## Introduction

This project is focused on analyzing clearance housing data gathered in 2016 in Melbourne, Australia. The given dataset was retrieved from Kaggle and it was created by user Tony Pino. (<https://www.kaggle.com/anthonypino/melbourne-housing-market>)

In the dataset, we have multiple numerical variables, such as:

* price (AUD),
* number of rooms,
* distance from CBD (city center district in Melbourne) (km),
* number of parking spots,
* number of bathrooms
* year of construction
* land size (m³)

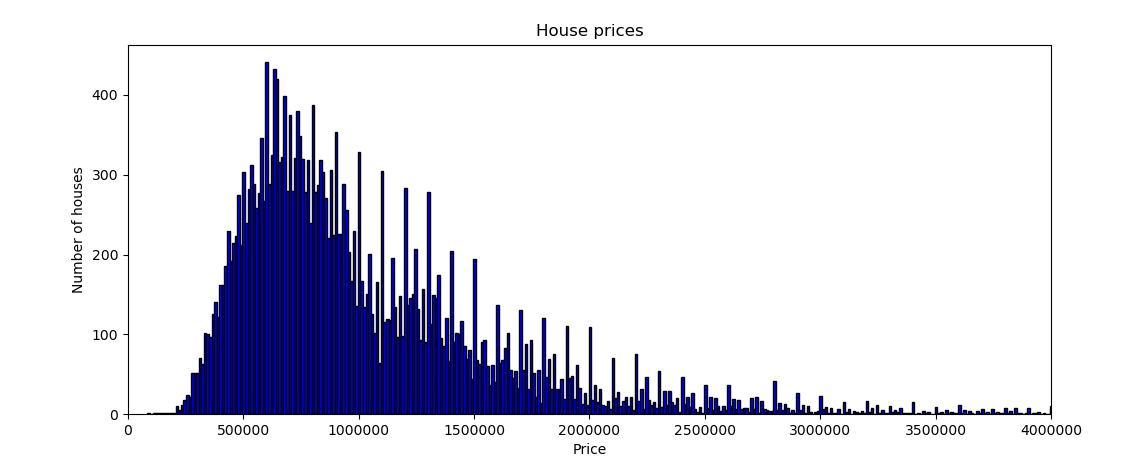
Many of these variables are sometimes not specified in advertisements, which is why we have very many null values in the dataset. We disregard these when working with this dataset.

Alongside numerical variables, we have categorical values, such as:

* type of home (house, unit, townhouse)
* method sold (e.g. sold after auction, withdrawn prior to auction, property sold)
* suburb
* real estate agent selling the property

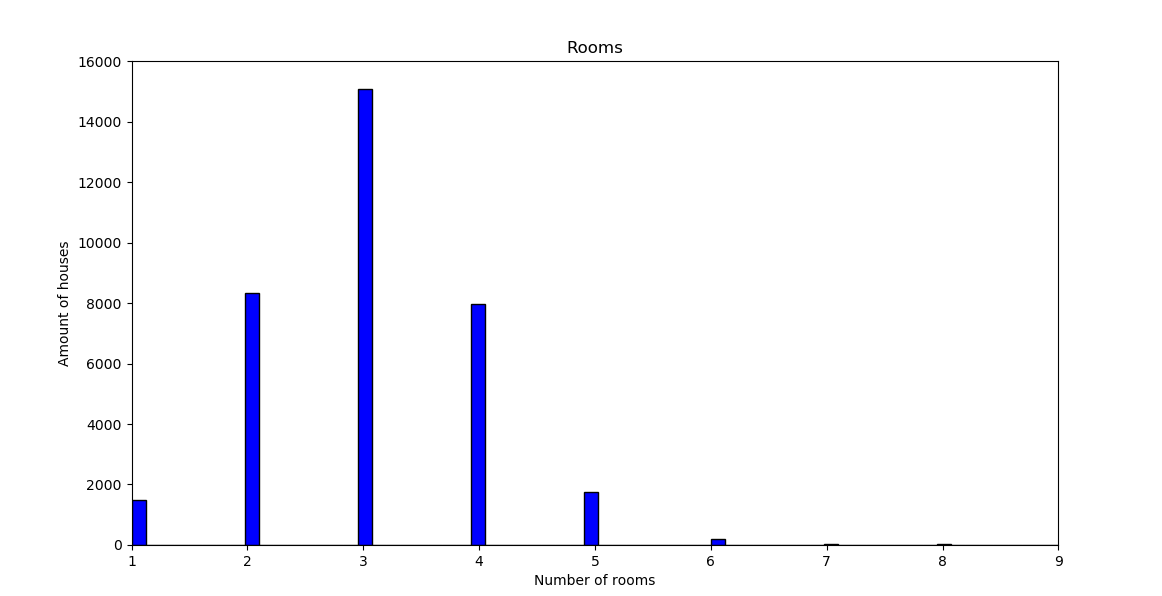
## Initial Distributions

The first graph shows a distribution of house prices, where we remove outliers (houses priced over 4 million AUD) because we notice a large drop-off of the number of houses on the market priced up to 11 million AUD. The resulting graph:



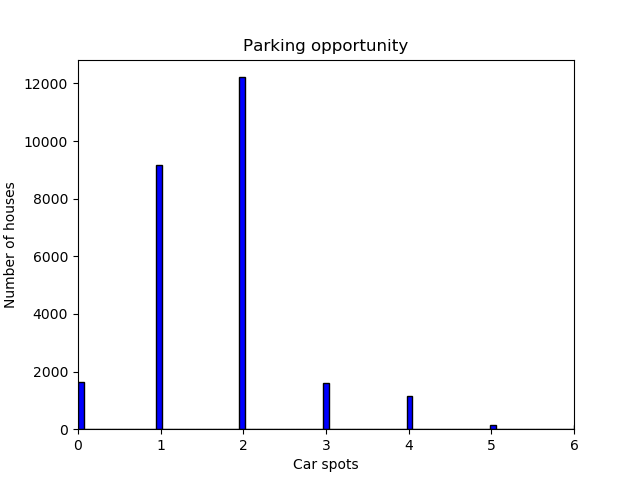
From this graph, we can securely say, that most houses on the market were priced between 400,000 and 1,500,000 AUD.

Next, we have a distribution of the number of rooms in the given dataset:



From this graph, we can gather that most houses on the market at the given time were between 2 and 4 bedrooms, 3 bedrooms being the most popular.

When looking at parking spots included with housing, upon initial review, any house with over 6 parking spots is an outlier, which is why 6 is the x limit on the next graph. Here we can see the graph displaying the amount of parking spots included with houses:

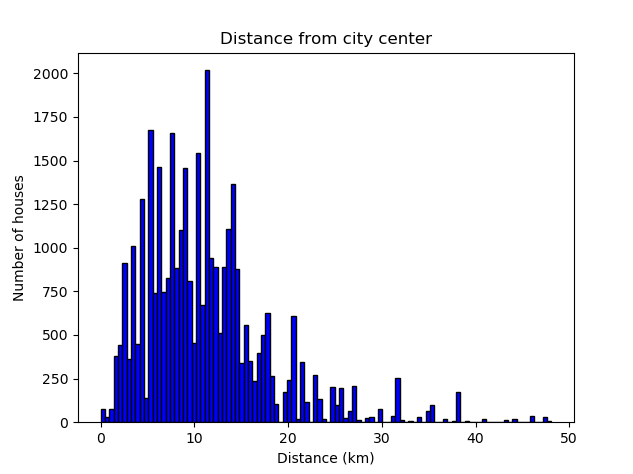


In the given graph, only existing data is shown (housing with unspecified parking is left out), but we can see, that the most housing would have had 1 or 2 parking spots included.

The following graph will display the amount of bathrooms in the given houses. Null values were omitted, so only values specified are shown:

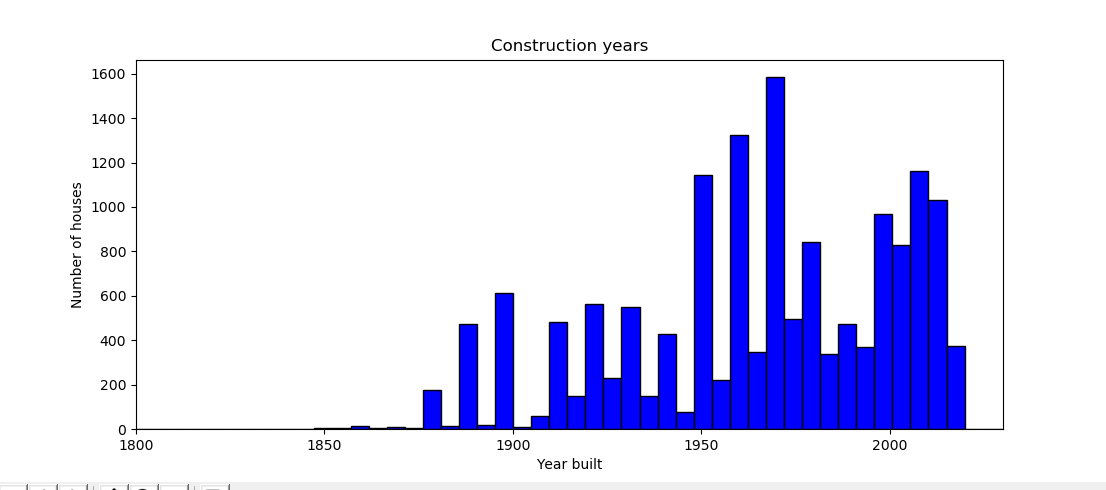


As we can see, most houses have 1 or 2 bathrooms, some having 3, but over that are mostly outliers. Interestingly, some housing was marked to have no bathrooms, as shown on the graph. This would probably be specified as such, because the house being sold is perhaps still under development and the bathroom is not yet complete.

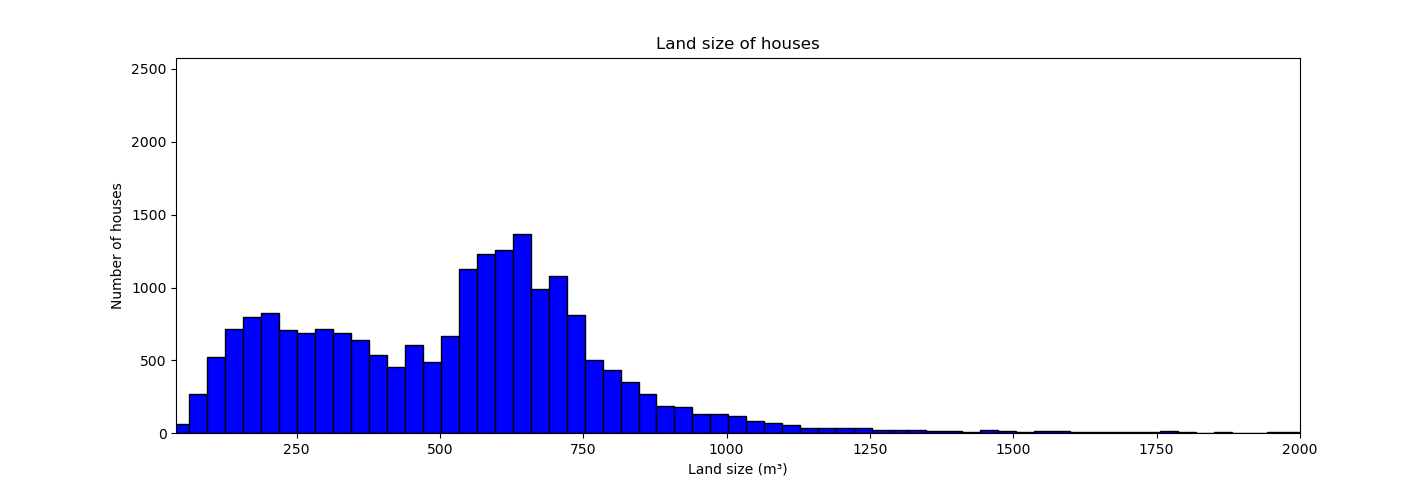
Another measure present in the dataset is the distance of houses from the Melbourne city center district. The distribution of which can be seen on the following graph:

From this graph, we can gather that most houses on the market at the time were located between 2km and 15km from the city center.

In the next graph, we will be looking at the construction year of houses. Houses with unassigned construction years were filtered out, also outliers were removed, which were present due to error or time travel (year built 1200 or 2100).



In this graph, we can see that most houses were built after 1950s. Before that, we see a trend of certain years for construction. After the 50s, construction became more uniform.

Another statistic to look at is land size. Again, we are looking at the ones with posted land size, removing outliers and zero values:

From this graph, we can see, that most common square meterage for houses on the market was 500-750m³. My guess for the square meterage under 200m³ would be for apartments, which have no included land beside the living area.

## Mean, Median, Mode

Price:

* Mean: 1050173.344955408
* Median: 870000.0
* ModeResult(mode=array([600000]), count=array([235]))

Rooms:

* Mean: 3.0310124221820582
* Median: 3.0
* ModeResult(mode=array([3]), count=array([15084]))

Parking spots:

* Mean: 1.7288453442535114
* Median: 2.0
* ModeResult(mode=array([2]), count=array([12214]))

Bathrooms:

* Mean: 1.624798167549097
* Median: 2.0
* ModeResult(mode=array([1]), count=array([12969]))

Distance from city center:

* Mean: 10.74561051182006
* Median: 10.0
* ModeResult(mode=array([11]), count=array([2815]))

Year built:

* Mean: 1965.289884894862
* Median: 1970.0
* ModeResult(mode=array([1970]), count=array([1490]))

## Variability

Price:

* Range: 11115000
* IQR: 660000.0
* Variance: 411480079004.6319
* Standard deviation: 641467.1301046001

Rooms:

* Range: 15
* IQR: 2.0
* Variance: 0.9407698981989097
* Standard deviation: 0.9699329348975163

Parking spots:

* Range: 26
* IQR: 1.0
* Variance: 1.0216575805280208
* Standard deviation: 1.0107707853554242

Bathrooms:

* Range: 12
* IQR: 1.0
* Variance: 0.5244830375572879
* Standard deviation: 0.7242120114699065

Distance from the city center:

* Range: 48
* IQR: 8.0
* Variance: 46.14090742815402
* Standard deviation: 6.792709873692091

Year built:

* Range: 910
* IQR: 60.0
* Variance: 1393.392874526943
* Standard deviation: 37.328178023136125

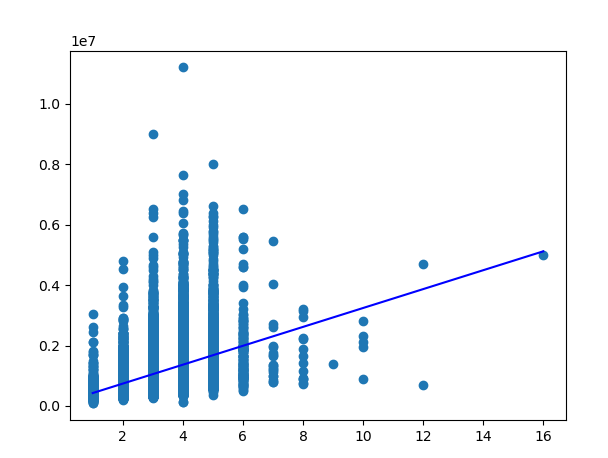
## Summary

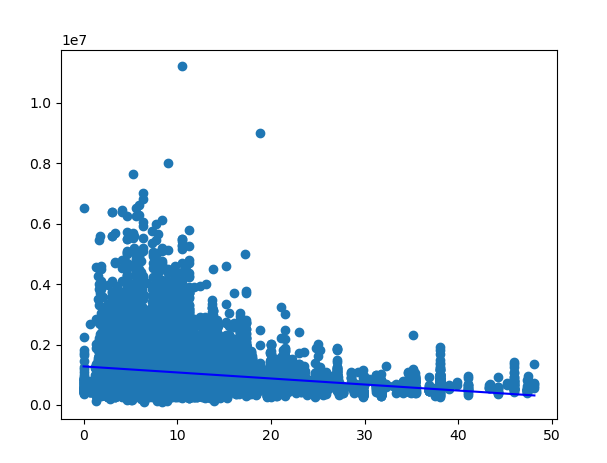
For calculations, Bessel’s correction was used, because most variables used were picked out from where they were present, with many fields being empty values.

Visual analysis mostly concludes the results of the variability. For example, year built and bathroom range values are quite high, due to outliers (house with 20 bathrooms or house built in 1200), but otherwise correct. The variability of house prices was shocking, but true. Other factors had most houses in specific values (bathrooms usually 1-3, others would be considered outliers), but house prices vary much more.

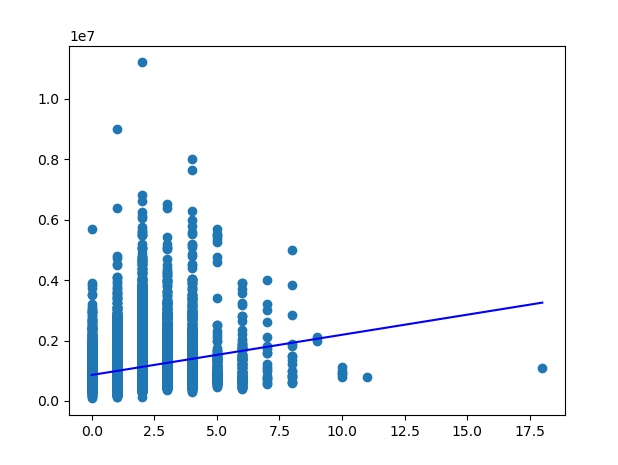
In conclusion, almost all variables had similar variances and ranges, which were skewed by outliers. House prices had the most drastic variance, because there are many more factors changing the price and price range is also much larger.

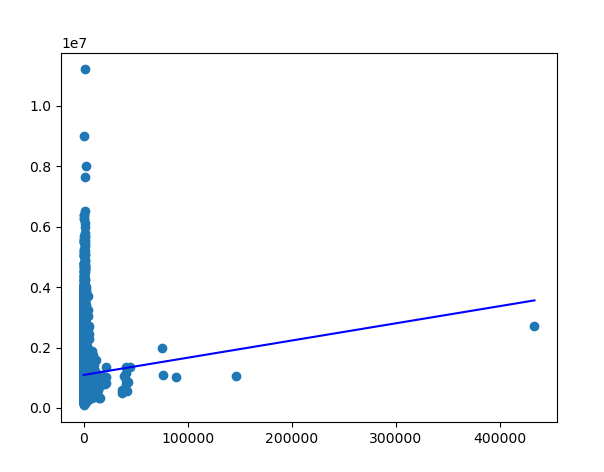
## 3.1.1 Correlations, linear relationships

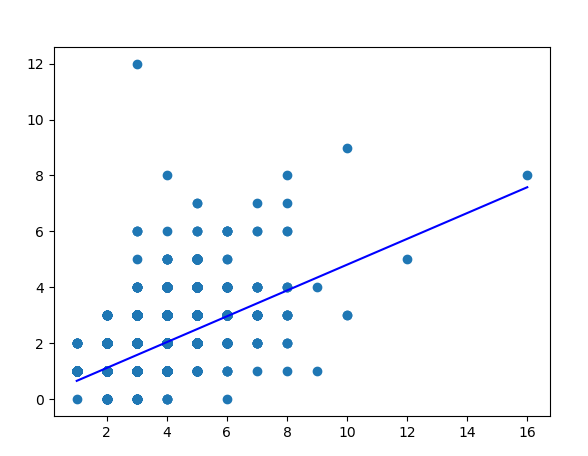
**Correlation between price and room count:**  
Pearsons: 0.46523835 Spearman: 0.504297  
Suggests a weak-moderate correlation.

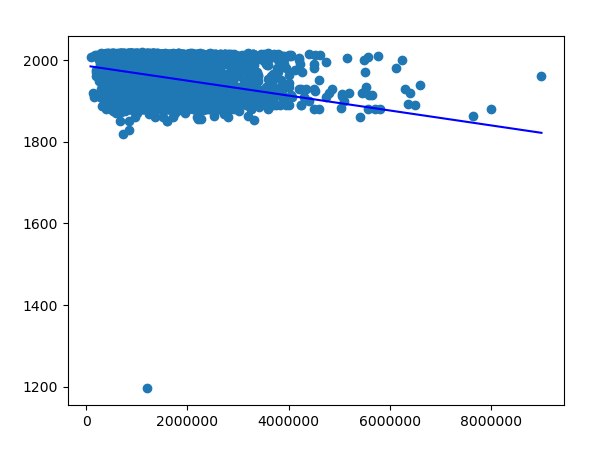
**Correlation between price and distance from city center:**Pearsons: -0.21138434 Spearman: -0.188093

Suggests a very weak correlation.

**Correlation between price and amount of parking spots:**Pearsons: 0.20180256 Spearman: 0.249667  
  
Suggests a very weak correlation (even after removing outliers.)

**Correlation between price and land size:**  
Pearsons: 0.03274837 Spearman: 0.276612  
  
Suggests a weak correlation.

**Correlation between amounts of rooms and bathrooms:**  
Pearsons: 0.61182586 Spearman: 0.615242  
  
Suggests a moderate correlation.

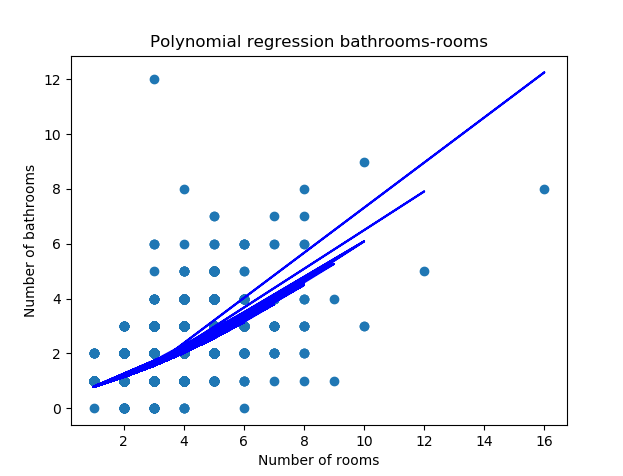
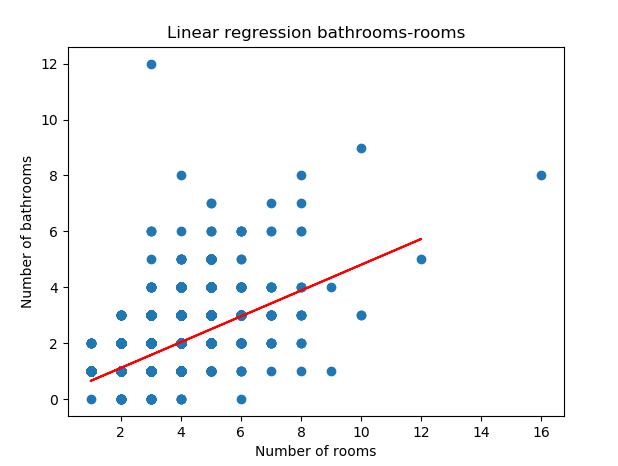
**Correlation between price and year built:**  
Pearsons: -0.33330556 Spearman: -0.370181  
  
Suggests a weak correlation.

## 3.1.2 Summary

After searching for correlations between variables, only one moderate correlation was found – correlation between number of rooms and number of bathrooms.  
This is most likely due to most other factors and how they all correlate. Also, when testing for correlations, interestingly removing outliers did not massively change the correlation matrix. I would suspect this to be due to a large dataset with many correlation points mapped, for which the outlier does not shift that much when removing some outliers.

## 3.2 Linear- and Polynomial Regression

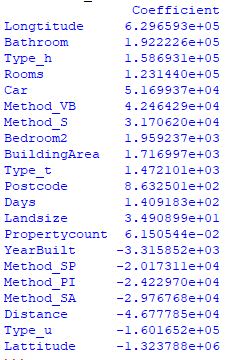
Because only one set of variables was shown to be of moderate correlation (number of rooms and number of bathrooms), I will use it to graph a scatter plot, linear- and polynomial regressions.

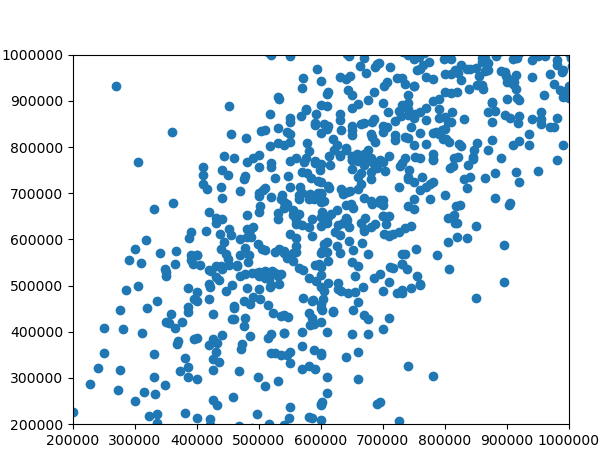


Sadly due to the nature of both the dependent and independent variables being only small integer values, most points overlap, not allowing us to understand how many points are actually in one position, but as we can see, the polynomial regression line is much more detailed in its curve. Both regressions used test data, which consisted of 20% of the two columns to predict values.

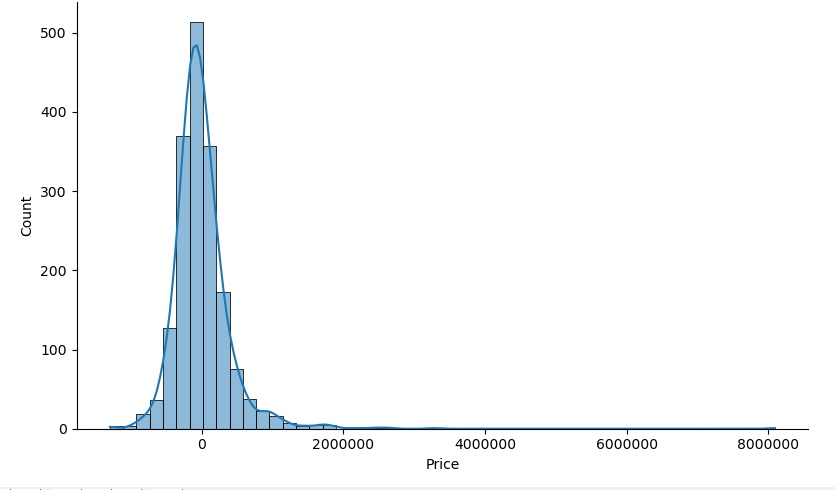
## 4.1 Training and testing regression model

As we already used test data and training in part 3.2, we will use similar methods to split the whole dataset into test and training data. This will be done in a 80%/20% split, 80% being the training data and 20% test data. Splitting was done with dates in mind, because dataset contains date, when advertisement was posted.

Set of features used is everything present in the dataset. Most correlations were made with location data, as we can see from this correlation table, where Long and Lat data show high correlation.  


Scatter plot of test data vs. predictions:  


Same thing as a distribution graph:



Metrics used to evaluate predictions:

MSE: 193414059041.92282

MAE: 267263.06127265695

RMSE: 439788.65269800083

## 4.2. Report

Using the large set of features showed the correlation between location and price. Making predictions with those features plus using all of the other features, which definitely trained the model better, showed us promising data with high correlation.