Discovering Patterns in the Russian Housing Market for Analysis and Prediction

3804ICT Assignment Part I | Data Investigation Report | Trimester 2, 2019

Joshua Russell s5057545

Joshua Mitchell s5055278

Hayden Flatley s5088623

{joshua.russell2, joshua.mitchell4, hayden.flatley}@griffithuni.edu.au

Introduction

The following data investigation report serves as a summary of the data investigation that was conducted for part I of this assignment. Data exploration, pre-processing and visualisation were the main topics covered in the investigation. Here, we will provide an overview of the dataset that was explored, the pre-processing steps that were implemented, and the interesting characteristics that were visualised. For details on the complete data investigation, please refer to the data investigation notebook which contains our code and a more in-depth analysis.

1. Data Exploration

Within this section we examine the number of data samples and attributes within the dataset, investigate the different types of attributes present in the dataset, perform feature selection for data pre-processing and visualisation, and lastly explore the statistical properties of the selected attributes.

In total, the dataset consisted of 38,133 data samples and 292 attributes. Nominal, ratio-scaled, discrete and continuous attributes were present in the dataset, whereas ordinal and interval-scaled attributes were not. Since the dataset contained a significant number of attributes, we decided to perform feature selection so that we could focus our data exploration, pre-processing and visualization.

Firstly, we performed manual feature selection in which we chose attributes based upon our domain knowledge of the data and what attributes would assumedly be useful for predicting sale price (which is the target variable of the dataset and the target variable for our regression and forecasting methods in part II). In short, we chose $price\_doc$ (the sale price of the property), $full\_sq$ (the total area of the property in square meters), sub_area (the name of the district that the property is within), num_room (the number of living rooms in the property), $build\_year$ (the year that the property was built) and $state$ (the condition that the property is in at the time of purchase). Following this, we performed metric-based feature selection methods to find attributes with strong predictive capabilities for property sale price. We used correlation analysis and extreme gradient boosting for this task. Nine additional attributes were selected based upon these methods.

The final part of this section explored the statistical information of the selected attributes. Specifically, this involved examining the measures of central tendency (mean, median, midrange, and mode) and the measures of spread (variance, standard deviation, and the five-number summary). This exploration of the selected attributes showed that there were unexpected, potentially noisy data points within the dataset, and that the dataset time range was between 20th August 2011 and 30th May 2016. Consequently, we concluded that we would need to perform data pre-processing before visualisation so that the visualisation techniques were not skewed by noisy values and would not give false representations of the attribute data distributions.

2. Data Pre-Processing

This section discusses problems of and solutions to poor data quality, implements methods for cleaning the dataset, and investigates the creation of new attributes through data transformation and discretization.

We begin by stating that raw data, which is typically poor in quality, can dramatically reduce the accuracy of data mining methods and lead to incorrect conclusions. This data quality issue may stem from factors such as noise or incorrect data entry, which in turn may skew the distribution of attributes in the dataset. A simple way of dealing with these outliers and missing values would be to delete them from the dataset. However, this approach provides convenience at the expense of statistical significance due to smaller sample sizes. As a result, we also explore other data cleaning techniques (such as filling in the missing data points where attributes are predictable). In our data pre-processing that follows, we assume that the data is "missing at random", meaning that missing values of an attribute are independent of other attributes.

Our first data cleaning investigation was of incomplete (missing) data. We observed the top 60 missing value percentage attributes in a bar plot as shown in *Figure 1*.

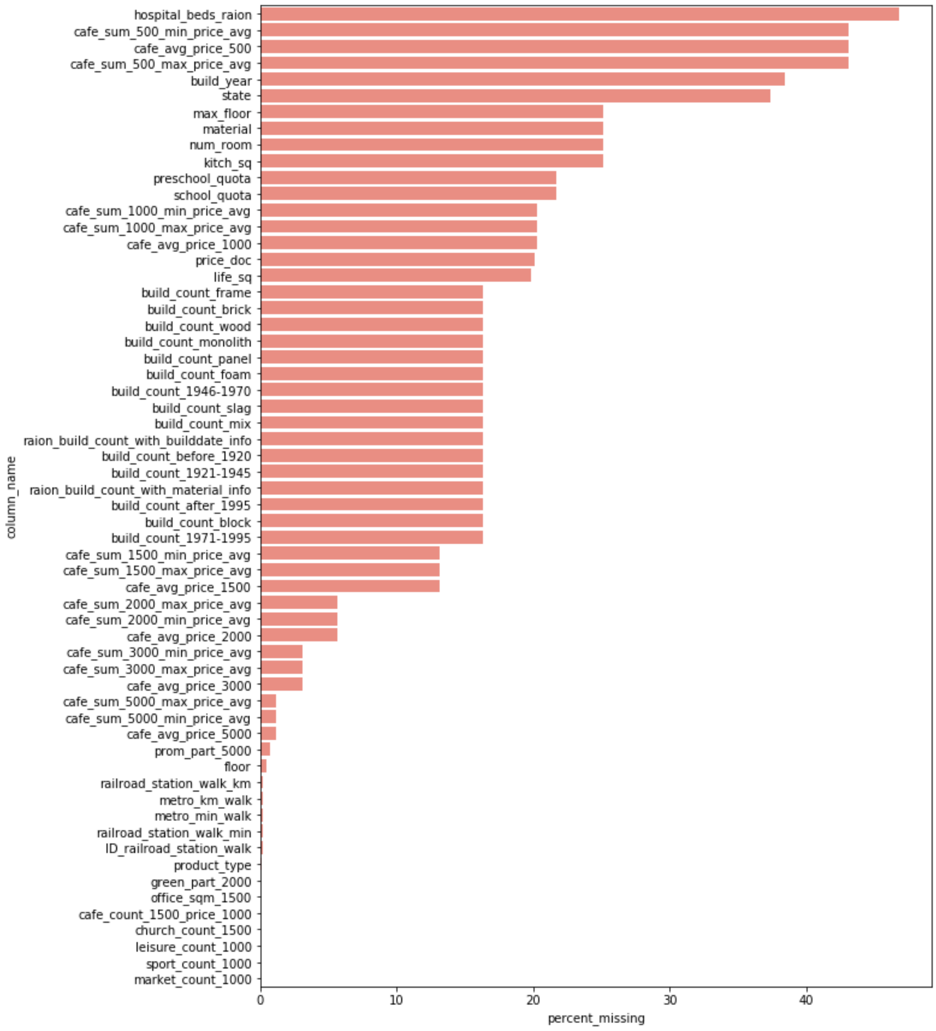

Figure 1. Bar plot of the Top 60 Percent Missing Value Attributes

There were 17 attributes missing more than 20 percent of data points. One of these attributes was $price\_doc$, although this was simply because we set the $price\_doc$ values in the test set to NaN for merging purposes. We checked that there were no missing values for $price\_doc$ in the training set and found that there were indeed no missing values.

Therefore, there were 16 attributes with more than 20 percent of their data missing. Since this is a small fraction of the 292 attributes within the primary housing market dataset, it would be inefficient to drop all of these data samples with missing values. As we would lose a significant number of data samples to base our analysis and algorithms off. Instead, we decided to disregard the attributes when required, and only disregard the data samples when investigating one of the particular attributes with a significant number of missing values in isolation (i.e. on its own). We later explored filling in the missing values instead of simply dropping them.

Next, we examined intentionally incorrect data entries. We found that the minimum value for $build\_year$ was 0. Intentionally incorrect data is data that is entered typically to hide particular information, such as January 1st for a birthday. This approach may have also been used to hide the property age by entering an uninformative value for $build\_year$. Through printing the value counts of the attribute, we observe that 0 was entered 899 times and 1 was entered 555 times. Both these value counts seem too frequent to be considered accidental or random noise. As a result, we believe that these values of 0 and 1 for $build\_year$ are intentionally incorrect data within the Russian housing market dataset. Such values will be dealt with when dealing with noisy data as a whole.

In our analysis on noisy data, we found that the $state$ attribute had a noisy value of 33 (where the typical values were from 1 to 4), that $build\_year$ had both extreme maximum and minimum values which were dropped, that $price\_doc$ (which represents property sale price) contained values which were significantly greater than the mean value but appeared to be expensive properties rather than noise, and that there were cases where the attribute $full\_sq$ (total area of property in square meters) was less than $life\_sq$ (living area of property in square meters), which by definition is incorrect, and therefore we corrected these values by setting $full\_sq$ to be equal to $life\_sq$.

We performed outlier detection by applying a basic statistical method of checking if a value is plus or minus three standard deviations away from the mean of the attribute, and if so, classifying it as an outlier. We found that there was not a significant number of outliers present (based upon this simple statistical method), and that the context of the attributes makes it understandable for there to be some relatively large differences in values. Thus, we did not clean any of the values and left them as interesting data points to explore.

Lastly, we explored data transformation and discretization. The reasoning for this is that for data mining methods such as association rule mining, we cannot use continuous numerical attributes as the total number of itemsets would be too large, and the meaning behind the implication would not be very significant. We discussed that one way of working around this is to perform attribute/feature construction and construct categorial attributes from the continuous numerical ones. For the initial data investigation, we performed this on the attributes $park\_km$ (distance to park) and $full\_sq$ (total property area), by assigning the continuous data points to classes of $[NEAR, MODERATE, FAR]$ and $[SMALL, MEDIUM, LARGE]$, respectively. We based the numerical ranges for these classes off our domain knowledge of distances, and statistical information of common house sizes within Russia.

3. Data Visualisation

After exploring the general characteristics of the dataset and performing necessary data pre-processing, we performed data visualisation to gain a better understanding of the attributes. Within this final section, we visualised the statistical information of selected attributes, examined the housing market share, looked into district property sales and prices, observed the relationship between property distance to utilities and property sale price, saw which house wall materials were more popular and expensive, investigated the significance of property condition on sale price, studied the time series nature of sale price, evaluated the influence of property size on property price, and finally looked at how the build year of the properties affected price. In the remaining pages of this data investigation report we summarise these studies. For a more detailed analysis of the results and their implications, please refer to the data investigation notebook.

3.1. Graphing Statistical Information

We firstly examined the attributes by plotting box plots of their five-number summaries. These box plots gave us an overview of the spread of the data and also displayed extreme points which suggested the need for further noise/outlier investigations. Due to the fact that two different distributions of data can be represented by the same box plot, we also displayed histograms of the attributes. Histograms divide the values of the attributes into bins and show a graph that is representative of the distribution of the data. This improved our understanding of the shape of the data distributions in the Russian housing market dataset.

We found that the investigated attributes had quite a significant number of data points beyond the whiskers of the box plots. For the example shown below in *Figure 2*, this told us that the majority of homes are sold at a price between 0 and 20,000,000 rubles, although there are still many properties more expensive than this. Furthermore, seeing a relatively steady increase in price among data samples, rather than a few random points beyond the whiskers, tells us that these points are unlikely noise and instead just expensive properties. The histograms showed us that the majority of the attributes had right-skewed distributions. Which was somewhat expected due to many of the attributes being ratio-scaled.

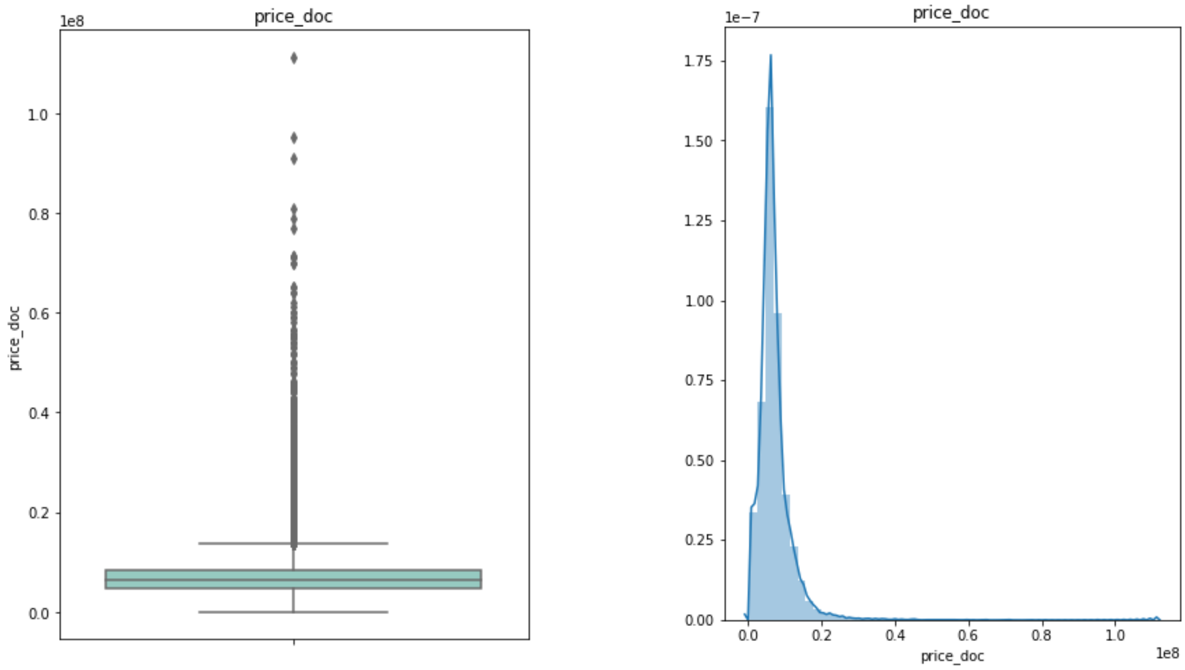

Figure 2. Box Plot and Histogram of the price_doc (Property Sale Price) Attribute

3.2. Housing Market Share

Besides examining the property ecological zone classification and the property districts with nuclear reactors, we also investigated the housing market share of the *product_type* attribute, which specifies whether the property was purchased for owner occupancy, or as an investment property. We found that two thirds of the purchased properties were purchased as investment properties, while the remaining one third were purchased for the home buyer. This provides us with insight into the state of real estate within Russia, that there are many investors buying properties rather than home buyers purchasing homes. This also supports the notion that the vast majority of properties within this dataset are apartments, rather than houses. Since apartments are generally purchased as investment properties for leasing. We assume that this information may prove useful in our forecasting and regression investigations.

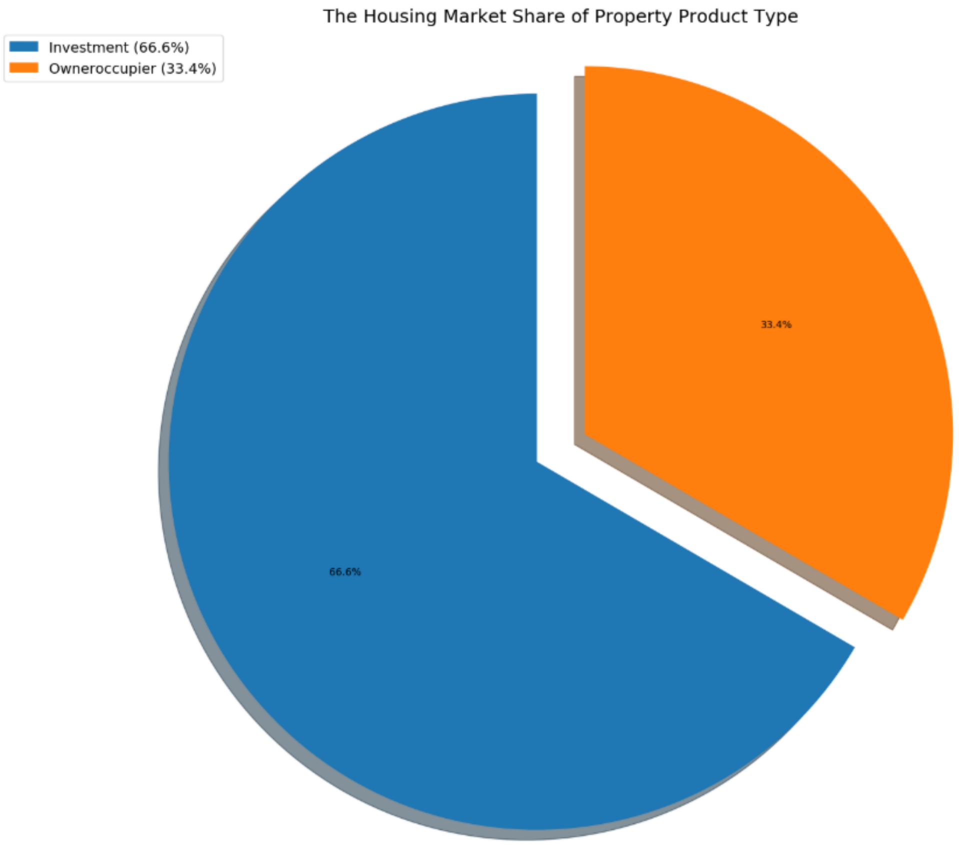

Figure 3. Pie Chart of the Housing Market Share of Property Product Type

Examining properties with varying floor quantities further supported our hypothesis that the majority of the properties within the dataset are urban apartments. As we saw that the most common value for the max_$floor$ attribute in *Figure 4* below was 17. Moreover, we observed that the distribution of the attribute $floor$ peaks at 4 floors and consists mainly of floor values that are above the expected value for a regular house.

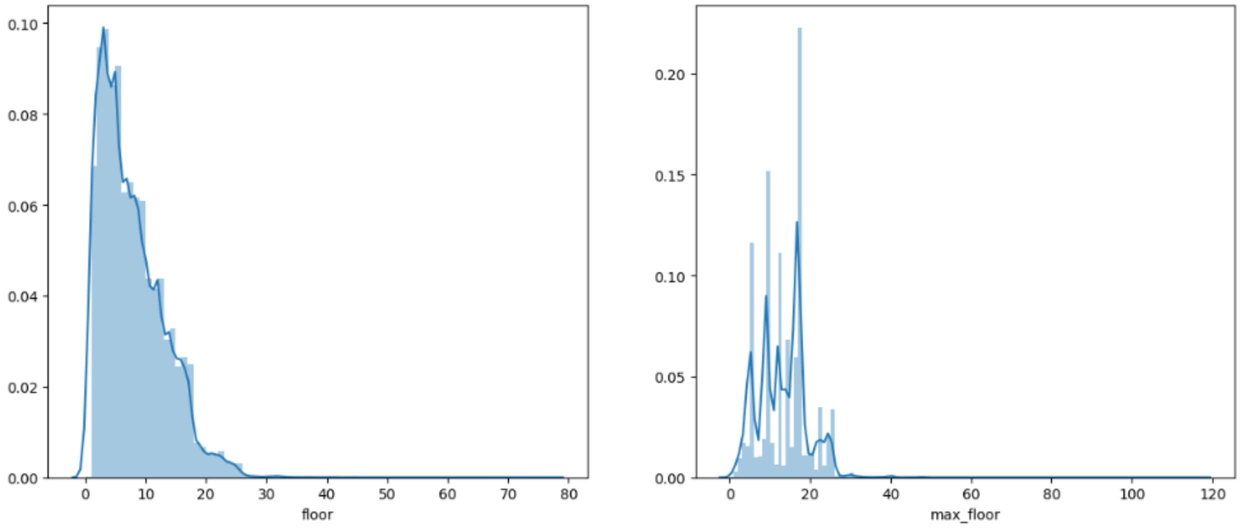

Figure 4. Histograms of the $floor$ and max_$floor$ Attributes

3.3. District Property Sales

Within this investigation we studied the number of property sales in each district. In order to gain an intuition behind where these districts were located, we decided to visualise this information spatially on a map. In *Figure 5* below, we display a world map zoomed into Moscow, Russia. Interestingly, we found that the majority of the districts within the dataset were districts within/around Moscow. We found that the total number of properties sold in districts around the city was much greater than those districts within the centre of the city.

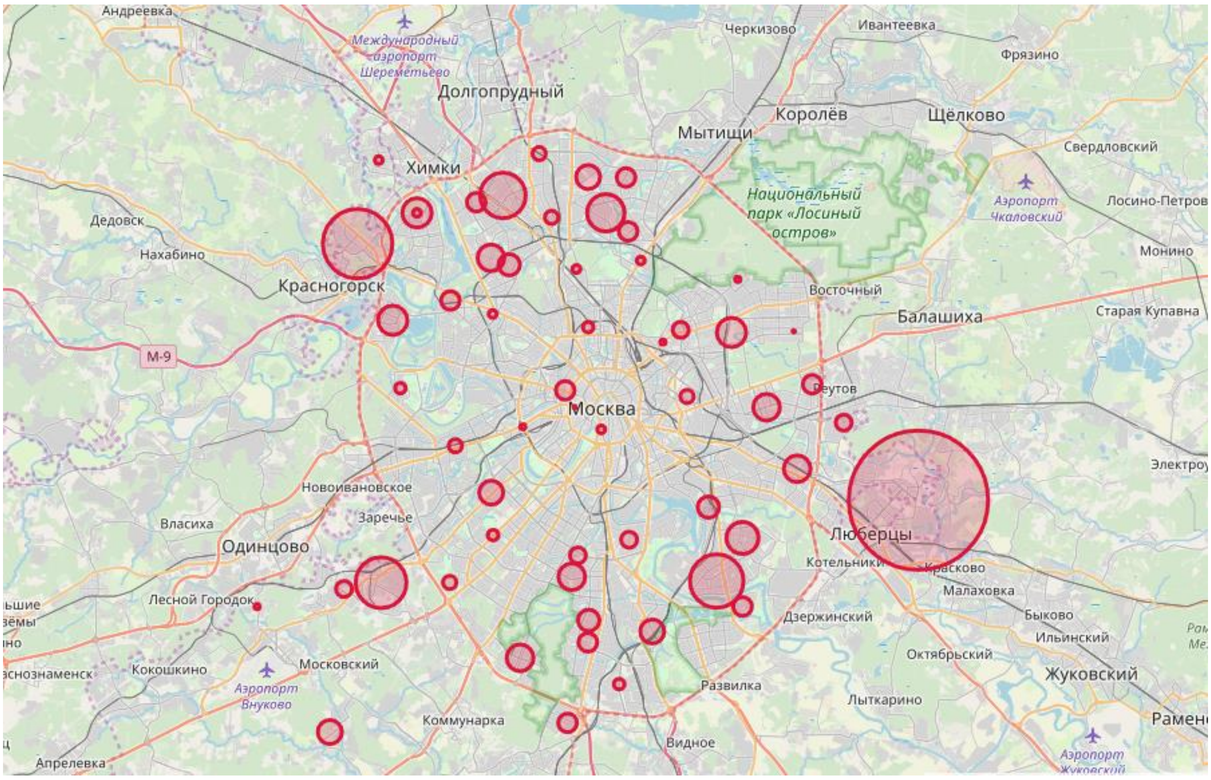

Figure 5. World Map Zoomed into Moscow, Russia Displaying Circles which are Scaled to represent the Number of Properties Sold within each District

3.4. District Property Prices

Due to the scatterplot used in this study being rather large in size, it was not included within this summary data investigation report.

The key findings of this analysis were that districts with the most expensive average property price did not have a wider, more expensive distribution of property price that is evenly spread out. Instead, like the other districts, they had a group of property prices which appeared to be around the average price, and then some expensive properties that seemed relatively spaced out. Additionally, we saw quite a difference in property sale price between different districts. From this information, we gather that using sale price for classifying properties into districts may be quite useful.

3.5. Price by Property Distance to Utilities

Within this study we visualized how distances to local public services effect the sale price of properties. The most apparent relationship found was that properties close to local public services were, as a whole, more expensive than properties that were located further away from these services. We saw this through the right-skewed nature of the data distributions. However, being the "closest" did not seem to mean that the property would be one of the most expensive properties. Furthermore, the definition of what was considered "close" seemed to change among attributes. This will need to be considered when conducting attribute construction of categorical attributes from continuous numerical attributes. Lastly, based upon the observed data distributions, this information was assumed to assist greatly in predicting sale price. As algorithms such as the multilayer perceptron should be able to learn for example that if a property is 40km away from the nearest school, then its price should be no greater than 20,000,000 rubles.

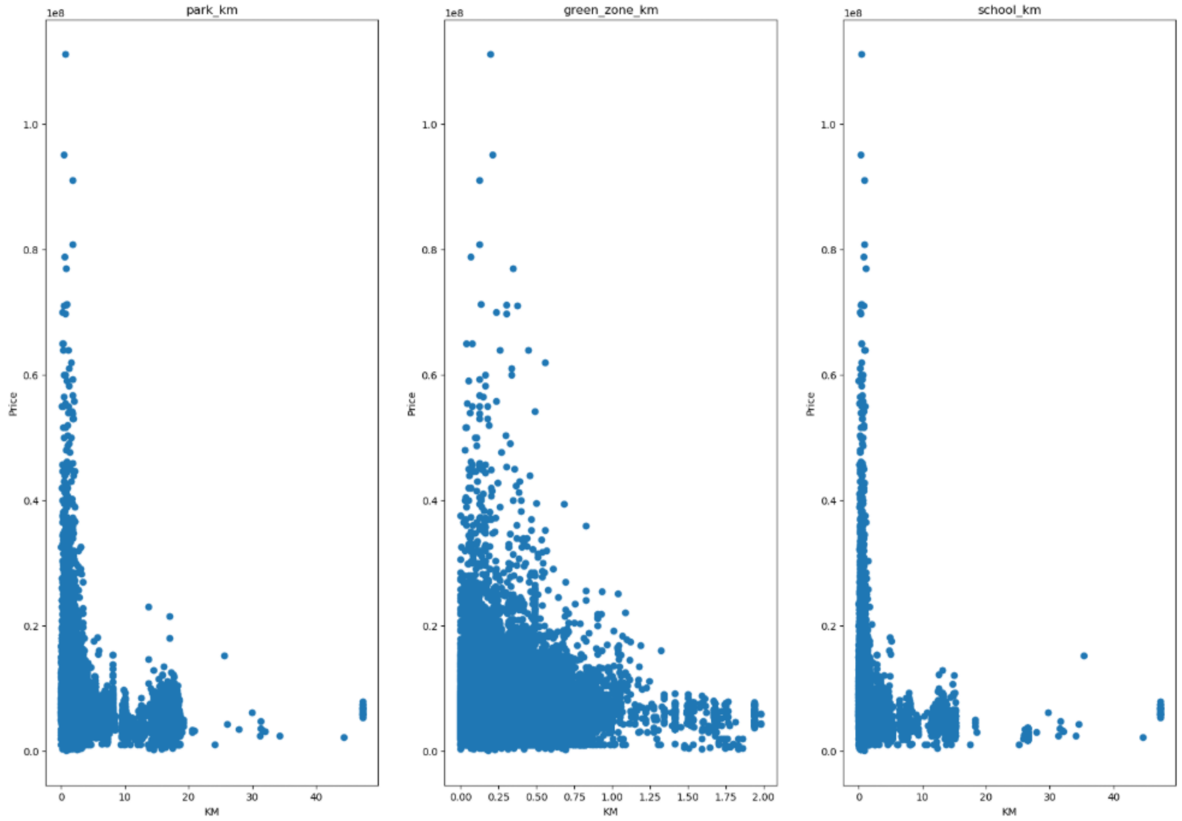

Figure 6. Scatter Plots of Attributes park_km (distance to nearest park), green_zone_km (distance to nearest green zone), and school_km (distance to nearest school) against Sale Price (i.e. price_doc)

3.6. House Material

This investigation looked at the different materials that properties in the dataset were built out of. By examining the overall material distribution in the dataset, we saw that panel was the most common material that purchased properties were made out of by a significant margin. Wood on the other hand appeared to be the wall material of a minimal number of properties, while the rest of the other materials seemed to have relatively the same counts. The violin plot of this study is shown in *Figure 7*. We also examined the distribution of property material per district, and the relationship between sale price and material. There appeared to be some limits in price of particular materials, which suggests to us that this attribute may also be useful for regression analysis in predicting sale price. Furthermore, we observed that different districts rather distinctively used different wall materials. This observation suggests that this attribute may contain strong predictive capabilities for the task of district classification.

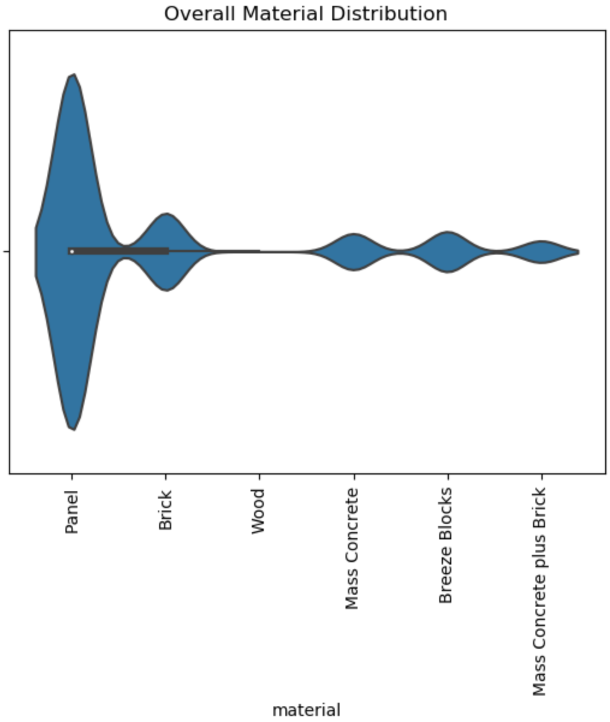

Figure 7. Violin Plot of the $material$ (property wall material) Attribute

3.7. Prices of Different Property States

This analysis examined the $state$ attribute, which represents the condition of the property (1 meaning the property is in a poor condition, and 4 meaning the property is in an immaculate condition). Intuitively, we thought that the poorer the condition of the property, the less its sale price would be. However, in *Figure 8* we saw that this was not necessarily the case, as the different states all shared rather similar sale price distributions (with only relatively small differences). From this analysis, we concluded that the $state$ attribute may not be as significant of an indicator for sale price as we previously had expected.

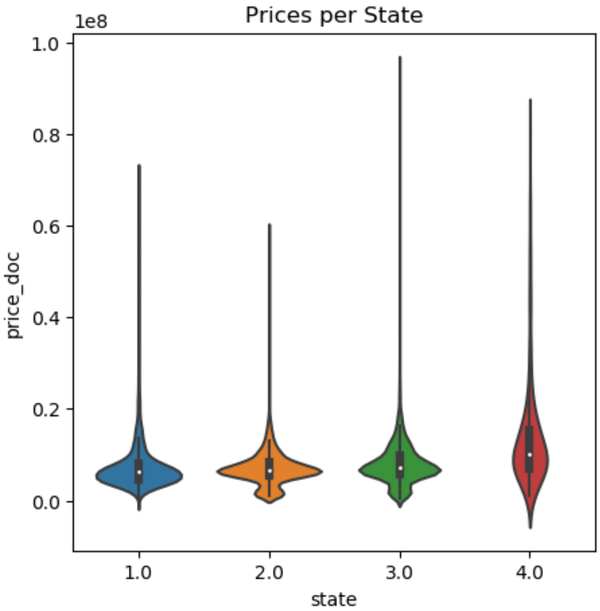

Figure 8. Violin Plots for Different Values of the $state$ Attribute against Property Sale Price ($price\_doc$)

3.8. Time Series Visualisation

This investigation analysed the general time series characteristics of the data. Specifically, we looked at the monthly average house sale price over time. Firstly, we were able to see that, after pre-processing, the dataset contains property sales from August 2011 to June 2015. In the first half of 2012 there was a rather large increase in property price, before it slowly dropped back down in the second half of the year. Since the beginning of 2013 however, property sale price has rather steadily increased. The fact that there is a trend present in the data may make forecasting methods quite accurate at predicting house price. Furthermore, investigating the performance of regression methods with and without time information may lead to interesting results we can examine in the final report.

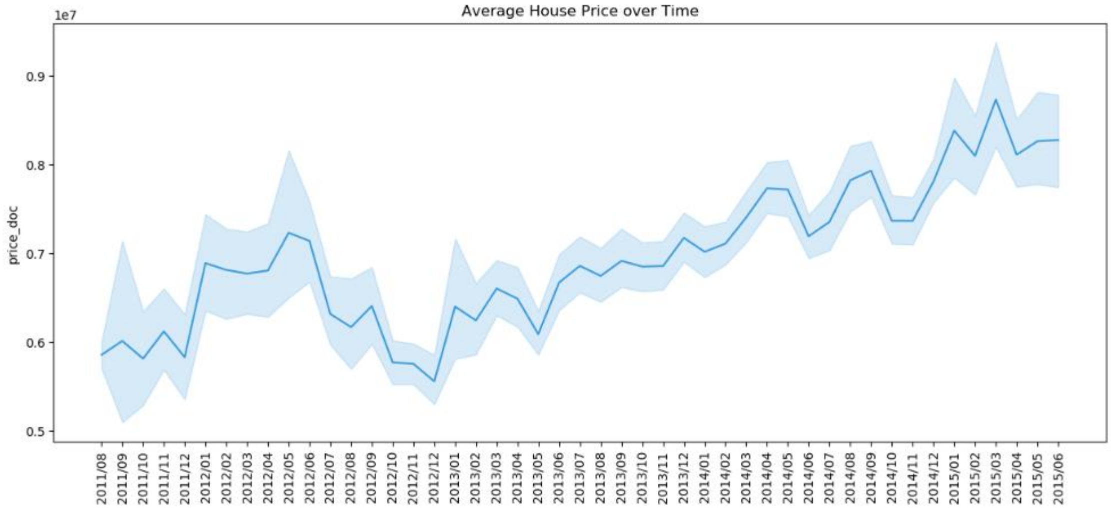

Figure 9. Line Plot of the Monthly Average Property Price over Time

3.9. Influence of Size on House Price

Within this study we examined the effects of the attributes $full\_sq$ and $life\_sq$ on property sale price. The difference between these attributes is that $full\_sq$ takes into consideration loggias, balconies and other non-residential areas whereas $life\_sq$ does not. We found that there was a gradual positive gradient in $full\_sq$ and $life\_sq$ as property price increased. Besides the general behavior of the data, there were a few points that appeared to be outliers (properties with large area and a low sale price). We then took the difference of these two attributes to extract the impact of non-residential areas on price. From this analysis we saw a far lower degree of influence over house prices as data points stayed relatively low and flat as price increased.

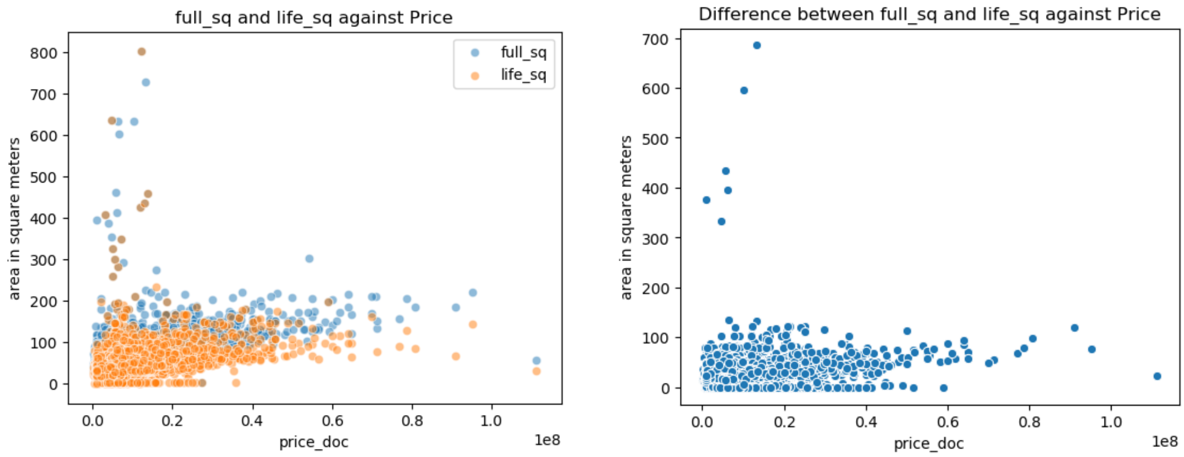

Figure 10. (Left) Scatter Plot of $full\_sq$ and $life\_sq$ against Price. (Right) Scatter Plot of the Difference between $full\_sq$ and $life\_sq$ against Price.

3.10. Insights into Housing Construction over Time

Lastly, we examine how sale price changes depending on the age of the building. The general trend of the time series data is that sale price declines as build year becomes more recent. After 1950 there is a steady decline, followed by an increase, and then another decline around 2008. We hypothesize that these drop offs are associated with the post-World War II building boom that occurred in the Soviet Union around 1954 and the 2008 financial crisis or Russian great recession, respectively. Cheaper materials would have been used to construct properties during this time which would now make the properties worth less. Before 1950, the majority of the properties would have been houses rather than apartment complexes, hence the sporadic nature of the price in years before. Since the build years appear to have rather unique sale price ranges, we predict that this attribute will be important in our regression analysis of sale price.

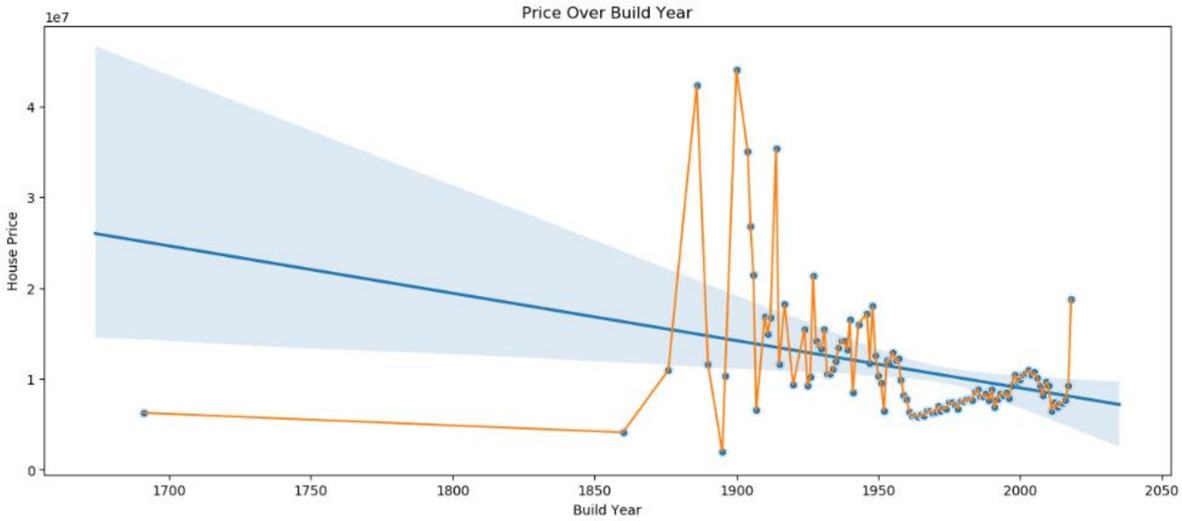

Figure 11. Scatter/Line Plot of the Property Sale Price over Build Year