059	1	0.06
poly(Floor, 3, raw = TRUE)1	0.040	1	0.04
TentYes	0.015	1	0.01
poly(Floor, 3, raw = TRUE)3	0.000	1	0.00
poly(Floor, 3, raw = TRUE)2	-0.003	1	-0.00
NeighborhoodΑίγλη	-0.271	1	-0.27

Fair Price

783 8.7

Price per SqMeter

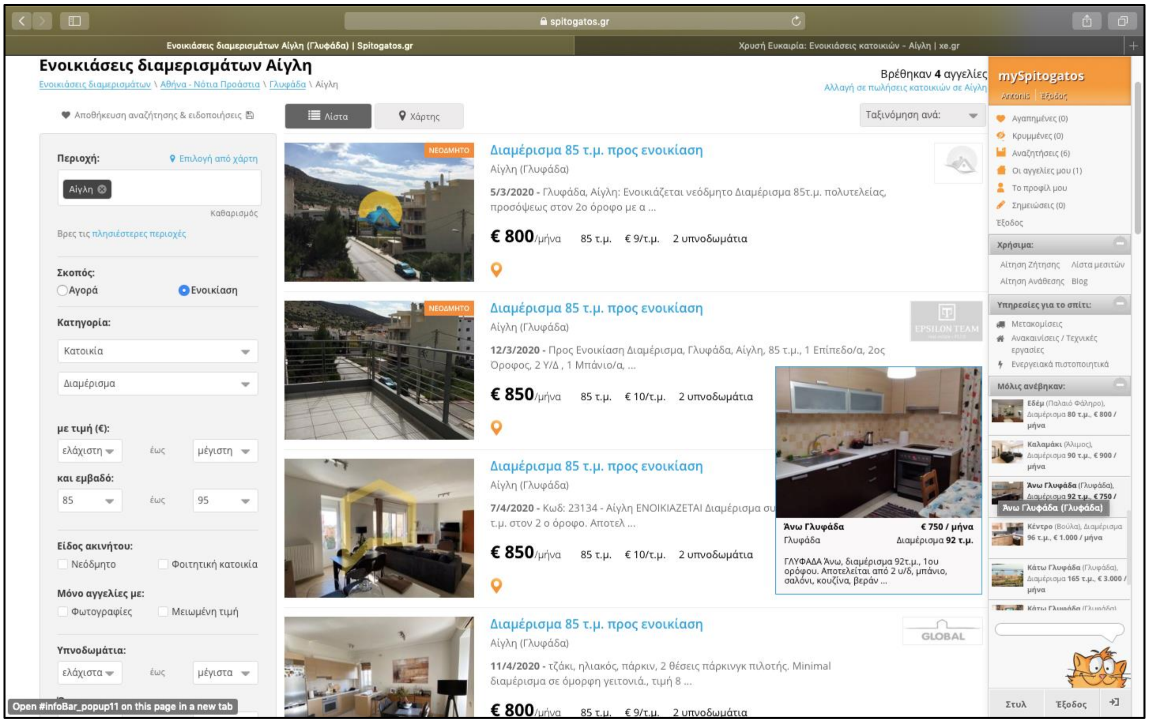
Image 15. Fair Price Calculator

The question that arises is whether the outcome of the model is accurate enough to be trusted.

There are two ways to answer this.

The first is the empirical method, where a search in the online ad platforms could indicate if the fair price suggested by the model is close to the reality, to the ads that are uploaded at the time being having the same or similar characteristics.

The following screenshots are the results of a search in the two most popular websites. Spitogatos.gr in Image 16 is also the website where the data were pulled from to build the dataset. Xe.gr in Image 17 is the market leader platform. Both sites display almost the same ads for a search in Egli for an apartment between 85 and 95m^2. Unfortunately, there is no apartment with exactly the specified characteristics. The results vary in prices between 800-900€, which would mean that the ads would be characterized overpriced from the model if the characteristics matched exactly. However, the increased prices in the sites' ads are acceptable by the model since the apartments are either much younger or in a higher floor or with parking, characteristics that increase the fair price of the model (even though the size is slightly smaller).



Ενοικιάσεις διαμερισμάτων Αίγλη

Ενοικιάσεις διαμερισμάτων \ Αθήνα - Νότια Προάστια \ Γλυφάδα \ Αίγλη

Βρέθηκαν **4** αγγελίες

Αλλαγή σε πωλήσεις κατοικιών σε Αίγλη

Διαμέρισμα 85 τ.μ. προς ενοικίαση
Αίγλη (Γλυφάδα)
5/3/2020 - Γλυφάδα, Αίγλη: Ενοικιάζεται νεόδμητο Διαμέρισμα 85τ.μ. πολυτελείας, προσόψεως στον 2ο όροφο με α ...
€ 800/μήνα 85 τ.μ. € 9/τ.μ. 2 υπνοδωμάτια

Διαμέρισμα 85 τ.μ. προς ενοικίαση
Αίγλη (Γλυφάδα)
12/3/2020 - Προς Ενοικίαση Διαμέρισμα, Γλυφάδα, Αίγλη, 85 τ.μ., 1 Επίπεδο/α, 2ος Όροφος, 2 Υ/Δ , 1 Μπάνιο/α, ...
€ 850/μήνα 85 τ.μ. € 10/τ.μ. 2 υπνοδωμάτια

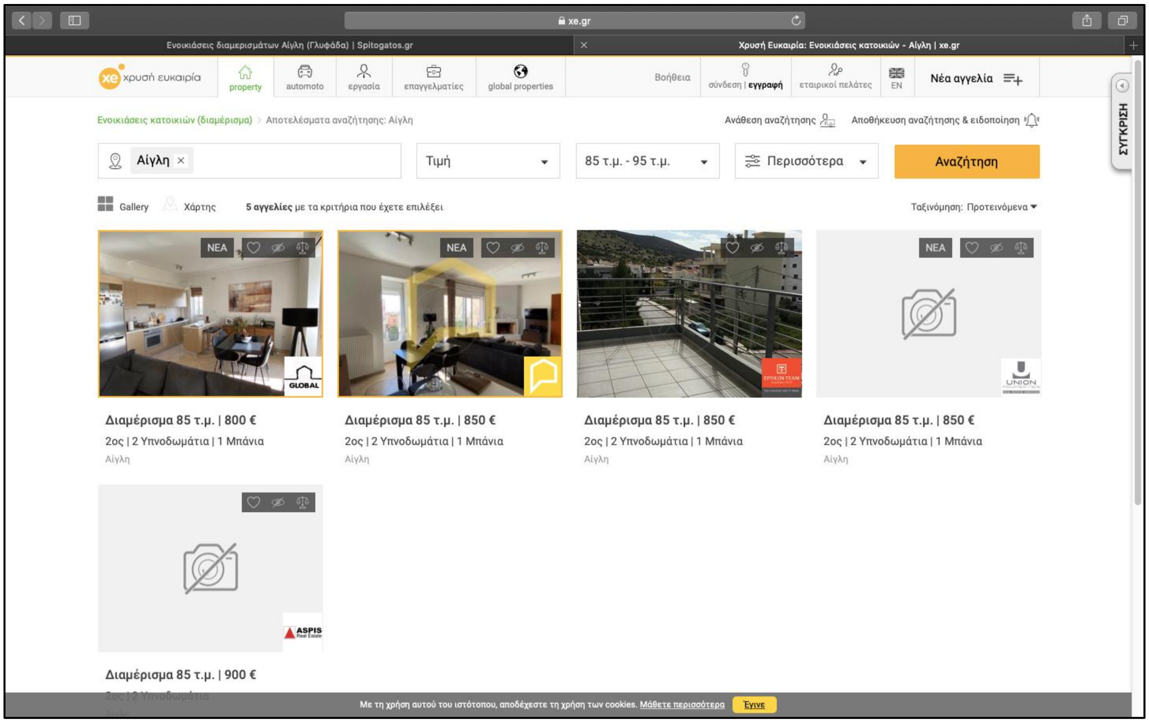
Διαμέρισμα 85 τ.μ. προς ενοικίαση
Αίγλη (Γλυφάδα)
7/4/2020 - Κωδ: 23134 - Αίγλη ΕΝΟΙΚΙΑΖΕΤΑΙ Διαμέρισμα σ
τ.μ. στον 2 ο όροφο. Αποτελ ...
€ 850/μήνα 85 τ.μ. € 10/τ.μ. 2 υπνοδωμάτια

Διαμέρισμα 85 τ.μ. προς ενοικίαση
Αίγλη (Γλυφάδα)
11/4/2020 - τζάκι, ηλιακός, πάρκιν, 2 θέσεις πάρκινγκ πιλοτής. Minimal διαμέρισμα σε όμορφη γειτονιά., τιμή 8 ...
€ 800/μήνα 85 τ.μ. € 9/τ.μ. 2 υπνοδωμάτια

Άνω Γλυφάδα **€ 750 / μήνα**
Γλυφάδα Διαμέρισμα **92 τ.μ.**
ΓΛΥΦΑΔΑ Άνω, διαμέρισμα 92τ.μ., 1ου ορόφου. Αποτελείται από 2 υ/δ, μπάνιο, σαλόνι, κουζίνα, βεράν ...

Εδέμ (Παλαιό Φάληρο), Διαμέρισμα **80 τ.μ., € 800 / μήνα**
Καλαμάκι (Άλιμος), Διαμέρισμα **90 τ.μ., € 900 / μήνα**
Άνω Γλυφάδα (Γλυφάδα), Διαμέρισμα **92 τ.μ., € 750 /**
Κέντρο (Βούλα), Διαμέρισμα **96 τ.μ., € 1.000 / μήνα**
Κάτω Γλυφάδα (Γλυφάδα), Διαμέρισμα **165 τ.μ., € 3.000 / μήνα**

Image 16. Spitogatos.gr Search Output for Fair Price



Ενοικιάσεις κατοικιών (διαμέρισμα) > Αποτελέσματα αναζήτησης: Αίγλη

5 αγγελίες με τα κριτήρια που έχετε επιλέξει

Διαμέρισμα 85 τ.μ. | 800 €
2ος | 2 Υπνοδωμάτια | 1 Μπάνια
Αίγλη

Διαμέρισμα 85 τ.μ. | 850 €
2ος | 2 Υπνοδωμάτια | 1 Μπάνια
Αίγλη

Διαμέρισμα 85 τ.μ. | 850 €
2ος | 2 Υπνοδωμάτια | 1 Μπάνια
Αίγλη

Διαμέρισμα 85 τ.μ. | 850 €
2ος | 2 Υπνοδωμάτια | 1 Μπάνια
Αίγλη

Διαμέρισμα 85 τ.μ. | 900 €
2ος | 2 Υπνοδωμάτια

Με τη χρήση αυτού του ιστότοπου, αποδέχεστε τη χρήση των cookies. Μάθετε περισσότερα

Image 17. Xe.gr Search Output for Fair Price

The second method to measure the predictive performance of the model, is the statistical one.

In order to perform this method, a new dataset was created with ads uploaded on Spitogatos.gr website in the first three weeks of March 2020. This dataset was used as 'test' in order the model which was built on another pool of ads to predict their price and then compared with the actual prices of the test dataset.

The results of this method can be seen in Figure 12.

The two axis of the diagram stand for the actual versus the predicted price. In a perfect model all the dots should lie on the red 45^{o} line. For apartments with actual prices up to approximately 1500€ the fair price of the model is pretty close and the dots gather near the red line. As the actual price increases, there more dots with a larger distance from the red line, which indicates that the model does not perform that well for the more expensive apartments.

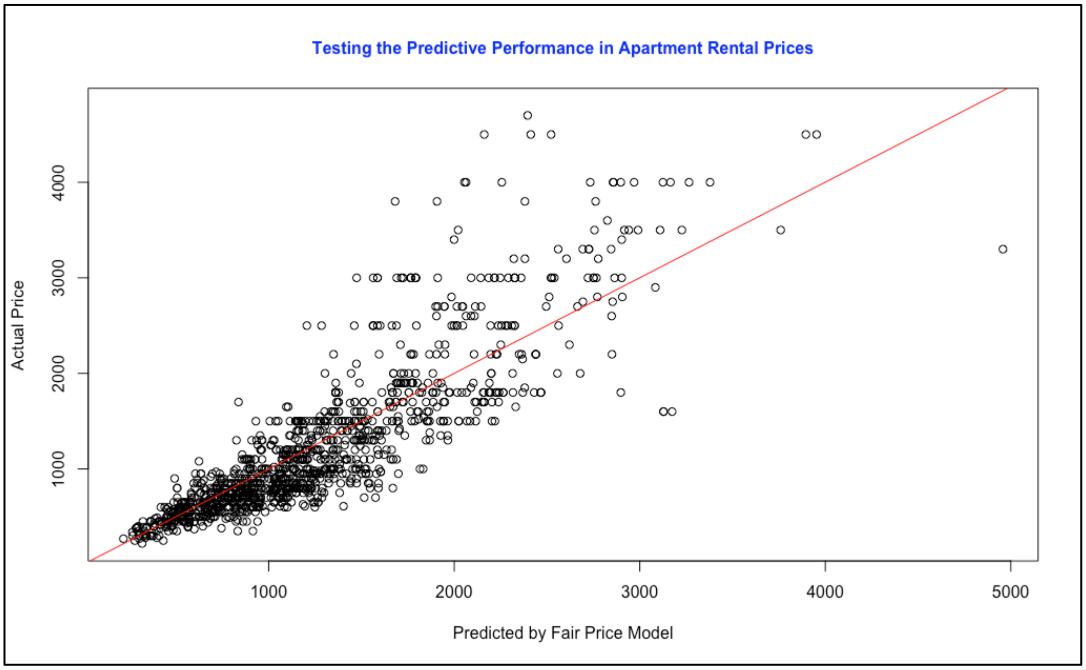


Figure 12. Model's Prediction Accuracy

However, the most appropriate method for testing the accuracy would be the first since in a real-time application the model would be built frequently upon the same adds that will be used in the users' search (with the hypothesis that more or less the same ads appear in both websites).

The most suitable role that the model would play is not the prediction of a price for a new point but the set of an exact point (which was used in the model development) in the price distribution in order to compare it with all other apartments that build the fair price and characterize it as over or underpriced.

Furthermore, a KPI will be established and monitored every month in order to measure the effectiveness of the model. It will measure the deviation of the predicted fair price to the actual values of the new ads that will be uploaded in the platform. A specific numerical target for the KPI will be set after the first months of the actual deployment.

VIII. BUSINESS APPLICATION – INTEGRATION WITH HOWS.GR

The scope of this study is to create a model that will be used as an extra feature in a start-up business idea that would be an online real estate agent for apartment rentals.

The model will run on every uploaded add as an extra feature in order to give a hint to the seeker whether the apartment is over or under-priced. There would also be a page where the i-frame of the QlikSense app will give the ability to the users to calculate the fair price of the ideal apartment applying their preferences in the characteristics. This page would be available for the apartment owners as well, to give them an idea when pricing their property in the ad uploading process.

The whole concept of the start-up would be based on transparency for both sides of the rental agreement, so the fair price calculator would be a feature totally aligned with the idea.

As for the technical part, the jobs that already described in this study will be more automated in order to handle the size of data and the daily refresh needed.

The data crawler and the data scraper will be adjusted to run on daily basis fetching all available ads at the specific date. For better convenience, a different process will be created for every different SubArea. This will prevent for failing the whole process that will run overnight.

The ETL phase will be changed in order to keep to handle the larger input of data since the historic ads will also be kept.

Since the size of the data will be increasing as the time passes, the final table will be replaced from a star schema data warehouse.

The fact table of the Schema will have almost the same design as the final table used in this study.

A key difference will be that all binary characteristics will be converted to dimensions and appear in the fact table as the foreign keys connected to these dimensions.

The numerical variable will remain as is and used as the metrics of the fact table.

An extra dimension that will show whether an ad is active or not at the most recent update will be added and will also be used to filter the data that will be used in the model as described earlier.

Two date keys will also be added to the fact table which will point the relevant date dimension that will also be created. The first date will refer to the opening date of the ad and the second to the closing date. This will offer the valuable insights of the time since uploaded and closed for every ad from which the aggregate turnovers in days for any grouping can be calculated.

Finally, an extra autogenerated fact_id in the fact table will be the primary key of the table. The code of the ad should also act as a primary key but it would be wiser to maintain a primary key that does not come from an external source. The max fact_id will equal at any time to the total number of ads, both active and inactive that are held in the data warehouse.

This study described the statistical analysis and the modelling through the RStudio console.

The plan is to work in an updated version of the SQL server (2017) that supports the R Language scripts.

This way, the model will be built upon an SSIS task and run automatically integrated with the data that will be kept in the SQL server.

The statistical methods used to be the model will be re-examined and the model will be optimized periodically. Probably, an extra explanatory variable that will measure the period effect will be used (a common sense logic indicates that rental prices tend to be a little higher in September especially in areas near universities).

At the end, the QlikSense will also be connected directly to the SQL data warehouse and will be automatically updated when the SQL processes are finished every night.

Extra analytics features concerning the historical trends of the number of ads or the prices could also be created and integrated with the platform.

REFERENCES

1. Law 1249/1982 – FEK 43/A/5-4-1982, Available online: https://www.e-nomothesia.gr/kat-oikonomia/n-1249-1982.html
2. Ανδρέου Ε., (2016), Ανάπτυξη Συστήματος Μαζικών Εκτιμήσεων Αξιών οικιστικών ακινήτων με χρήση τεχνολογιών G.I.S.: Εφαρμογή στο Δήμο Χαλανδρίου Αττικής», Διπλωματική εργασία, Τομέας Τοπογραφίας – Περιοχή Κτηματολογίου, Σχολή Αγρονόμων και Τοπογράφων Μηχανικών, Εθνικό Μετσόβιο Πολυτεχνείο
3. Βασιλός Π.; Διακοδημήτρης Φ., (2014), Εκτιμήσεις Ακινήτων, Διπλωματική εργασία, Σχολή Πολιτικών Μηχανικών του Εθνικού Μετσόβιου Πολυτεχνείο, Τομέας Δομοστατικής, Αθήνα
4. Παπακυριακου Μ., (2017), Συγκριτική μελέτη αντικειμενικών και εμπορικών αξιών των ακινήτων σε αστικά κέντρα της Ελλάδας κατά την περίοδο της δημοσιονομικής κρίσης, Available online: https://ikee.lib.auth.gr/record/292474/files/PAPAKYRIAKOY_DE532.pdf
5. Παπαευθυμίου Ι., (2013), Εκτίμηση αξίας Ακινήτων. Αξιοποίηση τοπικών μοντέλων παλινδρόμησης, Available online: http://gisc.gr/sac/docs/proceedings_sac1/2_Papaefthymiou_SAC1.pdf.
6. Ζεντέλης Π., (2001), Real Estate: Αξία - Εκτιμήσεις - Αναπ́ τυξη - Επενδύσεις - Διαχείριση, Εκδόσεις Παπασωτηρίου, Αθήνα
7. Καρανικόλας Ν., (2010), Η εκτίμηση των ακινήτων, Εκδοσ́εις Δίσιγμα, Θεσσαλονίκη
8. Σιωμόπουλος Ι. Κ., (2000), Τα Συστήματα Προσδιορισμού της Αξίας των Ακινήτων, Ipirotiki Software & Publications
9. Smith L.B.; Rosen K.T.; Fallis G. (1988), Recent developments in economic models of housing markets.
10. Sirmans G.S.; Macpherson D.A.; Zietz, (2005), The composition of hedonic pricing models. Available online: http://aresjournals.org/doi/abs/10.5555/reli.13.1.j03673877172w0w2

11. Mora-Garcia R. T.; Cespedes-Lopez M. F.; Perez-Sanchez V. R.; Marti P; Perez-Sanchez J. C., (2019), Determinants of the Price of Housing in the Province of Alicante (Spain): Analysis Using Quantile Regression
12. Bohl M.; Michels W.; Oelgemöller J., (2012), Determinanten von Wohnimmobilienpreisen, Das Beispiel der Stadt Münster. Jahrbuch für Regionalwissenschaft
13. Kaya A.; Atan M., (2017) Determination of the factors that affect house prices in Turkey by using Hedonic Pricing
14. Yayar R.; Demir D., (2014), Hedonic estimation of housing market prices in Turkey
15. Keskin B.; Watkins C., (2017), Defining spatial housing submarkets: Exploring the case for expert delineated boundaries
16. McGreal W.S.; Taltavull de la Paz, (2013), P. Implicit house prices: Variation over time and space in Spain
17. Chasco-Yrigoyen C.; Sánchez-Reyes B, (2012), Externalidades ambientales y precio de la vivienda en Madrid: Un análisis con regresión cuantílica espacial. Revista Galega de Economía, Available online: http://www.usc.es/econo/RGE/Vol21_2/castelan/bt4c.pdf
18. Hajime S.; Daiki S., A comparison of apartment rent price prediction using a large dataset: Kriging versus DNN, Available online: https://arxiv.org/pdf/1906.11099.pdf
19. Bourassa S.; Hoesli M.; Scognamiglio D., (2010), "International Articles: Housing Finance, Prices, and Tenure in Switzerland", Journal of Real Estate Literature, Vol. 18; No.2; p. 261-282
20. Fletcher M.; Gallimore P.; Mangan, J., (2000), Heteroscedasticity in hedonic house price models, Journal of Property Research, p. 93-108; Published online: 07 Feb 2011
21. Gustaffson A.; Wogenius S., Modelling Apartment Prices with the Multiple Linear Regression Model, Available online: http://www.diva-portal.org/smash/get/diva2:725045/FULLTEXT01.pdf
22. Yuanhang L.; Xuanyu Z.; Yulian Z., (2019), Predicting Airbnb Listing Price Across Different Cities
23. Hansen N. R., (2012), Regression with R, Available online: http://web.math.ku.dk/~richard/courses/regression2013/handoutWeek1.pdf
24. Gosiewska A.; Biecek P., (2018), An R Package for Model-Agnostic Visual Validation and Diagnostic, Available online:

https://www.researchgate.net/publication/327835619_auditor_an_R_Package_for_Model-Agnostic_Visual_Validation_and_Diagnostic

25. Ahuja M.S.; Bal J.S., (2014), “Web Crawler: Extracting the Web Data”, International Journal of Computer Trends and Technology (IJCTT), p. 132-137, Published: 7/2014
26. Zhao B., (2018), Web Scraping

APPENDIX

SQL Query – Data Transformation

```
SELECT
        [Code]
        ,[Code2]
        ,[Area1] as Area
        ,[Area2] as SubArea
        ,case
                when (F0_N = 'Περιοχή' and F0_V != '') then substring(F0_V,1,charindex(',',F0_V)-1)
                when (F1_N = 'Περιοχή' and F1_V != '') then substring(F1_V,1,charindex(',',F1_V)-1)
                when (F2_N = 'Περιοχή' and F2_V != '') then substring(F2_V,1,charindex(',',F2_V)-1)
                when (F3_N = 'Περιοχή' and F3_V != '') then substring(F3_V,1,charindex(',',F3_V)-1)
                when (F4_N = 'Περιοχή' and F4_V != '') then substring(F4_V,1,charindex(',',F4_V)-1)
                when (F5_N = 'Περιοχή' and F5_V != '') then substring(F5_V,1,charindex(',',F5_V)-1)
                when (F6_N = 'Περιοχή' and F6_V != '') then substring(F6_V,1,charindex(',',F6_V)-1)
                when (F7_N = 'Περιοχή' and F7_V != '') then substring(F7_V,1,charindex(',',F7_V)-1)
                when (F8_N = 'Περιοχή' and F8_V != '') then substring(F8_V,1,charindex(',',F8_V)-1)
                when (F9_N = 'Περιοχή' and F9_V != '') then substring(F9_V,1,charindex(',',F9_V)-1)
                when (F10_N = 'Περιοχή' and F10_V != '') then substring(F10_V,1,charindex(',',F10_V)-1)
                when (F11_N = 'Περιοχή' and F11_V != '') then substring(F11_V,1,charindex(',',F11_V)-1)
                when (F12_N = 'Περιοχή' and F12_V != '') then substring(F12_V,1,charindex(',',F12_V)-1)
                when (F13_N = 'Περιοχή' and F13_V != '') then substring(F13_V,1,charindex(',',F13_V)-1)
                when (F14_N = 'Περιοχή' and F14_V != '') then substring(F14_V,1,charindex(',',F14_V)-1)
                when (F15_N = 'Περιοχή' and F15_V != '') then substring(F15_V,1,charindex(',',F15_V)-1)
        end as Neighborhood
        ,case
                when Price != '' then replace(substring([Price],3,charindex('/', [Price])-3),'.','')
        end as Price
        ,case
                when (F0_N = 'Εμβαδό' and F0_V != '') then substring(F0_V,1,charindex(' ',F0_V)-1)
                when (F1_N = 'Εμβαδό' and F1_V != '') then substring(F1_V,1,charindex(' ',F1_V)-1)
                when (F2_N = 'Εμβαδό' and F2_V != '') then substring(F2_V,1,charindex(' ',F2_V)-1)
                when (F3_N = 'Εμβαδό' and F3_V != '') then substring(F3_V,1,charindex(' ',F3_V)-1)
                when (F4_N = 'Εμβαδό' and F4_V != '') then substring(F4_V,1,charindex(' ',F4_V)-1)
                when (F5_N = 'Εμβαδό' and F5_V != '') then substring(F5_V,1,charindex(' ',F5_V)-1)
                when (F6_N = 'Εμβαδό' and F6_V != '') then substring(F6_V,1,charindex(' ',F6_V)-1)
                when (F7_N = 'Εμβαδό' and F7_V != '') then substring(F7_V,1,charindex(' ',F7_V)-1)
                when (F8_N = 'Εμβαδό' and F8_V != '') then substring(F8_V,1,charindex(' ',F8_V)-1)
                when (F9_N = 'Εμβαδό' and F9_V != '') then substring(F9_V,1,charindex(' ',F9_V)-1)
                when (F10_N = 'Εμβαδό' and F10_V != '') then substring(F10_V,1,charindex(' ',F10_V)-1)
                when (F11_N = 'Εμβαδό' and F11_V != '') then substring(F11_V,1,charindex(' ',F11_V)-1)
                when (F12_N = 'Εμβαδό' and F12_V != '') then substring(F12_V,1,charindex(' ',F12_V)-1)
                when (F13_N = 'Εμβαδό' and F13_V != '') then substring(F13_V,1,charindex(' ',F13_V)-1)
                when (F14_N = 'Εμβαδό' and F14_V != '') then substring(F14_V,1,charindex(' ',F14_V)-1)
                when (F15_N = 'Εμβαδό' and F15_V != '') then substring(F15_V,1,charindex(' ',F15_V)-1)
        end as Size
        ,case
                when F0_N = 'Τύπος' then F0_V
                when F1_N = 'Τύπος' then F1_V
                when F2_N = 'Τύπος' then F2_V
                when F3_N = 'Τύπος' then F3_V
                when F4_N = 'Τύπος' then F4_V
                when F5_N = 'Τύπος' then F5_V
                when F6_N = 'Τύπος' then F6_V
                when F7_N = 'Τύπος' then F7_V
                when F8_N = 'Τύπος' then F8_V
                when F9_N = 'Τύπος' then F9_V
                when F10_N = 'Τύπος' then F10_V
                when F11_N = 'Τύπος' then F11_V
                when F12_N = 'Τύπος' then F12_V
                when F13_N = 'Τύπος' then F13_V
                when F14_N = 'Τύπος' then F14_V
                when F15_N = 'Τύπος' then F15_V
        end as Type
        ,case
                when F0_N = 'Σύστημα Θέρμανσης' then F0_V
                when F1_N = 'Σύστημα Θέρμανσης' then F1_V
                when F2_N = 'Σύστημα Θέρμανσης' then F2_V
                when F3_N = 'Σύστημα Θέρμανσης' then F3_V
                when F4_N = 'Σύστημα Θέρμανσης' then F4_V
                when F5_N = 'Σύστημα Θέρμανσης' then F5_V
                when F6_N = 'Σύστημα Θέρμανσης' then F6_V
                when F7_N = 'Σύστημα Θέρμανσης' then F7_V
                when F8_N = 'Σύστημα Θέρμανσης' then F8_V
                when F9_N = 'Σύστημα Θέρμανσης' then F9_V
                when F10_N = 'Σύστημα Θέρμανσης' then F10_V
                when F11_N = 'Σύστημα Θέρμανσης' then F11_V
                when F12_N = 'Σύστημα Θέρμανσης' then F12_V
                when F13_N = 'Σύστημα Θέρμανσης' then F13_V
                when F14_N = 'Σύστημα Θέρμανσης' then F14_V
```

Query 1. Data Transformation cont'd

```
        when F15_N = 'Σύστημα Θέρμανσης' then F15_V
end as Heating
,case
        when F0_N = 'Υπνοδωμάτια' then F0_V
        when F1_N = 'Υπνοδωμάτια' then F1_V
        when F2_N = 'Υπνοδωμάτια' then F2_V
        when F3_N = 'Υπνοδωμάτια' then F3_V
        when F4_N = 'Υπνοδωμάτια' then F4_V
        when F5_N = 'Υπνοδωμάτια' then F5_V
        when F6_N = 'Υπνοδωμάτια' then F6_V
        when F7_N = 'Υπνοδωμάτια' then F7_V
        when F8_N = 'Υπνοδωμάτια' then F8_V
        when F9_N = 'Υπνοδωμάτια' then F9_V
        when F10_N = 'Υπνοδωμάτια' then F10_V
        when F11_N = 'Υπνοδωμάτια' then F11_V
        when F12_N = 'Υπνοδωμάτια' then F12_V
        when F13_N = 'Υπνοδωμάτια' then F13_V
        when F14_N = 'Υπνοδωμάτια' then F14_V
        when F15_N = 'Υπνοδωμάτια' then F15_V
end as Bedrooms
,case
        when F0_N = 'Μπάνια' then F0_V
        when F1_N = 'Μπάνια' then F1_V
        when F2_N = 'Μπάνια' then F2_V
        when F3_N = 'Μπάνια' then F3_V
        when F4_N = 'Μπάνια' then F4_V
        when F5_N = 'Μπάνια' then F5_V
        when F6_N = 'Μπάνια' then F6_V
        when F7_N = 'Μπάνια' then F7_V
        when F8_N = 'Μπάνια' then F8_V
        when F9_N = 'Μπάνια' then F9_V
        when F10_N = 'Μπάνια' then F10_V
        when F11_N = 'Μπάνια' then F11_V
        when F12_N = 'Μπάνια' then F12_V
        when F13_N = 'Μπάνια' then F13_V
        when F14_N = 'Μπάνια' then F14_V
        when F15_N = 'Μπάνια' then F15_V
end as Bathrooms
,case
        when F0_N = 'Όροφος' then F0_V
        when F1_N = 'Όροφος' then F1_V
        when F2_N = 'Όροφος' then F2_V
        when F3_N = 'Όροφος' then F3_V
        when F4_N = 'Όροφος' then F4_V
        when F5_N = 'Όροφος' then F5_V
        when F6_N = 'Όροφος' then F6_V
        when F7_N = 'Όροφος' then F7_V
        when F8_N = 'Όροφος' then F8_V
        when F9_N = 'Όροφος' then F9_V
        when F10_N = 'Όροφος' then F10_V
        when F11_N = 'Όροφος' then F11_V
        when F12_N = 'Όροφος' then F12_V
        when F13_N = 'Όροφος' then F13_V
        when F14_N = 'Όροφος' then F14_V
        when F15_N = 'Όροφος' then F15_V
end as Floor
,case
        when F0_N = 'Θέση στάθμευσης' and F0_V = 'Ναι' then 'Yes'
        when F1_N = 'Θέση στάθμευσης' and F1_V = 'Ναι' then 'Yes'
        when F2_N = 'Θέση στάθμευσης' and F2_V = 'Ναι' then 'Yes'
        when F3_N = 'Θέση στάθμευσης' and F3_V = 'Ναι' then 'Yes'
        when F4_N = 'Θέση στάθμευσης' and F4_V = 'Ναι' then 'Yes'
        when F5_N = 'Θέση στάθμευσης' and F5_V = 'Ναι' then 'Yes'
        when F6_N = 'Θέση στάθμευσης' and F6_V = 'Ναι' then 'Yes'
        when F7_N = 'Θέση στάθμευσης' and F7_V = 'Ναι' then 'Yes'
        when F8_N = 'Θέση στάθμευσης' and F8_V = 'Ναι' then 'Yes'
        when F9_N = 'Θέση στάθμευσης' and F9_V = 'Ναι' then 'Yes'
        when F10_N = 'Θέση στάθμευσης' and F10_V = 'Ναι' then 'Yes'
        when F11_N = 'Θέση στάθμευσης' and F11_V = 'Ναι' then 'Yes'
        when F12_N = 'Θέση στάθμευσης' and F12_V = 'Ναι' then 'Yes'
        when F13_N = 'Θέση στάθμευσης' and F13_V = 'Ναι' then 'Yes'
        when F14_N = 'Θέση στάθμευσης' and F14_V = 'Ναι' then 'Yes'
        when F15_N = 'Θέση στάθμευσης' and F15_V = 'Ναι' then 'Yes'
        else 'No'
end as Parking
,case
        when (F0_N = 'Έτος κατασκευής' and F0_N != '') then F0_V
        when (F1_N = 'Έτος κατασκευής' and F1_N != '') then F1_V
        when (F2_N = 'Έτος κατασκευής' and F2_N != '') then F2_V
        when (F3_N = 'Έτος κατασκευής' and F3_N != '') then F3_V
        when (F4_N = 'Έτος κατασκευής' and F4_N != '') then F4_V
        when (F5_N = 'Έτος κατασκευής' and F5_N != '') then F5_V
        when (F6_N = 'Έτος κατασκευής' and F6_N != '') then F6_V
        when (F7_N = 'Έτος κατασκευής' and F7_N != '') then F7_V
        when (F8_N = 'Έτος κατασκευής' and F8_N != '') then F8_V
        when (F9_N = 'Έτος κατασκευής' and F9_N != '') then F9_V
        when (F10_N = 'Έτος κατασκευής' and F10_N != '') then F10_V
        when (F11_N = 'Έτος κατασκευής' and F11_N != '') then F11_V
```

Query 1. Data Transformation cont'd

```
        when (F12_N = 'Έτος κατασκευής' and F12_N != '') then F12_V
        when (F13_N = 'Έτος κατασκευής' and F13_N != '') then F13_V
        when (F14_N = 'Έτος κατασκευής' and F14_N != '') then F14_V
        when (F15_N = 'Έτος κατασκευής' and F15_N != '') then F15_V
end as YearBuilt
,case
        when F0_V like '%Ανακαινισμένο: Ναι%' then 'Yes'
        when F1_V like '%Ανακαινισμένο: Ναι%' then 'Yes'
        when F2_V like '%Ανακαινισμένο: Ναι%' then 'Yes'
        when F3_V like '%Ανακαινισμένο: Ναι%' then 'Yes'
        when F4_V like '%Ανακαινισμένο: Ναι%' then 'Yes'
        when F5_V like '%Ανακαινισμένο: Ναι%' then 'Yes'
        when F6_V like '%Ανακαινισμένο: Ναι%' then 'Yes'
        when F7_V like '%Ανακαινισμένο: Ναι%' then 'Yes'
        when F8_V like '%Ανακαινισμένο: Ναι%' then 'Yes'
        when F9_V like '%Ανακαινισμένο: Ναι%' then 'Yes'
        when F10_V like '%Ανακαινισμένο: Ναι%' then 'Yes'
        when F11_V like '%Ανακαινισμένο: Ναι%' then 'Yes'
        when F12_V like '%Ανακαινισμένο: Ναι%' then 'Yes'
        when F13_V like '%Ανακαινισμένο: Ναι%' then 'Yes'
        when F14_V like '%Ανακαινισμένο: Ναι%' then 'Yes'
        when F15_V like '%Ανακαινισμένο: Ναι%' then 'Yes'
        else 'No'
end as Renovated
,case
        when F0_V like '%Βαμμένο: Ναι%' then 'Yes'
        when F1_V like '%Βαμμένο: Ναι%' then 'Yes'
        when F2_V like '%Βαμμένο: Ναι%' then 'Yes'
        when F3_V like '%Βαμμένο: Ναι%' then 'Yes'
        when F4_V like '%Βαμμένο: Ναι%' then 'Yes'
        when F5_V like '%Βαμμένο: Ναι%' then 'Yes'
        when F6_V like '%Βαμμένο: Ναι%' then 'Yes'
        when F7_V like '%Βαμμένο: Ναι%' then 'Yes'
        when F8_V like '%Βαμμένο: Ναι%' then 'Yes'
        when F9_V like '%Βαμμένο: Ναι%' then 'Yes'
        when F10_V like '%Βαμμένο: Ναι%' then 'Yes'
        when F11_V like '%Βαμμένο: Ναι%' then 'Yes'
        when F12_V like '%Βαμμένο: Ναι%' then 'Yes'
        when F13_V like '%Βαμμένο: Ναι%' then 'Yes'
        when F14_V like '%Βαμμένο: Ναι%' then 'Yes'
        when F15_V like '%Βαμμένο: Ναι%' then 'Yes'
        else 'No'
end as Painted
,case
        when F0_V like '%Επιπλωμένο: Ναι%' then 'Yes'
        when F1_V like '%Επιπλωμένο: Ναι%' then 'Yes'
        when F2_V like '%Επιπλωμένο: Ναι%' then 'Yes'
        when F3_V like '%Επιπλωμένο: Ναι%' then 'Yes'
        when F4_V like '%Επιπλωμένο: Ναι%' then 'Yes'
        when F5_V like '%Επιπλωμένο: Ναι%' then 'Yes'
        when F6_V like '%Επιπλωμένο: Ναι%' then 'Yes'
        when F7_V like '%Επιπλωμένο: Ναι%' then 'Yes'
        when F8_V like '%Επιπλωμένο: Ναι%' then 'Yes'
        when F9_V like '%Επιπλωμένο: Ναι%' then 'Yes'
        when F10_V like '%Επιπλωμένο: Ναι%' then 'Yes'
        when F11_V like '%Επιπλωμένο: Ναι%' then 'Yes'
        when F12_V like '%Επιπλωμένο: Ναι%' then 'Yes'
        when F13_V like '%Επιπλωμένο: Ναι%' then 'Yes'
        when F14_V like '%Επιπλωμένο: Ναι%' then 'Yes'
        when F15_V like '%Επιπλωμένο: Ναι%' then 'Yes'
        else 'No'
end as Furnished
,case
        when F0_V like '%Πόρτα ασφαλείας: Ναι%' then 'Yes'
        when F1_V like '%Πόρτα ασφαλείας: Ναι%' then 'Yes'
        when F2_V like '%Πόρτα ασφαλείας: Ναι%' then 'Yes'
        when F3_V like '%Πόρτα ασφαλείας: Ναι%' then 'Yes'
        when F4_V like '%Πόρτα ασφαλείας: Ναι%' then 'Yes'
        when F5_V like '%Πόρτα ασφαλείας: Ναι%' then 'Yes'
        when F6_V like '%Πόρτα ασφαλείας: Ναι%' then 'Yes'
        when F7_V like '%Πόρτα ασφαλείας: Ναι%' then 'Yes'
        when F8_V like '%Πόρτα ασφαλείας: Ναι%' then 'Yes'
        when F9_V like '%Πόρτα ασφαλείας: Ναι%' then 'Yes'
        when F10_V like '%Πόρτα ασφαλείας: Ναι%' then 'Yes'
        when F11_V like '%Πόρτα ασφαλείας: Ναι%' then 'Yes'
        when F12_V like '%Πόρτα ασφαλείας: Ναι%' then 'Yes'
        when F13_V like '%Πόρτα ασφαλείας: Ναι%' then 'Yes'
        when F14_V like '%Πόρτα ασφαλείας: Ναι%' then 'Yes'
        when F15_V like '%Πόρτα ασφαλείας: Ναι%' then 'Yes'
        else 'No'
end as SafetyDoor
,case
        when F0_V like '%Τζάκι: Ναι%' then 'Yes'
        when F1_V like '%Τζάκι: Ναι%' then 'Yes'
        when F2_V like '%Τζάκι: Ναι%' then 'Yes'
        when F3_V like '%Τζάκι: Ναι%' then 'Yes'
        when F4_V like '%Τζάκι: Ναι%' then 'Yes'
        when F5_V like '%Τζάκι: Ναι%' then 'Yes'
```

Query 1. Data Transformation cont'd

```
            when F6_V like '%Τζάκι: Ναι%' then 'Yes'
            when F7_V like '%Τζάκι: Ναι%' then 'Yes'
            when F8_V like '%Τζάκι: Ναι%' then 'Yes'
            when F9_V like '%Τζάκι: Ναι%' then 'Yes'
            when F10_V like '%Τζάκι: Ναι%' then 'Yes'
            when F11_V like '%Τζάκι: Ναι%' then 'Yes'
            when F12_V like '%Τζάκι: Ναι%' then 'Yes'
            when F13_V like '%Τζάκι: Ναι%' then 'Yes'
            when F14_V like '%Τζάκι: Ναι%' then 'Yes'
            when F15_V like '%Τζάκι: Ναι%' then 'Yes'
            else 'No'
end as Fireplace
,case
            when F0_V like '%Κλιματισμός: Ναι%' then 'Yes'
            when F1_V like '%Κλιματισμός: Ναι%' then 'Yes'
            when F2_V like '%Κλιματισμός: Ναι%' then 'Yes'
            when F3_V like '%Κλιματισμός: Ναι%' then 'Yes'
            when F4_V like '%Κλιματισμός: Ναι%' then 'Yes'
            when F5_V like '%Κλιματισμός: Ναι%' then 'Yes'
            when F6_V like '%Κλιματισμός: Ναι%' then 'Yes'
            when F7_V like '%Κλιματισμός: Ναι%' then 'Yes'
            when F8_V like '%Κλιματισμός: Ναι%' then 'Yes'
            when F9_V like '%Κλιματισμός: Ναι%' then 'Yes'
            when F10_V like '%Κλιματισμός: Ναι%' then 'Yes'
            when F11_V like '%Κλιματισμός: Ναι%' then 'Yes'
            when F12_V like '%Κλιματισμός: Ναι%' then 'Yes'
            when F13_V like '%Κλιματισμός: Ναι%' then 'Yes'
            when F14_V like '%Κλιματισμός: Ναι%' then 'Yes'
            when F15_V like '%Κλιματισμός: Ναι%' then 'Yes'
            else 'No'
end as Aircondition
,case
            when F0_V like '%Ενδοδαπέδια θέρμανση: Ναι%' then 'Yes'
            when F1_V like '%Ενδοδαπέδια θέρμανση: Ναι%' then 'Yes'
            when F2_V like '%Ενδοδαπέδια θέρμανση: Ναι%' then 'Yes'
            when F3_V like '%Ενδοδαπέδια θέρμανση: Ναι%' then 'Yes'
            when F4_V like '%Ενδοδαπέδια θέρμανση: Ναι%' then 'Yes'
            when F5_V like '%Ενδοδαπέδια θέρμανση: Ναι%' then 'Yes'
            when F6_V like '%Ενδοδαπέδια θέρμανση: Ναι%' then 'Yes'
            when F7_V like '%Ενδοδαπέδια θέρμανση: Ναι%' then 'Yes'
            when F8_V like '%Ενδοδαπέδια θέρμανση: Ναι%' then 'Yes'
            when F9_V like '%Ενδοδαπέδια θέρμανση: Ναι%' then 'Yes'
            when F10_V like '%Ενδοδαπέδια θέρμανση: Ναι%' then 'Yes'
            when F11_V like '%Ενδοδαπέδια θέρμανση: Ναι%' then 'Yes'
            when F12_V like '%Ενδοδαπέδια θέρμανση: Ναι%' then 'Yes'
            when F13_V like '%Ενδοδαπέδια θέρμανση: Ναι%' then 'Yes'
            when F14_V like '%Ενδοδαπέδια θέρμανση: Ναι%' then 'Yes'
            when F15_V like '%Ενδοδαπέδια θέρμανση: Ναι%' then 'Yes'
            else 'No'
end as FloorHeating
,case
            when F0_V like '%Συναγερμός: Ναι%' then 'Yes'
            when F1_V like '%Συναγερμός: Ναι%' then 'Yes'
            when F2_V like '%Συναγερμός: Ναι%' then 'Yes'
            when F3_V like '%Συναγερμός: Ναι%' then 'Yes'
            when F4_V like '%Συναγερμός: Ναι%' then 'Yes'
            when F5_V like '%Συναγερμός: Ναι%' then 'Yes'
            when F6_V like '%Συναγερμός: Ναι%' then 'Yes'
            when F7_V like '%Συναγερμός: Ναι%' then 'Yes'
            when F8_V like '%Συναγερμός: Ναι%' then 'Yes'
            when F9_V like '%Συναγερμός: Ναι%' then 'Yes'
            when F10_V like '%Συναγερμός: Ναι%' then 'Yes'
            when F11_V like '%Συναγερμός: Ναι%' then 'Yes'
            when F12_V like '%Συναγερμός: Ναι%' then 'Yes'
            when F13_V like '%Συναγερμός: Ναι%' then 'Yes'
            when F14_V like '%Συναγερμός: Ναι%' then 'Yes'
            when F15_V like '%Συναγερμός: Ναι%' then 'Yes'
            else 'No'
end as Alarm
,case
            when F0_V like '%Ηλιακός θερμοσίφωνας: Ναι%' then 'Yes'
            when F1_V like '%Ηλιακός θερμοσίφωνας: Ναι%' then 'Yes'
            when F2_V like '%Ηλιακός θερμοσίφωνας: Ναι%' then 'Yes'
            when F3_V like '%Ηλιακός θερμοσίφωνας: Ναι%' then 'Yes'
            when F4_V like '%Ηλιακός θερμοσίφωνας: Ναι%' then 'Yes'
            when F5_V like '%Ηλιακός θερμοσίφωνας: Ναι%' then 'Yes'
            when F6_V like '%Ηλιακός θερμοσίφωνας: Ναι%' then 'Yes'
            when F7_V like '%Ηλιακός θερμοσίφωνας: Ναι%' then 'Yes'
            when F8_V like '%Ηλιακός θερμοσίφωνας: Ναι%' then 'Yes'
            when F9_V like '%Ηλιακός θερμοσίφωνας: Ναι%' then 'Yes'
            when F10_V like '%Ηλιακός θερμοσίφωνας: Ναι%' then 'Yes'
            when F11_V like '%Ηλιακός θερμοσίφωνας: Ναι%' then 'Yes'
            when F12_V like '%Ηλιακός θερμοσίφωνας: Ναι%' then 'Yes'
            when F13_V like '%Ηλιακός θερμοσίφωνας: Ναι%' then 'Yes'
            when F14_V like '%Ηλιακός θερμοσίφωνας: Ναι%' then 'Yes'
            when F15_V like '%Ηλιακός θερμοσίφωνας: Ναι%' then 'Yes'
            else 'No'
end as SolarBoiler
```

Query 1. Data Transformation cont'd

```
,case
        when F0_V like '%Τέντα: Ναι%' then 'Yes'
        when F1_V like '%Τέντα: Ναι%' then 'Yes'
        when F2_V like '%Τέντα: Ναι%' then 'Yes'
        when F3_V like '%Τέντα: Ναι%' then 'Yes'
        when F4_V like '%Τέντα: Ναι%' then 'Yes'
        when F5_V like '%Τέντα: Ναι%' then 'Yes'
        when F6_V like '%Τέντα: Ναι%' then 'Yes'
        when F7_V like '%Τέντα: Ναι%' then 'Yes'
        when F8_V like '%Τέντα: Ναι%' then 'Yes'
        when F9_V like '%Τέντα: Ναι%' then 'Yes'
        when F10_V like '%Τέντα: Ναι%' then 'Yes'
        when F11_V like '%Τέντα: Ναι%' then 'Yes'
        when F12_V like '%Τέντα: Ναι%' then 'Yes'
        when F13_V like '%Τέντα: Ναι%' then 'Yes'
        when F14_V like '%Τέντα: Ναι%' then 'Yes'
        when F15_V like '%Τέντα: Ναι%' then 'Yes'
        else 'No'
end as Tent
,case
        when F0_V like '%Βεράντα: Ναι%' then 'Yes'
        when F1_V like '%Βεράντα: Ναι%' then 'Yes'
        when F2_V like '%Βεράντα: Ναι%' then 'Yes'
        when F3_V like '%Βεράντα: Ναι%' then 'Yes'
        when F4_V like '%Βεράντα: Ναι%' then 'Yes'
        when F5_V like '%Βεράντα: Ναι%' then 'Yes'
        when F6_V like '%Βεράντα: Ναι%' then 'Yes'
        when F7_V like '%Βεράντα: Ναι%' then 'Yes'
        when F8_V like '%Βεράντα: Ναι%' then 'Yes'
        when F9_V like '%Βεράντα: Ναι%' then 'Yes'
        when F10_V like '%Βεράντα: Ναι%' then 'Yes'
        when F11_V like '%Βεράντα: Ναι%' then 'Yes'
        when F12_V like '%Βεράντα: Ναι%' then 'Yes'
        when F13_V like '%Βεράντα: Ναι%' then 'Yes'
        when F14_V like '%Βεράντα: Ναι%' then 'Yes'
        when F15_V like '%Βεράντα: Ναι%' then 'Yes'
        else 'No'
end as Veranda
,case
        when F0_V like '%Αποθήκη: Ναι%' then 'Yes'
        when F1_V like '%Αποθήκη: Ναι%' then 'Yes'
        when F2_V like '%Αποθήκη: Ναι%' then 'Yes'
        when F3_V like '%Αποθήκη: Ναι%' then 'Yes'
        when F4_V like '%Αποθήκη: Ναι%' then 'Yes'
        when F5_V like '%Αποθήκη: Ναι%' then 'Yes'
        when F6_V like '%Αποθήκη: Ναι%' then 'Yes'
        when F7_V like '%Αποθήκη: Ναι%' then 'Yes'
        when F8_V like '%Αποθήκη: Ναι%' then 'Yes'
        when F9_V like '%Αποθήκη: Ναι%' then 'Yes'
        when F10_V like '%Αποθήκη: Ναι%' then 'Yes'
        when F11_V like '%Αποθήκη: Ναι%' then 'Yes'
        when F12_V like '%Αποθήκη: Ναι%' then 'Yes'
        when F13_V like '%Αποθήκη: Ναι%' then 'Yes'
        when F14_V like '%Αποθήκη: Ναι%' then 'Yes'
        when F15_V like '%Αποθήκη: Ναι%' then 'Yes'
        else 'No'
end as Warehouse
,case
        when F0_V like '%Κήπος: Ναι%' then 'Yes'
        when F1_V like '%Κήπος: Ναι%' then 'Yes'
        when F2_V like '%Κήπος: Ναι%' then 'Yes'
        when F3_V like '%Κήπος: Ναι%' then 'Yes'
        when F4_V like '%Κήπος: Ναι%' then 'Yes'
        when F5_V like '%Κήπος: Ναι%' then 'Yes'
        when F6_V like '%Κήπος: Ναι%' then 'Yes'
        when F7_V like '%Κήπος: Ναι%' then 'Yes'
        when F8_V like '%Κήπος: Ναι%' then 'Yes'
        when F9_V like '%Κήπος: Ναι%' then 'Yes'
        when F10_V like '%Κήπος: Ναι%' then 'Yes'
        when F11_V like '%Κήπος: Ναι%' then 'Yes'
        when F12_V like '%Κήπος: Ναι%' then 'Yes'
        when F13_V like '%Κήπος: Ναι%' then 'Yes'
        when F14_V like '%Κήπος: Ναι%' then 'Yes'
        when F15_V like '%Κήπος: Ναι%' then 'Yes'
        else 'No'
end as Garden
,case
        when F0_V like '%Πισίνα: Ναι%' then 'Yes'
        when F1_V like '%Πισίνα: Ναι%' then 'Yes'
        when F2_V like '%Πισίνα: Ναι%' then 'Yes'
        when F3_V like '%Πισίνα: Ναι%' then 'Yes'
        when F4_V like '%Πισίνα: Ναι%' then 'Yes'
        when F5_V like '%Πισίνα: Ναι%' then 'Yes'
        when F6_V like '%Πισίνα: Ναι%' then 'Yes'
        when F7_V like '%Πισίνα: Ναι%' then 'Yes'
        when F8_V like '%Πισίνα: Ναι%' then 'Yes'
        when F9_V like '%Πισίνα: Ναι%' then 'Yes'
        when F10_V like '%Πισίνα: Ναι%' then 'Yes'
```

Query 1. Data Transformation cont'd

```
        when F11_V like '%Πισίνα: Ναι%' then 'Yes'
        when F12_V like '%Πισίνα: Ναι%' then 'Yes'
        when F13_V like '%Πισίνα: Ναι%' then 'Yes'
        when F14_V like '%Πισίνα: Ναι%' then 'Yes'
        when F15_V like '%Πισίνα: Ναι%' then 'Yes'
        else 'No'
end as Pool
,case
        when F0_V like '%Ασανσέρ: Οχι%' then 'No'
        when F1_V like '%Ασανσέρ: Οχι%' then 'No'
        when F2_V like '%Ασανσέρ: Οχι%' then 'No'
        when F3_V like '%Ασανσέρ: Οχι%' then 'No'
        when F4_V like '%Ασανσέρ: Οχι%' then 'No'
        when F5_V like '%Ασανσέρ: Οχι%' then 'No'
        when F6_V like '%Ασανσέρ: Οχι%' then 'No'
        when F7_V like '%Ασανσέρ: Οχι%' then 'No'
        when F8_V like '%Ασανσέρ: Οχι%' then 'No'
        when F9_V like '%Ασανσέρ: Οχι%' then 'No'
        when F10_V like '%Ασανσέρ: Οχι%' then 'No'
        when F11_V like '%Ασανσέρ: Οχι%' then 'No'
        when F12_V like '%Ασανσέρ: Οχι%' then 'No'
        when F13_V like '%Ασανσέρ: Οχι%' then 'No'
        when F14_V like '%Ασανσέρ: Οχι%' then 'No'
        when F15_V like '%Ασανσέρ: Οχι%' then 'No'
        else 'Yes'
end as Elevator
,case
        when F0_V like '%Πρόσοψης: Ναι%' then 'Yes'
        when F1_V like '%Πρόσοψης: Ναι%' then 'Yes'
        when F2_V like '%Πρόσοψης: Ναι%' then 'Yes'
        when F3_V like '%Πρόσοψης: Ναι%' then 'Yes'
        when F4_V like '%Πρόσοψης: Ναι%' then 'Yes'
        when F5_V like '%Πρόσοψης: Ναι%' then 'Yes'
        when F6_V like '%Πρόσοψης: Ναι%' then 'Yes'
        when F7_V like '%Πρόσοψης: Ναι%' then 'Yes'
        when F8_V like '%Πρόσοψης: Ναι%' then 'Yes'
        when F9_V like '%Πρόσοψης: Ναι%' then 'Yes'
        when F10_V like '%Πρόσοψης: Ναι%' then 'Yes'
        when F11_V like '%Πρόσοψης: Ναι%' then 'Yes'
        when F12_V like '%Πρόσοψης: Ναι%' then 'Yes'
        when F13_V like '%Πρόσοψης: Ναι%' then 'Yes'
        when F14_V like '%Πρόσοψης: Ναι%' then 'Yes'
        when F15_V like '%Πρόσοψης: Ναι%' then 'Yes'
        else 'No'
end as FrontStreet
,case
        when F0_V like '%Θέα: Ναι%' then 'Yes'
        when F1_V like '%Θέα: Ναι%' then 'Yes'
        when F2_V like '%Θέα: Ναι%' then 'Yes'
        when F3_V like '%Θέα: Ναι%' then 'Yes'
        when F4_V like '%Θέα: Ναι%' then 'Yes'
        when F5_V like '%Θέα: Ναι%' then 'Yes'
        when F6_V like '%Θέα: Ναι%' then 'Yes'
        when F7_V like '%Θέα: Ναι%' then 'Yes'
        when F8_V like '%Θέα: Ναι%' then 'Yes'
        when F9_V like '%Θέα: Ναι%' then 'Yes'
        when F10_V like '%Θέα: Ναι%' then 'Yes'
        when F11_V like '%Θέα: Ναι%' then 'Yes'
        when F12_V like '%Θέα: Ναι%' then 'Yes'
        when F13_V like '%Θέα: Ναι%' then 'Yes'
        when F14_V like '%Θέα: Ναι%' then 'Yes'
        when F15_V like '%Θέα: Ναι%' then 'Yes'
        else 'No'
end as WithView
,case
        when F0_V like '%Κατοικίδια ευπρόσδεκτα: Ναι%' then 'Yes'
        when F1_V like '%Κατοικίδια ευπρόσδεκτα: Ναι%' then 'Yes'
        when F2_V like '%Κατοικίδια ευπρόσδεκτα: Ναι%' then 'Yes'
        when F3_V like '%Κατοικίδια ευπρόσδεκτα: Ναι%' then 'Yes'
        when F4_V like '%Κατοικίδια ευπρόσδεκτα: Ναι%' then 'Yes'
        when F5_V like '%Κατοικίδια ευπρόσδεκτα: Ναι%' then 'Yes'
        when F6_V like '%Κατοικίδια ευπρόσδεκτα: Ναι%' then 'Yes'
        when F7_V like '%Κατοικίδια ευπρόσδεκτα: Ναι%' then 'Yes'
        when F8_V like '%Κατοικίδια ευπρόσδεκτα: Ναι%' then 'Yes'
        when F9_V like '%Κατοικίδια ευπρόσδεκτα: Ναι%' then 'Yes'
        when F10_V like '%Κατοικίδια ευπρόσδεκτα: Ναι%' then 'Yes'
        when F11_V like '%Κατοικίδια ευπρόσδεκτα: Ναι%' then 'Yes'
        when F12_V like '%Κατοικίδια ευπρόσδεκτα: Ναι%' then 'Yes'
        when F13_V like '%Κατοικίδια ευπρόσδεκτα: Ναι%' then 'Yes'
        when F14_V like '%Κατοικίδια ευπρόσδεκτα: Ναι%' then 'Yes'
        when F15_V like '%Κατοικίδια ευπρόσδεκτα: Ναι%' then 'Yes'
        else 'No'
end as PetsAllowed
,case
        when F0_V like '%Αυτόνομη%' then 'Yes'
        when F1_V like '%Αυτόνομη%' then 'Yes'
        when F2_V like '%Αυτόνομη%' then 'Yes'
        when F3_V like '%Αυτόνομη%' then 'Yes'
```

Query 1. Data Transformation cont'd

```
                when F4_V like '%Αυτόνομη%' then 'Yes'
                when F5_V like '%Αυτόνομη%' then 'Yes'
                when F6_V like '%Αυτόνομη%' then 'Yes'
                when F7_V like '%Αυτόνομη%' then 'Yes'
                when F8_V like '%Αυτόνομη%' then 'Yes'
                when F9_V like '%Αυτόνομη%' then 'Yes'
                when F10_V like '%Αυτόνομη%' then 'Yes'
                when F11_V like '%Αυτόνομη%' then 'Yes'
                when F12_V like '%Αυτόνομη%' then 'Yes'
                when F13_V like '%Αυτόνομη%' then 'Yes'
                when F14_V like '%Αυτόνομη%' then 'Yes'
                when F15_V like '%Αυτόνομη%' then 'Yes'
                else 'No'
        end as HeatingAutonomous
        ,case
                when F0_V like '%Φυσικό αέριο%' then 'Yes'
                when F1_V like '%Φυσικό αέριο%' then 'Yes'
                when F2_V like '%Φυσικό αέριο%' then 'Yes'
                when F3_V like '%Φυσικό αέριο%' then 'Yes'
                when F4_V like '%Φυσικό αέριο%' then 'Yes'
                when F5_V like '%Φυσικό αέριο%' then 'Yes'
                when F6_V like '%Φυσικό αέριο%' then 'Yes'
                when F7_V like '%Φυσικό αέριο%' then 'Yes'
                when F8_V like '%Φυσικό αέριο%' then 'Yes'
                when F9_V like '%Φυσικό αέριο%' then 'Yes'
                when F10_V like '%Φυσικό αέριο%' then 'Yes'
                when F11_V like '%Φυσικό αέριο%' then 'Yes'
                when F12_V like '%Φυσικό αέριο%' then 'Yes'
                when F13_V like '%Φυσικό αέριο%' then 'Yes'
                when F14_V like '%Φυσικό αέριο%' then 'Yes'
                when F15_V like '%Φυσικό αέριο%' then 'Yes'
                else 'No'
        end as HeatingNaturalGas
FROM [Hows_II].[dbo].[AdFeatures_Export]
WHERE 1=1
        AND Code != ''
        AND Code not in (SELECT Distinct Code FROM [Hows_II].[dbo].[AdFeatures_List])
```

Query 1. Data Transformation

SQL Query – Data Transformation Updates

```
UPDATE [Hows_II].[dbo].[AdFeatures_List]
  set [Floor] = 0
  where [Floor] = 'Ισόγειο'

UPDATE [Hows_II].[dbo].[AdFeatures_List]
  set [Floor] = 0.5
  where [Floor] = 'Ημιόροφος'


UPDATE [Hows_II].[dbo].[AdFeatures_List]
  set [Floor] = -1
  where [Floor] in ('Ημιυπόγειο', 'Υπόγειο')

  -------------------

UPDATE [Hows_II].[dbo].[AdFeatures_List]
  set YearBuilt = NULL
  where YearBuilt = 'Υπό κατασκευή'

UPDATE [Hows_II].[dbo].[AdFeatures_List]
  set YearBuilt = NULL
  where YearBuilt = '-'

UPDATE [Hows_II].[dbo].[AdFeatures_List]
  set YearBuilt = NULL
  where YearBuilt < 1920

  -------------------

UPDATE [Hows_II].[dbo].[AdFeatures_List]
  set Bathrooms = 1
  where Bathrooms is null
```

Query 2. Data Transformation II – Updates cont'd

```
 -----------------------


UPDATE [Hows_II].[dbo].[AdFeatures_List]
set Age =
case
        when YearBuilt < 1980 then '>40'
        when YearBuilt < 2000 then '20-40'
        when YearBuilt < 2020 then '<20'
end

-----------------------

UPDATE [Hows_II].[dbo].[AdFeatures_List]
set AgeD =
case
        when YearBuilt >= 2010 then 1
        when YearBuilt >= 2000 then 2
        when YearBuilt >= 1990 then 3
        when YearBuilt >= 1980 then 4
        when YearBuilt >= 1970 then 5
        when YearBuilt >= 1960 then 6
        when YearBuilt >= 1950 then 7
        when YearBuilt >= 1940 then 8
        when YearBuilt >= 1930 then 9
end
```

Query 2. Data Transformation Updates

SQL Query – Data Cleansing

```
DELETE  [Hows_II].[dbo].[AdFeatures_List]
where PpM > 50
```

Query 3. Data Cleansing

R Code – Descriptive Statistics

```
#### Import Data and Transform Data Types

con <- DBI::dbConnect(odbc::odbc(),
                      Driver = "SQL Server",
                      Server = "localhost\\SQLEXPRESS",
                      Database = "HOWS",
                      Trusted_Connection = "True")

str(hows)

hows$Price <- as.numeric(hows$Price)
hows$Size <- as.numeric(as.character(hows$Size))
hows$ppm <- as.numeric(hows$Price/hows$Size)
hows$Floor <- as.character(hows$Floor)
hows$Floor <- as.numeric(hows$Floor)
hows$Bathrooms <- as.character(hows$Bathrooms)
hows$Bathrooms <- as.numeric(hows$Bathrooms)
hows$Bedrooms <- as.numeric(as.character(hows$Bedrooms))
hows$YearBuilt <- as.character(hows$YearBuilt)
hows$YearBuilt <- as.numeric(hows$YearBuilt)
hows$AgeD <- as.numeric(hows$AgeD)

hows <- hows[,-c(1,2,8,35)]



#### Group Numerical & Non-Numerical Variables

library(psych)
numeric.only <- sapply(hows,class)=='numeric'
hows.num <-hows[,numeric.only]
hows.num <- hows.num[,c(8,1,2,5,3,4,6,7)]
hows.fac <- hows[ , !numeric.only]



#### Numerical Variables - Histograms & Correlation Matrix

par(mfrow=c(3,3))
hist(hows.num[,1], main=names(hows.num)[1])
hist(hows.num[,2], main=names(hows.num)[2])
hist(hows.num[,3], main=names(hows.num)[3])
hist(hows.num[,4], main=names(hows.num)[4])
hist(hows.num[,5], main=names(hows.num)[5])
hist(hows.num[,6], main=names(hows.num)[6])
hist(hows.num[,7], main=names(hows.num)[7])
hist(hows.num[,8], main=names(hows.num)[8])

par(mfrow=c(1,1))
library(corrplot)
corrplot(cor(hows.num, use = "complete.obs"), method ="number")
```

Code 1. Descriptive Statistics cont'd

```
#### Non-Numerical Variables - BoxPlots

par(mfrow=c(1,5))

plot(ppm ~ hows.fac[,1], main=names(hows.fac)[1], data = hows)
plot(ppm ~ hows.fac[,2], main=names(hows.fac)[2], data = hows)
plot(ppm ~ hows.fac[,3], main=names(hows.fac)[3], data = hows)
plot(ppm ~ hows.fac[,4], main=names(hows.fac)[4], data = hows)
plot(ppm ~ hows.fac[,5], main=names(hows.fac)[5], data = hows)

plot(ppm ~ hows.fac[,6], main=names(hows.fac)[6], data = hows)
plot(ppm ~ hows.fac[,7], main=names(hows.fac)[7], data = hows)
plot(ppm ~ hows.fac[,8], main=names(hows.fac)[8], data = hows)
plot(ppm ~ hows.fac[,9], main=names(hows.fac)[9], data = hows)
plot(ppm ~ hows.fac[,10], main=names(hows.fac)[10], data = hows)

plot(ppm ~ hows.fac[,11], main=names(hows.fac)[11], data = hows)
plot(ppm ~ hows.fac[,12], main=names(hows.fac)[12], data = hows)
plot(ppm ~ hows.fac[,13], main=names(hows.fac)[13], data = hows)
plot(ppm ~ hows.fac[,14], main=names(hows.fac)[14], data = hows)
plot(ppm ~ hows.fac[,15], main=names(hows.fac)[15], data = hows)

plot(ppm ~ hows.fac[,16], main=names(hows.fac)[16], data = hows)
plot(ppm ~ hows.fac[,17], main=names(hows.fac)[17], data = hows)
plot(ppm ~ hows.fac[,18], main=names(hows.fac)[18], data = hows)
plot(ppm ~ hows.fac[,19], main=names(hows.fac)[19], data = hows)
plot(ppm ~ hows.fac[,20], main=names(hows.fac)[20], data = hows)

plot(ppm ~ hows.fac[,21], main=names(hows.fac)[21], data = hows)
plot(ppm ~ hows.fac[,22], main=names(hows.fac)[22], data = hows)
plot(ppm ~ hows.fac[,23], main=names(hows.fac)[23], data = hows)
plot(ppm ~ hows.fac[,24], main=names(hows.fac)[24], data = hows)
plot(ppm ~ hows.fac[,25], main=names(hows.fac)[25], data = hows)
```

Code 1. Descriptive Statistics

R Code – Model

```
#### Final Model and Residuals Assumptions Testing


model_5final<- lm(log(ppm) ~ log(Size) + Aircondition + poly(Floor,3,raw=TRUE)  +  Furnished + HeatingAutonomous
                + HeatingNaturalGas + Parking + Tent +  poly(AgeD,3,raw=TRUE) + SubArea + Neighborhood,
                data = hows)
summary(model_5final)


library(ggfortify)
autoplot(model_5final)


#1. The normality of errors
plot(model_5final, which = 2)


#2. Constant Variance
plot(model_5final, which = 3)

Stud.residualsLog1 <-rstudent(model_5final)
FittedLog1 <- fitted(model_5final)
plot(FittedLog1, Stud.residualsLog1) ; abline(h=c(-2,2), col=2, lty=2)
plot(FittedLog1, Stud.residualsLog1^2) ; abline(h=4, col=2, lty=2)


#3. Non Linearity of residuals
library(car)
residualPlot(model_5final, type='rstudent')


#4. Influential Points - Outliers
plot(rstudent(model_5final), type='l')

cutoff <- 4/((nrow(model_5final)-length(model_5final$coefficients)-2))
plot(model_5final, which=4, cook.levels=cutoff)


#5. Indipendence of errors
plot(rstudent(model_5final), type='l')
title(main = list("Independence of Errors - Final Model", cex = 1.5,col = "red", font = 2))
```

Code 2. Model & Residuals Tests