## 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)

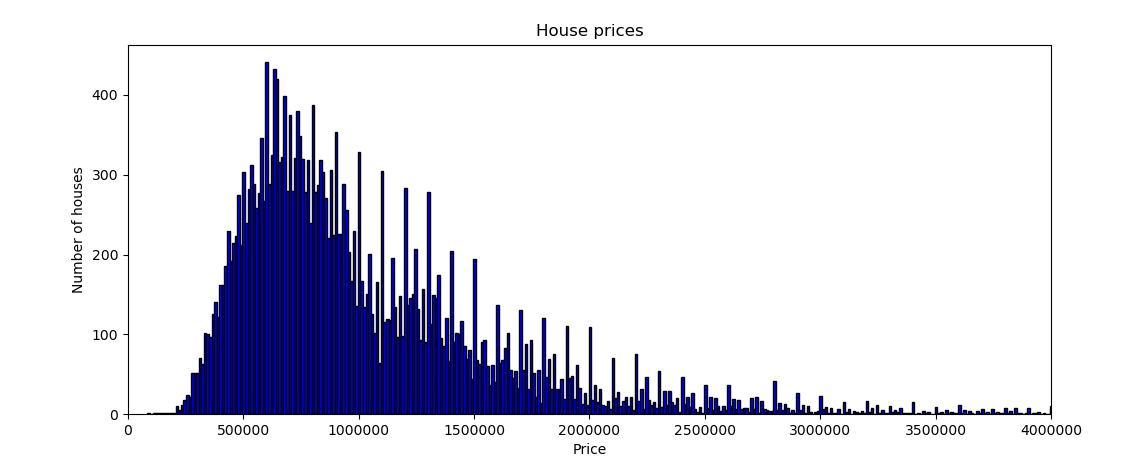
Many of these variables are sometimes not specified in advertisements, which is why we have very many null values in the dataset. We disregard these when working with this dataset.

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* method sold (e.g. sold after auction, withdrawn prior to auction, property sold)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

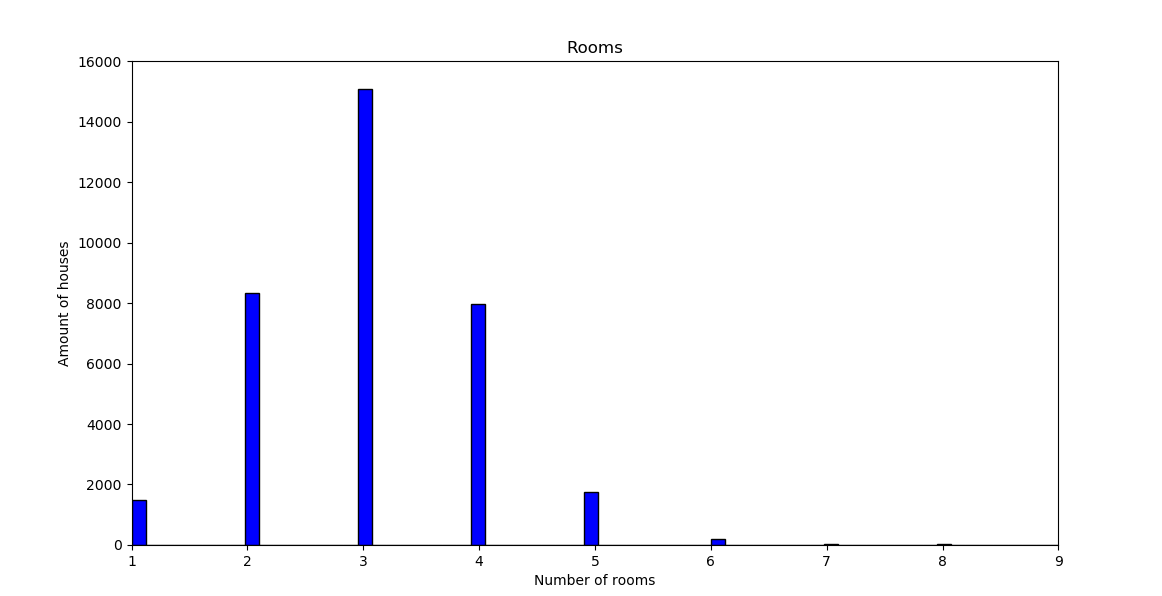
## Initial Distributions

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:



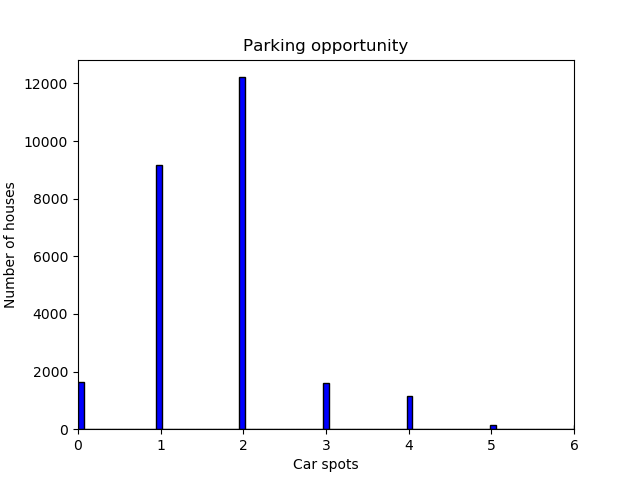
From this graph, we can securely say, that most houses on the market were priced between 400,000 and 1,500,000 AUD.

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:

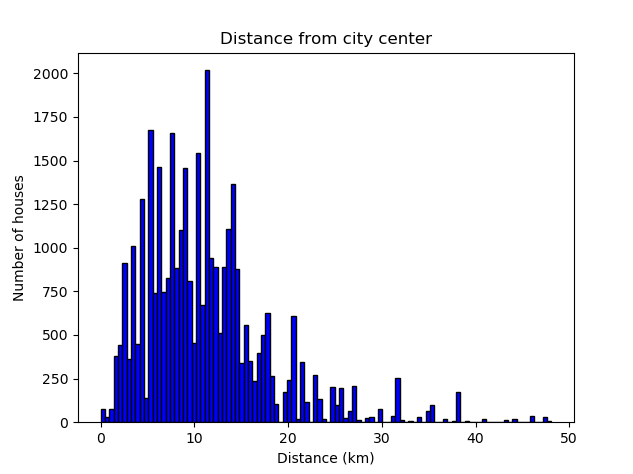


In the given graph, only existing data is shown (housing with unspecified parking is left out), but we can see, that the most housing would have had 1 or 2 parking spots included.

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

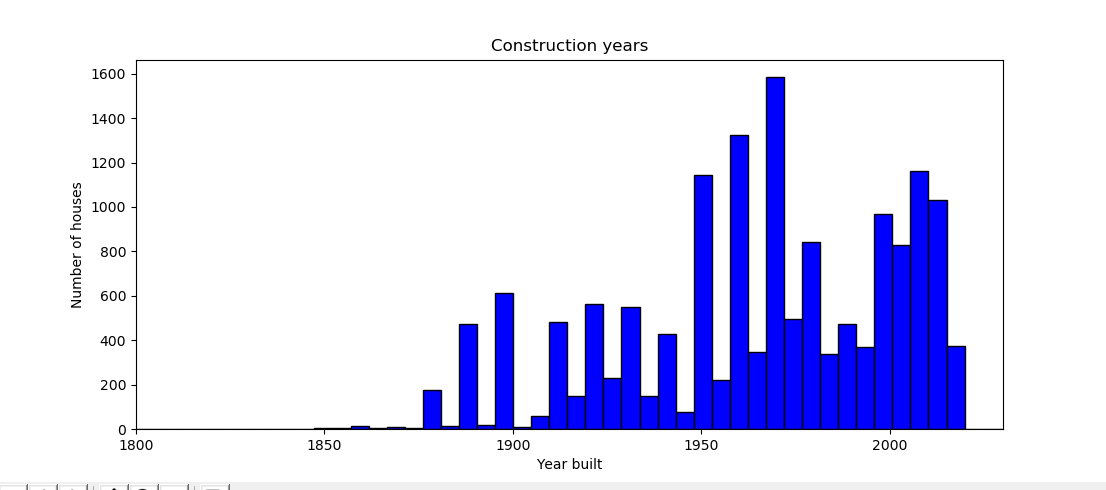


As we can see, most houses have 1 or 2 bathrooms, some having 3, but over that are mostly outliers. Interestingly, some housing was marked to have no bathrooms, as shown on the graph. This would probably be specified as such, because the house being sold is perhaps still under development and the bathroom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

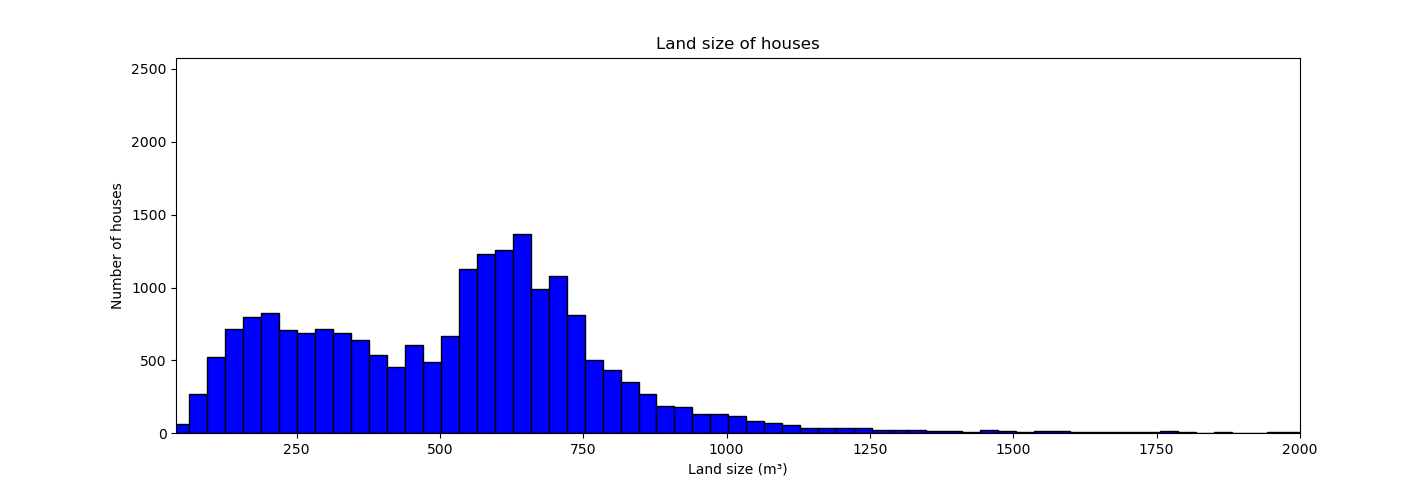
Another measure present in the dataset is the distance of houses from the Melbourne city center district. The distribution of which can be seen on the following graph:

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