# Predicting Melbourne House Prices from 2016

MATH2319 Machine Learning Applied Project Phase 1

Sangyoon Lee (s3727080)

March 31, 2018

# Contents

## 1. Introduction

The objective of this project is to build classifiers to explore relationships of Melbourne house prices which are over 1,000,000 AUD (1 million AUD) or less in a year from January of 2016 Melbourne, Australia. House Prices in Melbourne have been greatly changing over past few years. The data sets were sourced publicly available results posted from Domain.com.au, collected by Tony Pino at kaggle.com. This project contains of two phases. Phase 1 focuses on data preprocessing and detailed descriptive statistical analysis of the data. And phase 2 shall include applications of machine learning techniques.

Section 2 describes data set. Section 3 describes preprocessing and section 4 describes descriptive statistical analysis by exploring inter relationships between each attributes.

## 2. Data set

Tony Pino at kaggle.com provides one data set, Melbourne housing FULL, is going to be considered. This data is collected from Domain.com.au, specifically Melbourne house prices data. The training data set has 34857 observations and consist of 21 variables collected from 2016. In this project, this data set will be manipulated to target feature and hence showing inter relationships between attributes.

### 2.1 Target Feature

$$\text{House Price} = \begin{cases} \geq\ 1M\ AUD, if\ housing\ price\ exceeds\ 1M\ AUID \\ < 1M\ AUD, otherwise \end{cases}$$

The target feature has two classes '>= 1M AUD' and '< 1M AUD'. Our goal is to predict whether new houses on lease in Melbourne will be over $1 Million AUD or below.

### 2.2 Descriptive Features

| **Attributes** | **Details** |
|---|---|
| **Suburb** | Suburb |
| **Address** | Address |
| **Rooms** | Number of rooms |
| **Price** | Price in Australian dollars |

| | |
|---|---|
| **Method** | 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PI - property passed in; PN - sold prior not disclosed; SN - sold not disclosed; NB - no bid; VB - vendor bid; W - withdrawn prior to auction; SA - sold after auction; SS - sold after auction price not disclosed.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. |
| **Type** | br - bedroom(s); h - house, cottage, villa, semi, terrace; u - unit, duplex; t - townhouse; dev site - development site; o res - other residential. |
| **SellerG** | Real Estate Agent |
| **Date** | Date sold |
| **Distance** | Distance from CBD |
| **Postcode** | Post code |
| **Regionname** | General Region (West, North West, North, North east ...etc) |
| **Propertycount** | Number of properties that exist in the suburb |
| **Bedroom2** | Scraped # of Bedrooms (from different source) |
| **Bathroom** | Number of Bathrooms |
| **Car** | Number of car spots |
| **Landsize** | Land Size |
| **BuildingArea** | Building Size |
| **YearBuilt** | Year the house was built |
| **CouncilArea** | Governing council for the area |
| **Lattitude** | Latitude |
| **Longitude** | Longitude |

Most of the attributes are named as it is.  There may exist some ambiguity in classification of attributes, such as Room and Bedroom2. Next section of this project is about manipulating ambiguous attributes and further collect useful data.

## 3. Data Pre-processing

### 3.1 Preliminaries

In this project, we used following R packages.

```r
library(magrittr)
library(knitr)
library(mlr)
library(GGally)
library(cowplot)
library(ggmap)
```

We read dataset and keep string as string.

```r
housing <- read.csv('Melbourne_housing_FULL.csv', stringsAsFactors = FALSE)
```

## 3.2 Data Cleaning and Transformation

```
str(housing)

## 'data.frame':    34857 obs. of  21 variables:
##  $ Suburb       : chr  "Abbotsford" "Abbotsford" "Abbotsford" "Abbotsford"
 ...
##  $ Address      : chr  "68 Studley St" "85 Turner St" "25 Bloomburg St" "1
8/659 Victoria St" ...
##  $ Rooms        : int  2 2 2 3 3 3 4 4 2 2 ...
##  $ Type         : chr  "h" "h" "h" "u" ...
##  $ Price        : int  NA 1480000 1035000 NA 1465000 850000 1600000 NA NA
NA ...
##  $ Method       : chr  "SS" "S" "S" "VB" ...
##  $ SellerG      : chr  "Jellis" "Biggin" "Biggin" "Rounds" ...
##  $ Date         : chr  "3/09/2016" "3/12/2016" "4/02/2016" "4/02/2016" ...
##  $ Distance     : chr  "2.5" "2.5" "2.5" "2.5" ...
##  $ Postcode     : chr  "3067" "3067" "3067" "3067" ...
##  $ Bedroom2     : int  2 2 2 3 3 3 3 3 4 3 ...
##  $ Bathroom     : int  1 1 1 2 2 2 1 2 1 2 ...
##  $ Car          : int  1 1 0 1 0 1 2 2 2 1 ...
##  $ Landsize     : int  126 202 156 0 134 94 120 400 201 202 ...
##  $ BuildingArea : num  NA NA 79 NA 150 NA 142 220 NA NA ...
##  $ YearBuilt    : int  NA NA 1900 NA 1900 NA 2014 2006 1900 1900 ...
##  $ CouncilArea  : chr  "Yarra City Council" "Yarra City Council" "Yarra Ci
ty Council" "Yarra City Council" ...
##  $ Lattitude    : num  -37.8 -37.8 -37.8 -37.8 -37.8 ...
##  $ Longtitude   : num  145 145 145 145 145 ...
##  $ Regionname   : chr  "Northern Metropolitan" "Northern Metropolitan" "No
rthern Metropolitan" "Northern Metropolitan" ...
##  $ Propertycount: chr  "4019" "4019" "4019" "4019" ...
```

```
summarizeColumns(housing)
```

*Table: Feature summary of each columns*

| name | type | na | mean | disp | median | mad | min | max | nlevs |
|---|---|---|---|---|---|---|---|---|---|
| <chr> | <chr> | <int> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <int> |
| Suburb | character | 0 | *NA* | 9.757868e-01 | *NA* | *NA* | 1.00000 | 844.0000 | 351 |
| Address | character | 0 | *NA* | 9.998279e-01 | *NA* | *NA* | 1.00000 | 6.0000 | 34009 |
| Rooms | integer | 0 | 3.031012e+00 | 9.699329e-01 | 3.0000 | 1.482600e+00 | 1.00000 | 16.0000 | 0 |
| Type | character | 0 | *NA* | 3.120464e-01 | *NA* | *NA* | 3580.00000 | 23980.0000 | 3 |
| Price | integer | 7610 | 1.050173e+06 | 6.414671e+05 | 870000.0000 | 4.299540e+05 | 85000.00000 | 11200000.0000 | 0 |
| Method | character | 0 | *NA* | 4.335714e-01 | *NA* | *NA* | 36.00000 | 19744.0000 | 9 |
| SellerG | character | 0 | *NA* | 9.036349e-01 | *NA* | *NA* | 1.00000 | 3359.0000 | 388 |

| name | type | na | mean | disp | median | mad | min | max | nlevs |
|---|---|---|---|---|---|---|---|---|---|
| <chr> | <chr> | <int> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <int> |
| Date | character | 0 | *NA* | 9.678974e-01 | *NA* | *NA* | 3.00000 | 1119.0000 | 78 |
| Distance | character | 0 | *NA* | 9.592621e-01 | *NA* | *NA* | 1.00000 | 1420.0000 | 216 |
| Postcode | character | 0 | *NA* | 9.757868e-01 | *NA* | *NA* | 1.00000 | 844.0000 | 212 |
| Bedroom2 | integer | 8217 | 3.084647e+00 | 9.806897e-01 | 3.0000 | 1.482600e+00 | 0.00000 | 30.0000 | 0 |
| Bathroom | integer | 8226 | 1.624798e+00 | 7.242120e-01 | 2.0000 | 1.482600e+00 | 0.00000 | 12.0000 | 0 |
| Car | integer | 8728 | 1.728845e+00 | 1.010771e+00 | 2.0000 | 1.482600e+00 | 0.00000 | 26.0000 | 0 |
| Landsize | integer | 11810 | 5.935990e+02 | 3.398842e+03 | 521.0000 | 3.113460e+02 | 0.00000 | 433014.0000 | 0 |
| BuildingArea | numeric | 21115 | 1.602564e+02 | 4.012671e+02 | 136.0000 | 6.078660e+01 | 0.00000 | 44515.0000 | 0 |
| YearBuilt | integer | 19306 | 1.965290e+03 | 3.732818e+01 | 1970.0000 | 4.447800e+01 | 1196.00000 | 2106.0000 | 0 |
| CouncilArea | character | 0 | *NA* | 8.945692e-01 | *NA* | *NA* | 3.00000 | 3675.0000 | 34 |
| Lattitude | numeric | 7976 | -3.781063e+01 | 9.027890e-02 | -37.8076 | 8.077205e-02 | -38.19043 | -37.3902 | 0 |
| Longtitude | numeric | 7976 | 1.450019e+02 | 1.201688e-01 | 145.0078 | 1.012912e-01 | 144.42379 | 145.5264 | 0 |
| Regionname | character | 0 | *NA* | 6.604412e-01 | *NA* | *NA* | 3.00000 | 11836.0000 | 9 |
| Propertycount | character | 0 | *NA* | 9.757868e-01 | *NA* | *NA* | 1.00000 | 844.0000 | 343 |

With structure and summarize columns above, we were able to see facts that:

- There are partial number of missing values for Price, Bedroom2, Bathroom, Car, Landsize, BuildingArea, YearBuilt, Lattitude and Longtitude. We assume that samples are normally distributed then we analyze with the given enough size of samples.
- We are specifically interested in House price, as our main target feature, there exist 7610 samples with no price given in data.
- There are attributes providing same information under sub-groups, such as suburb, post code council area and region name. We will be only considering region names for later part.
- Address is considered to be unique ID, so address does not provide any useful information.
- Seller is not objective feature to compare, so we exclude this attribute.
- Distance and Propertycount which have format ‘character’, are supposed to be integers.

To sum up, we delete columns which are not useful to discuss in this project.

```
housing$Address = NULL
housing$SellerG = NULL
housing$Suburb = NULL
housing$CouncilArea = NULL
housing$Bedroom2 = NULL
housing$Postcode = NULL
```

And delete rows with no price data specified.

```
housing <- housing[!is.na(housing$Price), ]
```

Remaining number of samples and attributes are: 27247 samples and 15 attributes.

```r
dim(housing)

## [1] 27247    15
```

Futher, we classify price into binary classes, which are over 1 Million AUD and below 1 Million AUD. We define classes in new attribute, PriceMillion.

```r
housing$PriceMillion <- ifelse(housing$Price >= 1000000, '>= 1M', '< 1M')
nrow(housing[housing$PriceMillion==">= 1M",])

## [1] 10955

nrow(housing[housing$PriceMillion=="< 1M",])

## [1] 16292

nrow(housing)

## [1] 27247
```

*Table: Frequency of House Price Classes*

| Class | Frequency | Percentage |
|---|---|---|
| < 1M | 16292 | 59.79% |
| ≥ 1M | 10955 | 40.21% |

Around 60% of houses in Melbourne are shown below 1 Million AUD. Note that, there are more houses with below 1 Million AUD, but mean of house price is over 1 Million AUD while median of house price is 870,000 AUD. This is because samples are collected with no upper boundaries for house price and the most expensive house contained in data frame is 11,200,000 AUD, and the cheapest house price is 85,000 AUD.

From above, we mentioned Distance is supposed to be an integer, so we change format of distance to integers as well as propertycount. For distance, we will be analyzing rounded values.

```r
housing$Distance <- as.numeric(as.character(housing$Distance))
housing$Distance <- round(housing$Distance,0)
housing$Propertycount <- as.numeric(as.character(housing$Propertycount))
```

Full date information is providing too many factors. We simply dropping days to simplify Date factors 12 per annum (12 months) from 365 factors per annum (365 days).

```
housing$Date <- substring(housing$Date,3)
housing$Date <- sub("^[^0-9]*","",housing$Date)
```

In order to exclude outliers from the graph for Section 4: Data Exploration, we group data values in certain intervals. We keep original data and manipulate in new column of data frame.

```
housing$BuildingArea2 <- cut(housing$BuildingArea, seq(0,700,by=70))

housing$Propertycount2 <- cut(housing$Propertycount, seq(0,22500,by=1500), di
g.lab=10)

housing$YearBuilt2 <- cut(housing$YearBuilt, seq(1820,2020,by=20),
dig.lab=10)
housing$Landsize2 <- cut(housing$Landsize, seq(0,2500,by=250), dig.lab=10);
```

We computed level table and organized results into new table.

- There are only 5 methods: S (property sold), SP (property sold prior), PI (property passed in), VB (vendor bid), SA (sold after auction)
- There are only 3 types: h (house, cottage), t (town house), u (unit, duplex)

```
sapply( housing[ sapply(housing, is.character)], table)
```

*Table: Level table for House*

| Attributes | Contents | | |
|---|---|---|---|
| **Type** | h: 18470 | | |
| | t: 2866 | | |
| | u: 5908 | | |
| **Method** | PI: 3255 | | |
| | S: 17514 | | |
| | SA: 190 | | |
| | SP: 3602 | | |
| | VB: 2683 | | |
| **Date** | 01/2016: 2 | 02/2017: 526 | 01/2018: 646 |
| | 02/2016: 35 | 03/2017: 841 | 02/2018: 1506 |
| | 03/2016: 401 | 04/2017: 805 | 03/2018: 1521 |
| | 05/2016: 1167 | 05/2017: 1453 | |
| | 06/2016: 962 | 06/2017: 1463 | |
| | 07/2016: 553 | 07/2017: 1805 | |
| | 08/2016: 911 | 09/2017: 2053 | |
| | 09/2016: 1166 | 10/2017: 2441 | |
| | 10/2016: 677 | 11/2017: 1994 | |
| | 11/2016: 1413 | 12/2017: 723 | |
| | 12/2016: 767 | | |

| | |
|---|---|
| **Regionname** | Eastern Metropolitan: 3722<br>Eastern Victoria: 166<br>Northern Metropolitan: 7864<br>Northern Victoria: 166<br>South-Eastern Metropolitan: 1341<br>Southern Metropolitan: 8524<br>Western Metropolitan: 5815<br>Western Victoria: 96 |
| **PriceMillion** | < 1M: 16289<br>>= 1M: 10955 |

Finally, below table presents the summary statistics after data preprocessing in section 3.

```
summarizeColumns(housing)
```

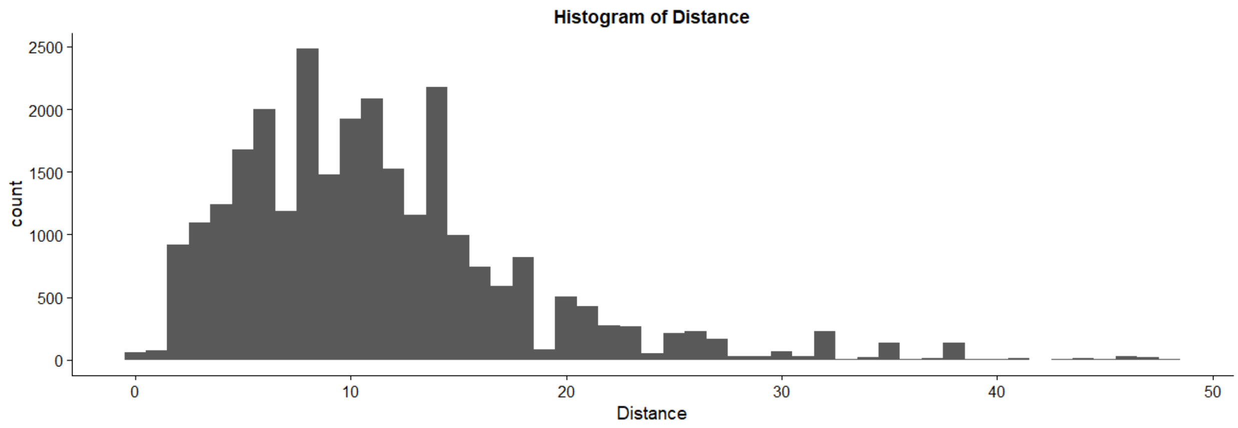
*Table: Summary Statistics after preprocessing*

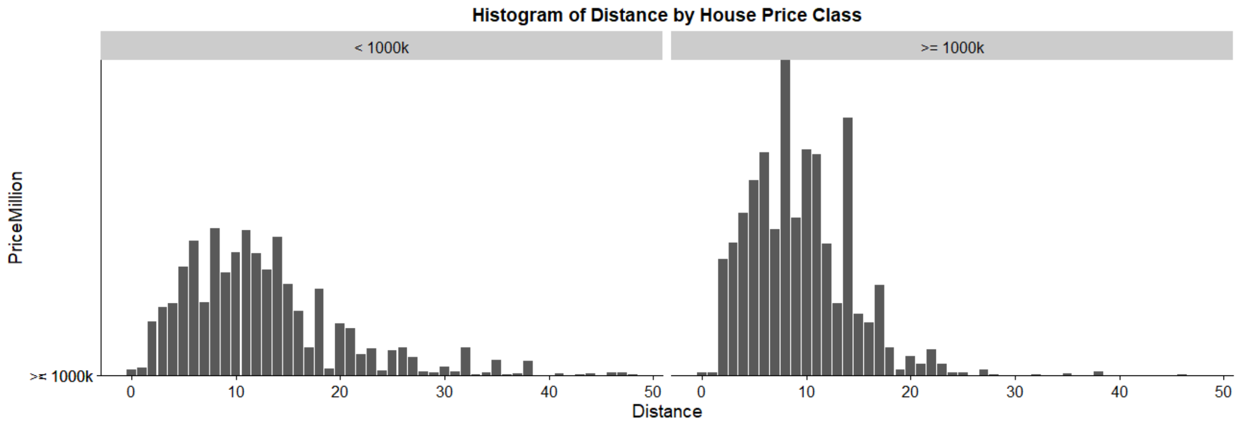
| name | type | na | mean | disp | median | mad | min | max | nlevs |
|---|---|---:|---:|---:|---:|---:|---:|---:|---:|
| <chr> | <chr> | <int> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <int> |
| Rooms | integer | 0 | 2.992365e+00 | 9.548100e-01 | 3.00000 | 1.482600e+00 | 1.00000 | 16.0000 | 0 |
| Type | character | 0 | *NA* | 3.220526e-01 | *NA* | *NA* | 2866.00000 | 18470.0000 | 3 |
| Price | integer | 0 | 1.050210e+06 | 6.414923e+05 | 870000.00000 | 4.299540e+05 | 85000.00000 | 11200000.0000 | 0 |
| Method | character | 0 | *NA* | 3.571429e-01 | *NA* | *NA* | 190.00000 | 17514.0000 | 5 |
| Date | character | 0 | *NA* | 9.104023e-01 | *NA* | *NA* | 2.00000 | 2441.0000 | 25 |
| Distance | numeric | 0 | 1.126633e+01 | 6.817567e+00 | 10.00000 | 5.930400e+00 | 0.00000 | 48.0000 | 0 |
| Bathroom | integer | 6444 | 1.591683e+00 | 7.008804e-01 | 1.00000 | 0.000000e+00 | 0.00000 | 9.0000 | 0 |
| Car | integer | 6821 | 1.715370e+00 | 9.942161e-01 | 2.00000 | 1.482600e+00 | 0.00000 | 18.0000 | 0 |
| Landsize | integer | 9262 | 5.934889e+02 | 3.757266e+03 | 512.00000 | 3.113460e+02 | 0.00000 | 433014.0000 | 0 |
| BuildingArea | numeric | 16588 | 1.568346e+02 | 4.492228e+02 | 133.00000 | 5.782140e+01 | 0.00000 | 44515.0000 | 0 |
| YearBuilt | integer | 15160 | 1.966609e+03 | 3.676237e+01 | 1970.00000 | 4.299540e+01 | 1196.00000 | 2019.0000 | 0 |
| Lattitude | numeric | 6251 | -3.780696e+01 | 9.161945e-02 | -37.80046 | 8.352968e-02 | -38.19043 | -37.3978 | 0 |
| Longtitude | numeric | 6251 | 1.449967e+02 | 1.206803e-01 | 145.00320 | 1.055018e-01 | 144.42379 | 145.5264 | 0 |
| Regionname | character | 0 | *NA* | 6.871238e-01 | *NA* | *NA* | 96.00000 | 8524.0000 | 8 |
| Propertycount | numeric | 0 | 7.566781e+03 | 4.492382e+03 | 6567.00000 | 3.998572e+03 | 83.00000 | 21650.0000 | 0 |
| PriceMillion | character | 0 | *NA* | 4.021069e-01 | *NA* | *NA* | 10955.00000 | 16289.0000 | 2 |
| Landsize2 | factor | 11411 | *NA* | *NA* | *NA* | *NA* | 33.00000 | 6655.0000 | 10 |
| BuildingArea2 | factor | 16677 | *NA* | *NA* | *NA* | *NA* | 13.00000 | 4887.0000 | 10 |
| Propertycount2 | factor | 0 | *NA* | 8.219792e-01 | *NA* | *NA* | 0.00000 | 4850.0000 | 13 |
| YearBuilt2 | factor | 15162 | *NA* | *NA* | *NA* | *NA* | 1.00000 | 2691.0000 | 10 |

# 4. Data Exploration

## 4.1 Numerical Features

### 4.1.1 Distance





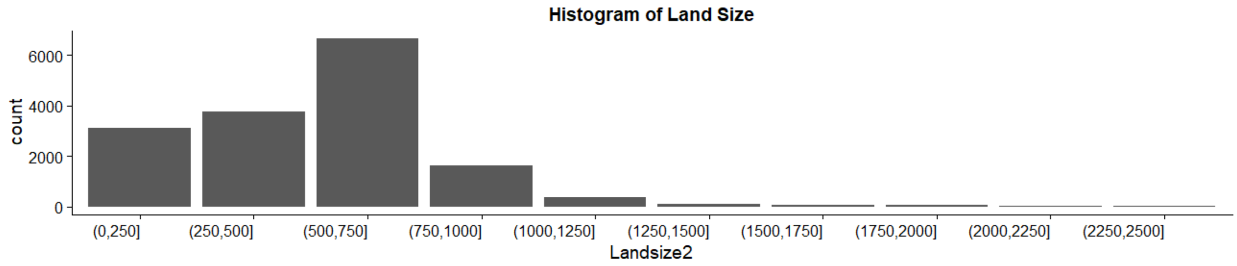
```
mean(housing$Distance);
## [1] 11.26633
mean(housing$Distance[housing$Price >= 1000000]);
## [1] 9.444728
mean(housing$Distance[housing$Price < 1000000]);
## [1] 12.49144
```

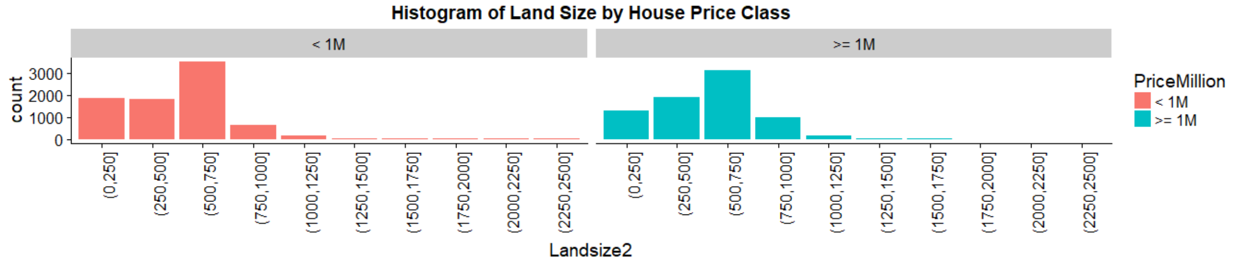
*Table: Mean of Distance by House Price classes*

| Class | Total | < 1 Million | >= 1 Million |
|---|---|---|---|
| **Distance** | 11.26633 | 9.444728 | 12.49144 |

Mean distance of over 1 million dollars houses are more distance from CBD compare to below 1 million dollar houses. Two distributions show similar left-skewness, except distribution of data. Most of houses over 1 million dollars are concentrated within 20km away from CBD. However, below 1 million dollars houses are less concentrated as over 1 million dollars. This implies the fact that houses over 25km away from CBD are generally below 1 million dollars.

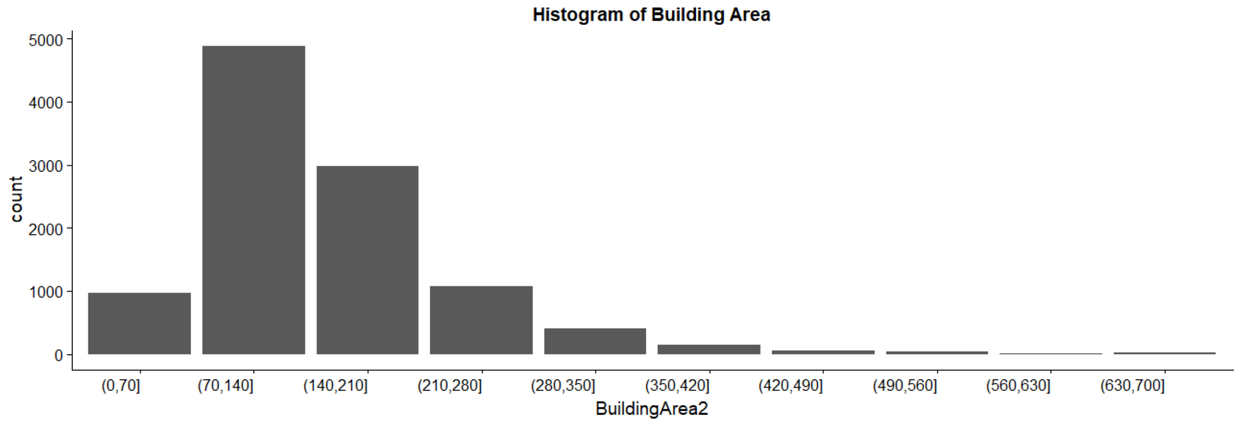
### 4.1.2 Land Size

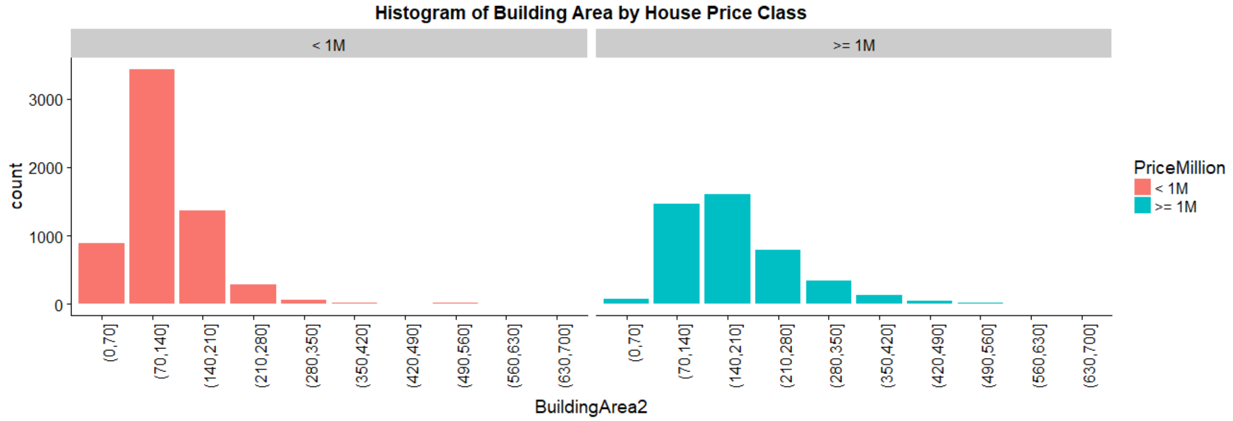




Two distributions by house price class show similar skewness. Distribution of land size is very similar, but taking note that if land size is too large, house price is more likely to be less than 1 million dollars. Link to histogram of distance infer that, premises which is far apart from CBD, are more likely to have large hand size.

### 4.1.3 Building Area

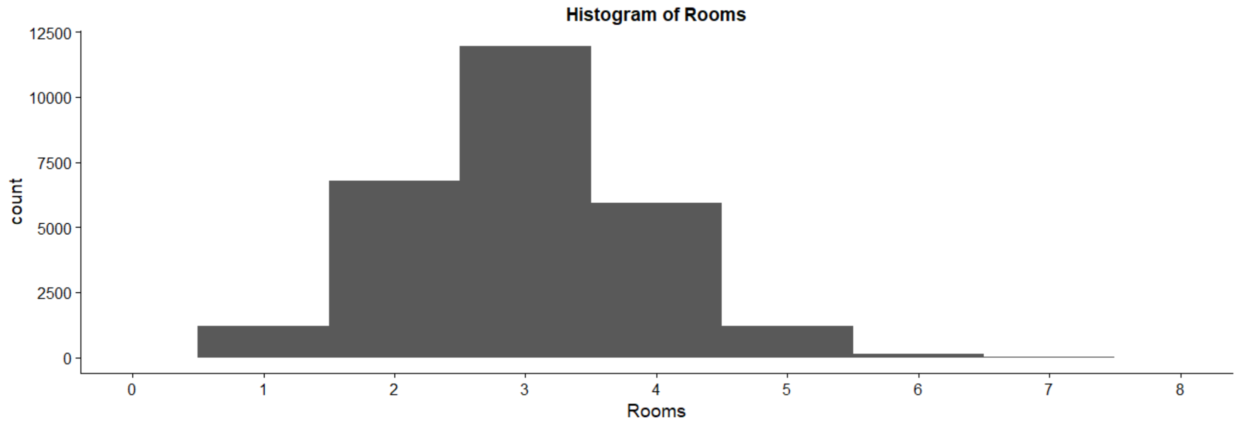


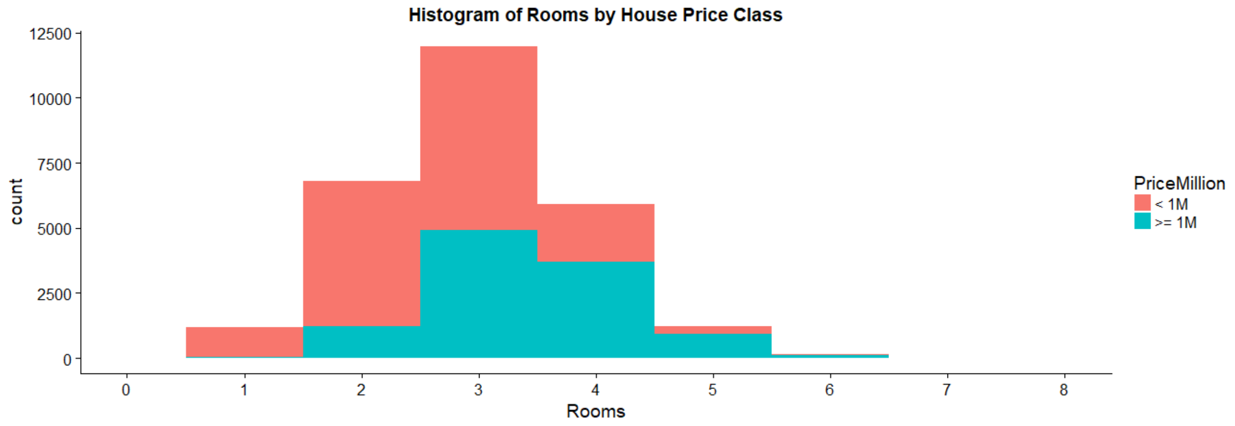


For below 1 million dollars houses, majority of the houses are below 210 meter squared. But over 1 million dollars houses have more variety in bigger size of building area.

Compare to histogram of land size, below 1 million dollars houses have bigger number of large land, but smaller number of large building area.

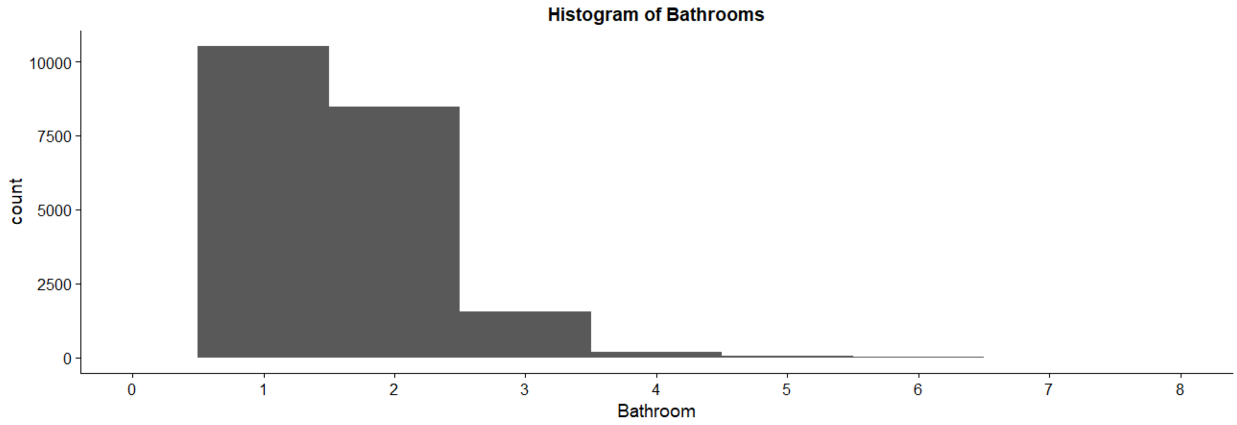
### 4.1.4 Room

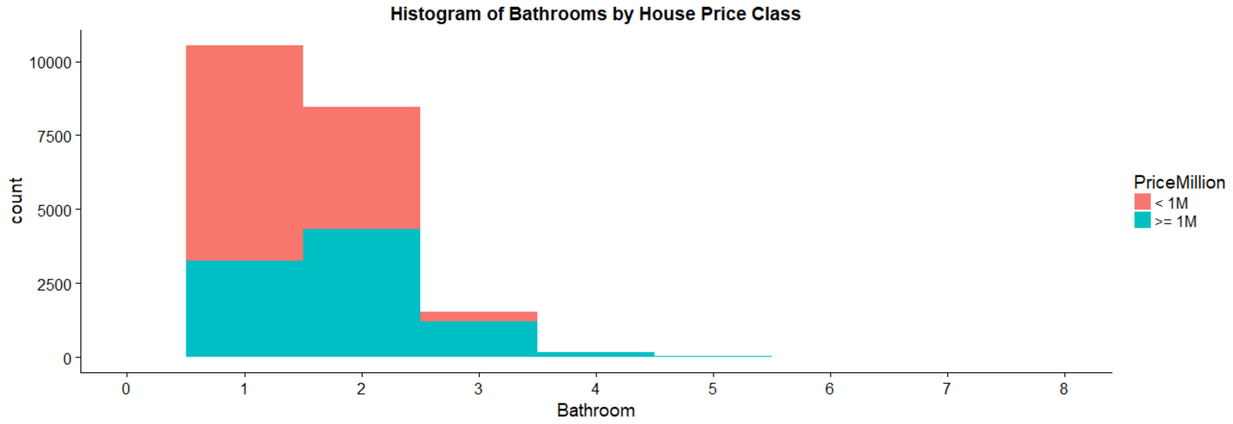




As the size of building area for over a million dollars houses are larger, the houses are expected to have more rooms and bathrooms. Histogram of bathrooms is prepared next. In order to find a house with 4 or more rooms, the price of house is generally over one million dollars in Melbourne.

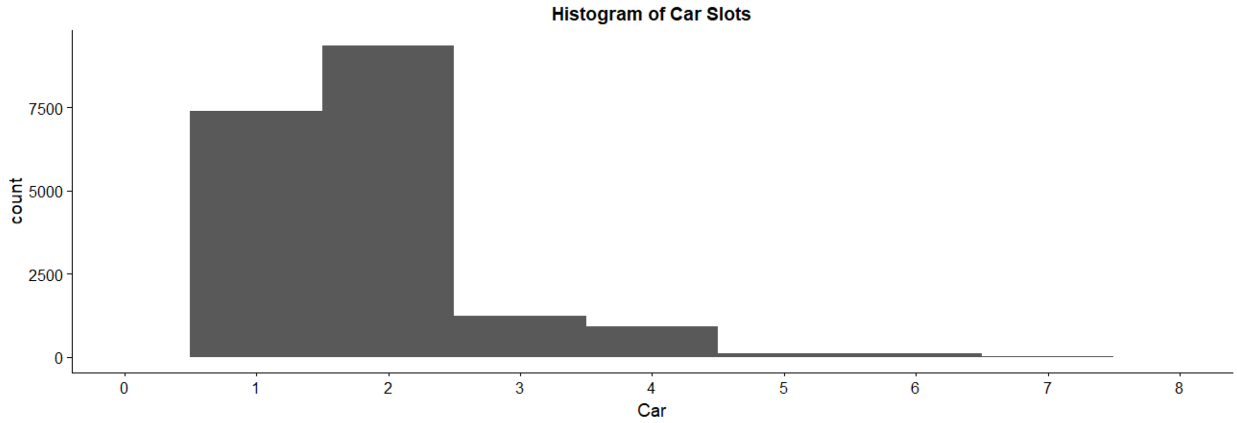
### 4.1.5 Bathroom

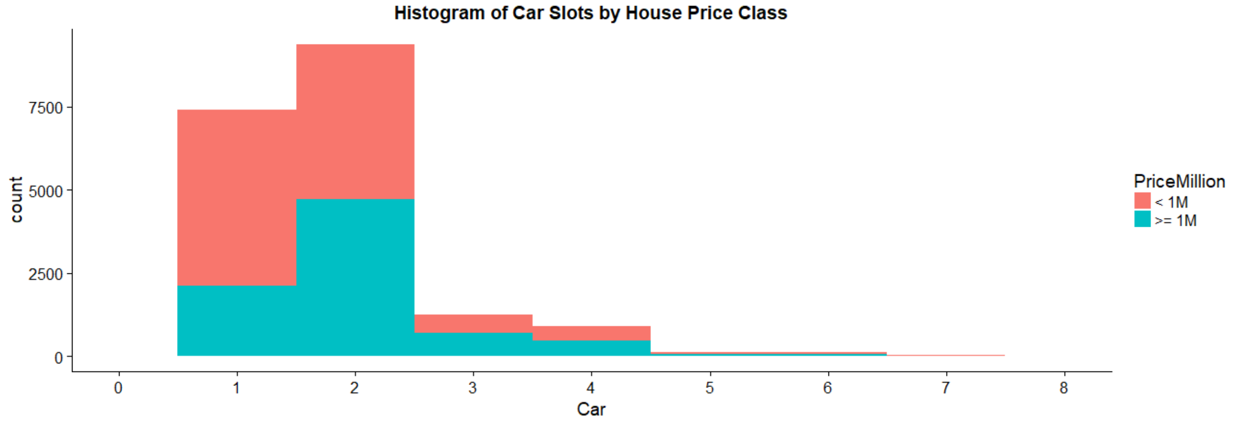




Linked from previous histograms, the houses are common to have more bathrooms with more rooms.

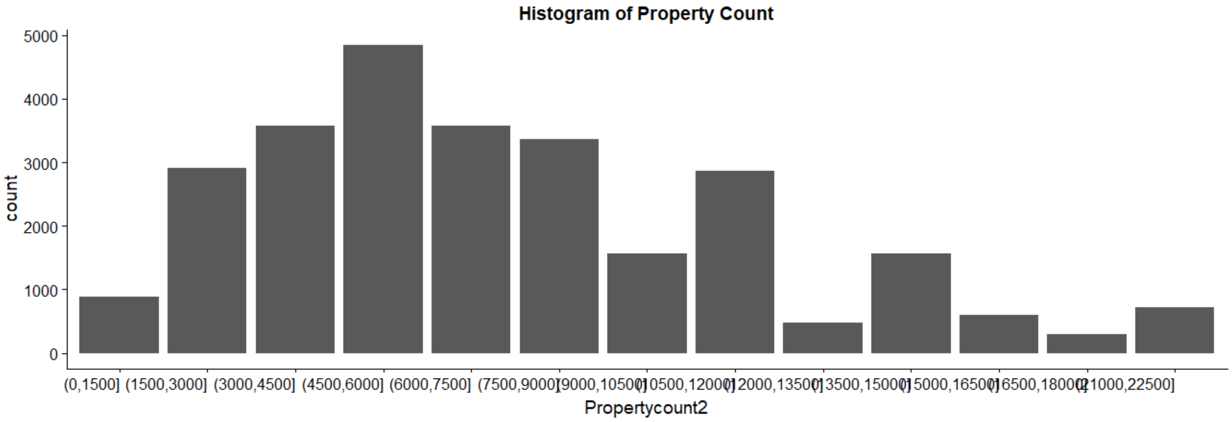
## 4.1.6 Car slots

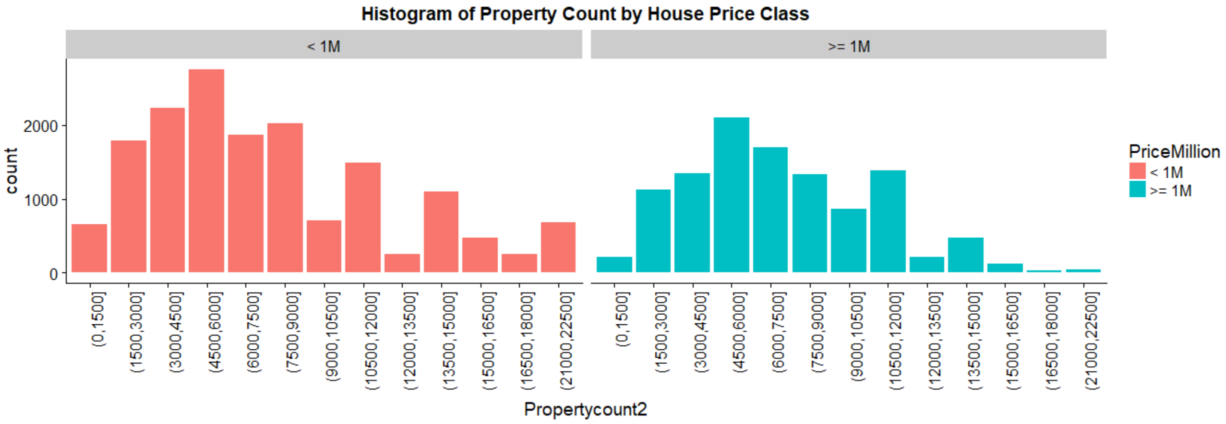

Histogram of Car Slots


Histogram of Car Slots by House Price Class

Number of car slots have inter relationship with land size. Since large land size is generally distributed below one billion dollars, and large building area is distributed over one million dollars, large car slots does not require any aspects other than land size.

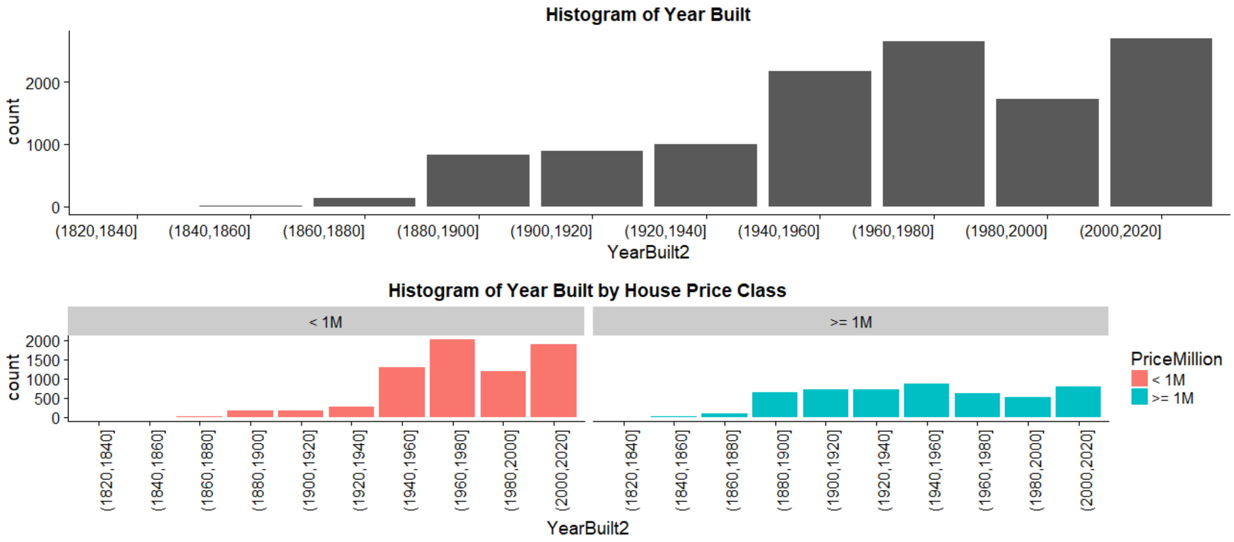
### 4.1.7 Property Count





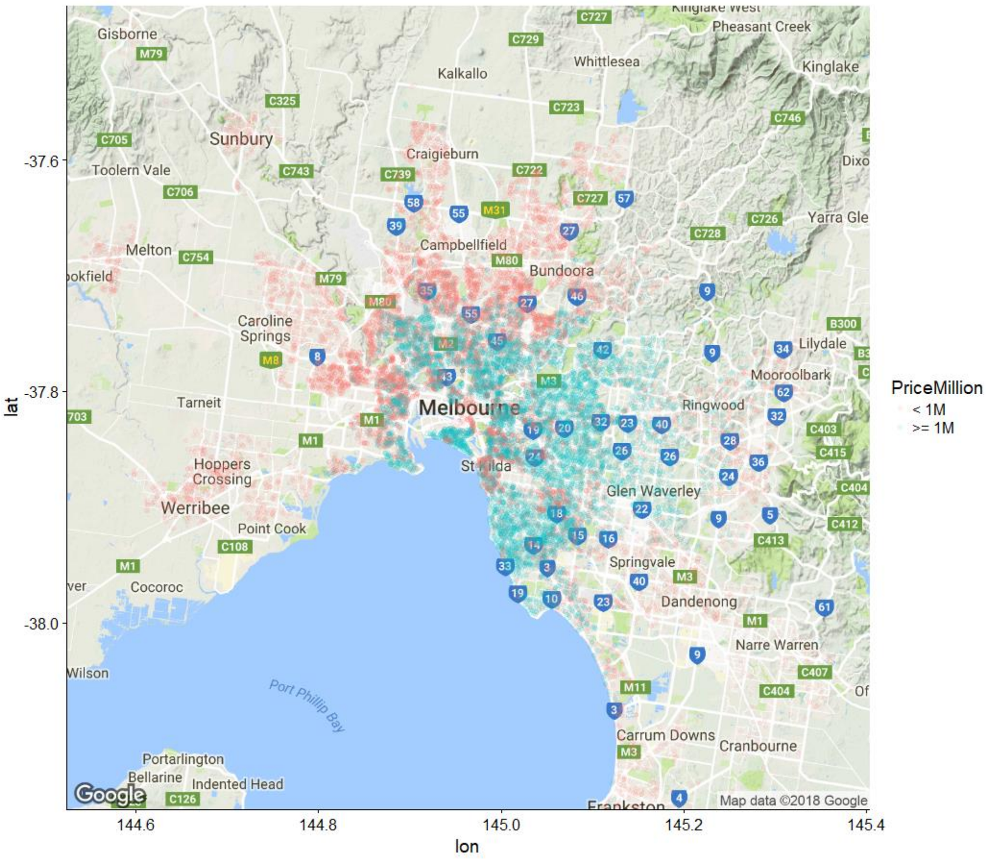
Number of properties in a suburb have one significant feature, in which if there exist 'huge' number of properties in a suburb, then the class of house is more likely to be below one million dollars houses. Two distribution of property counts show very similar shape for left part of property counts.

### 4.1.8 Year Built



Old houses are generally more expensive, and as the houses are newer, price is expected to be cheaper.

### 4.1.9 Longitude & Latitude

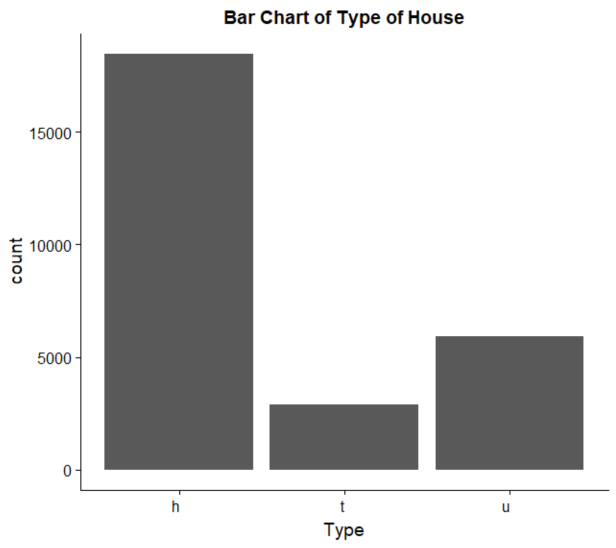


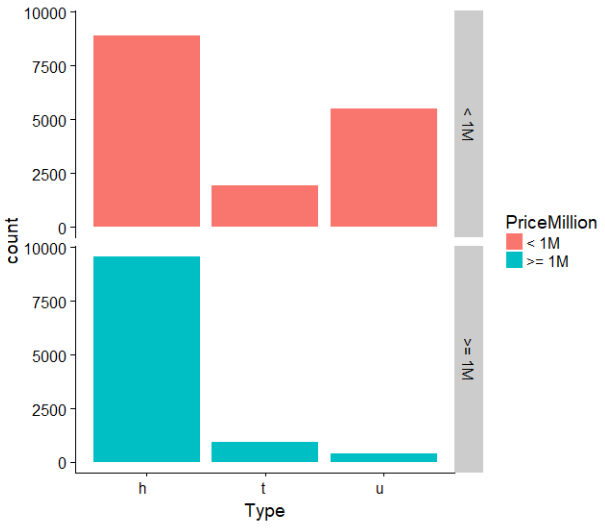
Blue dots are over 1 million dollars houses and red dots are below 1 million dollars houses.

We will discuss this map in categorical features, Region Name.

## 4.2 Categorical Features

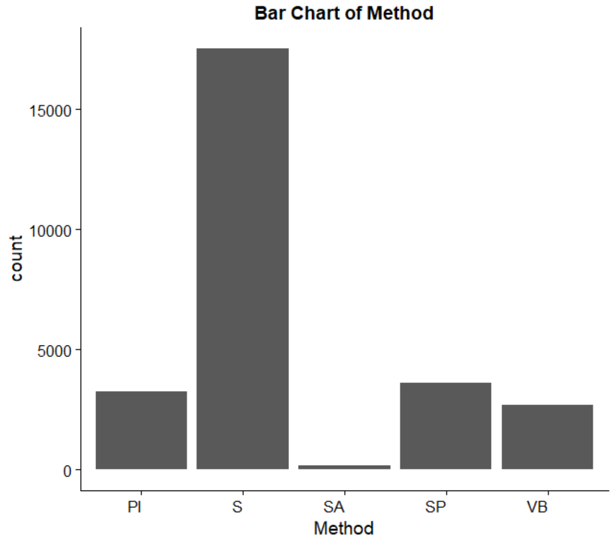
### 4.2.1 Type

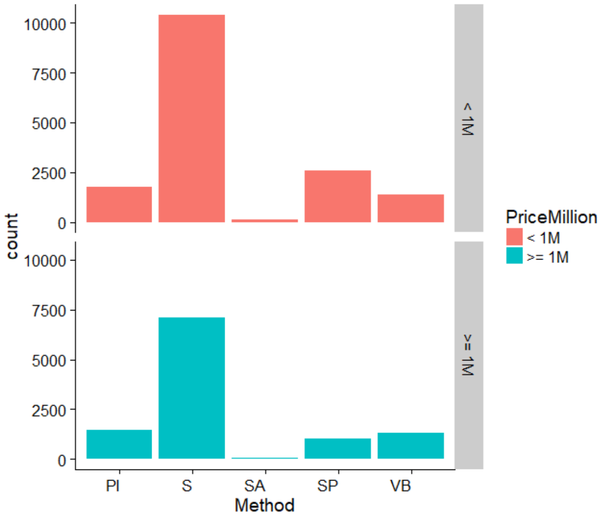

Bar Chart of Type of House



There are small number of unit houses which are over a million dollars, hence unit houses are mostly below one million dollars. Town houses and general type of houses show similar counts, but considering number of over 1 million dollar houses are about 40%, general type of houses are more likely to be over one million dollars.

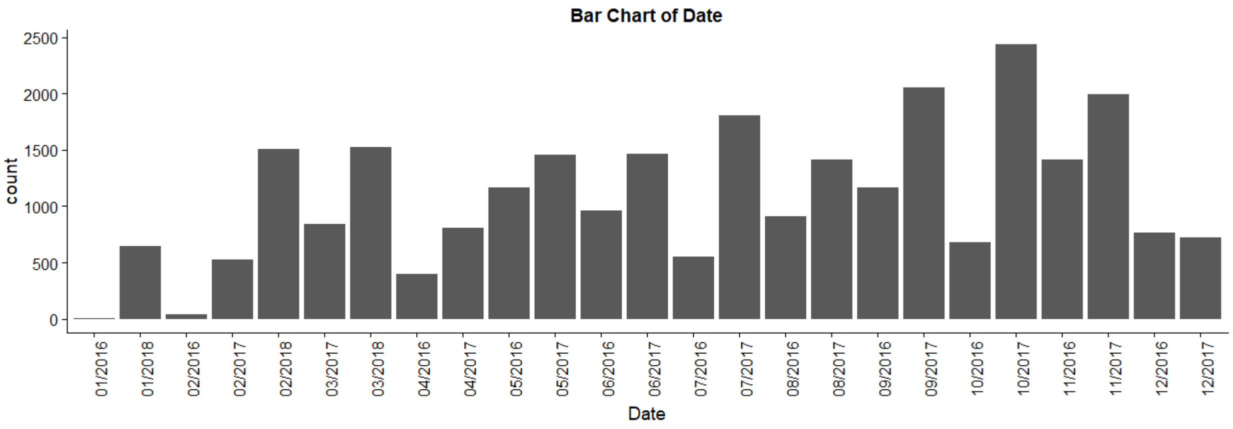
## 4.2.2 Method

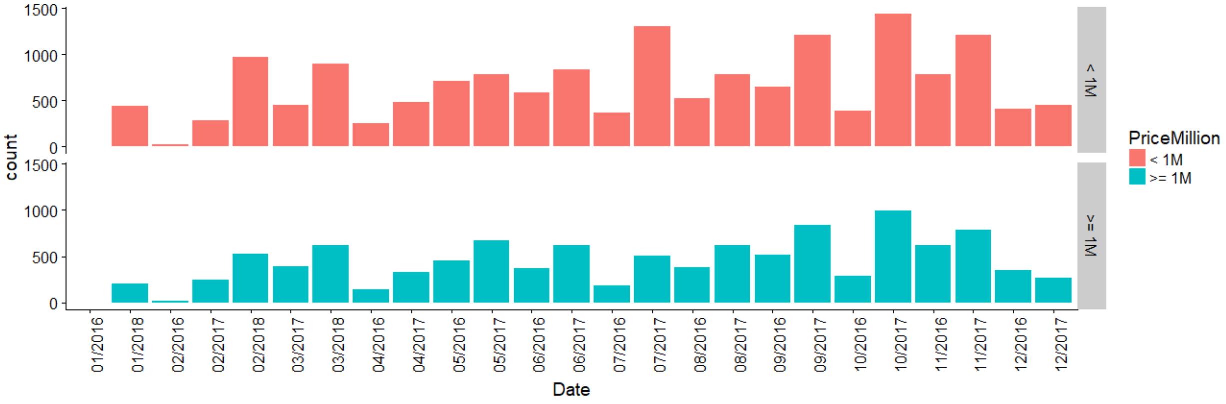




Relationships between price class and method seems to be very weak. Only feature to mention in above bar chart is, SP (property sold prior) is lower for houses with over one million dollars.

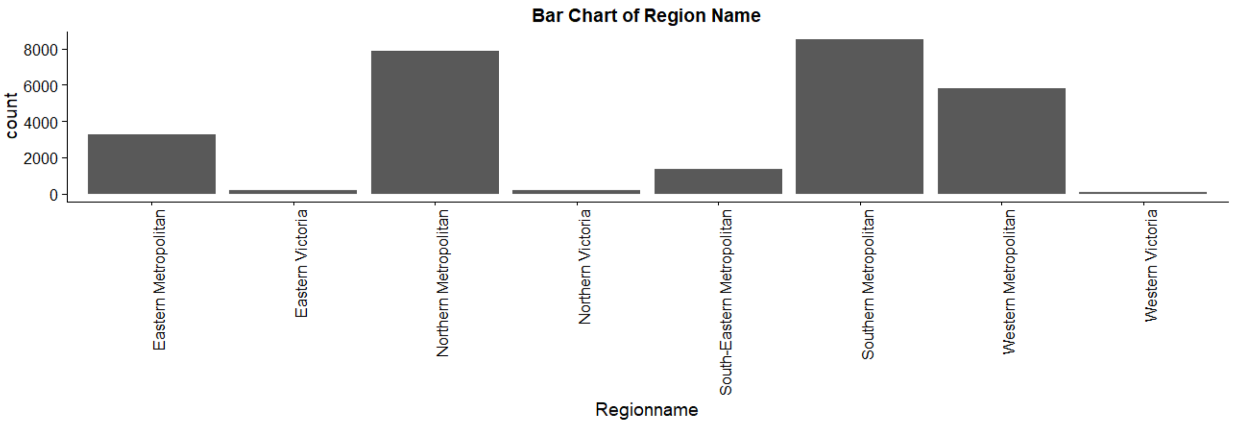
### 4.2.3 Date

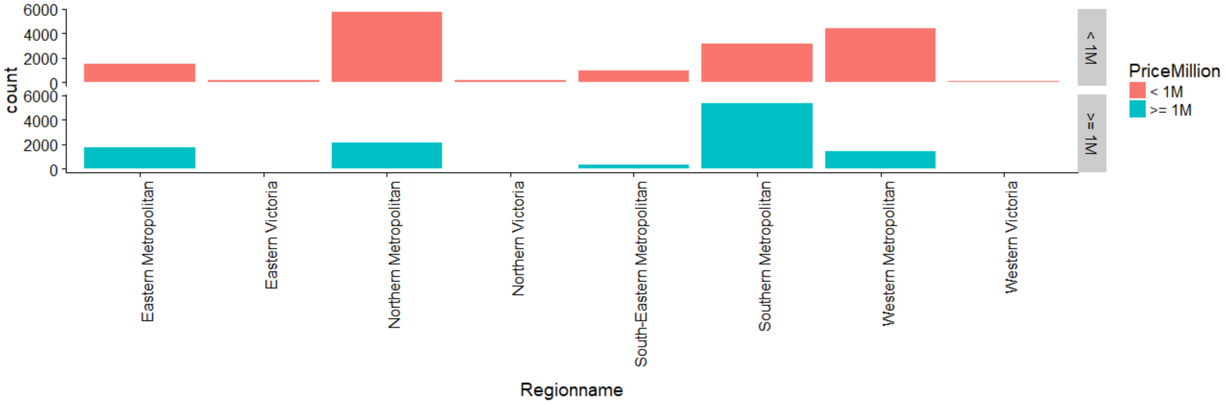




Bar chart of Date provides trend of house prices in Melbourne. Trend of house prices did not change much over past 2 years. Ratio of below and over one million dollars houses is kept.

### 4.2.4 Region name





Houses with over one million dollars are mostly distributed in metropolitan areas. Especially, Southern Metropolitan area is concentrated. Houses in Eastern Metropolitan and Southern Metropolitan tends to have prices over one million dollars. Majority house price of other areas are below one million dollars.

Western Metropolitan area and Northern Metropolitan have few number of houses over one million dollars. Going outside of metropolitan areas, it is rare to find houses over one million dollars.

## 5. Summary

At first, we delete samples with no price value then we divided price into two classes, over one million dollars and below one million dollars to compare. For numerical features, we split numerical values into certain intervals for building area, land size, property count and year built to exclude outliers from the graph. We explored inter relationships between land size and car slots, then building area and number of rooms. Categorical features, method and date did not show significant features in our format. However, region names and longitude/latitude explored concentration of expensive houses in Melbourne and its type of houses.