DAF 1

Housing Market Behaviour analysis using the Melbourne Housing Market

Strategic Thinking
Higher Diploma in Science in Data Analytics for Business
Graham Glanville and Mark Morrison

By Conor Sheehan (sba 20128), Eric Parfrey (sba 20129),
Giuliano Silva (sba 20130) and Hasan Aziz (sba 20140)

CCT College Dublin

Table of Contents

1. Business Understanding

Having discussed various possibilities, our group came to the conclusion that a compelling area of study and implementation of Machine Learning tools would be to apply them to the resolution of the current Dublin housing crisis. One of the reasons for this is that one of our members has a particularly detailed understanding of the domain and as a result of this, we were confident we would be able to address the issues relating to the Housing crisis using a regression model.

Our group consists of four members and we structured the group according to each member's individual skillset. It is our intention to garner experience in the initial phase through the allocation of roles and to rotate these roles in due course so that everyone can develop their skills in all roles and have the same experience during this project. Within this structure, all members are working in two roles simultaneously for Phase 1:

Phase 1		
Role	**Name**	**Task**
Project Manager	Giuliano Silva	Planning, executing, monitoring, controlling and closing projects
Business Analyst	Eric / Giuliano	Creating a detailed business analysis, outlining problems, opportunities and solutions for a business.
Communications Manager	Conor / Hasan	Ensuring that communications are sent, received, and (to the degree possible) understood.
Researcher	Conor / Eric / Hasan	Identifying trends and patterns, conducting fieldwork and tests when required

Table 1. Roles and tasks of each individual

It was our initial conviction that the issues of the current situation related most notably to the cost of construction. On this basis, we set about analysing four separate data sets:

- Ireland Census (2010 - 2020)
- National House Construction Cost Index (2010 - 2016)
- Marketing Information Annual Indices (1990 - 2018)
- Property Price Register for Ireland (2010 - 2018)

These datasets relate to house prices, the cost of construction in the capital city, market indices such as mortgage rates and consumer prices from 1990 to 2019. We also used data from the census department which allowed us to factor in population growth.

Having done so, it was our conclusion that the real issue was not the cost of construction because although it had increased over the years, it had not done so in a way that had adversely affected the purchase price of property in general. We also found through our studies that the mortgage rate for borrowing had decreased over the same period of time. Although house purchase prices have increased we believe that the issue of demand does not lie primarily in the private sector where our Exploration Data Analysis shows that prices have been rising. We found that the main issues relate to the fact that the primary demand for homes in Ireland is in areas where supply has not increased in the past ten years and that the manner in which the government has been subsidising and in general administering the construction of dwellings has been misguided.

It became apparent to us that the most important form of dwellings to first time buyers and lower income members of the populace looking to secure habitation are apartments and second hand dwellings coming onto the market based on our Exploration Data Analysis.

Under current initiatives, such as the House Assistance Payment (HAP) and Rental Accommodation Scheme (RAS) schemes, the government facilitates subsidized buy to rent schemes which facilitate landlords and a general property as profit generating asset philosophy.

Based on our references, Dr Rory Hearne from Maynooth University in his article from 2018, 'Why fixing Ireland's housing crisis requires a change of policy', *RTE*, affirms that if the government implemented not for profit construction projects led by local councils, as are found in continental European countries such as Denmark, then the current need for dwellings could be most effectively alleviated.

With this in mind we analysed various data sets that related to this subject matter and our initial idea was to test Dr Hearn’s hypothesis via Machine Learning as a means of predicting whether such an increase in the number of houses built would work and supply the growing demand in Ireland today. Unfortunately we could not find sufficient data with the right amount of information which was sufficiently diverse and applicable. As a result of this we decided to change direction towards the Private Market where we could analyze and apply Machine Learning processes. It was our conviction that we needed to find data which related sufficiently to housing prices and thus facilitate Machine Learning tools such as regression models.

With this in mind we looked at datasets which related to other cities such as Melbourne which were more proficient in supplying us with the data we needed. Our idea is to create a template model at first using a Melbourne dataset and figure out how we could generalize this model to apply to the Irish Housing Market.

1.1 Phase 1 Individual thoughts

1.1.1 Conor Sheehan

My initial thoughts were that the idea of trying to draw conclusions which might shed light on how to better administer the housing crisis would be a fascinating area of analysis not only because it would provide a compelling means through which Machine Learning tools could be implemented, but also because it would shed light on and issue of great social and economic significance.

I knew the subject would provide challenges but with my personal domain knowledge stemming from the fact that architecture is my area of work and the fact that I have lived in Dublin my whole life, I felt I would be able to contribute significantly to the group effort.

The main challenge facing the group was initially clarifying what the primary variable would be in terms of our primary focus. After initial analysis we became aware that our attention was best focused on government policies which then led us to look further afield in ensuring a compelling use of Machine Learning tools. Working in the team has been extremely satisfying not only as a result of each individual's commitment to the task, but also the diversity of skills and abilities each member brings to the group endeavour.

1.1.2 Eric Parfrey

My initial experience working on this group research project has been quite positive and enjoyable. Coming from a Physics background I have experience in academic research projects. My main role has been to ask the right research questions for our business problem and data set to ensure we are analysing the problem from all possible angles as well as analyzing the data in conjunction with Giuliano. I have enjoyed the group work so far and found it quite interesting to be paired with individuals that have varying backgrounds from myself and each bring their own sort of skills and expertise .

Initially we had looked at analysing the Irish housing crisis to propose solutions based off Machine Learning algorithms. This problem is of a personal nature as I've first hand experience with this issue as do my colleagues. After our analysis of the problem it became clear that we would not be able to propose solutions based on ML. From my understanding, the majority of solutions would base on changing government policy and would involve statistical analysis but not ML. Additionally we could not find readily available datasets to help with our analysis. Because of this I suggested we steer our efforts towards building a model that can predict the housing market as there are available datasets online. With the aim to apply the model to the Irish market at a later date.

1.1.3 Giuliano Silva

Participating in this project is very interesting because of our different backgrounds and the combination of our various skill sets has been a great experience. I have a degree in Economic Sciences with experience as a Financial Analyst and coder. As a result of this my main task in the group has been analyzing the databases and also administering as Project Manager guidance to all members what is necessary so that we are always able to fulfill what is indicated in our Gantt Chart.

Our initial proposal as described above was to try to analyze the Irish situation because one of our members has extensive knowledge in this subject but after our first analysis we encountered a lack of concise data which we could apply to Machine Learning models. As a result, we decided to try to create a model that would be generalized and if possible garner enough data from the next phases to apply for the Irish case.

1.1.4 Hasan Aziz

Participating in this project has so far been an excellent educational experience as I feel many of my pre-existing notions about the housing crisis such as how it began and particularly what could potentially put an end to it were challenged through research. Many of the already established policies the government have in place sound excellent on paper but when put through a thorough analysis, seem completely unnecessary and even at times appear to be an inadequate attempt to satisfy the public and cause them to overlook real solutions. Perhaps this intentional lack of understanding is due to the government needing more manpower, research and in some areas funds to truly tackle the issue at hand. I look forward to working on the coming phases, and building a better understanding of the subject.

1.2 Data Selection

Our choice to analyse the behavior of prices in the house market was through the Melbourne dataset available on the Kaggle website. The reason for this choice was that there were two datasets containing various information about the properties and it is our intention to explore all possible variables so that we can generalize our model.

These two datasets are called 'Full' and 'Less'. The 'Full' dataset contains 34857 rows and 21 columns and the 'Less' dataset contains 63023 rows and 13 columns.

The format with fewer columns contains the same columns as the 'Full' but does not have specific information. As a result of this we decided to first analyze the 'Full' format so that we can then answer our initial questions and also get to know the Australian housing market better by exploring all the features that really matter in order to predict the price as accurately as possible.

Our main focus is on social, logistical and economic factors which could have an influence on property prices in the central and greater Melbourne area. Based on our initial research of Melbourne we wanted to see if proximity to the Central Business District (CBD) had a definitive influence on property prices or if more nuanced logistical elements such as car parking spaces and bedroom numbers per property were of greater significance.

The Figure 1 above shows how Melbourne is distributed by Region:

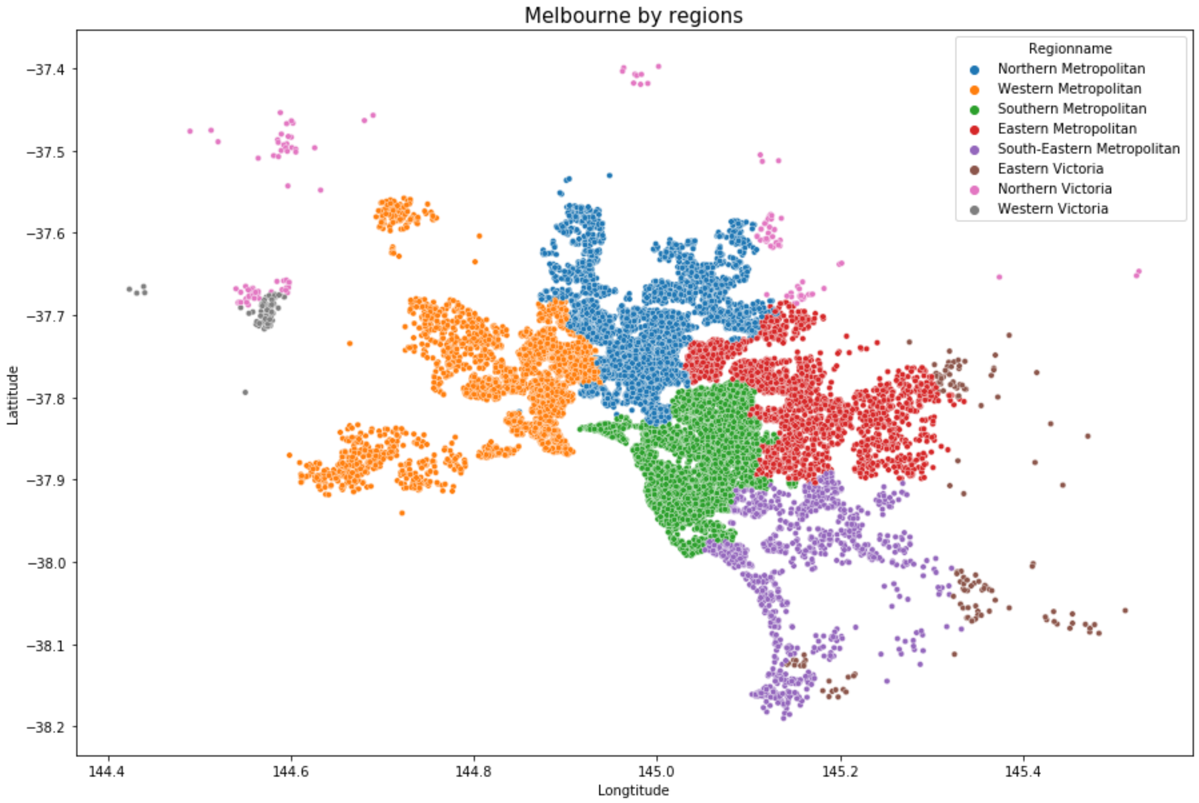


Figure 1. Melbourne by regions.

Greater Melbourne is where the South-Eastern Metropolitan area is located as well as the Eastern Metropolitan area. The Central Business District (CBD) is the heart of Melbourne where all the businesses are located as well as hospitality sectors such as beachfront restaurants, pubs and shops surrounded by suburban art. It is also where hi rise buildings and the Docklands are located.

2. Data Understanding

In our first EDA we will focus on the 'Full' dataset which contains 21 columns and 34,857 rows in total. Below is Table 1 which illustrates the type of variables and also the count of non-null values. As we mentioned before, the maximum count of non-null values is 34,857 using this we can consider which variables are containing missing values and consequently treat each one separately in order to understand them better. By analysing them separately we can investigate which approach would be most suitable to fill the missing values in with.

Data Dictionary

Data Item	Non-null count	Data Type	Description
Suburb	34,857	object	Suburb
Address	34,857	object	Address
Rooms	34,857	int	Number of rooms
Type	34,857	object	h: house,cottage,villa, semi,terrace u: unit, duplex; t: townhouse
Price	27,247	float	Price in Australian

			dollars

Method	34,857	object	S: property sold;
			SP: property sold prior;
			PI: property passed in;
			PN: sold prior not disclosed;
			SN: sold not disclosed;
			NB: no bid;
			VB: vendor bid;
			W: withdrawn prior to auction;
			SA: sold after auction;
			SS: sold after auction price not disclosed.
			N/A: price or highest bid not available.

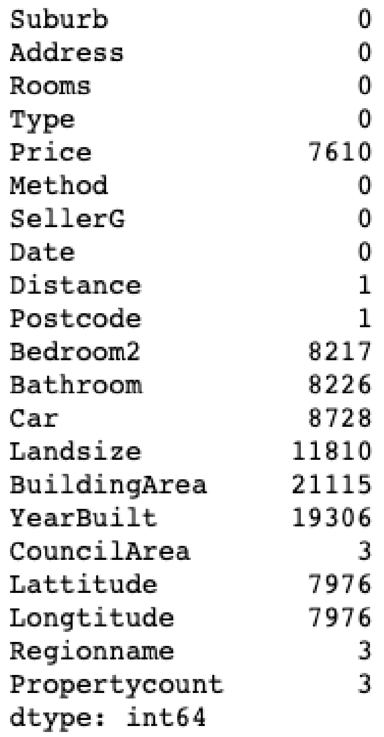
SellerG	34,857	object	Real State Agent
Date	34,857	object	Date Sold
Distance	34,856	float	Distance from CBD in Kilometers
Postcode	34,856	float	Postcode
Bedroom2	26,64	float	Scraped # of Bedrooms (from different source)
Bathroom	26,631	float	Number of Bathrooms
Car	26,129	float	Number of carspots
Land Size	23,047	float	Land Size in meters
Building Area	13,742	float	Building Size in meters

Year Built	15,551	float	Year the house was built
Council Area	34,854	object	Governing Council for the Area
Latitude	26,881	float	Self explanatory
Longitude	26,881	float	Self explanatory
Region Name	34,854	object	General Region (West, North West, North, North east …etc)
Property Count	34,854	float	Number of properties that exist in the Suburb

Table 2. Data Dictionary.

The data types are significant for our Data preparation stage and will require changes to be made to them to ensure our Data Preparation can work as intended. These changes will be made to variables that will need to be treated as objects. By making these changes to the data type of our variables we will see a positive impact on our analysis while gaining an accurate and insightful understanding of our data set. We will also make use of Feature Engineering to create new features from the data such as seasons and we plan to use the months to discover a trend of seasonality in the property sales.

Figure 2 shows a count of missing values in this dataset. We encountered a lot of missing values as see above in Table 2:

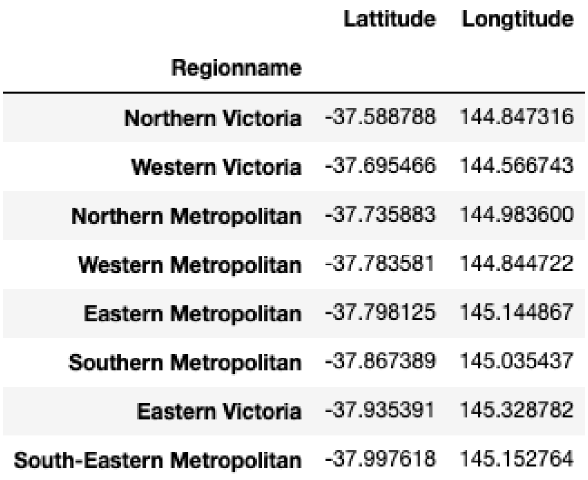


```
Suburb               0
Address              0
Rooms                0
Type                 0
Price             7610
Method               0
SellerG              0
Date                 0
Distance             1
Postcode             1
Bedroom2          8217
Bathroom          8226
Car               8728
Landsize         11810
BuildingArea     21115
YearBuilt        19306
CouncilArea          3
Lattitude         7976
Longtitude        7976
Regionname           3
Propertycount        3
dtype: int64
```

Figure 2. Count of missing values.

To handle our missing values throughout our data we will split & sort by region name, to calculate the mean values for the variable within that specific region. We will see if this method will work to complete our imputation. Before imputation we will assess the data sorted by region name, as imputation methods need to be handled carefully.

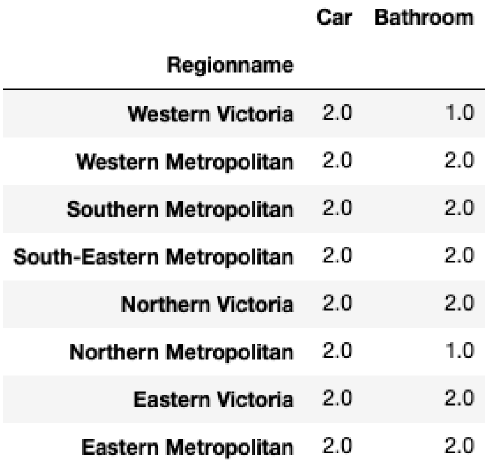
Figure 3 shows the mean values for Latitude & Longitude by each region. Knowing that those values reflect the average of the Latitude and Longitude of the region it is our intention to see if they will be useful to our Data Visualization as shown in Figure 1 and later for our model in seeing if they have an impact on predicting price.



Regionname	Lattitude	Longtitude
Northern Victoria	-37.588788	144.847316
Western Victoria	-37.695466	144.566743
Northern Metropolitan	-37.735883	144.983600
Western Metropolitan	-37.783581	144.844722
Eastern Metropolitan	-37.798125	145.144867
Southern Metropolitan	-37.867389	145.035437
Eastern Victoria	-37.935391	145.328782
South-Eastern Metropolitan	-37.997618	145.152764

Figure 3. Mean of Latitude and Longitude by region name.

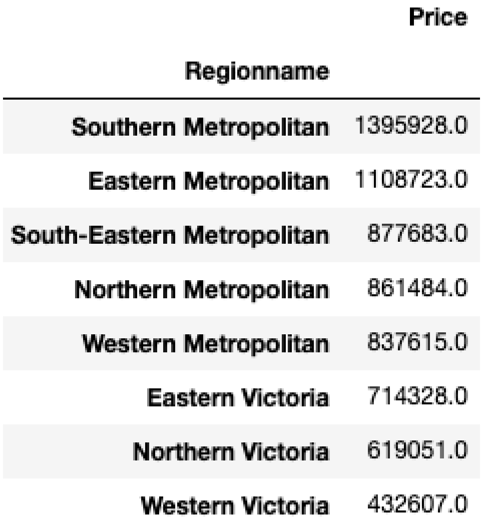
Figure 4 shows the mean of the number of car parking spaces and bathrooms for each property by each region. As we can see all the regions are very similar in these two aspects.



Regionname	Car	Bathroom
Western Victoria	2.0	1.0
Western Metropolitan	2.0	2.0
Southern Metropolitan	2.0	2.0
South-Eastern Metropolitan	2.0	2.0
Northern Victoria	2.0	2.0
Northern Metropolitan	2.0	1.0
Eastern Victoria	2.0	2.0
Eastern Metropolitan	2.0	2.0

Figure 4. Mean of Car park and Bathroom by region name

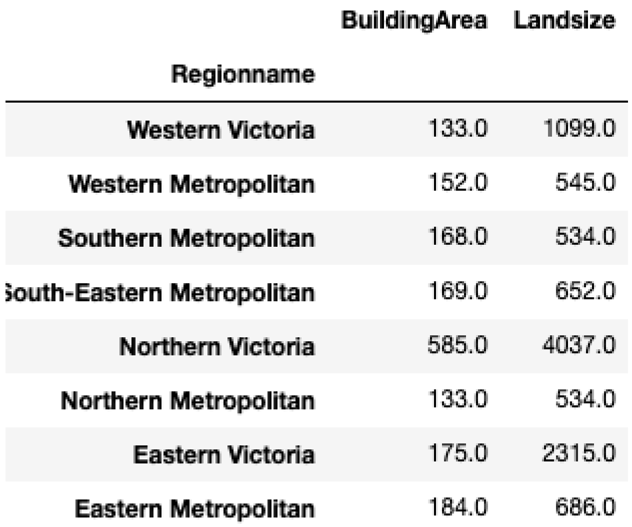
Price is our target variable and as a result, it is our intention to understand the average price of each region. Figure 5 shows us that the Southern Metropolitan area followed by the Eastern Metropolitan area are the most expensive regions based on their means. We showed in Figure 1 that these regions are also part of the more built up and serviced areas in Melbourne. One of the reasons for this is that the Central Businesses District and Docklands are located in this region. The presence of hi rise buildings and all the beachfront restaurants, pubs and shops also indicate a reason for the high price. Conversely, Western Victoria prices are cheaper and checking Figure 1, we can see that the region is further away and separated from those urban areas that are more populated.



	Price
Regionname	
Southern Metropolitan	1395928.0
Eastern Metropolitan	1108723.0
South-Eastern Metropolitan	877683.0
Northern Metropolitan	861484.0
Western Metropolitan	837615.0
Eastern Victoria	714328.0
Northern Victoria	619051.0
Western Victoria	432607.0

Figure 5. Mean of Price by region name.

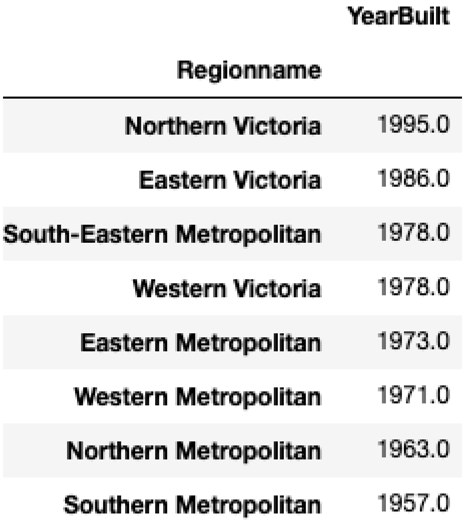
Figure 6 shows the mean values of building area and land size by region for each property. As expected the metropolitan areas have a low land size compared to the other regions. The mean value for the building area is mostly the same for each region bar Northern Victoria.



	BuildingArea	Landsize
Regionname		
Western Victoria	133.0	1099.0
Western Metropolitan	152.0	545.0
Southern Metropolitan	168.0	534.0
South-Eastern Metropolitan	169.0	652.0
Northern Victoria	585.0	4037.0
Northern Metropolitan	133.0	534.0
Eastern Victoria	175.0	2315.0
Eastern Metropolitan	184.0	686.0

Figure 6. Mean of Building Area and Land Size by region name.

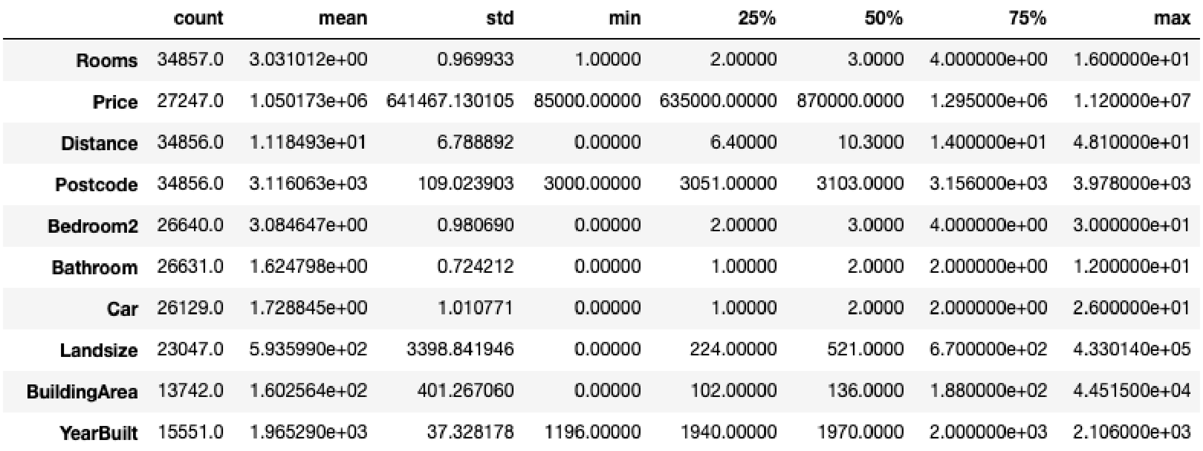
Finally, we analysed the Year built based on the mean as shown in Figure 7. The average year that the houses in Northern and Eastern Victoria were built was very low compared to other areas so it is possible that they became established as the newest regions in Melbourne as the population grew and as a result the city started to spread out. Looking again at Figure 6 and comparing to Figure 7, we could make the assumption that the regions Northern Victoria & Eastern Victoria are the affluent suburbs due to high land size and high building area. Using Figure 5, the lowest average prices are the regions mentioned and Western Victoria, from this basic analysis, land size and building area may not have as much impact on price predictions as we had previously thought. The main influence that we can see from Figures 5, 6 & 7 on the price is the properties proximity to the central business district / metropolitan areas of inner city Melbourne.



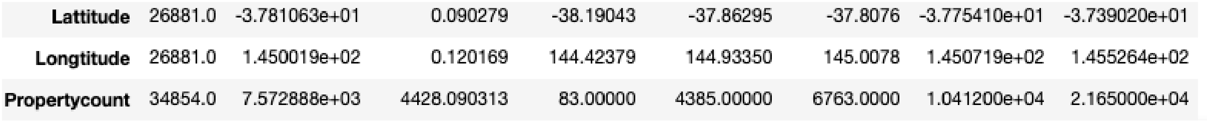
	YearBuilt
Regionname	
Northern Victoria	1995.0
Eastern Victoria	1986.0
South-Eastern Metropolitan	1978.0
Western Victoria	1978.0
Eastern Metropolitan	1973.0
Western Metropolitan	1971.0
Northern Metropolitan	1963.0
Southern Metropolitan	1957.0

Figure 7. Mean of Year Built by region name.

Another check that is required is to view all the statistical metrics such as the percentiles, count, mean and standard deviation of each numerical variable. Using this information we can easily see how the data is distributed, if there's outliers, if there's zero values, or if there is any discrepancy in the dataset as shown in Figure 8.



	count	mean	std	min	25%	50%	75%	max
Rooms	34857.0	3.031012e+00	0.969933	1.00000	2.00000	3.0000	4.000000e+00	1.600000e+01
Price	27247.0	1.050173e+06	641467.130105	85000.00000	635000.00000	870000.0000	1.295000e+06	1.120000e+07
Distance	34856.0	1.118493e+01	6.788892	0.00000	6.40000	10.3000	1.400000e+01	4.810000e+01
Postcode	34856.0	3.116063e+03	109.023903	3000.00000	3051.00000	3103.0000	3.156000e+03	3.978000e+03
Bedroom2	26640.0	3.084647e+00	0.980690	0.00000	2.00000	3.0000	4.000000e+00	3.000000e+01
Bathroom	26631.0	1.624798e+00	0.724212	0.00000	1.00000	2.0000	2.000000e+00	1.200000e+01
Car	26129.0	1.728845e+00	1.010771	0.00000	1.00000	2.0000	2.000000e+00	2.600000e+01
Landsize	23047.0	5.935990e+02	3398.841946	0.00000	224.00000	521.0000	6.700000e+02	4.330140e+05
BuildingArea	13742.0	1.602564e+02	401.267060	0.00000	102.00000	136.0000	1.880000e+02	4.451500e+04
YearBuilt	15551.0	1.965290e+03	37.328178	1196.00000	1940.00000	1970.0000	2.000000e+03	2.106000e+03



Lattitude	26881.0	-3.781063e+01	0.090279	-38.19043	-37.86295	-37.8076	-3.775410e+01	-3.739020e+01
Longtitude	26881.0	1.450019e+02	0.120169	144.42379	144.93350	145.0078	1.450719e+02	1.455264e+02
Propertycount	34854.0	7.572888e+03	4428.090313	83.00000	4385.00000	6763.0000	1.041200e+04	2.165000e+04

Figure 8. Central Tendency Measures for the numerical variables

We can easily spot that in Year Built we got a value of 2106 instead of 2016 so it must be a situation where someone typed the word incorrectly. As a result we will need to fix this value in our Data Preparation section. Another thing of note is the amount of zeros contained in numerical variables such as Bathroom, Bedroom2, Building Area, Car, Distance and Land Size. We will probably need to ignore these zero values because they will affect our analysis as it is most likely safe to conclude that it is very rare to find a house that doesn't have a bathroom or bedroom.

As we are going to test Regression models on the data we need to transform the categorical variables into numerical ones. This demands that we make sure that we convert the right variables thus avoiding the creation of thousands of new features from the unique categorical values there are in the variables. Table 3 shows us that we have a lot of classes for the same variables which means that we need to make decisions about each variable we are going to use. Address, Suburb and Postcode may be telling us the same thing so once we get to the modelling stage we will be able to have a better understanding of the features.

Number of unique classes for categorical variables 'Full' dataset		
Data Item	**Number of unique classes**	**Data Type**
Suburb	351	object
Address	34,004	object
Type	3	object

Method	9	object
SellerG	388	object
Postcode	211	object
Council Area	33	object
Region Name	8	object

Table 3. Number of unique classes for categorical variables.

2.1 Exploration Data Analysis

This dataset is telling us that some regions tend to be more expensive than the others. It could be for multiple reasons including those that relate to how the city of Melbourne itself was built. Figure 9 confirms exactly what we suspected in our first section. Southern Metropolitan and South-Eastern Metropolitan are the regions where the prices are higher. Additionally however we can also see that the prices vary a lot in the regions which means that they have a mix of house types and prices. Indicating that there are other variables that have a strong impact on the price.

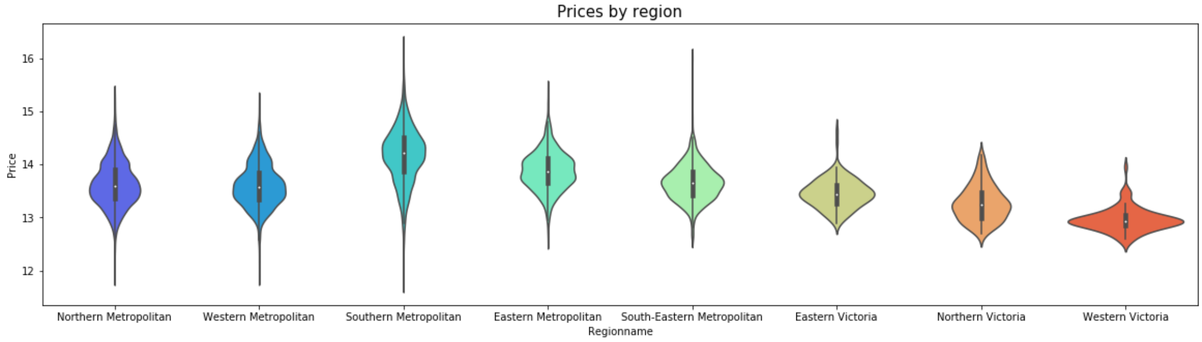


Figure 9. Distribution of prices across regions

Another way to visualize the price distribution by region is to plot a map of Melbourne and use price as an additional variable. As we have seen, the heart of Melbourne is located close to the Docklands which is the most affluent area in the city. Figure 15 shows us just what we discussed earlier which is a situation where houses with a higher price are concentrated in the same area namely, the Southern Metropolitan and South-Eastern Metropolitan areas.

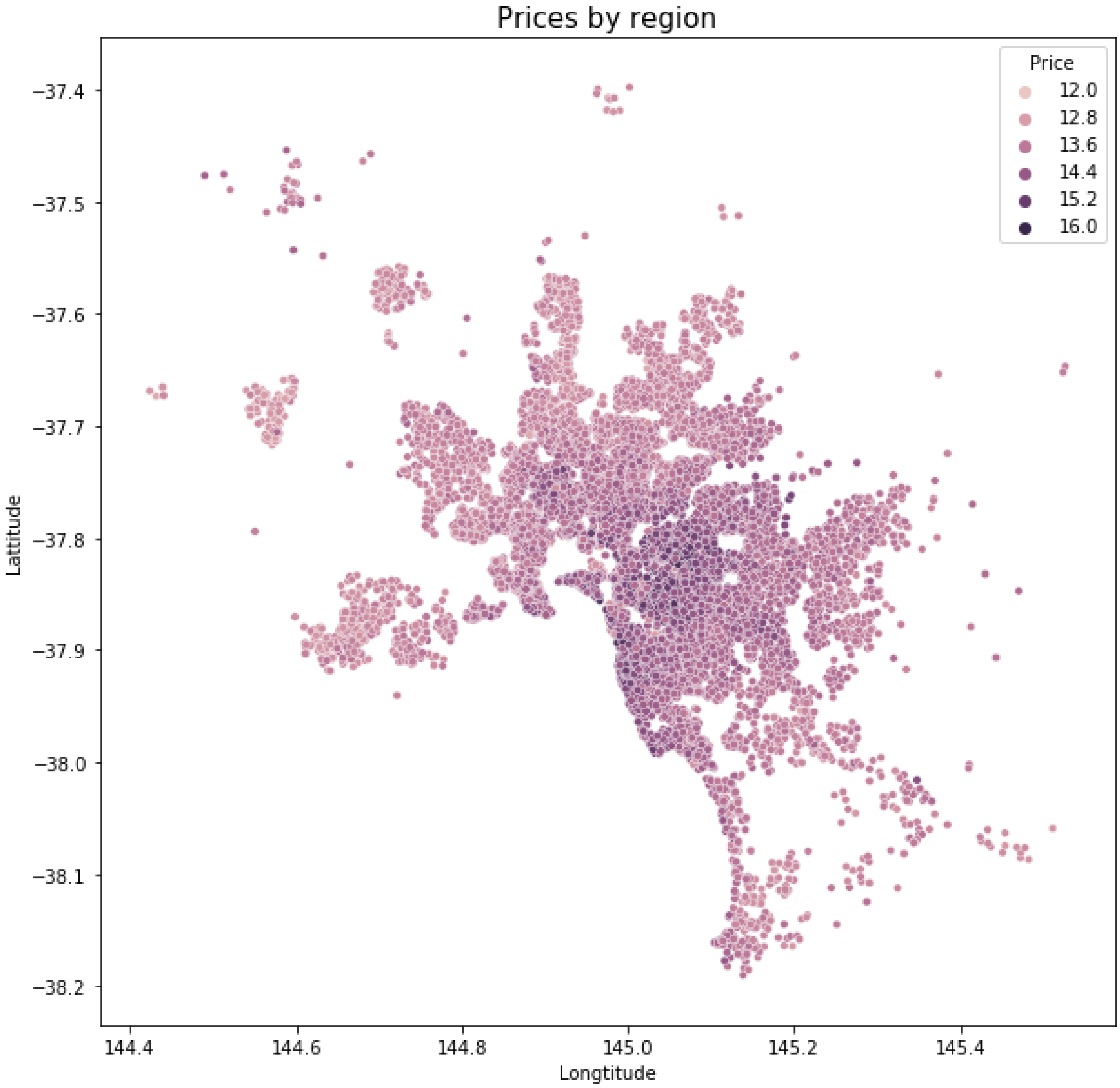


Figure 10. Prices by region.

The property types present in this dataset are divided into three classes. Figure 10 shows us that the distribution of these types are not normally distributed and most of them are concentrated into the house, cottage, village, semi and terrace types which are denominated as 'h'.

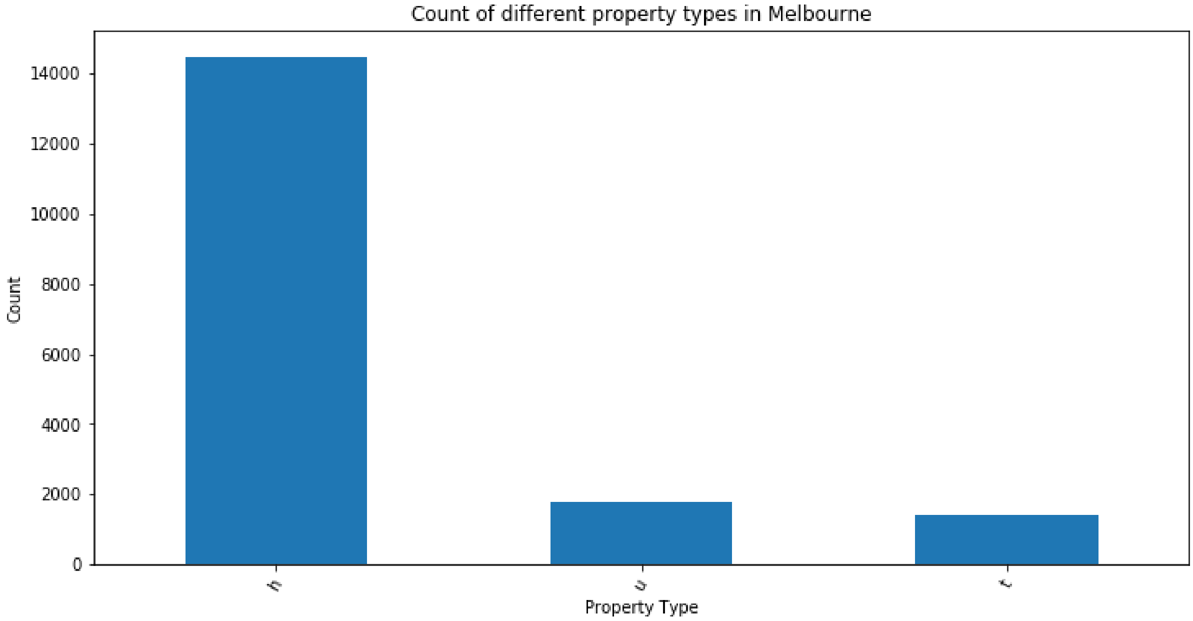


Figure 11. Count of different property types in Melbourne.

Figure 12 shows the distribution of properties by region. Eastern Victoria is followed by Northern Metropolitan as the most populated region in Melbourne based on this metric and are highlighted in brown & green respectively. From Figure 5 it can be seen that Eastern Victoria is the most expensive of the non urban areas while Northern Metropolitan is the second least expensive metropolitan area. From this initial analysis, Northern Metropolitan has the best ratio for price vs closeness to the CBD and that may be a reason for its high property count.

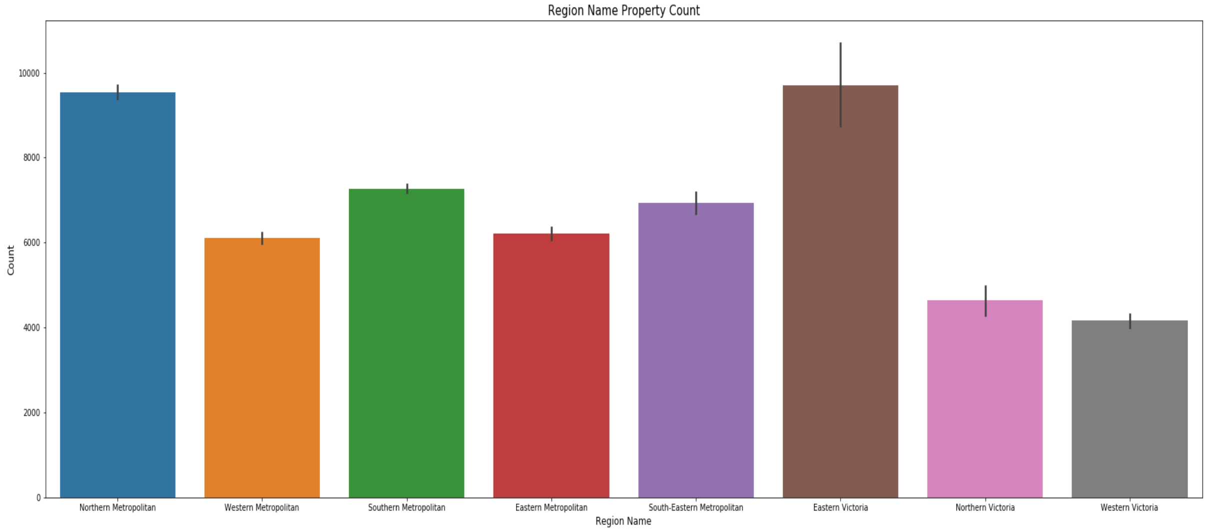


Figure 12. Region name property count.

When we talk about seasonality of sales, some types of business may get good sales in summer more than other seasons for example. Therefore, we are going to use a Boxplot to see if there's any sort of seasonality in our dataset as shown in Figure 13.

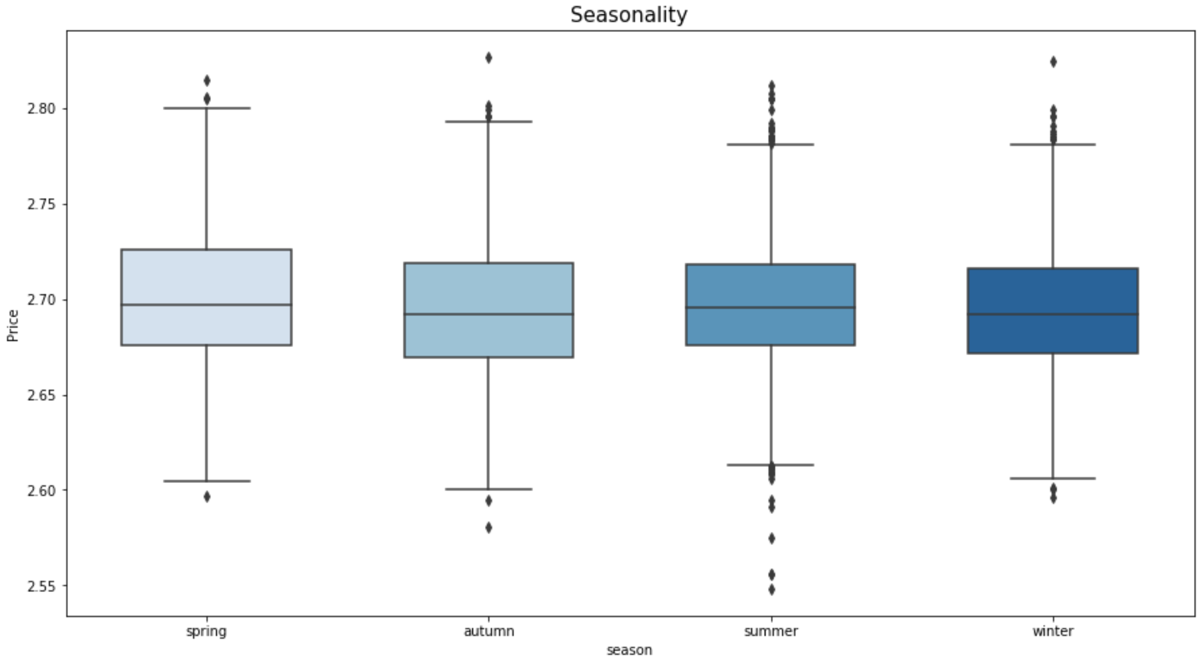


Figure 13. Seasonality.

All the seasons seem to be very constant and don’t have different types of Price. One thing which is clear is that for example, in the summer season we got more outliers than in other seasons. These outliers tell us that price can move towards a lower range. Based on the information provided, we can conduct further analysis but we may not use it in our model. It is very interesting to see that there's no evidence of seasonality and it appears the Housing Market in Melbourne is robust all seasons of the year. We analysed the years as well to see if there is some change over the course of time but there did not seem to be a change in the behaviour of the data.

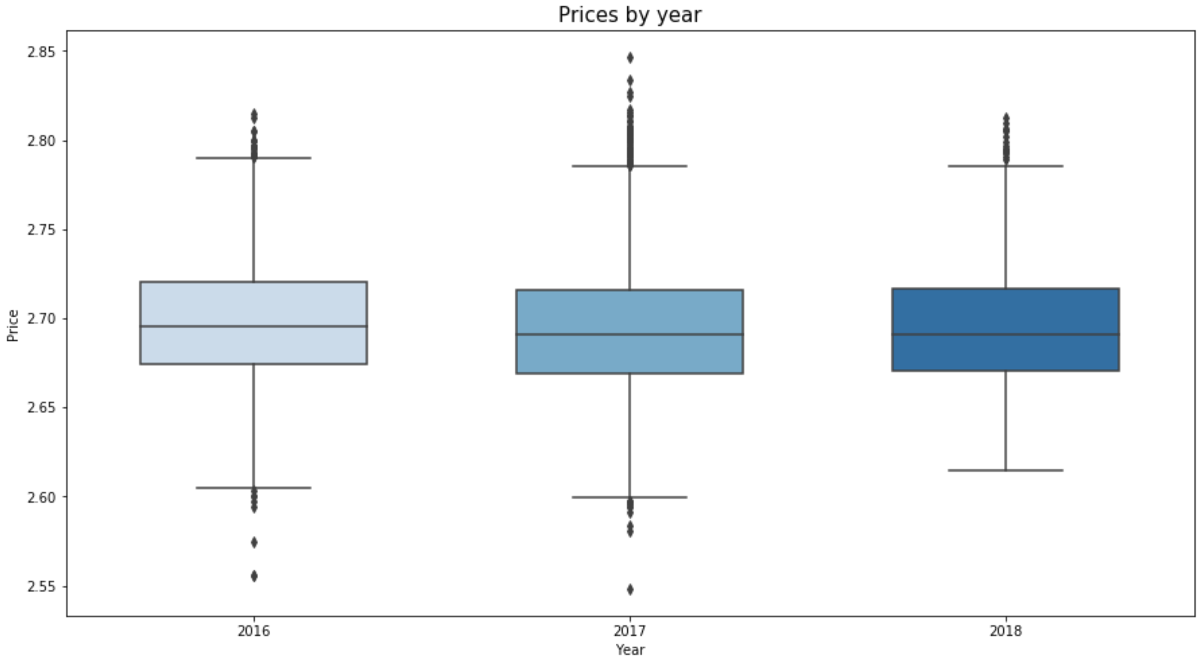


Figure 14. Prices by year.

The scatterplot below shows us a relationship between Price and number of rooms which makes sense because as the room number increases, the price increases which leads us to conclude that having more rooms means you probably would need a bigger building area which in turn means you would possibly have to spend more money on the construction of the property.

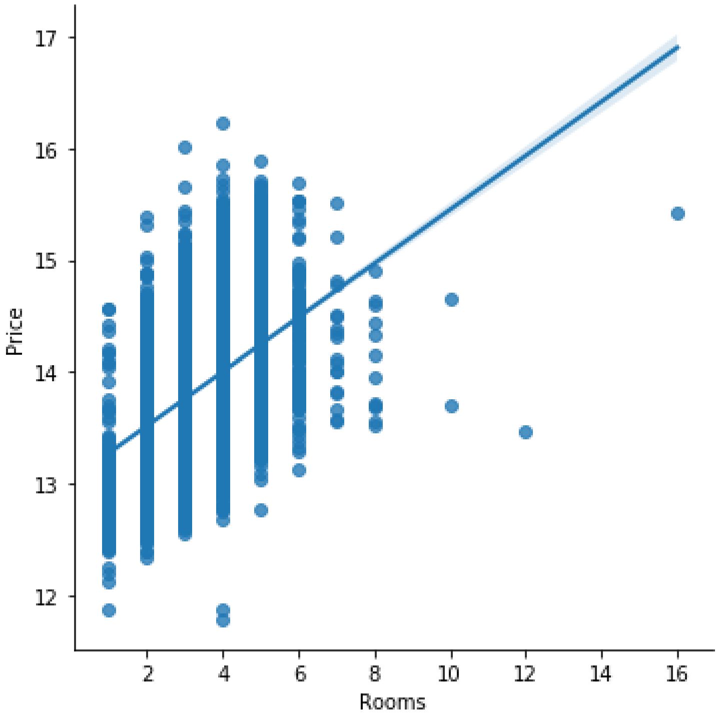


Figure 15. Scatterplot Price x Rooms

.

The same applies to Bathrooms as we saw in Figure 16:

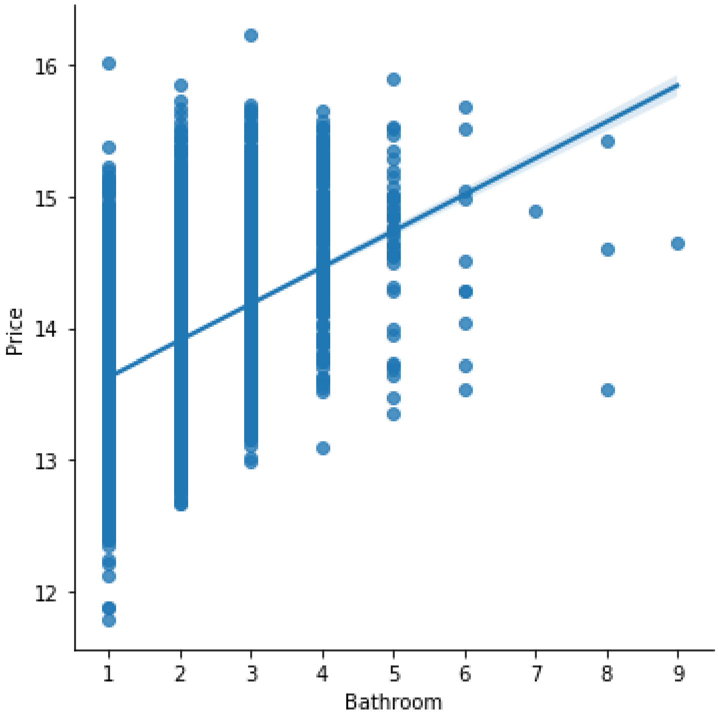


Figure 16. Scatterplot Price x Bathroom

House age doesn't seem to be correlated to the price which means that at this first stage of the analysis we can't draw conclusions on this variable.

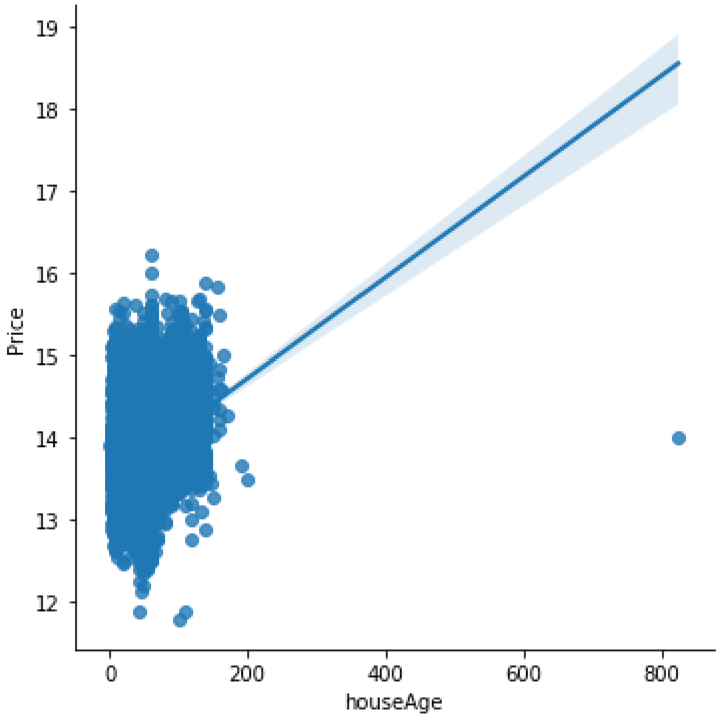


Figure 17. Scatterplot Price x House age

The number of carspots in the properties seems to have an influence on the price as the scatterplot shows us a straight line shown in Figure 18:

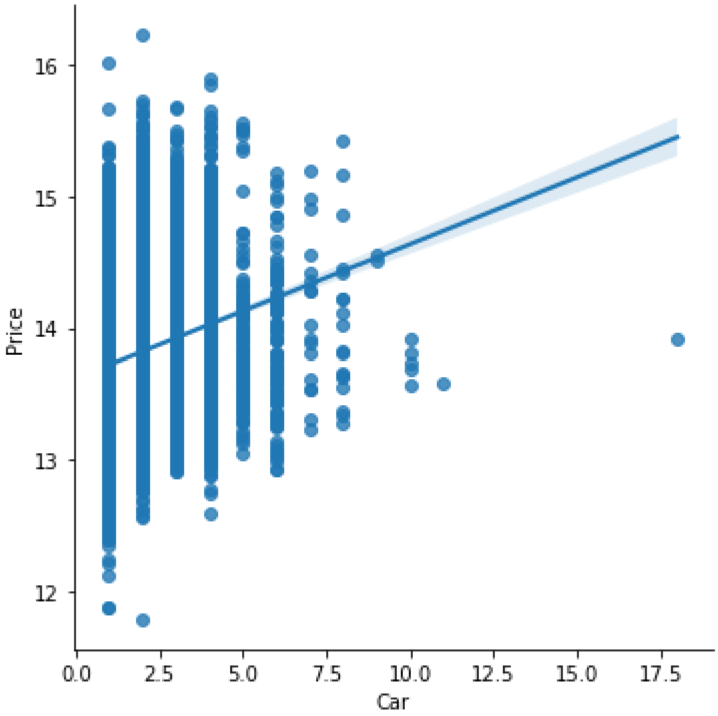
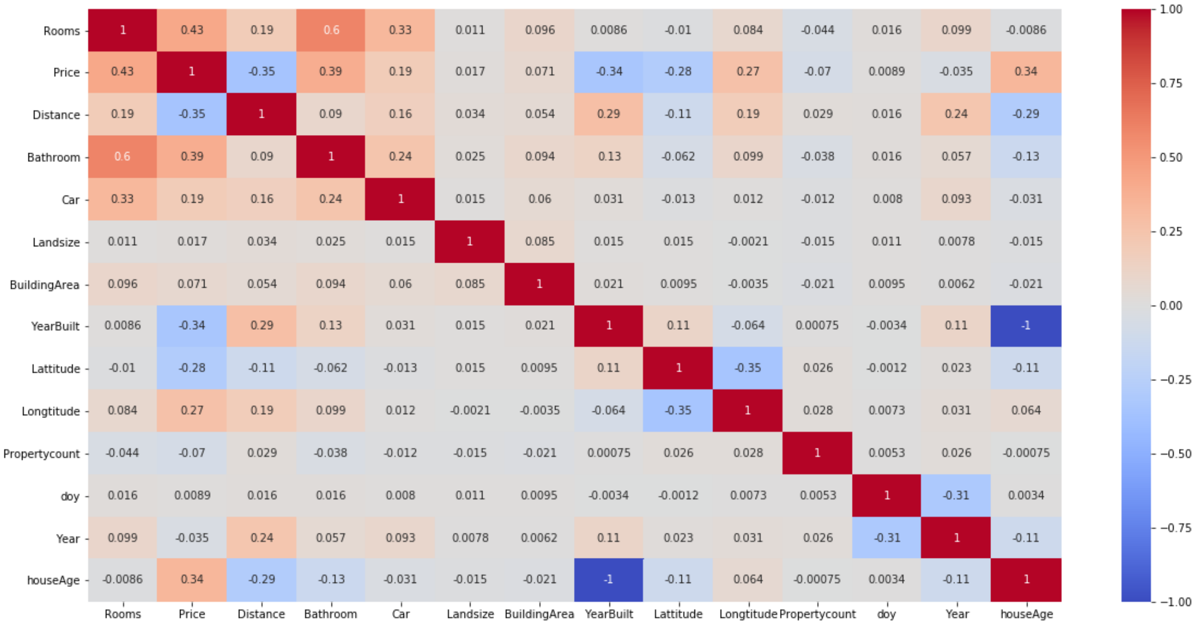


Figure 18. Scatterplot Price x Car

After exploring the variables independently, we are going to analyse the correlation between all the variables. The Multiple Linear Regression assumption says that there is no presence of multicollinearity and through this visualization map we can easily see that this is true. Year Built has a strong correlation to House Age which makes sense because it was established from it. All the other variables seem to not have multicollinearity which is good and as a result, we can go ahead with our final analysis.



	Rooms	Price	Distance	Bathroom	Car	Landsize	BuildingArea	YearBuilt	Lattitude	Longtitude	Propertycount	doy	Year	houseAge
Rooms	1	0.43	0.19	0.6	0.33	0.011	0.096	0.0086	-0.01	0.084	-0.044	0.016	0.099	-0.0086
Price	0.43	1	-0.35	0.39	0.19	0.017	0.071	-0.34	-0.28	0.27	-0.07	0.0089	-0.035	0.34
Distance	0.19	-0.35	1	0.09	0.16	0.034	0.054	0.29	-0.11	0.19	0.029	0.016	0.24	-0.29
Bathroom	0.6	0.39	0.09	1	0.24	0.025	0.094	0.13	-0.062	0.099	-0.038	0.016	0.057	-0.13
Car	0.33	0.19	0.16	0.24	1	0.015	0.06	0.031	-0.013	0.012	-0.012	0.008	0.093	-0.031
Landsize	0.011	0.017	0.034	0.025	0.015	1	0.085	0.015	0.015	-0.0021	-0.015	0.011	0.0078	-0.015
BuildingArea	0.096	0.071	0.054	0.094	0.06	0.085	1	0.021	0.0095	-0.0035	-0.021	0.0095	0.0062	-0.021
YearBuilt	0.0086	-0.34	0.29	0.13	0.031	0.015	0.021	1	0.11	-0.064	0.00075	-0.0034	0.11	-1
Lattitude	-0.01	-0.28	-0.11	-0.062	-0.013	0.015	0.0095	0.11	1	-0.35	0.026	-0.0012	0.023	-0.11
Longtitude	0.084	0.27	0.19	0.099	0.012	-0.0021	-0.0035	-0.064	-0.35	1	0.028	0.0073	0.031	0.064
Propertycount	-0.044	-0.07	0.029	-0.038	-0.012	-0.015	-0.021	0.00075	0.026	0.028	1	0.0053	0.026	-0.00075
doy	0.016	0.0089	0.016	0.016	0.008	0.011	0.0095	-0.0034	-0.0012	0.0073	0.0053	1	-0.31	0.0034
Year	0.099	-0.035	0.24	0.057	0.093	0.0078	0.0062	0.11	0.023	0.031	0.026	-0.31	1	-0.11
houseAge	-0.0086	0.34	-0.29	-0.13	-0.031	-0.015	-0.021	-1	-0.11	0.064	-0.00075	0.0034	-0.11	1

Figure 19. Correlation map

3. Data Preparation

3.1 Data Cleaning / Feature Engineering

It is our objective in this section to elaborate on all the data preparation phases that have been used and in doing so, clarify the data analysis encountered in the previous chapter. The first step is to clean the null values or alternatively to apply an imputation method which we will then use to do feature engineering. We will split the date into months and years in order to see how old the house is from today. The first section is very important because we have done a thorough analysis up to now and at this point we have all the understanding of the dataset we need to see what is the best option moving forward.

Our target variable Price will not be applied to any imputation method because we tried to use the mean Price by region but it did not prove conclusive. The distribution manifested around the same region and as a result the distribution was not ideal for our purposes. One of the assumptions needed for Multiple Linear regression is a normal distribution. As a result, we decided to not continue with this imputation method for our target variable. We still however need to make some adjustments to its distribution as the Figure 20 shows us:

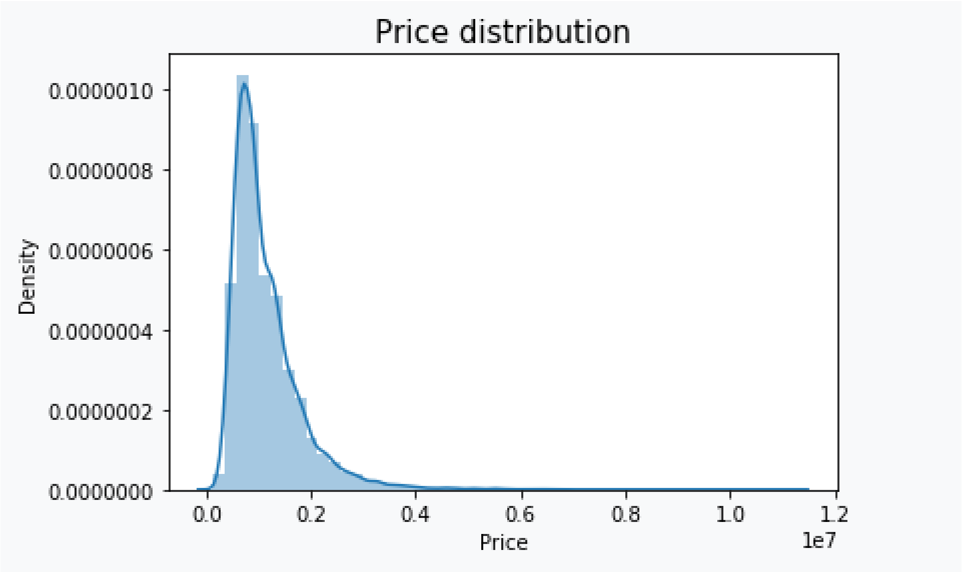


Figure 20. Original price distribution

As seen above the price distribution still has some positive skewness and in order to create a model that follows Multiple Linear Regression assumptions, we need to be able to shape the curve so that it is as normally distributed as possible. A natural logarithm can be used to transform the data and in doing so bring its shape as close as possible to that of a normal distribution or alternatively, to achieve a bell shape. One of the reasons that we chose to use this technique was to simplify the model but we also used it to deal with outliers and in doing so achieved a more desirable shape. As a result of this the results from this analysis had greater validity. Consequently, the log transformation reduces or removes the skewness of the continuous variable which can be seen in Figure 20. The first step that we took to complete the transformation was to firstly use the original Price distribution and fit the 'bell shape' to it to see how it would look. As we can see in Figure 21 it has changed a lot from the original and has a more flattened peak.

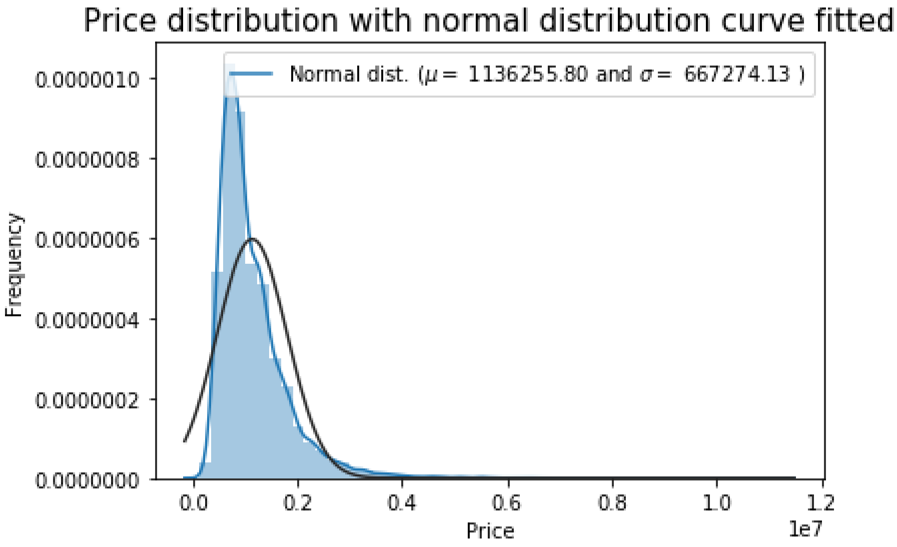


Figure 21. Normal distribution curve fitted.

The second step was to analyse the Probability Plot from the original distribution. This plot generates a probability plot of price against the quantiles of a specified theoretical distribution such as normal or exponential; in this case the normal distribution. In Figure 22, it does not have the desired appearance and does not seem to fit the line when we try to use the original distribution of the price variable.

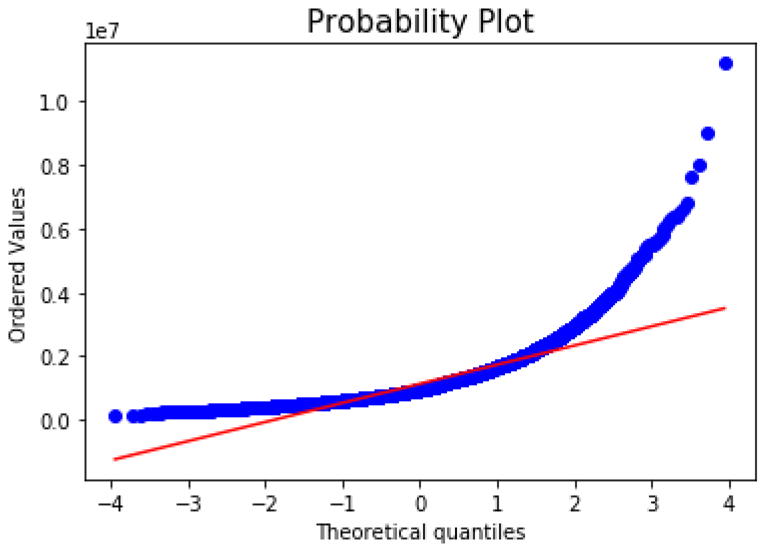


Figure 22. Probability plot

The final step was to apply the natural logarithm to the price variable. Figure 23, shows us its new distribution which looks much better and closer than the original one to a normal distribution. The positive skewness that we got initially has been removed.

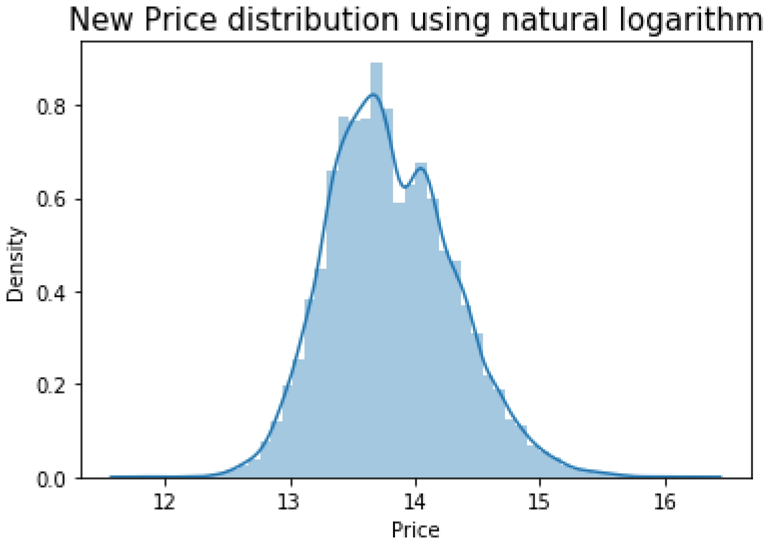


Figure 23. New distribution using natural logarithm

As expected the new Probability Plot after the log transformation is a straight line as shown in Figure 24 above:

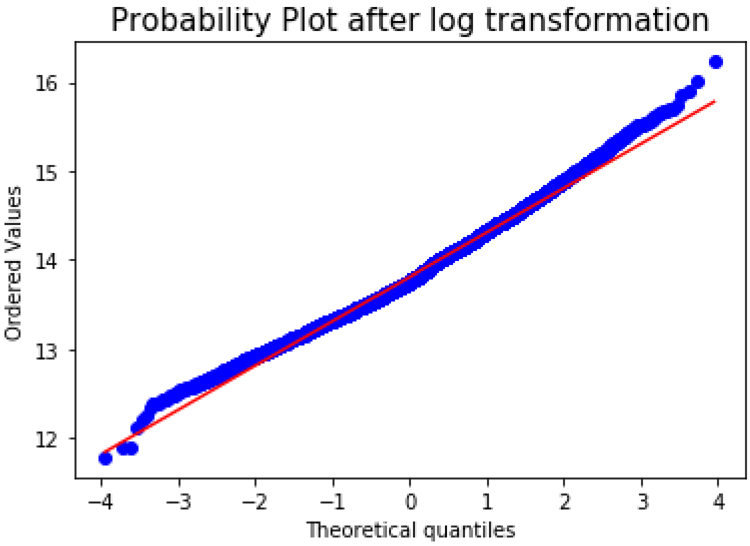
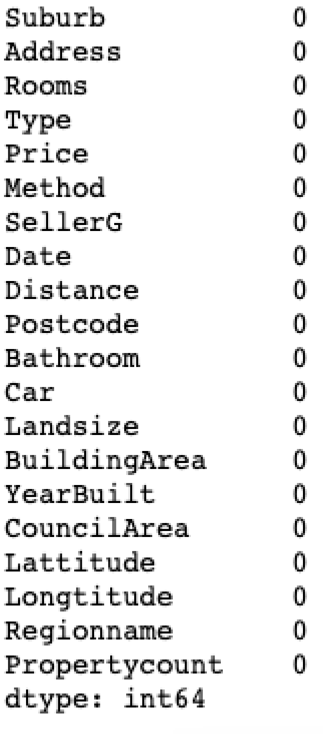


Figure 24. Probability plot after log transformation

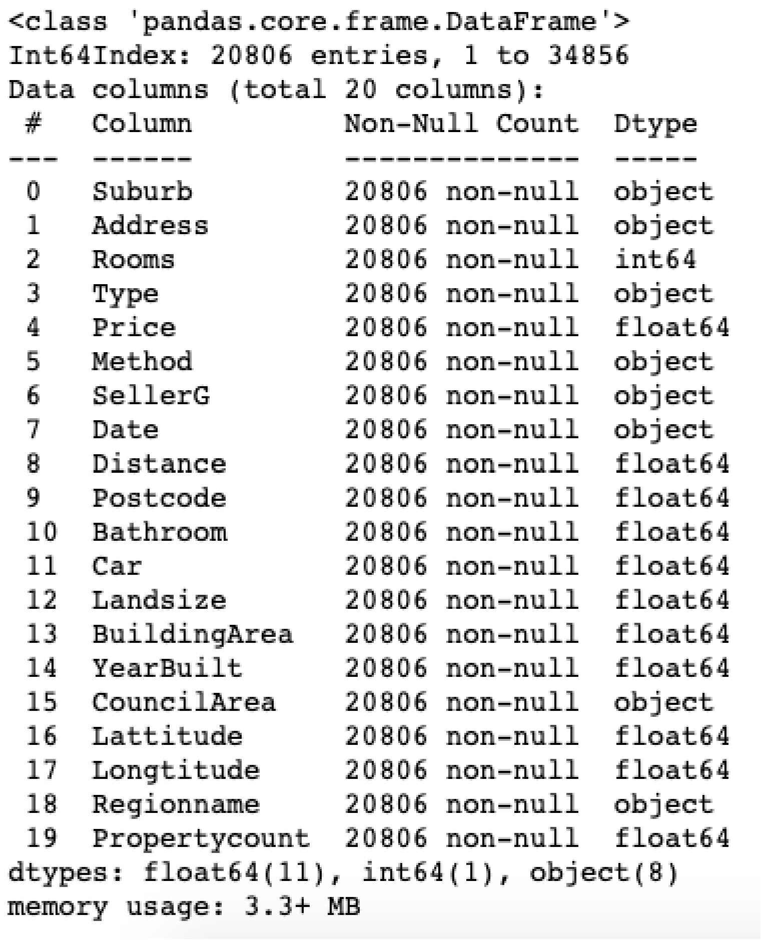
For all the other variables that have missing values we used an imputation method based on analysis that we presented by region. The way that we came upon this method was to use a loop method in order to fill out all the missing values. After implementing this imputational method we still had some missing values that we could make use of so we decided to drop them because they wouldn't add much to our analysis once we had the main features completed. We agreed to drop the column Bedroom2 because as we have Rooms in our dataset and it would be redundant to carry this variable further. Figure 25 shows the end results after completing the imputation.



```
Suburb           0
Address          0
Rooms            0
Type             0
Price            0
Method           0
SellerG          0
Date             0
Distance         0
Postcode         0
Bathroom         0
Car              0
Landsize         0
BuildingArea     0
YearBuilt        0
CouncilArea      0
Lattitude        0
Longtitude       0
Regionname       0
Propertycount    0
dtype: int64
```

Figure 25. Count of missing values after cleaning.

As we mentioned previously in Page 16, we will now need to change the data types.



```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 20806 entries, 1 to 34856
Data columns (total 20 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Suburb         20806 non-null  object
 1   Address        20806 non-null  object
 2   Rooms          20806 non-null  int64
 3   Type           20806 non-null  object
 4   Price          20806 non-null  float64
 5   Method         20806 non-null  object
 6   SellerG        20806 non-null  object
 7   Date           20806 non-null  object
 8   Distance       20806 non-null  float64
 9   Postcode       20806 non-null  float64
 10  Bathroom       20806 non-null  float64
 11  Car            20806 non-null  float64
 12  Landsize       20806 non-null  float64
 13  BuildingArea   20806 non-null  float64
 14  YearBuilt      20806 non-null  float64
 15  CouncilArea    20806 non-null  object
 16  Lattitude      20806 non-null  float64
 17  Longtitude     20806 non-null  float64
 18  Regionname     20806 non-null  object
 19  Propertycount  20806 non-null  float64
dtypes: float64(11), int64(1), object(8)
memory usage: 3.3+ MB
```

Figure 26. Data types after cleaning.

Based on Figure 26 we need to transform the variables that we might use in our modelling section. Bathroom, Car, Property Count and Year Built will be transformed into integer types. Postcodes will be transformed into objects. Table 4 summarizes our actions.

Data type transformations		
Data Item	**Data type**	**Converted to**
Postcode	object	integer

Property Count	float	integer
YearBuilt	float	integer
Bathroom	float	integer
Car	float	integer

Table 4. Data type transformations.

Before we start feature engineering we need to filter those zero values that we found analysing the Central Tendency and we also need to fix the Year Built value that tells us 2106 instead of 2016.

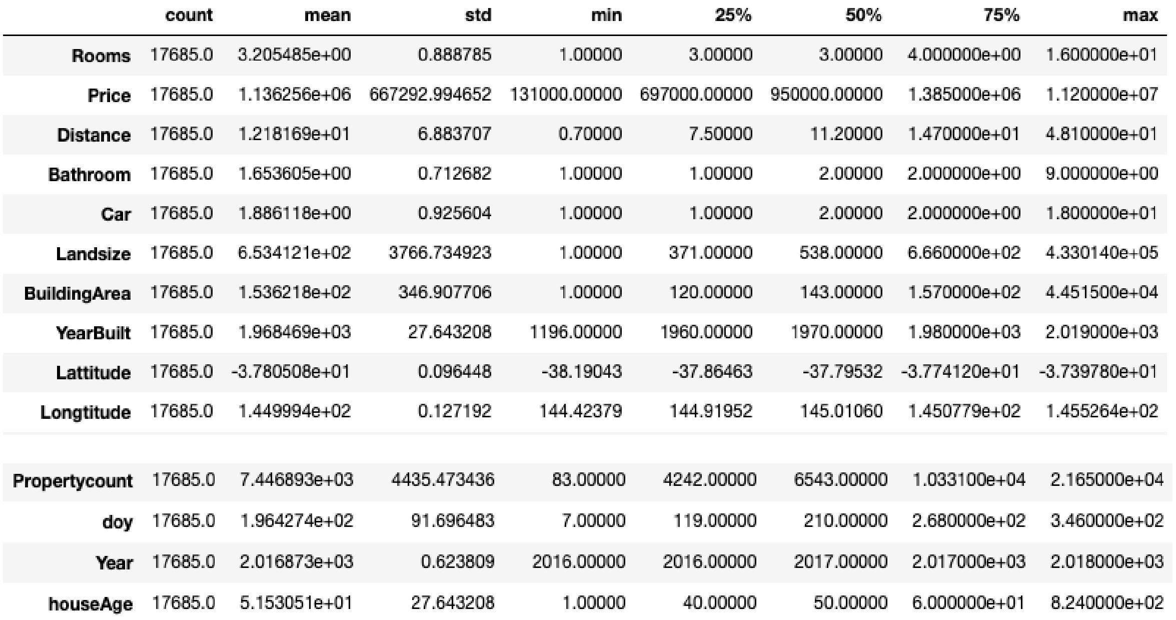
Lastly, we are going to create a new column called 'Data' which will be based on the 'Date' column but in this case we are transforming it as date time because by doing so we can extract the year and month. Days of the year will be an important variable for us as well because from that we can use information to define the seasons of the year such as summer, autumn, winter and spring. This will further allow us to understand if the seasonality affects the property sales. By subtracting the actual year from the Year Built we can get the House age which will be the new feature we will use instead of Year Built as shown in Table 5 below:

Feature Engineering			
New data	**New Data type**	**From**	**Original Data type**
Data	date/time	Date	object
Doy (days of year)	integer	Data	object
Year	date/time	Data	object

Season	object	Doy (days of year)	integer
House age	integer	2020 - Year Built	integer

Table 5. Feature Engineering summary.

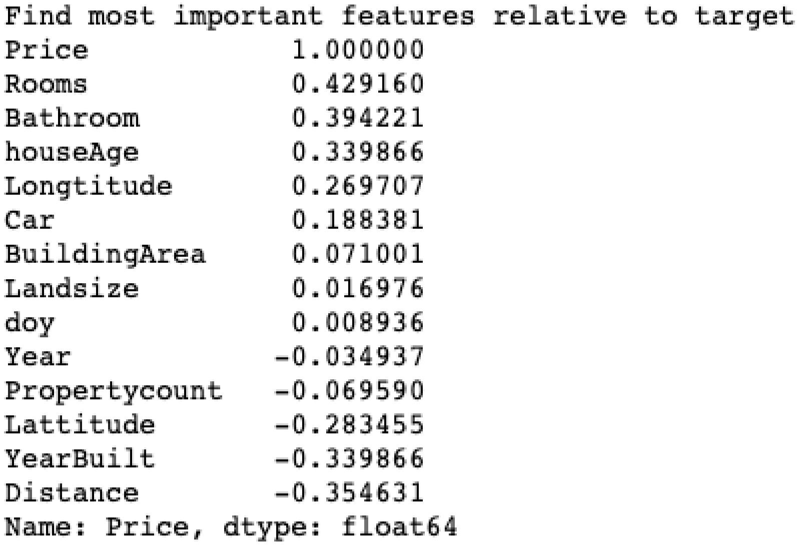
After implementing these steps we have our dataset ready to be manipulated. The results of our actions and transformations can be seen in Figure 27 which displays the Central Tendency of each numerical variable. Those zero values are no longer evident in this dataset and there is no evidence of the error typing in Year Built.



	count	mean	std	min	25%	50%	75%	max
Rooms	17685.0	3.205485e+00	0.888785	1.00000	3.00000	3.00000	4.000000e+00	1.600000e+01
Price	17685.0	1.136256e+06	667292.994652	131000.00000	697000.00000	950000.00000	1.385000e+06	1.120000e+07
Distance	17685.0	1.218169e+01	6.883707	0.70000	7.50000	11.20000	1.470000e+01	4.810000e+01
Bathroom	17685.0	1.653605e+00	0.712682	1.00000	1.00000	2.00000	2.000000e+00	9.000000e+00
Car	17685.0	1.886118e+00	0.925604	1.00000	1.00000	2.00000	2.000000e+00	1.800000e+01
Landsize	17685.0	6.534121e+02	3766.734923	1.00000	371.00000	538.00000	6.660000e+02	4.330140e+05
BuildingArea	17685.0	1.536218e+02	346.907706	1.00000	120.00000	143.00000	1.570000e+02	4.451500e+04
YearBuilt	17685.0	1.968469e+03	27.643208	1196.00000	1960.00000	1970.00000	1.980000e+03	2.019000e+03
Lattitude	17685.0	-3.780508e+01	0.096448	-38.19043	-37.86463	-37.79532	-3.774120e+01	-3.739780e+01
Longtitude	17685.0	1.449994e+02	0.127192	144.42379	144.91952	145.01060	1.450779e+02	1.455264e+02
Propertycount	17685.0	7.446893e+03	4435.473436	83.00000	4242.00000	6543.00000	1.033100e+04	2.165000e+04
doy	17685.0	1.964274e+02	91.696483	7.00000	119.00000	210.00000	2.680000e+02	3.460000e+02
Year	17685.0	2.016873e+03	0.623809	2016.00000	2016.00000	2017.00000	2.017000e+03	2.018000e+03
houseAge	17685.0	5.153051e+01	27.643208	1.00000	40.00000	50.00000	6.000000e+01	8.240000e+02

Figure 27. Central Tendency Measures after the Data Preparation.

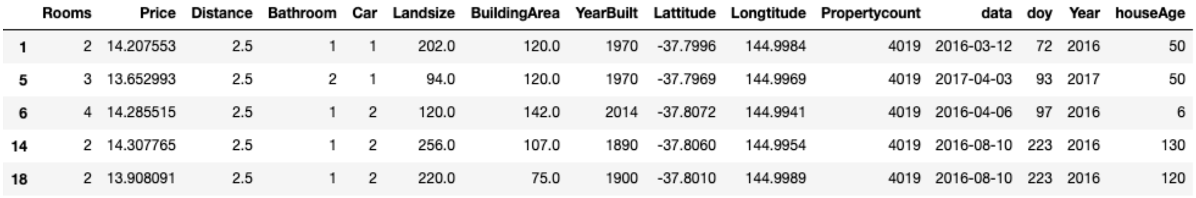
Another way to use correlation is to find the greatest number of features relative to our target variable, in this case, price. Of course, price is totally correlated to itself but we can nonetheless see that we have some features that are correlated to it positively such as Rooms, Bathroom, House Age, Longitude. At the same time, other features are correlated to price negatively such as Distance, Year Built and Latitude.



```
Find most important features relative to target
Price            1.000000
Rooms            0.429160
Bathroom         0.394221
houseAge         0.339866
Longtitude       0.269707
Car              0.188381
BuildingArea     0.071001
Landsize         0.016976
doy              0.008936
Year            -0.034937
Propertycount   -0.069590
Lattitude       -0.283455
YearBuilt       -0.339866
Distance        -0.354631
Name: Price, dtype: float64
```

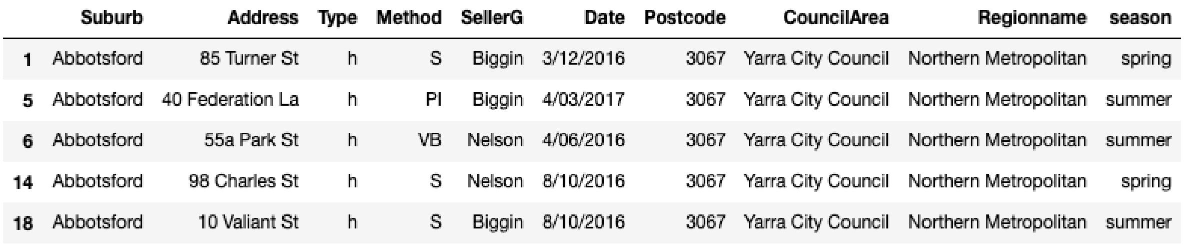
Figure 28. Most important features relative to Price.

This is how the top 5 rows of our dataset look split into categorical and numerical:



	Rooms	Price	Distance	Bathroom	Car	Landsize	BuildingArea	YearBuilt	Lattitude	Longtitude	Propertycount	data	doy	Year	houseAge
1	2	14.207553	2.5	1	1	202.0	120.0	1970	-37.7996	144.9984	4019	2016-03-12	72	2016	50
5	3	13.652993	2.5	2	1	94.0	120.0	1970	-37.7969	144.9969	4019	2017-04-03	93	2017	50
6	4	14.285515	2.5	1	2	120.0	142.0	2014	-37.8072	144.9941	4019	2016-04-06	97	2016	6
14	2	14.307765	2.5	1	2	256.0	107.0	1890	-37.8060	144.9954	4019	2016-08-10	223	2016	130
18	2	13.908091	2.5	1	2	220.0	75.0	1900	-37.8010	144.9989	4019	2016-08-10	223	2016	120

Figure 29. Dataset top 5 numerical variables.



	Suburb	Address	Type	Method	SellerG	Date	Postcode	CouncilArea	Regionname	season
1	Abbotsford	85 Turner St	h	S	Biggin	3/12/2016	3067	Yarra City Council	Northern Metropolitan	spring
5	Abbotsford	40 Federation La	h	PI	Biggin	4/03/2017	3067	Yarra City Council	Northern Metropolitan	summer
6	Abbotsford	55a Park St	h	VB	Nelson	4/06/2016	3067	Yarra City Council	Northern Metropolitan	summer
14	Abbotsford	98 Charles St	h	S	Nelson	8/10/2016	3067	Yarra City Council	Northern Metropolitan	spring
18	Abbotsford	10 Valiant St	h	S	Biggin	8/10/2016	3067	Yarra City Council	Northern Metropolitan	summer

Figure 30. Dataset top 5 categorical variables.

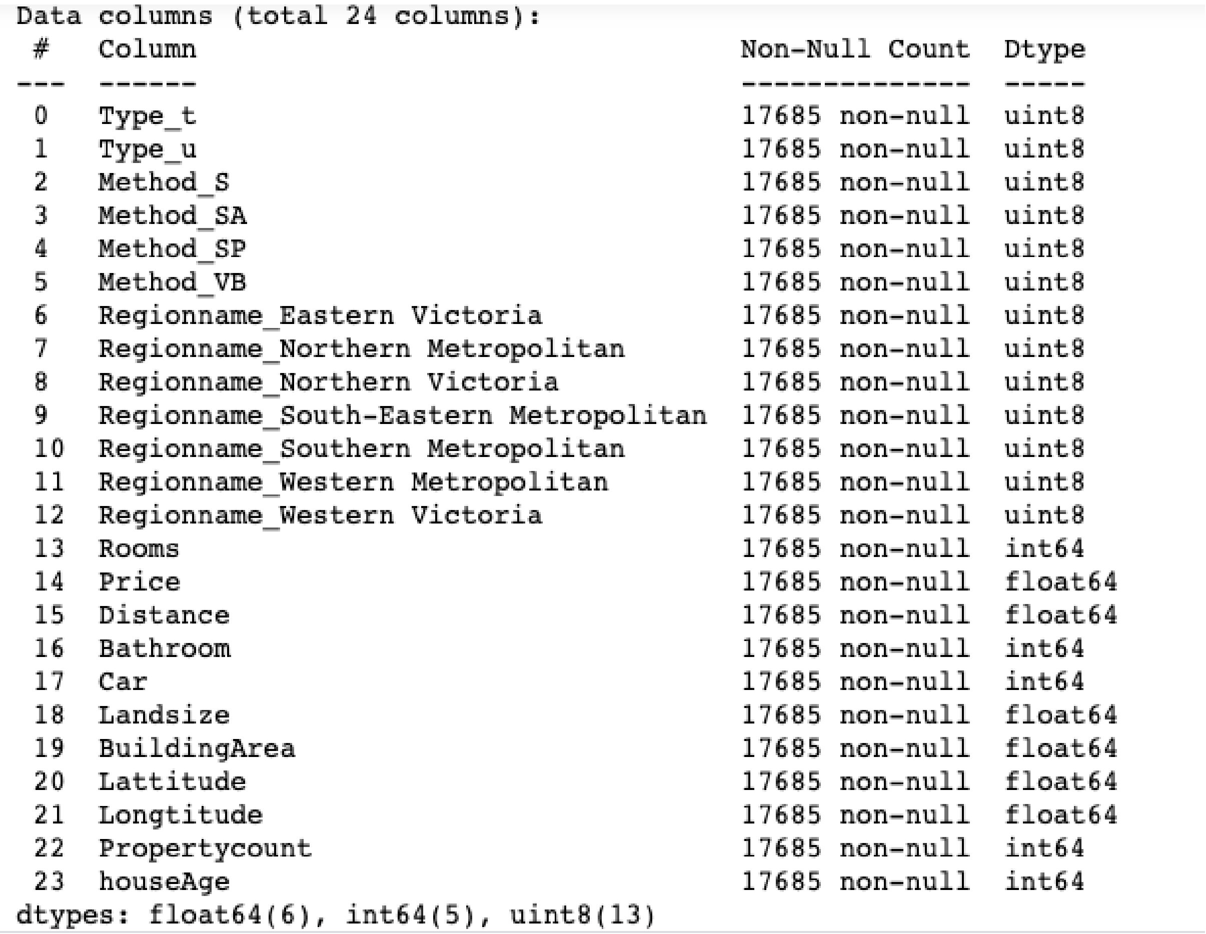
To conclude this section we will drop the features that we will not use as shown in Table 5.

Reasons to drop the columns			
Data Item	**Number of classes**	**Data Type**	**Reason**
Suburb	351	object	Number of classes
Address	34.004	object	Number of classes
SellerG	388	object	Number of classes
Date	-	object	House age instead
Postcode	211	float	Number of classes
Year Built	-	integer	House age instead
Year	-	integer	House age instead
Council Area	33	object	Number of classes
Season	4	object	Contributes nothing to the analysis
Daa	-	object	House age instead
Days of the year	-	object	Created to define seasons

Table 6. Reasons to drop the columns.

Few of them were very useful in creating new variables which were used in relation to Data Visualization. The variable 'Seasons' did not impact in any way on our analysis and additionally it does not impact on the price of properties in Melbourne as shown in Figure 13. As a result of this we can drop this variable and also avoid introducing multicollinearity as it is a categorical variable.

The use of numerical variables (discrete or continuous) in regression models is required. However in our dataset there are still 3 categorical variables (Type, Method and Region name) present after dropping the features that we found or agreed would not impact on our modeling. The method used to handle these remaining categorical variables was Dummy Variable Encoding. This method transforms the categorical values into binary values meaning that it only produces a value 0 or 1 based on the categorical value in the column. When doing so it is necessary to drop one factor level when encoding for linear regression to avoid introducing multicollinearity. We then applied it to the 3 remaining categorical variables and our results are shown in Figure 31.



```
Data columns (total 24 columns):
 #   Column                                 Non-Null Count  Dtype  
---  ------                                 --------------  -----  
 0   Type_t                                 17685 non-null  uint8  
 1   Type_u                                 17685 non-null  uint8  
 2   Method_S                               17685 non-null  uint8  
 3   Method_SA                              17685 non-null  uint8  
 4   Method_SP                              17685 non-null  uint8  
 5   Method_VB                              17685 non-null  uint8  
 6   Regionname_Eastern Victoria            17685 non-null  uint8  
 7   Regionname_Northern Metropolitan       17685 non-null  uint8  
 8   Regionname_Northern Victoria           17685 non-null  uint8  
 9   Regionname_South-Eastern Metropolitan  17685 non-null  uint8  
 10  Regionname_Southern Metropolitan       17685 non-null  uint8  
 11  Regionname_Western Metropolitan        17685 non-null  uint8  
 12  Regionname_Western Victoria            17685 non-null  uint8  
 13  Rooms                                  17685 non-null  int64  
 14  Price                                  17685 non-null  float64
 15  Distance                               17685 non-null  float64
 16  Bathroom                               17685 non-null  int64  
 17  Car                                    17685 non-null  int64  
 18  Landsize                               17685 non-null  float64
 19  BuildingArea                           17685 non-null  float64
 20  Lattitude                              17685 non-null  float64
 21  Longtitude                             17685 non-null  float64
 22  Propertycount                          17685 non-null  int64  
 23  houseAge                               17685 non-null  int64  
dtypes: float64(6), int64(5), uint8(13)
```

Figure 31. Final dataset.

3.3 Phase 2 Individual thoughts

3.3.1 Conor Sheehan

My primary initial task in the group was to conduct research which provided information and data on the centre of Melbourne and its surrounding districts and suburbs. Myself and Hasan garnered information on the status of each area and how amenities, historical significance and geographical position in relation to the Central Business District had an influence on their contribution to our research.

We also investigated various Machine Learning models to see which would contribute most substantially to our analysis. These included Tweedie, XG Boost, Gradient Boost and Random Forest models.

I also aided the non native english speaking members of the group with and writing and documentation the needed assistance with.

3.3.2 Eric Parfrey

My initial work involved analysing the Melbourne data set with Guiliano to find and detect any patterns or relationships that could give us insights into Housing Market behaviors. These patterns or discoveries were sent to Conor & Hasan for further analysis. I feel we have a strong team in place that work well together. We are very focused on our goal and objective. Additionally our dataset has been analysed and cleaned with machine learning models already chosen to use to model our data. During phase 2 I feel we have gained valuable insights into housing market behaviors that can help with our prediction model when we start analyzing the Dublin market. I feel very confident going into phase 3 with the team and plan we have in place.

3.3.3 Giuliano Silva

Everyone contributed to their roles since the Researchers, giving support to the Analysts with precise information about Melbourne and regression models to be tested in the next phase and the Analysts that were able to deliver all the metrics and visualizations. I believe that we've done a good job so far and we are on our way to phase 3 just waiting for feedback to improve where it is necessary. This phase 2 was great because we have developed a solid knowledge about Melbourne and also we are aware of which models we are going to use in the modelling stage.

3.3.4 Hasan Aziz

I believe that at this stage we have moved from the theoretical part to the practical part, by presenting our data set from many different aspects, especially our understanding of it, and by analyzing the dataset by providing visual maps, diagrams, Data Dictionary and by presenting the missing values. However, of course, our collective work gave the project different ideas to reach the required development during this stage, and I think that each stage has its own challenges and I look forward to the next stage.

3.3.5 Group thoughts

It has been more than two months working together on our project and everyone knows their role in the group which helps a lot. Our group created a very good chemistry where we can flow with our research and also our schedules always coincide so we can always keep our weekly meeting. As for the research, we believe that after phase 2 we are in a good position to be able to elaborate the model and obtain good predictions about the data.

References

Pettinger, T. (2018). 'Factors that affect the housing market', *Economics Help.* Available at: https://www.economicshelp.org/blog/377/housing/factors-that-affect-the-housing-market/ (Accessed 03 November 2020)

De, Parnika (2020). 'Housing Market Crash Prediction Using Machine Learning and Historical Data'. *Master's Project San Jose State University.* Available at: https://scholarworks.sjsu.edu/cgi/viewcontent.cgi?article=1928&context=etd_projects (Accessed 05 November 2020)

Hearne, R. (2018). 'Why fixing Ireland's housing crisis requires a change of policy', *RTE.* Available at: https://www.maynoothuniversity.ie/research/spotlight-research/why-fixing-irelands-housing-crisis-requires-change-policy (Accessed 05 November 2020)

Pino, T. (2020). *Melbourne Housing Market.* Available at: https://www.kaggle.com/anthonypino/melbourne-housing-marketne-housing-market (Accessed 11 November 2020).

Chen, B. (2018). What is One-Hot Encoding and how to use Pandas get_dummies function. Available at:

https://towardsdatascience.com/what-is-one-hot-encoding-and-how-to-use-pandas-get-dummies-function-922eb9bd4970 (Accessed 14 December 2020).

Ford, C. (2015). Understanding Q-Q plots. Available at: https://data.library.virginia.edu/understanding-q-q-plots/ (Accessed 14 December 2020)

Htoon, K. S. (2020). Log Transformation: Purpose and Interpretation. Available at: https://medium.com/@kyawsawhtoon/log-transformation-purpose-and-interpretation-9444b4b049c9 (Accessed 12 December 2020)

Allison, P. (2012). *When Can You Safely Ignore Multicollinearity?*. Available at: https://statisticalhorizons.com/multicollinearity (Accessed 30 November 2020).

Garg, R. (2018). A Primer to Ensemble Learning - Bagging and Boosting. Available at: https://analyticsindiamag.com/primer-ensemble-learning-bagging-boosting/#:~:text=Bagging%20is%20a%20way%20to,based%20on%20the%20last%20classification (Accessed 15 December 2020).

D'Agostino, R. B. (1971). An omnibus test of normality for moderate and large sample size. *Biometrika, 58*, 341-348

D'Agostino, R. and Pearson, E. S. (1973). Tests for departure from normality. *Biometrika, 60,* 613-622

Little, M.A., et al. (2017). Using and understanding cross-validation strategies. Perspectives on Saeb et al. *GigaScience, 6*(5), p.gix020.

Pedregosa, F. et al. (2011). Scikit-learn: Machine Learning in Python. *JMLR, 12*, pp. 2825-2830.

Reitermanova, Z. (2010). Data splitting. In *WDS*, *10*, pp. 31-36.

Ruginski, I. (2016). Checking the assumptions of linear regression. *Accessed, 11*, p. 2018