THE HOUSING MARKET IN MELBOURNE, AUSTRALIA

By Paul Croome, Rodrigo Rosales Alvarez, and Viktoriia Shcherbachuk
DATA 602 - Project
Assistant Professor: Dr. Thuntida Ngamkham
University of Calgary
Calgary, Alberta
October 18, 2022

Introduction

Purchasing and owning a home is an important part of many people's lives. Diverse research has shown that home ownership is positively related to one's life satisfaction, via such factors as happiness, self-esteem, confidence, and security (Dietz and Haurin, 2003). However, purchasing a home is also a very difficult decision, as it is one of the most expensive purchases that most people will make in their lifetime and it can lead to significant amounts of debt. In fact, 51% of all wealth in Australian households is in real estate (Australian Council of Social Service, n.d.). In a time when the cost of living is continually rising in many areas of the world, it is increasingly important for people to make informed and economical decisions when purchasing a home.

Many factors can influence housing prices, including factors which are both internal and external to the house. For example, an external factor that is often negatively related to the price of a house is the distance of the house from its city's central business district (CBD). That is, the further a house is from the CBD, the lower its price (Gaolu, 2015; Teye, de Haan and Elsinga, 2017).

In our project, we examined real estate data from Melbourne, Australia, collected between January 28, 2016 and September 23, 2017, in order to determine how several internal and external factors of housing in Australia's second most populous city relate to housing prices.

Topics to Investigate

For our project, we will investigate the following hypotheses:

1. There is a negative correlation between housing prices and the distance of a property from Melbourne's CBD.
   Research from several large cities has found a negative relationship between housing prices and the distance of the house from its city's central business district (CBD) or city center (Gaolu, 2015; Teye, de Haan and Elsinga, 2017). We investigated whether this trend holds in Melbourne as well.

2. The different features of a house – e.g., the year it was built, the number of bedrooms/bathrooms, the property size, etc. – are related to the price of the property. We sought to determine the variables which are most powerfully related to housing prices in order to help identify the variables that make housing most expensive.
   To achieve this, we sought to build a regression model with the top 2 independent variables which affect housing prices the most in Melbourne's real estate market.

3. There is a difference in housing prices between the 'Metropolitan' (city area) and 'Victoria' (suburbs) regions of Melbourne.

Project planning

1.Data management

For this project, we selected the database "Melbourne housing snapshot" (Source: https://www.kaggle.com/datasets/dansbecker/melbourne-housing-snapshot), created by Tony Pino and posted in Kaggle for public use and research. This data was scraped from publicly available results posted every week from domain.com.au and cleaned by him. This is a well-structured, tabular database with 13580 rows and 21 columns which was collected between January 28, 2016, and September 23, 2017. The population of this dataset includes houses that were sold within the aforementioned timeframe. Our variables of interest include: house's selling price, date the house was sold, distance of the house from Melbourne's Central Business District (CBD), number of bedrooms and bathrooms in the house, house's number of parking spots, size of the land where the property stands, size of the building, year when the property was built, housing type, and name of region in Melbourne.

Data gathering

Below is a list of libraries which we used in our research project.

In order to select the variables for our analyses, we first downloaded and previewed our dataset. We then took into consideration the type of variables with which we would be working.

```
melbourne <- read_csv('melb_data.csv',show_col_types = FALSE)
head(melbourne,4)
```

```
## # A tibble: 4 x 21
##   Suburb Address Rooms Type   Price Method SellerG Date  Dista~1 Postc~2 Bedro~3
##   <chr>  <chr>   <dbl> <chr>  <dbl> <chr>  <chr>   <chr>   <dbl>   <dbl>   <dbl>
## 1 Abbot~ 85 Tur~     2 h     1.48e6 S      Biggin  3/12~     2.5    3067       2
## 2 Abbot~ 25 Blo~     2 h     1.03e6 S      Biggin  4/2/~     2.5    3067       2
## 3 Abbot~ 5 Char~     3 h     1.46e6 SP     Biggin  4/3/~     2.5    3067       3
## 4 Abbot~ 40 Fed~     3 h     8.5 e5 PI     Biggin  4/3/~     2.5    3067       3
## # ... with 10 more variables: Bathroom <dbl>, Car <dbl>, Landsize <dbl>,
## #   BuildingArea <dbl>, YearBuilt <dbl>, CouncilArea <chr>, Lattitude <dbl>,
## #   Longtitude <dbl>, Regionname <chr>, Propertycount <dbl>, and abbreviated
## #   variable names 1: Distance, 2: Postcode, 3: Bedroom2
```

To focus more on our dependent variable "Price", and the other independent variables with which we are interested for this project, we created another data frame including only our desired variables, and considered the types of each of these variables.

```
mel_research<-select(melbourne,Rooms,Type,Price,Date,Distance,
                  Bathroom,Car,Landsize,BuildingArea,YearBuilt,
                  Regionname)
str(mel_research)
```

```
## tibble [13,580 x 11] (S3: tbl_df/tbl/data.frame)
##  $ Rooms       : num [1:13580] 2 2 3 3 4 2 3 2 1 2 ...
##  $ Type        : chr [1:13580] "h" "h" "h" "h" ...
##  $ Price       : num [1:13580] 1480000 1035000 1465000 850000 1600000 ...
##  $ Date        : chr [1:13580] "3/12/2016" "4/2/2016" "4/3/2017" "4/3/2017" ...
##  $ Distance    : num [1:13580] 2.5 2.5 2.5 2.5 2.5 2.5 2.5 2.5 2.5 2.5 ...
```

```
##  $ Bathroom    : num [1:13580] 1 1 2 2 1 1 2 1 1 1 ...
##  $ Car         : num [1:13580] 1 0 0 1 2 0 0 2 1 2 ...
##  $ Landsize    : num [1:13580] 202 156 134 94 120 181 245 256 0 220 ...
##  $ BuildingArea: num [1:13580] NA 79 150 NA 142 NA 210 107 NA 75 ...
##  $ YearBuilt   : num [1:13580] NA 1900 1900 NA 2014 ...
##  $ Regionname  : chr [1:13580] "Northern Metropolitan" "Northern Metropolitan" "Northern Metropolitan
```

Data cleaning

The next step we took was to define the missing values in our selected data base.

```
sapply(mel_research,function(x)  sum(is.na(x)))
```

```
##        Rooms         Type        Price         Date     Distance     Bathroom
##            0            0            0            0            0            0
##          Car     Landsize BuildingArea    YearBuilt   Regionname
##           62            0         6450         5375            0
```

The table above shows the immense missing values in the 'BuildingArea' and 'YearBuilt' variables. Because these factors are important for our analyses, we decided to delete these missing values in order to include these variables going forward.

```
mel_final<-na.omit(mel_research)
```

During the data observation process, outliers were located in the variables "Price", "BuildingArea", and "YearBuilt". To ensure the most accurate results from our analyses, we transformed the necessary variables and deleted the outliers identified.

```
qtl <- quantile(mel_final$Price, probs=c(.25, .75))
iqr <- IQR(mel_final$Price)
mel_final <- subset(mel_final, mel_final$Price >
                      (qtl[1] - 1.5*iqr) & mel_final$Price <
                      (qtl[2]+1.5*iqr))

qtl <- quantile(mel_final$BuildingArea, probs=c(.25, .75))
iqr <- IQR(mel_final$BuildingArea)
mel_final <- subset(mel_final, mel_final$BuildingArea >
                      (qtl[1] - 1.5*iqr) & mel_final$BuildingArea
                    < (qtl[2]+1.5*iqr))

qtl <- quantile(mel_final$YearBuilt, probs=c(.25, .75))
iqr <- IQR(mel_final$YearBuilt)
mel_final <- subset(mel_final, mel_final$YearBuilt >
                      (qtl[1] - 1.5*iqr) & mel_final$YearBuilt <
                      (qtl[2]+1.5*iqr))
```

```
dim(mel_final)
```

```
## [1] 6247   11
```

After deleting the missing values and outliers, 6247 cases remained for our the research. This is still a considerable number of cases, and was deemed sufficient for project continuation.

Data wrangling

Upon investigating the types of variables in the data set, we determined it was necessary to conduct certain transformations with the several of the variables in order for the correct analysis, presentation, and interpretation of the data.

First, we created a new variable "date" in year-month-day format:

```
mel_final <- mel_final %>%
  mutate(date = as.Date(Date, format = "%d/%m/%Y"))
mel_final$Month<-months(mel_final$date)
mel_final$Year<-format(mel_final$date,format="%y")
```

Next, we transformed the 'Type' variable to make its possible values more easily readable and descriptive.

```
mel_final$Type <- mapvalues(mel_final$Type,
          from=c("h","u","t"),
          to=c("house","unit","townhouse"))
mel_final$Type<-as.factor(mel_final$Type)
str(mel_final$Type)
```

```
##  Factor w/ 3 levels "house","townhouse",..: 1 1 1 1 1 1 1 3 1 1 ...
```

```
Regionname<-as.factor(mel_final$Regionname)
summary(Regionname)
```

```
##       Eastern Metropolitan           Eastern Victoria
##                        608                         23
##      Northern Metropolitan          Northern Victoria
##                       1944                         25
## South-Eastern Metropolitan      Southern Metropolitan
##                        202                       1955
##       Western Metropolitan           Western Victoria
##                       1471                         19
```

In general, the vast majority of the properties in our data set are located in the Metropolitan areas of Melbourne, while only 67 properties are located in the suburbs (or Victorian regions) of Melbourne (Western, Easter and Norther Victoria).

Outlook of dependent variable Price

Before beginning our statistical analyses, it was important to describe our dependent variable, Price:

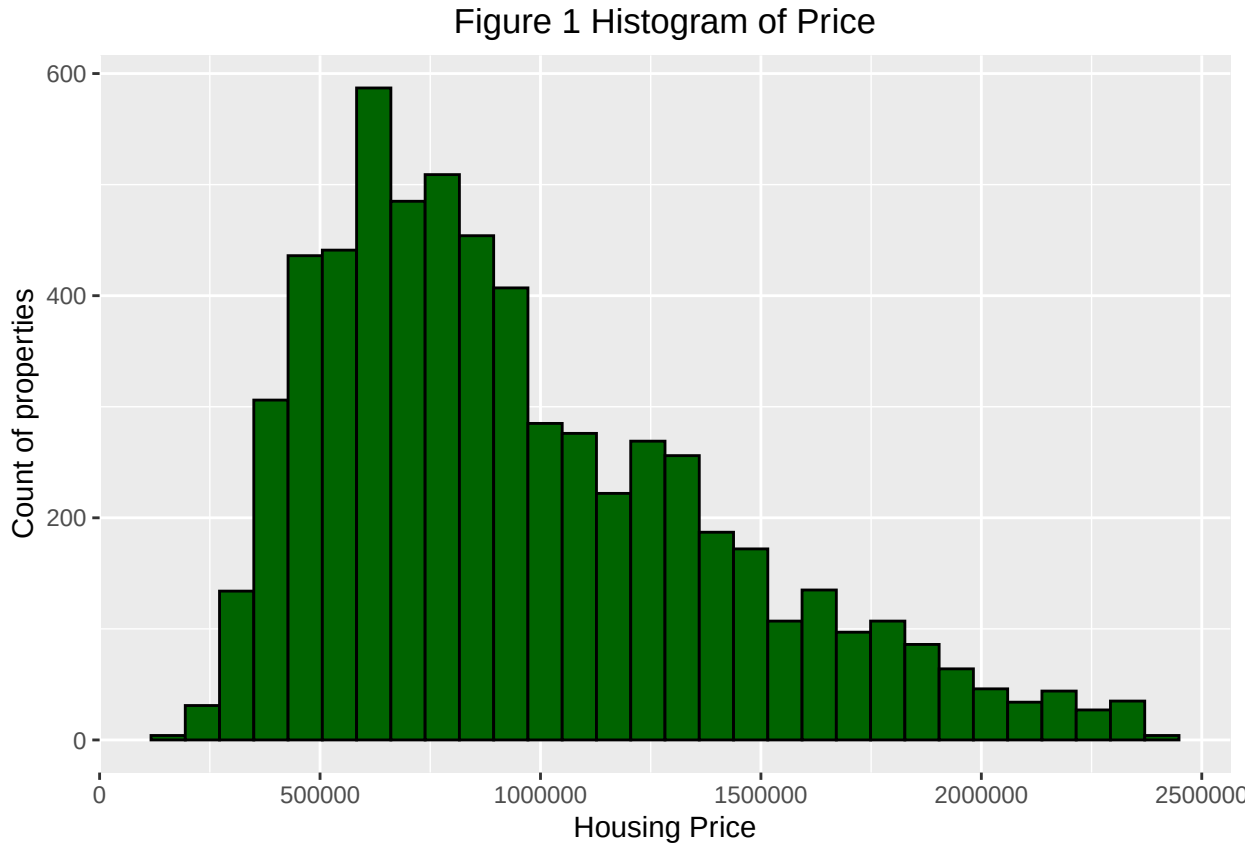
```
favstats(~Price, data = mel_final)
```

```
##     min     Q1 median      Q3     max     mean       sd    n missing
##  131000 610000 850000 1230000 2385000 953559.4 450804.9 6247       0
```

From the code above, we could see that mean price of the property in Melbourne was 953,559.40 AUD with standard deviation is 450,804.90 AUD. These values represent the considerable spread in property prices in Melbourne in general, even after removing outliers. Following this realization, we decided to illustrate the distribution of the price variable.

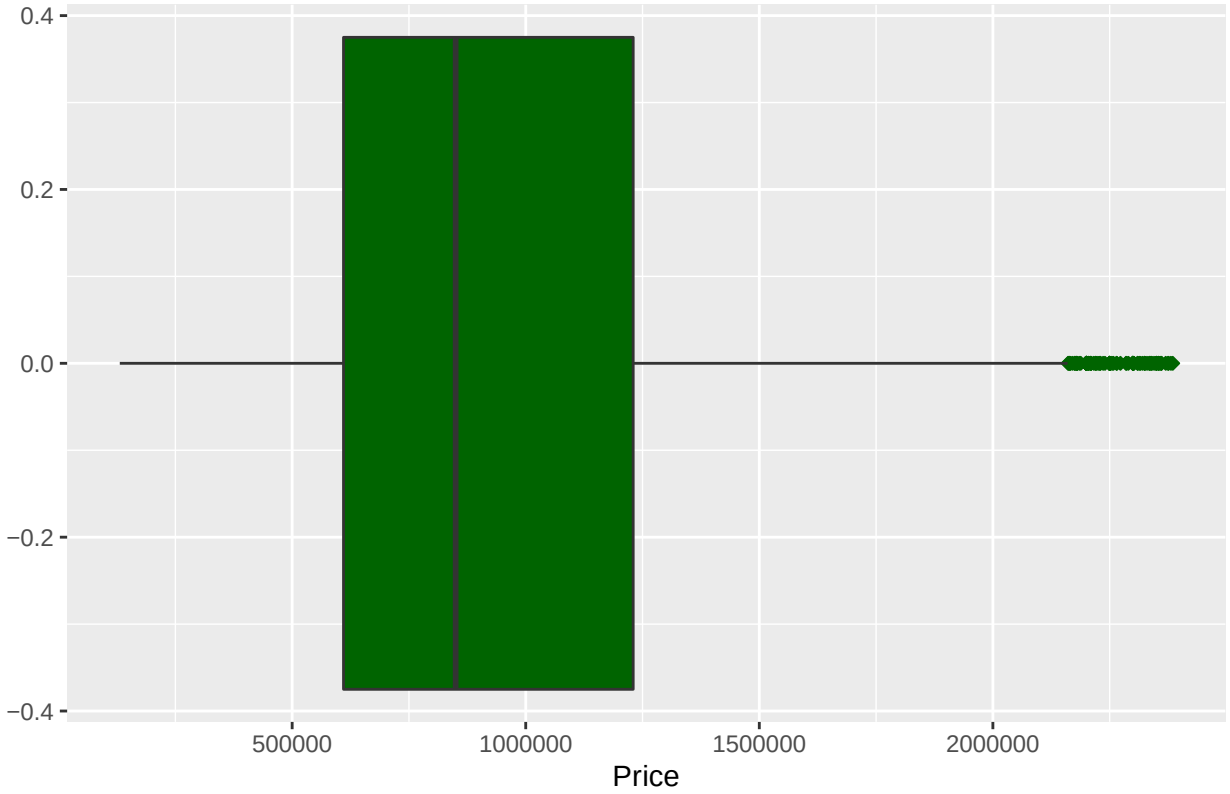
```
options(scipen=10000)
ggplot(mel_final, aes(x = Price, fill = ..count..)) +
  geom_histogram(color="black", fill="darkgreen") +
  ggtitle("Figure 1 Histogram of Price") +
  ylab("Count of properties") +
  xlab("Housing Price") +
  theme(plot.title = element_text(hjust = 0.5))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(mel_final, aes(x=Price)) +
  geom_boxplot(fill="darkgreen", outlier.color = 'darkgreen',outlier.shape = 23)+
  ggtitle("Figure 2 Boxplot of Price variable")
```

Figure 2 Boxplot of Price variable



From the histogram and boxplot above, it is clear that the distribution of our target variable Price is skewed to the right. After we removed outliers, there are still a considerable number of properties in Melbourne which are remarkably costly, exceeding even the maximum interval of the price distribution in our boxplot.

2. Statistical Analysis

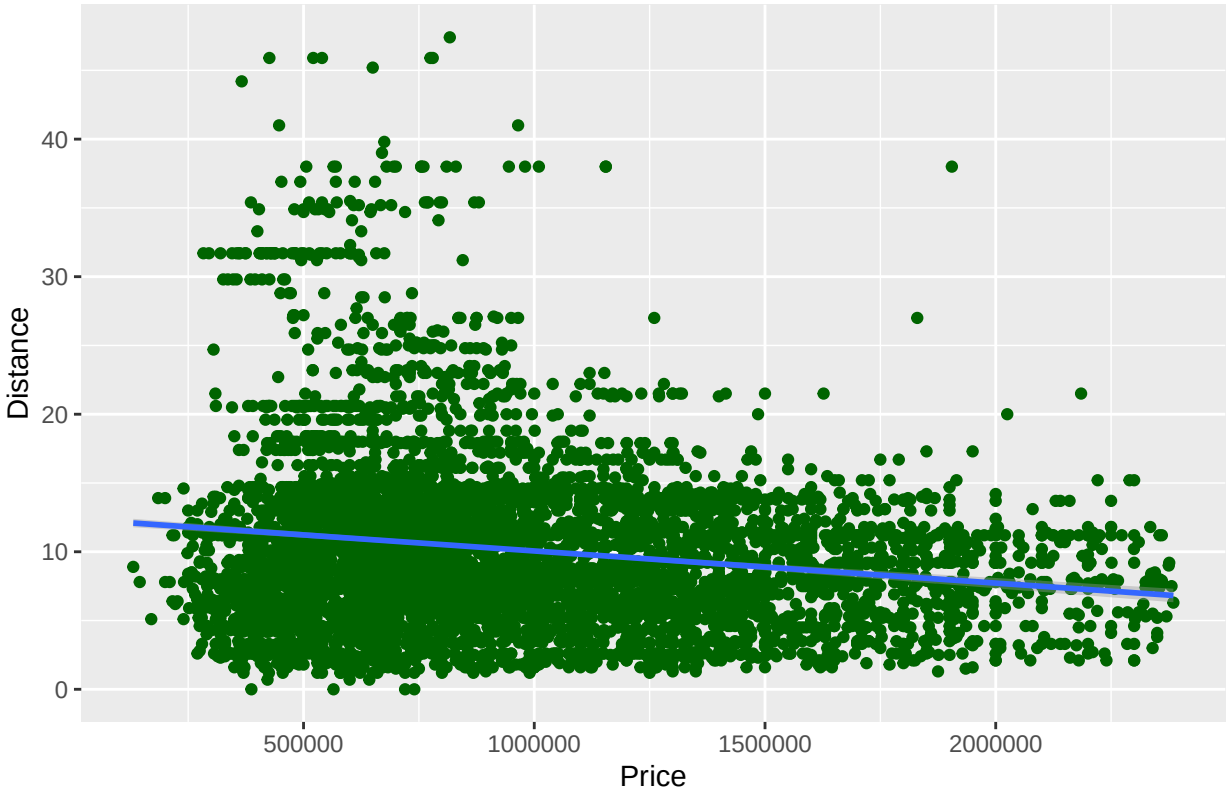
1.There is a negative correlation between housing prices and the distance of a property from Melbourne's CBD.

In order to answer this question, we created a correlation plot. The scatter plot below represents the relationship between housing prices and the 'Distance' variable, which is the distance of a house from the city center (CBD) of Melbourne.

```
ggplot(data = mel_final, aes(x = Price, y = Distance)) +
  geom_point(color = 'darkgreen') +
  stat_smooth(method = 'lm')+ggtitle("Figure 3 Correlation plot between Price and Distance variables")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

Figure 3 Correlation plot between Price and Distance variables



From the above scatter plot, we can see that there is a weak negative relationship between these two variables. To determine the exact strength of this relationship, we calculated the Pearson correlation coefficient below.

```
cor(mel_final$Price, mel_final$Distance)
```

```
## [1] -0.1738758
```

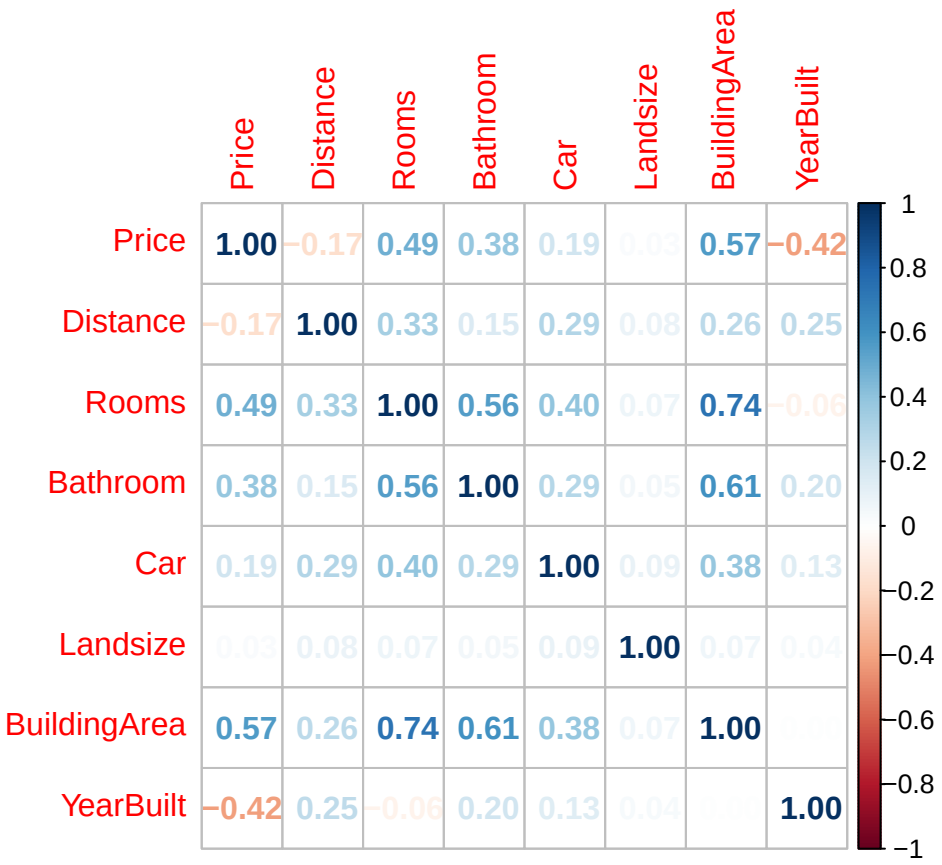
Conclusion: calculating the correlation coefficient using the 'cor()' function above, we can see that the correlation is about = -0.1739. It is weak correlation but still significant. Therefore, we can confirm our first hypothesis that there is a negative relationship between the price of a property and its distance from the CBD of Melbourne. In other words, the closer housing is located to the city center of Melbourne, the more costly the property is.

2.The different features of a house are related to the price of the property.

To investigate this question, we first created a data frame including only our numeric variables. Then, we constructed a correlation plot to explore the relationships between these variables.

```
mel_cor<-select(mel_final,Price,Distance,Rooms,Bathroom,Car,Landsize,BuildingArea,YearBuilt)
M<-cor(mel_cor)
```

```
corrplot(M, method="number")
```



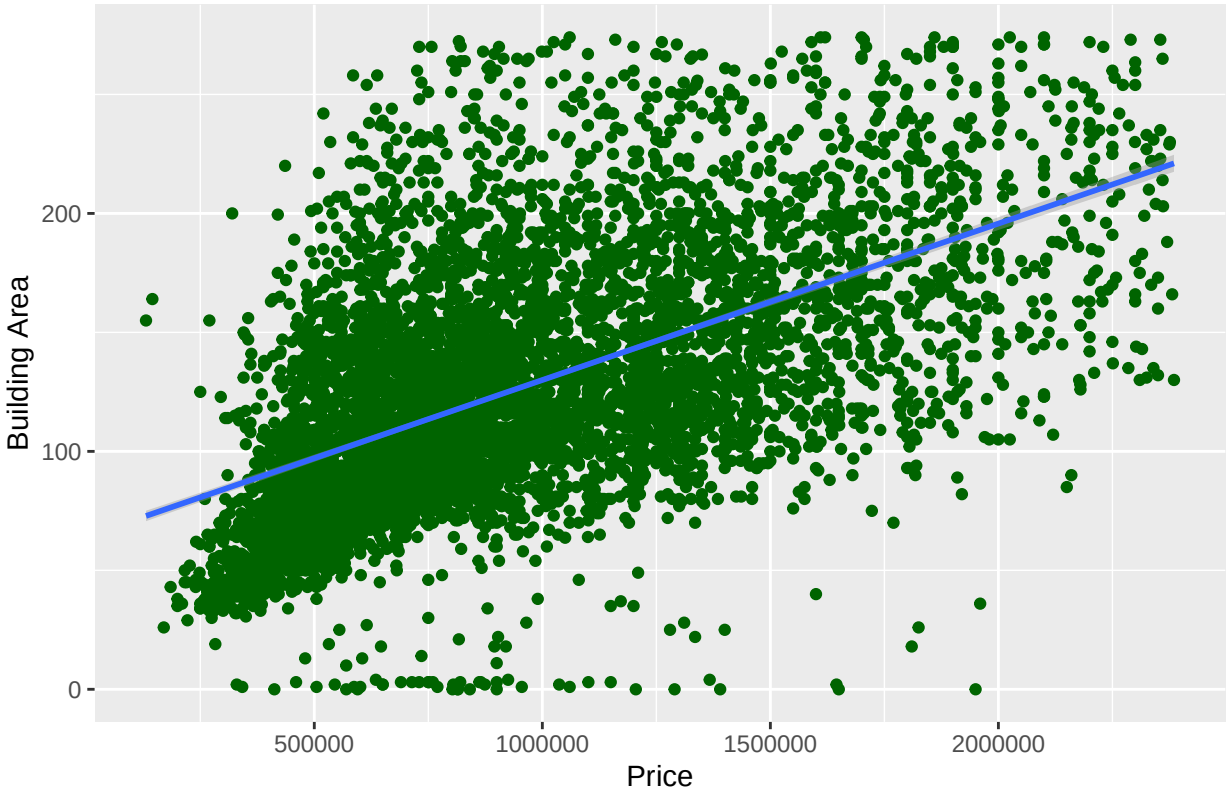
From the correlation plot, we can see that Building area (0.57), Year of Build (-0.42), and Rooms (0.49) are each significantly related to housing prices. However, there is a strong positive correlation (0.74) between the BuildingArea and Rooms variables, and a moderate positive correlation (0.61) between BuildingArea and Bathrooms. For that reason, Building Area was chosen over the Rooms and Bathrooms variables for further analysis.

Regression models: *1. Price and Building Area*

```
ggplot(data = mel_final, aes(x = Price, y = BuildingArea)) +
  geom_point(color = 'darkgreen')+
  stat_smooth(method = 'lm') +
  ggtitle('Figure 5 Regression Plot: Price - Building Area') +
  xlab("Price") + ylab("Building Area")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

Figure 5 Regression Plot: Price − Building Area



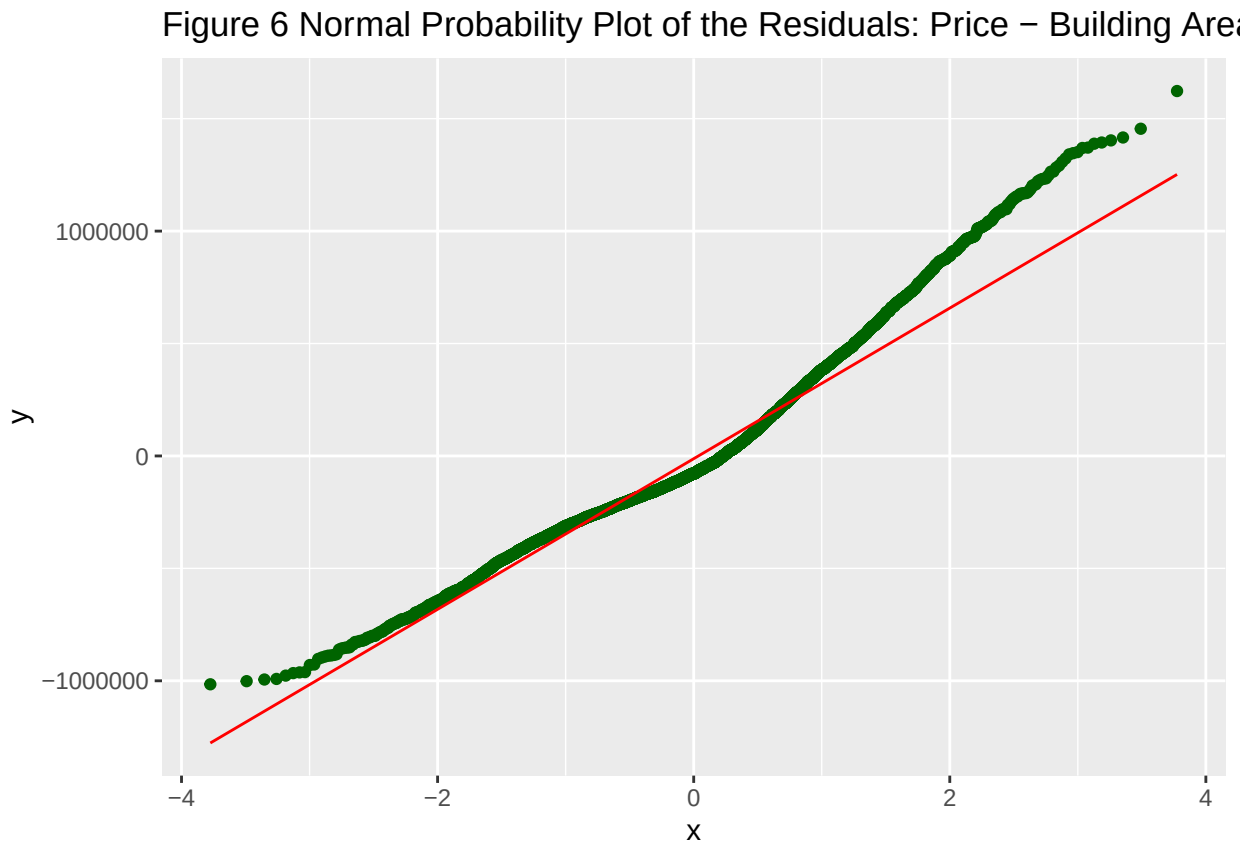

```
lm_price_building_area <- lm(Price ~ BuildingArea, data=mel_final)
summary(lm_price_building_area)
```

```
##
## Call:
## lm(formula = Price ~ BuildingArea, data = mel_final)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1015469  -238166   -78086   213637  1622876
##
## Coefficients:
##              Estimate Std. Error t value            Pr(>|t|)
## (Intercept)  327124.32   12361.88   26.46 <0.0000000000000002 ***
## BuildingArea   4935.44      90.12   54.77 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 370600 on 6245 degrees of freedom
## Multiple R-squared:  0.3245, Adjusted R-squared:  0.3243
## F-statistic:  2999 on 1 and 6245 DF,  p-value: < 0.00000000000000022
```

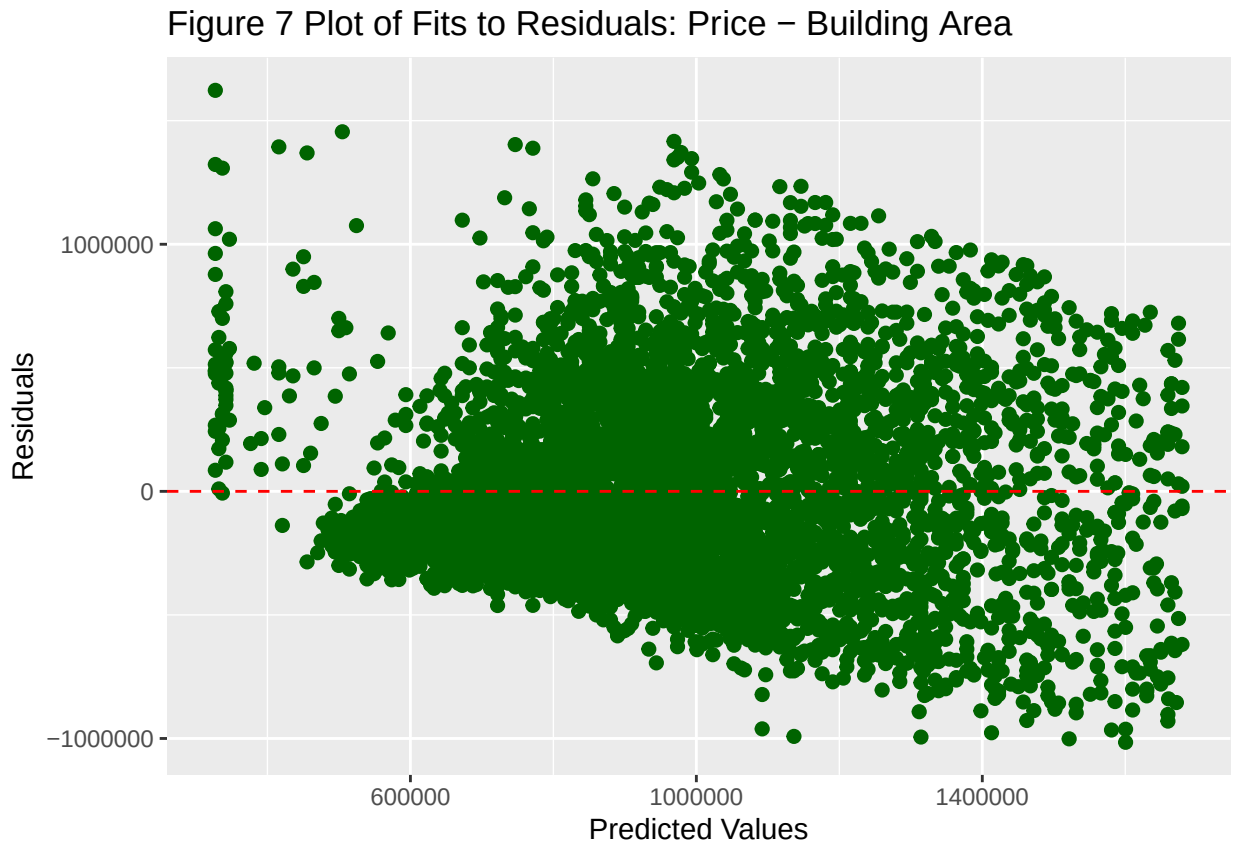
Before analyzing these results, we had to ensure that our model satisfies the necessary conditions - the normality of residuals condition, and the homoscedasticity condition. To check these, we created a Normal Probability Plot with the residuals and then plotted the predicted values of the model versus the residuals

to check for homoscedasticity. When both of these conditions are met, the residuals will follow the line of normality in a Normal Probability Plot, and the points in the plot checking for homoscedasticity (plotting standardized residuals against predicted values) will be randomly scattered. The same steps will be conducted for the next model as well.

```
predicted_value = lm_price_building_area$fitted.values
real_value = mel_final$Price
residual = lm_price_building_area$residuals

df = data.frame(real_value, predicted_value, residual)

ggplot(df, aes(sample=residual)) + stat_qq(col='darkgreen') + stat_qq_line(col='red') +
ggtitle("Figure 6 Normal Probability Plot of the Residuals: Price - Building Area")
```




```
# Homoscedasticity
ggplot(df, aes(x = predicted_value, y = residual)) +
  geom_point(size=2, col='darkgreen', position="jitter") +
  xlab("Predicted Values") + ylab("Residuals") +
  ggtitle("Figure 7 Plot of Fits to Residuals: Price - Building Area")+
  geom_hline(yintercept=0, color="red", linetype="dashed")
```

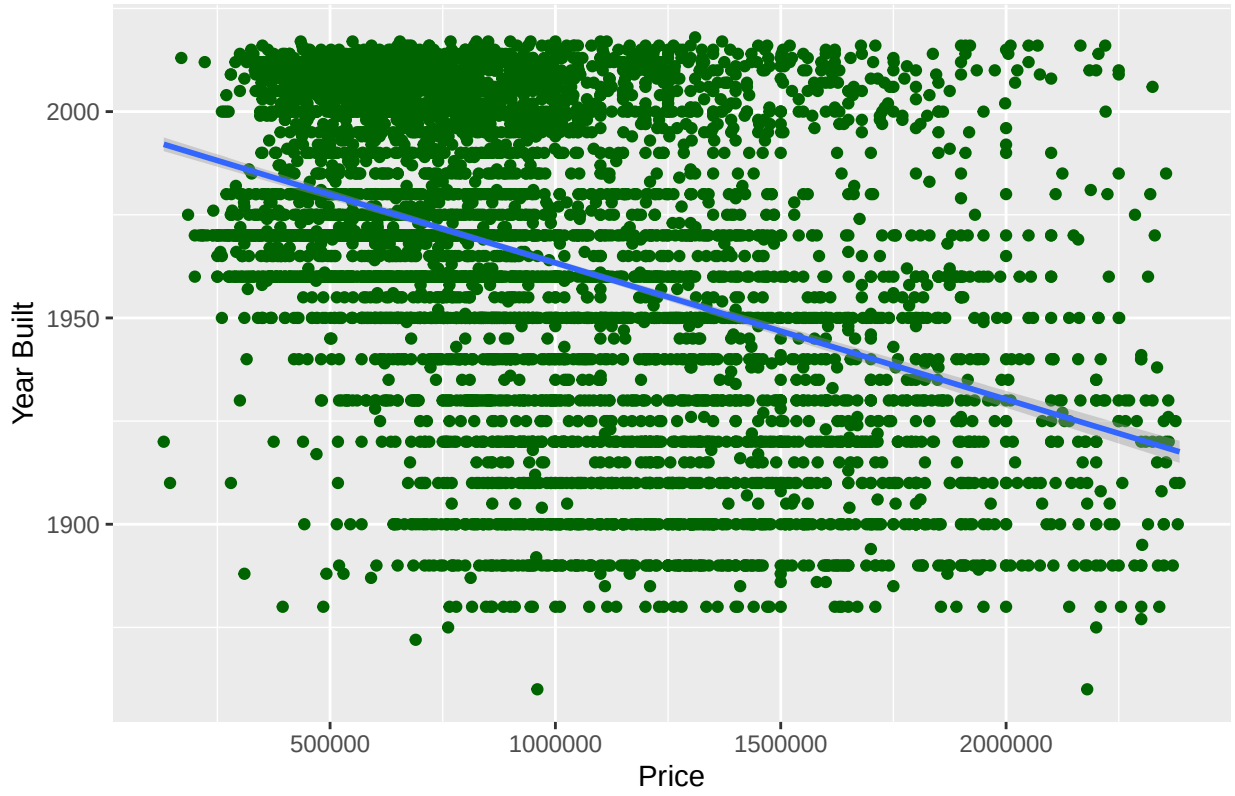


2. Price and Year Built

```
ggplot(data = mel_final, aes(x = Price, y = YearBuilt)) +
  geom_point(color = 'darkgreen') + stat_smooth(method = 'lm')+
  ggtitle('Figure 8 Regression Plot: Price - Year Built') +
  xlab("Price") + ylab("Year Built")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

Figure 8 Regression Plot: Price − Year Built

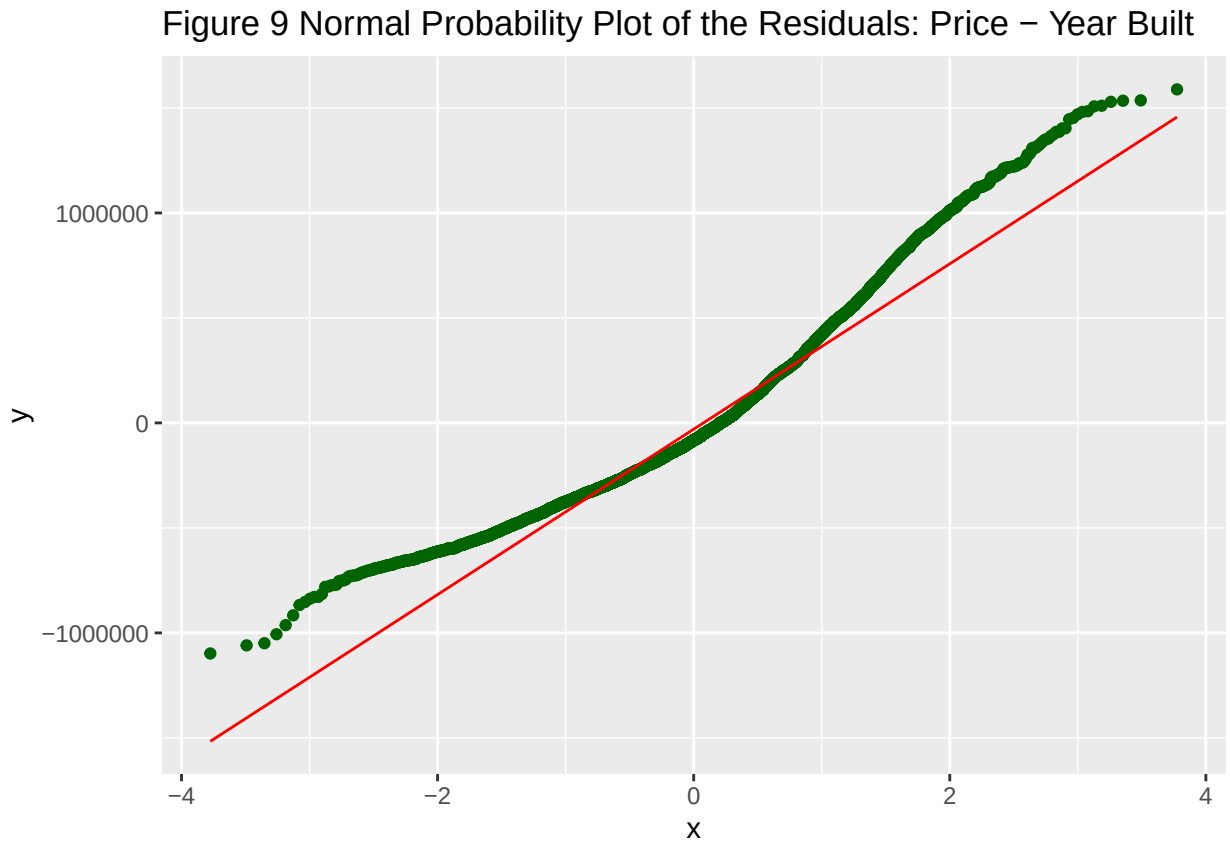



```
lm_price_year_built <- lm(Price ~ YearBuilt, data=mel_final)
summary(lm_price_year_built)
```

```
##
## Call:
## lm(formula = Price ~ YearBuilt, data = mel_final)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1098019  -295565   -82223   235612  1588326
##
## Coefficients:
##               Estimate Std. Error t value            Pr(>|t|)
## (Intercept) 11317179.3   285384.1   39.66 <0.0000000000000002 ***
## YearBuilt      -5274.4      145.2  -36.32 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 409600 on 6245 degrees of freedom
## Multiple R-squared:  0.1744, Adjusted R-squared:  0.1743
## F-statistic:  1319 on 1 and 6245 DF,  p-value: < 0.00000000000000022
```

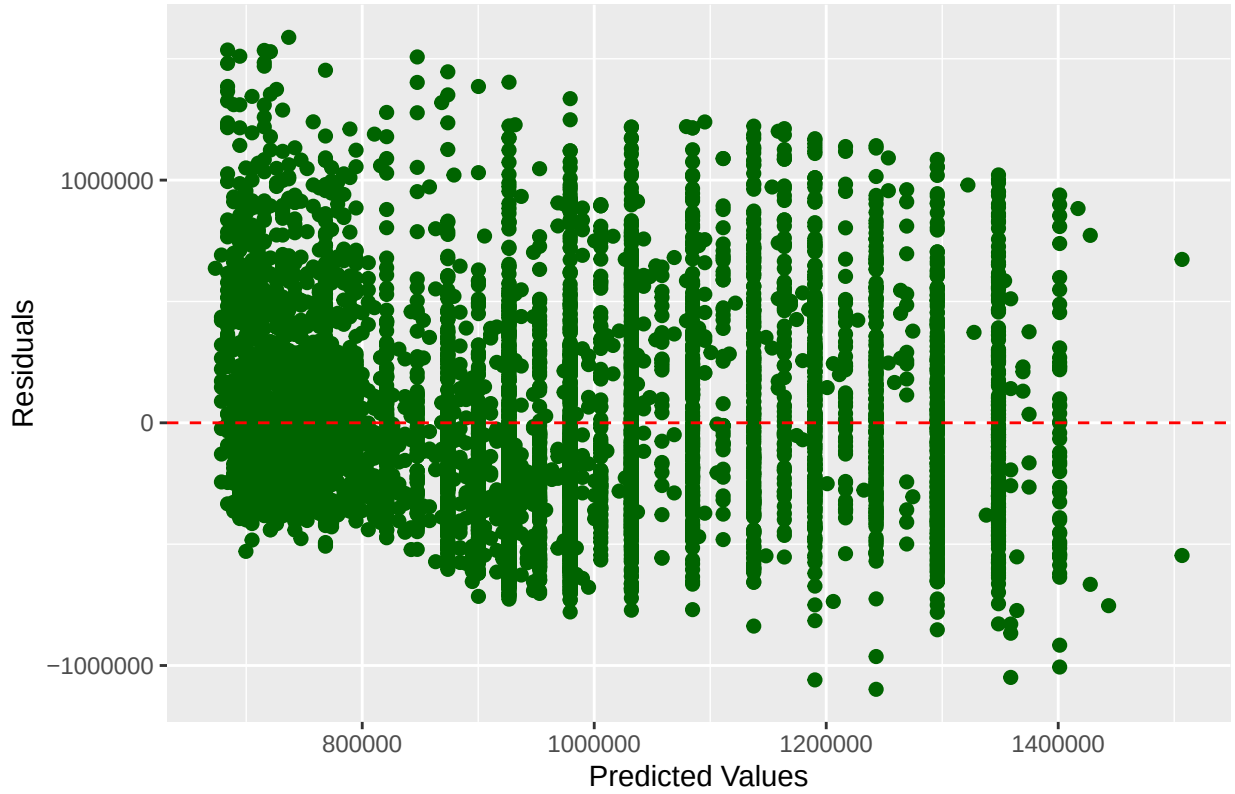
```
# Normality of Residuals
predicted_value = lm_price_year_built$fitted.values
real_value = mel_final$Price
```

```
residual = lm_price_year_built$residuals

df = data.frame(real_value, predicted_value, residual)

ggplot(df, aes(sample=residual)) + stat_qq(col='darkgreen') +
  stat_qq_line(col='red') +
  ggtitle("Figure 9 Normal Probability Plot of the Residuals: Price - Year Built")
```




```
# Heteroskedasticity
ggplot(df, aes(x = predicted_value, y = residual)) +
  geom_point(size=2, col='darkgreen', position="jitter") +
  xlab("Predicted Values") + ylab("Residuals") +
  ggtitle("Figure 10 Plot of Fits to Residuals: Price - Year Built")+
  geom_hline(yintercept=0, color="red", linetype="dashed")
```

Figure 10 Plot of Fits to Residuals: Price − Year Built



Based on our visual inspection of the plots above, testing the conditions for linear regression models, we determined that a more precise test for homoscedasticty was warranted. Therefore, we used the Breusch-Pagan test via the bptest() function (from the 'lmtest' library) with either of our models. This test uses a null hypothesis that homoscedasticity exists in the model, and an alternative hypothesis that heteroscedasticity is present. Therefore, we determined that we would not be able to continue with our models in their current state if the p-value from either of these tests were below 0.05.

```
bptest(lm_price_building_area)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  lm_price_building_area
## BP = 325.57, df = 1, p-value < 0.00000000000000022
```

```
bptest(lm_price_year_built)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  lm_price_year_built
## BP = 34.467, df = 1, p-value = 0.000000004335
```

Conclusions:

For this question we investigated the features of housing that have the greatest impact on housing prices. Firstly, the heat map with the correlations between our dependent variable (Price) and all of our independent variables was computed.

Secondly, we took it one step further and performed two linear regression models to try to find which variable of those two had the greatest R-squared value. In other words, we sought the variable which would have the biggest coefficient of determination, and would thus best explain the linear relationship with Price.

From the results of our Breusch-Pagan tests, we computed p-values less than 0.05 for either of our models. Therefore, we concluded that heteroscedasticity is present in both of our models. In other words, we could not rely on the findings in the model summary as being valid.

In order to construct a more valid linear regression model that would produce accurate pricing predictions, it would be necessary to transform the dependent variable 'Price', and to include other independent variables which could affect property prices. Even if these two models met the necessary conditions for a linear model, it is noteworthy that the greatest R-squared value = 0.3245 (model 1 Price~BuildingArea). Thus, this model would explain only 32% of the variance in the Price variable, which is a very low R-squared value.

3. There is a difference in housing prices between the 'Metropolitan' (city area) and 'Victoria' (suburbs) regions of Melbourne.

In order to test whether there is a difference in average prices between properties in the Metropolitan regions of Melbourne and the Victorian regions of Melbourne, we first set up a new variable which funneled the 8 regions into two. That is, we created a new column in the data set which identified each property as being located in either (a) one of the Metropolitan regions of Melbourne or (b) one of the Victorian regions of Melbourne. To do this, we used the 'mutate()' function:

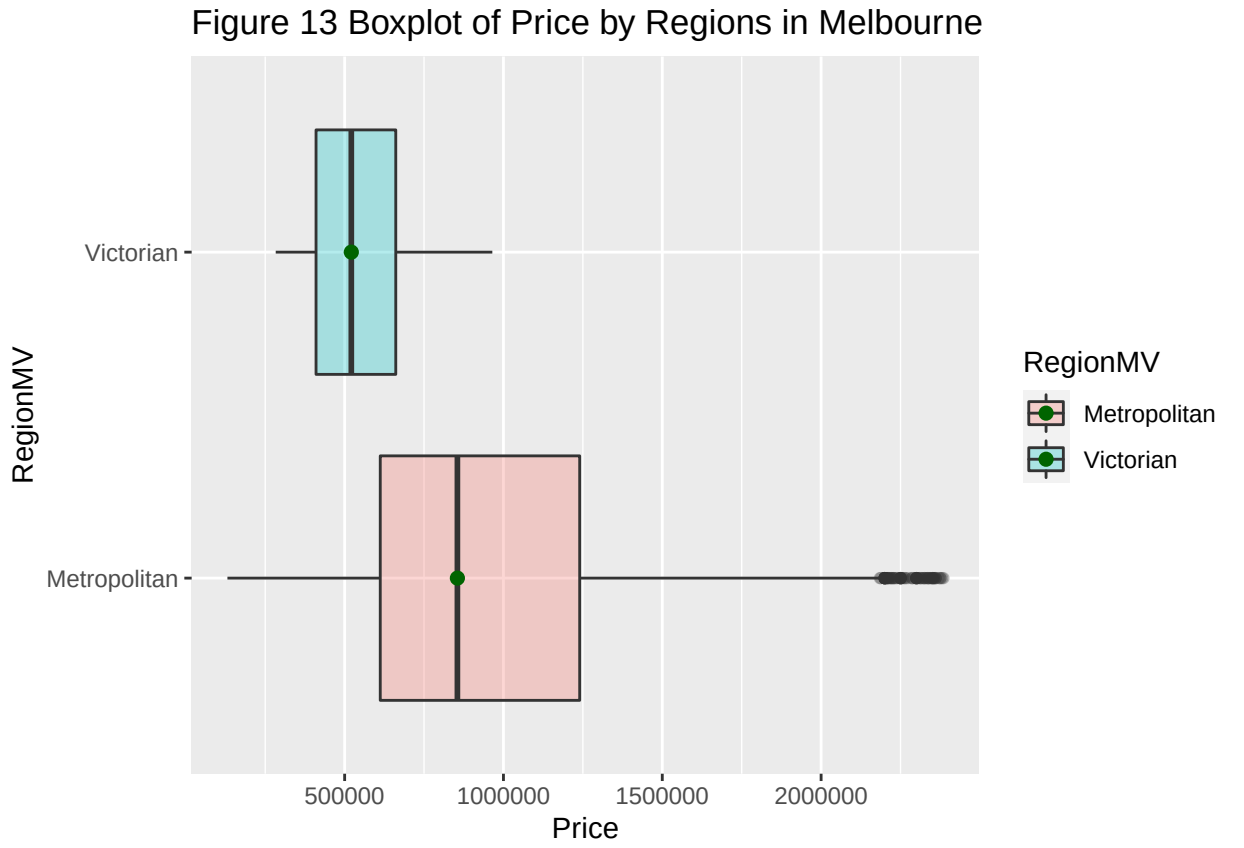
```
mel_final <- mel_final %>%
  mutate(RegionMV = case_when(
    grepl("Metropolitan", Regionname) ~ "Metropolitan",
    grepl("Victoria", Regionname) ~ "Victorian"
    ))

# Displaying head of the data set to exhibit the new column:
head(mel_final,4)
```

```
## # A tibble: 4 x 15
##   Rooms Type   Price Date  Dista~1 Bathr~2   Car Lands~3 Build~4 YearB~5 Regio~6
##   <dbl> <fct>  <dbl> <chr>   <dbl>   <dbl> <dbl>   <dbl>   <dbl>   <dbl> <chr>
## 1     2 house 1.03e6 4/2/~     2.5       1     0     156      79    1900 Northe~
## 2     3 house 1.46e6 4/3/~     2.5       2     0     134     150    1900 Northe~
## 3     4 house 1.6 e6 4/6/~     2.5       1     2     120     142    2014 Northe~
## 4     3 house 1.88e6 7/5/~     2.5       2     0     245     210    1910 Northe~
## # ... with 4 more variables: date <date>, Month <chr>, Year <chr>,
## #   RegionMV <chr>, and abbreviated variable names 1: Distance, 2: Bathroom,
## #   3: Landsize, 4: BuildingArea, 5: YearBuilt, 6: Regionname
```

Now that we had a variable appropriately established to identify the two groups of regions, we could begin our comparison of the average housing prices in the region groupings. To do this, we first visualized the distribution of housing prices for these two groups using box plots:

```
ggplot(mel_final, aes(x=RegionMV, y=Price, fill=RegionMV)) +
geom_boxplot(alpha=0.3)+
  coord_flip()+
```

```
  stat_summary(fun = "median", colour = "darkgreen", size = 2, geom = "point")+
  ggtitle("Figure 13 Boxplot of Price by Regions in Melbourne")
```



From the box plot above, housing prices in Metropolitan regions of Melbourne appear to be larger on average than housing prices in Victorian regions of Melbourne. In addition, housing prices in Metropolitan Melbourne are far more widely distributed, with a large number of outliers skewing the distribution to the right.

In order to test whether the mean housing prices in Metropolitan regions of Melbourne are greater than mean housing prices in Victorian regions of Melbourne, at a significance level of $\alpha = 0.05$, we conducted a right-tailed hypothesis test. In order to determine whether we could use the t-test method for this, we first checked whether there were sufficient observations in either group (Metropolitan and Victorian). If there were at least n = 25 observations in either group, then we would be able to assume that these data came from populations that could be modeled by the Normal distribution, thus allowing us to use a t-test.

```
mel_vic = as.factor(mel_final$RegionMV)
summary(mel_vic)
```

```
## Metropolitan    Victorian
##         6180           67
```

Evidently, for houses sold in both Victorian and Metropolitan regions, n >+ 25. Therefore, we were able to use a t-test to determine whether there was a difference in mean housing prices for houses sold in Metropolitan Melbourne and houses sold in Victorian Melbourne. Before conducting this test, we stated the following hypotheses at a significance level of $\alpha = 0.05$:

$H_0 : \mu_{met} = \mu_{vic} | \mu_{met} - \mu_{vic} = 0$

$H_a : \mu_{met} > \mu_{vic} | \mu_{met} - \mu_{vic} > 0$

To test these hypotheses, we will use the 't.test()' function:

```
t.test(~ Price | RegionMV, data=mel_final, alternative="greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  Price by RegionMV
## t = 18.859, df = 76.173, p-value < 0.00000000000000022
## alternative hypothesis: true difference in means between group Metropolitan and group Victorian is g
## 95 percent confidence interval:
##  374896    Inf
## sample estimates:
## mean in group Metropolitan    mean in group Victorian
##                   957969.6                   546768.4
```

P(|t| >= 18.859) = p-value = 0.0000... Therefore, p-value < alpha, and we thus reject H0. In other words, we can infer that the mean price of a house in the Metropolitan regions of Melbourne is greater than the mean price of a house in the Victorian regions of Melbourne.

Because the box plots created before conducting the t-test were quite heavily right-skewed, we decided to conduct this test again (with the same hypotheses and significance level) using permutation testing.

```
# First, to determine how many items should be sampled in either group for the
# permutations, we will determine how many houses are in either group of regions:
n_met = sum(mel_final$RegionMV=='Metropolitan')
n_vic = sum(mel_final$RegionMV=='Victorian')

# Next, we will compute the observed difference in the sample means (the test-statistic):
ob_dif = mean(~ Price, data=filter(mel_final, RegionMV=="Metropolitan")) -
  mean(~ Price, data=filter(mel_final, RegionMV=="Victorian"))

# Now, we will compute 10000 permutations of these data:
N = 10000
diffs = numeric(N)
for(i in 1:N)
{ index = sample((n_met+n_vic), n_met, replace=FALSE)
  diffs[i] = mean(mel_final$Price[index]) - mean(mel_final$Price[-index])
}

# Finally, we will compute the p-value for the right-tailed permutation test:
sum(diffs >= ob_dif) / N
```

```
## [1] 0
```

In accord with the previously conducted t-test, this permutation test resulted in a p-value near 0 as well. P-value < alpha. Therefore, we reject H0 once again and infer instead that the mean price of houses in Metropolitan regions of Melbourne is greater than the mean price of houses in the Victorian regions of Melbourne.

To further analyze and quantify the difference noted above, we constructed a 95% Bootstrap confidence interval of the difference in means, using 1000 resampling iterations:

```
nsims = 1000 # the number of simulations to conduct
xbar_met = numeric(nsims) # holds the mean of each resample from Metropolitan houses
xbar_vic = numeric(nsims) # holds the mean of each resample from Victorian houses

# holds the difference between the resample means
xbar_diffs = numeric(nsims)

# parse out data for houses in Metropolitan regions
met_data = filter(mel_final, RegionMV=="Metropolitan")

# parse out data for houses in Victorian regions
vic_data = filter(mel_final, RegionMV=="Victorian")
for (i in 1:nsims)
{   xbar_met[i] = mean(sample(met_data$Price, n_met, replace=TRUE))
    xbar_vic[i] = mean(sample(vic_data$Price, n_vic, replace=TRUE))
    xbar_diffs[i] = xbar_met[i] - xbar_vic[i]  # Compute the difference between the
    # sample means for each bootstrap resample
}

# Create a new data frame for the bootstrap distribution of differences in means:
boot_diffs = data.frame(xbar_met, xbar_vic, xbar_diffs)
```

From the bootstrap distribution created above for the differences in mean housing prices between the two groups of regions in Melbourne, we then used 'qdata()' to construct a 95% confidence interval:

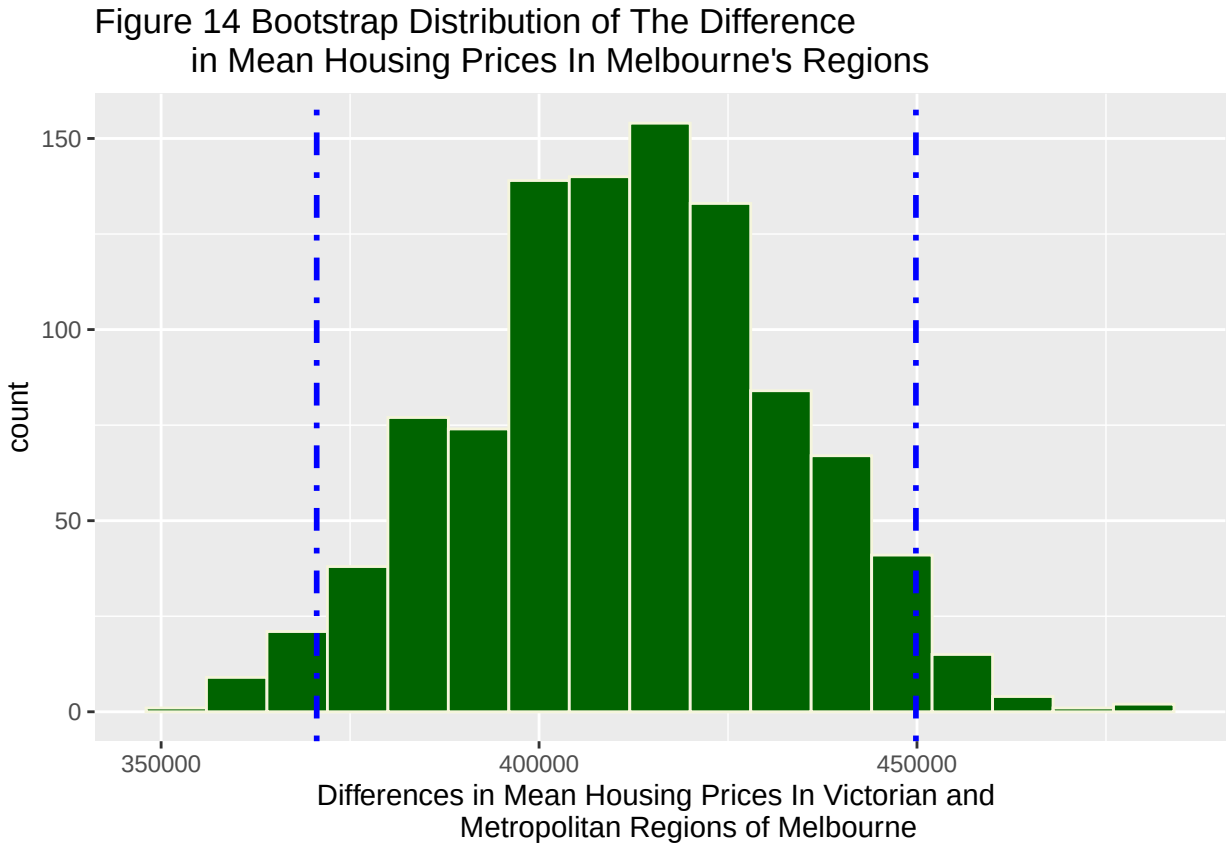
```
qdata(~ xbar_diffs, c(0.025, 0.975), data=boot_diffs)
```

```
##     2.5%    97.5%
## 370580.2 449858.0
```

Via bootstrap sampling, we can infer at the 95% confidence level that the difference between the mean price of houses in Metropolitan regions of Melbourne is between $369,549 and $453,145.3 greater than the mean price of houses in Victorian regions of Melbourne.

To visualize the above finding, we plotted the Bootstrap distribution on a histogram, with vertical lines denoting the upper and lower bounds of the 95% confidence interval:

```
ggplot(data=boot_diffs, aes(x = xbar_diffs)) +
  geom_histogram(fill='darkgreen', col='beige', binwidth=8000) +
  xlab("Differences in Mean Housing Prices In Victorian and
       Metropolitan Regions of Melbourne") +
  ggtitle("Figure 14 Bootstrap Distribution of The Difference
          in Mean Housing Prices In Melbourne's Regions") +
  geom_vline(xintercept=qdata(~ xbar_diffs, c(0.025, 0.975), data=boot_diffs), linetype='dotdash',
             color='blue',size=1.0)
```



3. Additional Analysis

Together with the previous statistical analyses, the following additional research was conducted in order to answer one question regarding the busiest month/season for real estate in Melbourne, together with additional visualizations which cover these interesting findings.

By answering this question, we would gain a better understanding of the real estate market in Melbourne, Australia as the seasons in Australia are different (in fact, opposite) from the seasons in Canada.

```
mel_f_2016<-as.data.frame(filter(mel_final, Year == "16"))
table(mel_f_2016$Month)
```

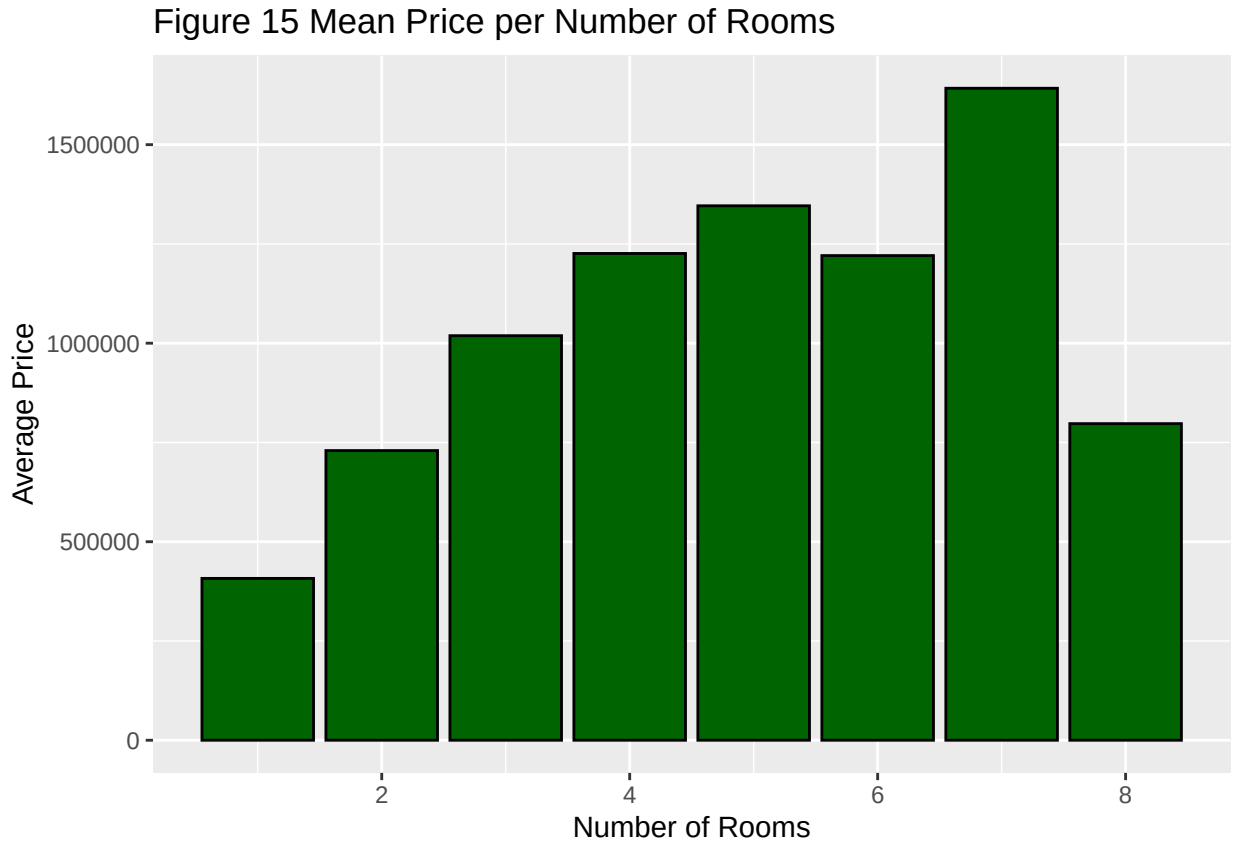
```
##
##     April    August  December  February      July      June       May  November
##       171       351       288        16       212       362       457       559
##   October September
##       276       472
```

```
mel_f_2017<-as.data.frame(filter(mel_final, Year == "17"))
table(mel_f_2017$Month)
```

```
##
##     April    August  February      July      June     March       May September
##       297       354       209       730       429       261       450       353
```

Conclusion: to investigate the tendency regarding the busiest month for real estate in Melbourne, we first separated the database by year, as – in our case – the years 2016 and 2017 (in part) were presented. After looking at the frequency of every month in the year, we can conclude that November, with 559 properties sold, is the "hottest" month for real estate in Melbourne in 2016. On the other hand, July saw 730 properties sold, and thus takes over this category in 2017. Therefore, we can conclude that Spring (November) and Winter (July) are the busiest time for real estate in Melbourne, Australia.

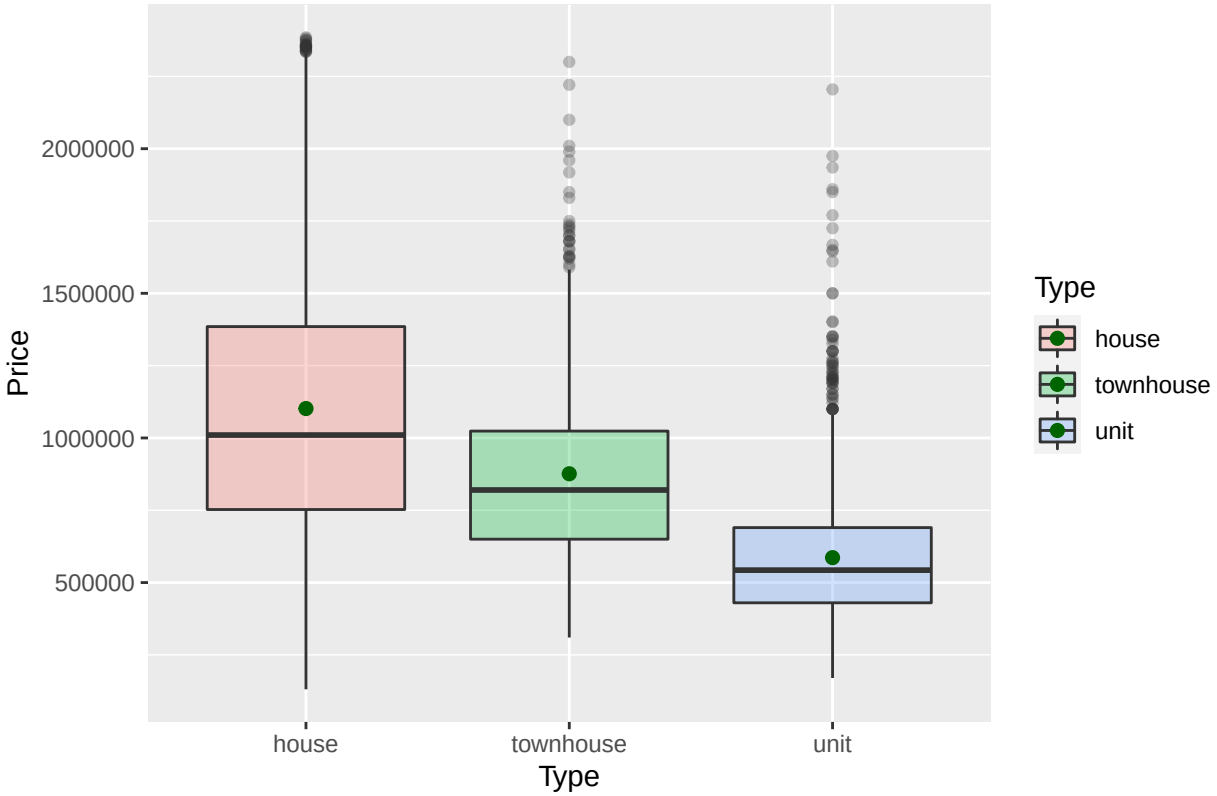
```
ggplot(data=mel_final, mapping=aes(x=Rooms, y=Price)) +
  stat_summary(fun.data=mean_sdl, geom="bar", color='black', fill="darkgreen")+
  labs(title="Mean Price per Number of Rooms", x="Number of Rooms", y="Average Price") +
  ggtitle('Figure 15 Mean Price per Number of Rooms')
```



We can observe interesting findings by looking at how the average price fluctuated depending on how many rooms the property had. For example, the highest average price prevails for properties with seven rooms. In contrast, the average price considerably declines for properties with eight rooms, lowering to roughly the same average property price as properties with two rooms.

```
ggplot(mel_final, aes(x=Type, y=Price, fill=Type)) +
  geom_boxplot(alpha=0.3)+stat_summary(fun = "mean", colour = "darkgreen", size = 2, geom = "point")+
  ggtitle("Figure 16 Boxplot of Mean Price by Type of the property in Melbourne")
```

Figure 16 Boxplot of Mean Price by Type of the property in Melbourne



By looking at the boxplot above, the average price of a property is greatest among houses, lower among townhouses, and lowest among units in Melbourne. At the same time, the minimum price for a house is the smallest compared with the other 2 categories. This illustrates that houses have the greatest variability in pricing. While units have the lowest average price in Melbourne, there are many outliers in the unit category. As a result, on average, the most expensive type of property in Melbourne is a house.

Conclusion

To finalized all the information above, we can infer that there is a negative relationship between the price of a property and its distance from the city center of Melbourne. The correlation indicator shows the weak and negative number (-0.17), however, it is significant as we are having more than 6000 cases. In other words, the price of the property increases when the distance from the city center of Melbourne is decreases.

In order to construct the better model for explaining the price of the property, it is necessary to transform the dependent variable 'Price', and to include other independent variables which could affect property prices. Our dependent variable 'Price' distribution plot was right-skewed, it leaded to the difficulties in our project and followed with the challenges in the linearity rule assumptions.

The real estate market of Metropolitan regions is more costly on average than the Victorian regions of Melbourne (suburbs). With 95% confidence interval the mean price of the properties in Metropolitan areas of Melbourne is higher on around 360,000-450,000 AU dollars, than the same properties in the suburbs of Melbourne.

References

Australian Council of Social Service. (n.d.). Components of Australia's wealth: Data and Figures. Poverty and Inequality in Australia. Retrieved September 16, 2022 from https://povertyandinequality.acoss.org.au/

inequality/components-of-australias-wealth/

Dietz, R. D., & Haurin, D. R. (2003). The social and private micro-level consequences of homeownership. Journal of Urban Economics, 54(3), 401–450. https://doi.org/10.1016/S0094-1190(03)00080-9

Gaolu, Z. (2015). The Effect of Central Business District on House Prices in Chengdu Metropolitan Area: A Hedonic Approach. [online] www.atlantis-press.com. doi:10.2991/cas-15.2015.83.

Teye, A.L., de Haan, J. and Elsinga, M.G. (2017). Risks and interrelationships of subdistrict house prices: the case of Amsterdam. Journal of Housing and the Built Environment, 33(2), pp.209–226. doi:10.1007/s10901-017-9568-z.

Dataset

https://www.kaggle.com/datasets/dansbecker/melbourne-housing-snapshot