# Data Processing Project

## Introduction

This data set is a snapshot of Melbourne Housing, including the variables such as `Rooms` meaning number of rooms, `Price` meaning price in dollars, `Regionname` meaning general region, `Propertycount` meaning number of properties that exist in the suburb, `Distance` meaning the distance from CBD and so on. Melbourne is experiencing a housing bubble as well as the cities in Chine such as Beijing, Shanghai and so on. We want use this data set to find a trend and give a prediction for the price of house.

In this project, we try to clean and process the data firsly. Then we will do some descriptive statistics analysis through tables and plots. Finally we will build a linear regression model to predict the price of the house using these predictors.

## Data processing

We use the data in https://www.kaggle.com/anthonypino/melbourne-housing-market. We download the data set from the above website and import the data set into R:

```r
# read the data
data = read.csv("melb_data.csv")
```

Since we are interested in the above five variables `Rooms`, `Price`, `Regionname`, `Propertycount`,`Distance`, we delete the observations, which has missing values in these five variables and extract these five variables for the following analysis:

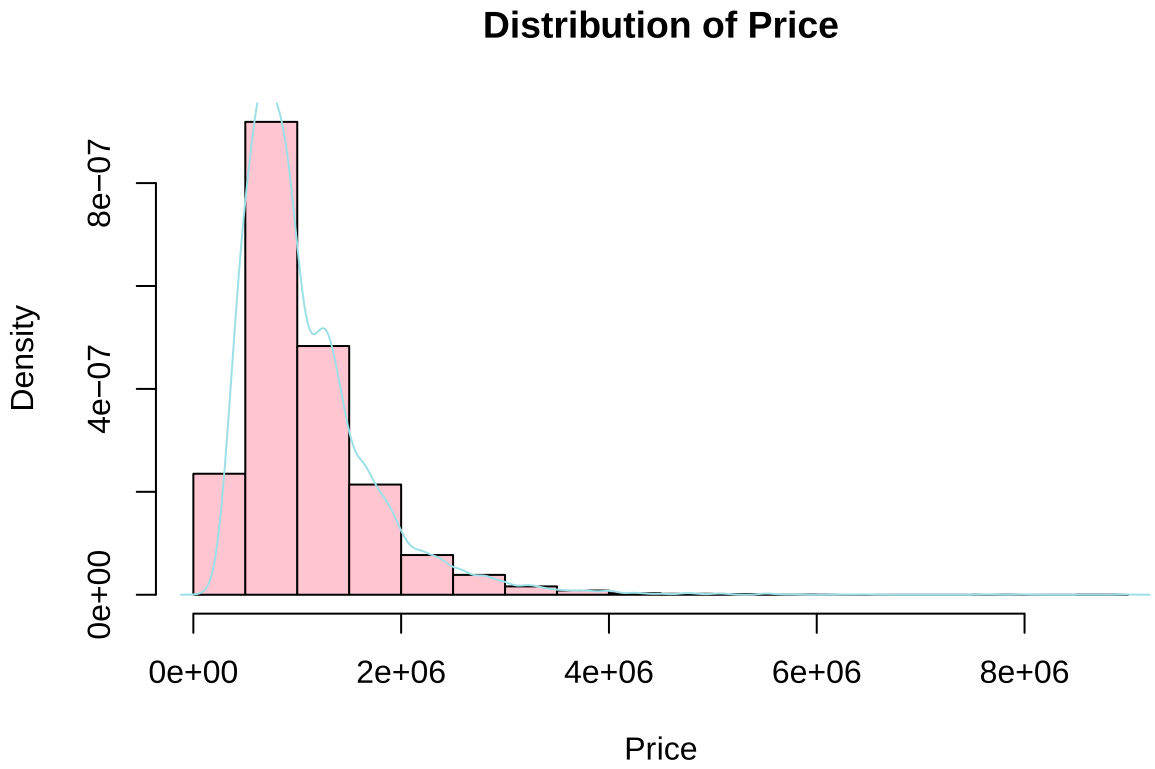
```r
# delete the missing values
index = !is.na(data$Price) & !is.na(data$Rooms) &
  !is.na(data$Regionname) & !is.na(data$Propertycount) &
  !is.na(data$Distance)
# extract these five variables
data = data[index,
            c("Price","Rooms","Regionname","Propertycount","Distance")]
```

Now the data set consists 13580 oservations.

## Descriptive statistics analysis

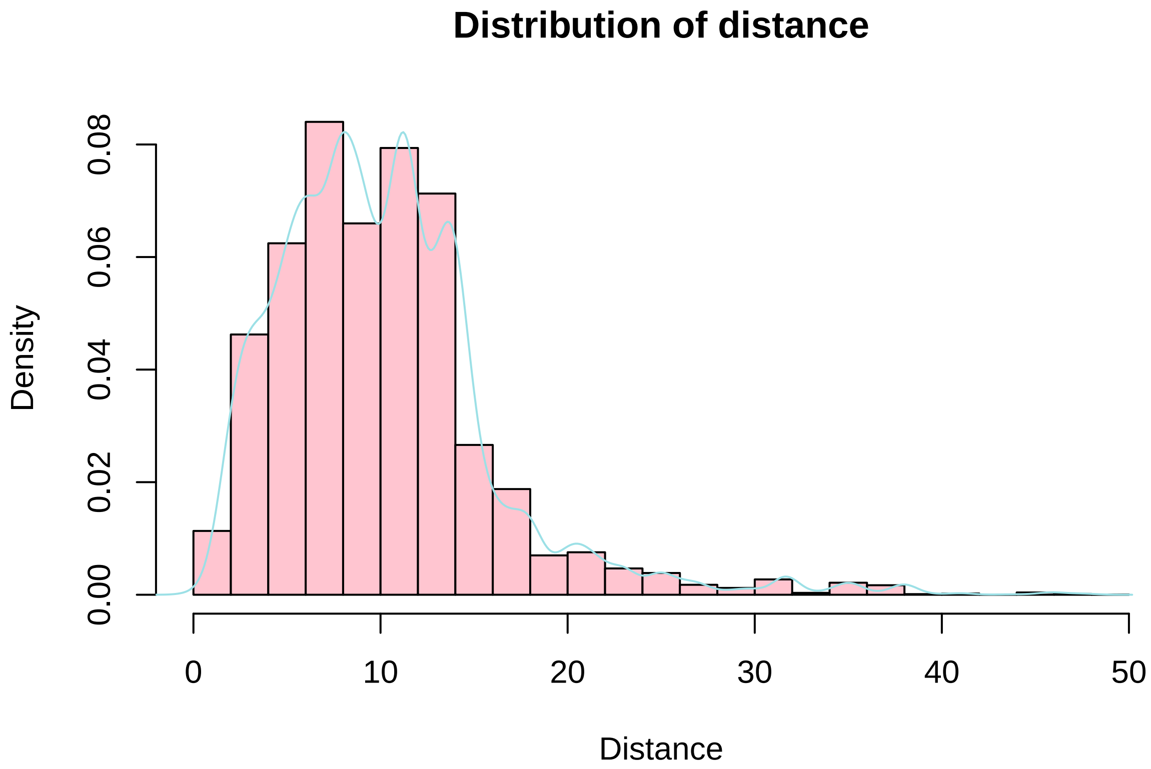
Firstly, we make a histogram for the `Price` to see the distribution:

```r
# histogram of price
hist(data$Price,prob=T,main="Distribution of Price",
     xlab="Price",col=hcl(0),breaks=20)
lines(density(data$Price),col=hcl(200))
```



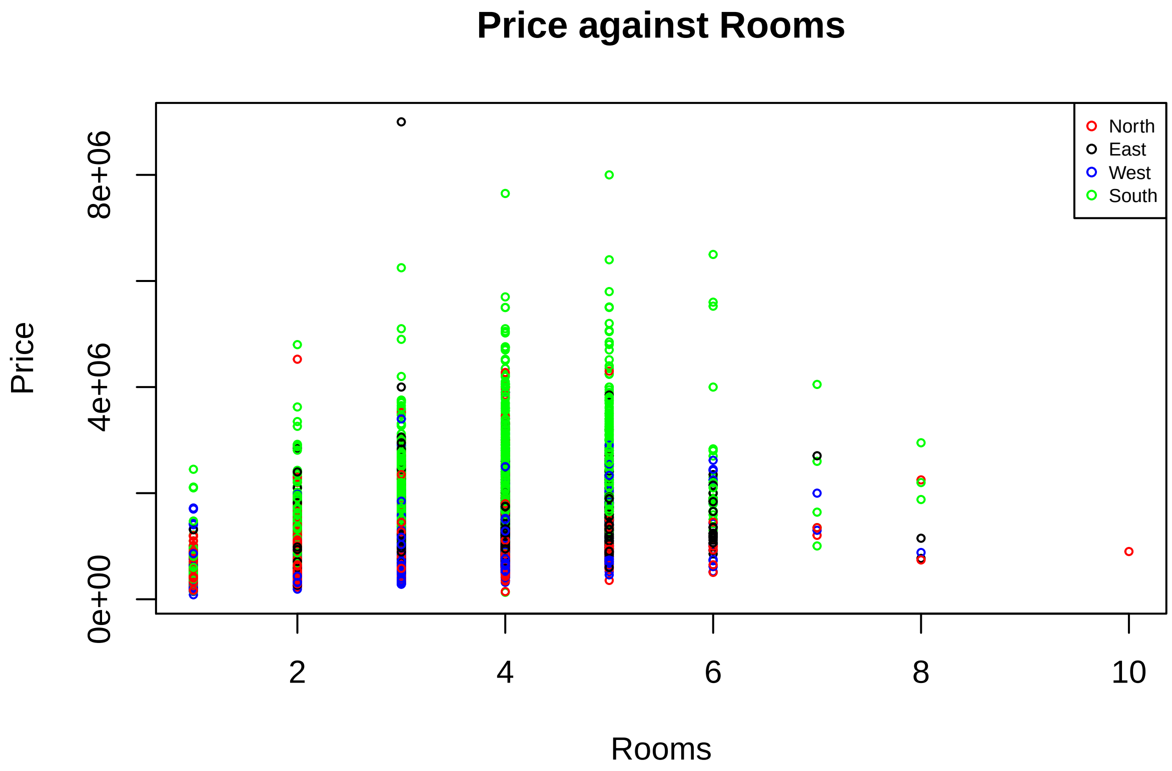
From the plot, we can see that `Price` is right skewed. Secondly, We make a histogram for the `Distance` to see the distribution:

```
# histogram of distance
hist(data$Distance,prob=T,xlab='Distance',
     main='Distribution of distance',col=hcl(0),breaks=20)
lines(density(data$Distance),col=hcl(200))
```



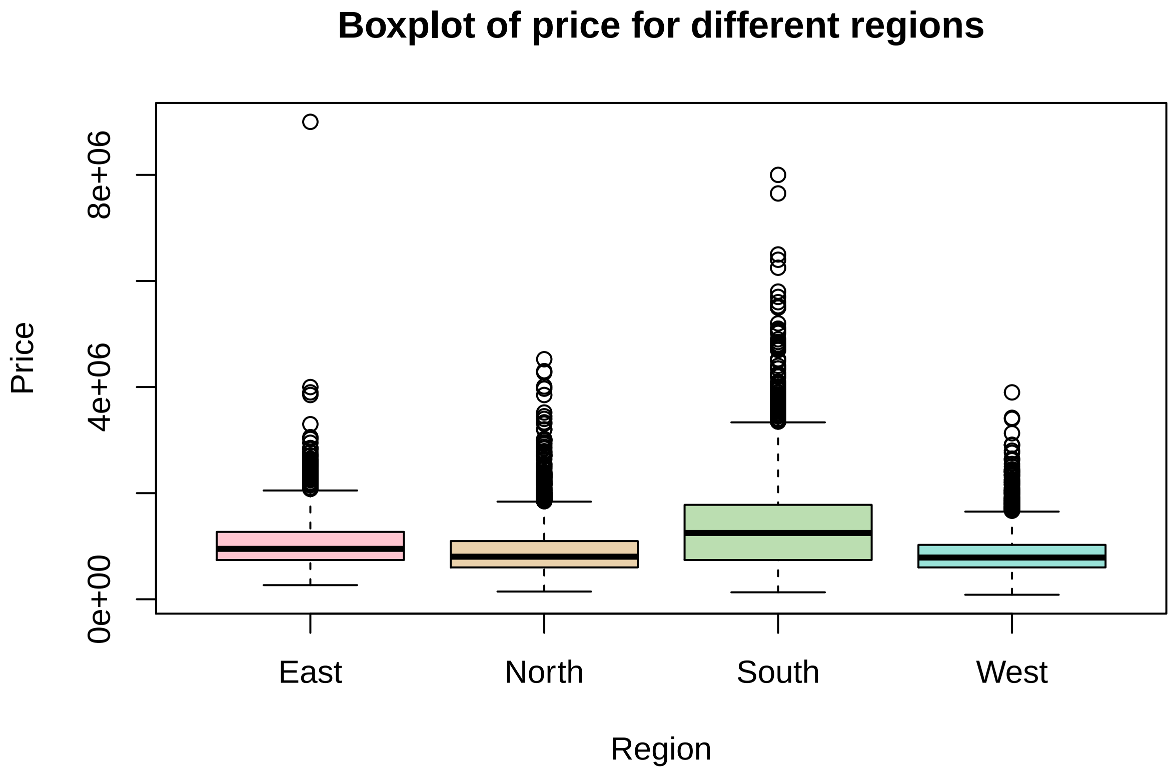
which is also a little right skewed. Thirdly, we make a scatter plot for `Price` and `Rooms` with `Price` at the vertical axis and color the data by `Regionname`

```
# combine the Region into 4 classes
data[data$Regionname=="Northern Metropolitan" |
       data$Regionname=="Northern Victoria","Regionname"] = "North"
data[data$Regionname=="Eastern Metropolitan" |
       data$Regionname=="Eastern Victoria" |
       data$Regionname=="South-Eastern Metropolitan",
     "Regionname"] = "East"
data[data$Regionname=="Western Metropolitan" |
       data$Regionname=="Western Victoria","Regionname"] = "West"
data[data$Regionname=="Southern Metropolitan" |
       data$Regionname=="Southern Victoria","Regionname"] = "South"
# define the color
data$col = NULL
for (i in 1:nrow(data)){
  if (data[i,"Regionname"]=="North"){data[i,"col"]="red"}
  else if (data[i,"Regionname"]=="East"){data[i,"col"]="black"}
  else if (data[i,"Regionname"]=="West"){data[i,"col"]="blue"}
  else if (data[i,"Regionname"]=="South"){data[i,"col"]="green"}
}
# scatter plot
plot(Price~Rooms,data=data,cex=0.5,main="Price against Rooms",
     col =data$col)
legend("topright",legend=c("North","East","West","South"),
       col=c("red","black","blue","green"),cex=0.6,pch=c(1,1,1,1))
```



From the scatter plot, we can see that as the number of rooms changes from 1 to 5, the price increases gradually. Furthermore, we can see that the price of house in southern area is the highest among the four regions while the prices of house in northern and western area are lower. We can also obtain the same conclusion through a boxplot for the `Price` grouped by `Region`

```
# boxplot for price grouped by region
boxplot(data$Price~data$Regionname,
     col=hcl(c(0,60,120,180,240,300,360,420)),
     xlab="Region",ylab="Price",
     main="Boxplot of price for different regions")
```



### Linear regression

Finally, we build a linear regression model using `Price` as dependent variable and other four variables as predictors.

```r
library(knitr)
# linear regression model
model = lm(Price~Rooms+Regionname+Propertycount+Distance,data=data)
```

The regression result is as follows:

```r
# regression result
kable(summary(model)$coef,
      caption="Coefficients estimates of linear regression")
```

Table 1: Coefficients estimates of linear regression

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---:|---:|---:|---:|
| (Intercept) | 3.552257e+05 | 2.096700e+04 | 16.942132 | 0.0000000 |
| Rooms | 3.983395e+05 | 4.385979e+03 | 90.821102 | 0.0000000 |
| RegionnameNorth | -2.318262e+05 | 1.485028e+04 | -15.610896 | 0.0000000 |
| RegionnameSouth | 2.186858e+05 | 1.404608e+04 | 15.569167 | 0.0000000 |
| RegionnameWest | -3.058005e+05 | 1.436864e+04 | -21.282501 | 0.0000000 |
| Propertycount | -2.205875e+00 | 9.722249e-01 | -2.268893 | 0.0232904 |
| Distance | -3.697188e+04 | 7.981737e+02 | -46.320588 | 0.0000000 |

From the results, we can see that the number of `Rooms` has a positive impact on the house price. The house in southern area is highest and then the house in eastern area and then the house in northern area and then the house in western area. The distance has a negative impact on the house price, which means that the more closed the house with CBD, the higher price the house has.

Finally, we export the data and save it as “Melbourne_Housing.csv”.

```
write.table(data,"Melbourne_Housing.csv",
            row.names=FALSE,col.names=TRUE,sep=",")
```