I. Abstract :

We're looking at six key variables to see how they influence home prices. Transaction date, house age, distance to nearest MRT station, number of convenience stores nearby, latitude, and longitude are among the factors. We're looking at this data to see if we can use it to estimate the house price. We will examine the correlation between the two variables, the R squared model, the f statistic, RSE, and P value, as well as create a linear regression model between a variable and the price of the house, in order to discover which of the variables does in fact influence the price of the house. We'll also use the Q-Q plot and the plot for the density of the residual to assess the data's normality. R squared indicates how well the model fits the data and runs from 0 to 1, with 0 indicating that the model did not explain the outcome and 1 indicating that the linear regression model explained a significant percentage of the outcome. The F-statistic indicates the model's overall significance, the higher the f statistic, the higher the model's quality. RSE aids in the calculation of the percentage of error (percent error = RSE/Y(Mean), with the lower the value, the better. Finally, if the p value is less than 0.05, we should reject the null hypothesis, implying that the data is statistically significant, or accept it if the p value is more than 0.05 and the data is not statistically significant. Based on all these factors, the third variable will be the best match for our data, as it has the highest R-squared of all the other models.

II.

1. Transaction date vs House price:

The graph for our first variable, Transaction date, is flat. That is, we know the correlation rate between transaction date and house price is practically 0 before we even calculate it. We obtain 0.0875 when we use R to calculate the correlation rate, which means the date the house was put on the market had no effect on the price. Next, we'll look at the p-value, which is 0.07537 >.05, indicating that the null hypothesis is not rejected and that the result is not statistically significant. When we look at the RSE, we can see that the calculated percent error is roughly 35%, which is a significant inaccuracy. Our f statistic is 3.178 and our R squared is 0.005246 both prove that our model is not a good fit for our data. Also, the coefficient β_1 for the transaction date is not significant because its p-value is less than 0.05 in other word the transaction date has no effect on the price of the house.

2. House age vs house price:

Our next variable is house age. I first calculated the correlation rate between the house age and price of house and got -0.210567 meaning that there is a small negative relationship between the age and the price.

The calculated P value for this variable is 1.56*e^-05 < .05 meaning that we would reject the null hypothesis and that the information is statistically significant at 95%. Also, the coefficient β_1 for the House age is significant because its p-value is less than 0.05 in other word the House age influences the price of the house, with one unite increase in the house age the price decrease by 0.25149. However, the calculated R-squared is 0.04434 which is very low meaning the model does not fit the data very well. Our calculated percent error was also very high at 35%, meaning our model was not very accurate. Because our R-squared was very low and percent error was high we can say that our model did not fit the data very well.

3. The distance to the nearest MRT station vs House price:

Our third variable is the distance to the nearest MRT station. The correlation between the distance to the nearest MRT station and price of house had the highest correlation rate at -0.67 meaning that there a negative relationship.

Our p value 2.2e^-16 < .05 meaning that we should reject the null hypothesis and it is statistically significant at 95%. Also, the coefficient β_1 for the distance to the nearest MTR station is significant because its p-value is less than 0.05 in other word The distance to the nearest influences the price of the house, with one unite increase in the distance to the nearest MTR station the price decrease by 0.0072621.

R-squared is 0.4538 meaning its fit is ok but not great meaning the model is kind of accurate. The calculated percent error was 26.5 which is relatively big. Finally, the F-statistic was 342 which is high meaning that our model is significant.

4. number of convenience stores vs House Price:

The fourth variable is the number of convenience stores. The correlation between the number of convenience stores and price of house is 0.5710049 meaning that there a weak positive relationship between the number of convenience stores and the price.

The p value 2.2e-16 < 0.05 meaning that we should reject the null hypothesis and it is statistically significant at 95%. Also, the coefficient β_1 for the number of convenience stores is significant because its p-value is less than 0.05 in other word the number of convenience stores influences the price of the house, with one unite increase in the number of conveniences stores the price increases by 2.6377.

R-squared is 0.326 meaning its fit is ok but not great meaning the model is kind of accurate. Looking at the F-statistic = 199.3 which is relatively high meaning that the model is relatively accurate. Finally, based on the Q-Q plot and the plot for the density of the residual we can tell that the data is normally distributed.

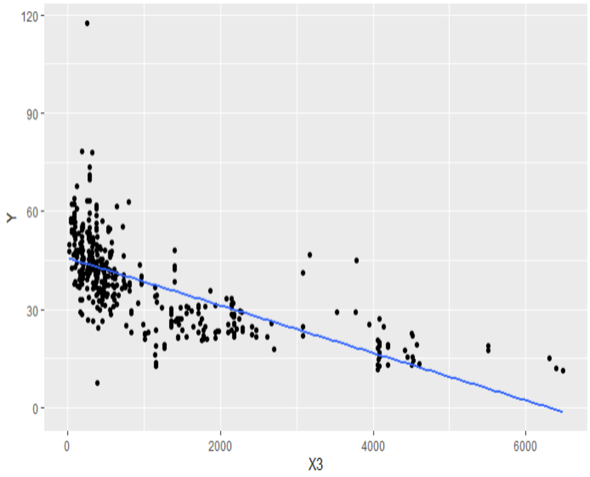
5. Latitude vs House price:

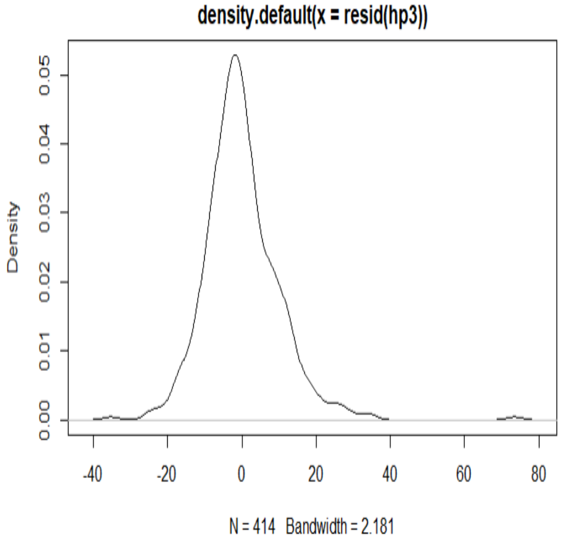
Next variable is the latitude of the house. Based on its correlation = 0.5463067 with the price we can see that there is a relationship between them, but it is not quite strong. The P value for this variable is 2.2e-16< .05 meaning that we would reject the null hypothesis and that the information is statistically significant at 95%. Also, the coefficient β_1 for the latitude is significant because its p-value is less than 0.05 in other word the House latitude influences the price of the house, with one unite increase in the house age the price gets high by 598.97. However, the calculated R-squared is 0.2985 which is low meaning the model does not fit the data very well.

6. Longitude vs house price:

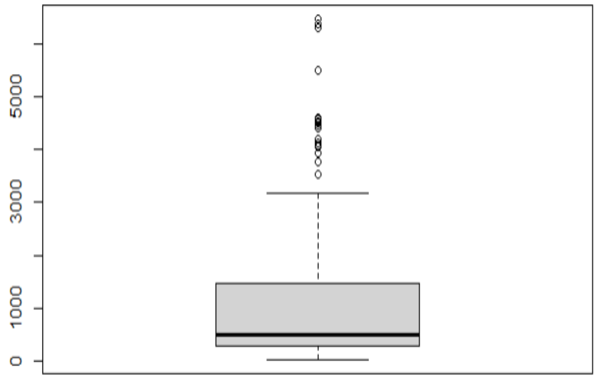
The last variable is the longitude, its correlation with price is 0.5232865 meaning that there is a positive relationship but not very strong. The P value for this variable is 2.2e-16< .05 meaning that we would reject the null hypothesis and that the information is statistically significant at 95%. Also, the coefficient β_1 for the longitude is significant because its p-value is less than 0.05 in other word the House longitude influences the price of the house, with one unite increase in the house longitude the price gets high by 463.93. However, the calculated R-squared is 0.27385 which is low meaning that the model does not fit the data very well.

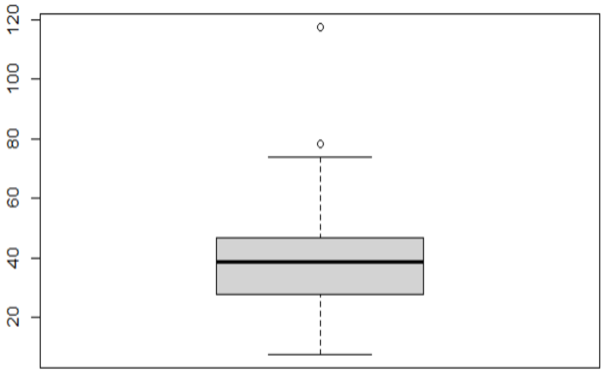
Compared to all the SLRs model The third variable has the best overall linear regression model therefore the distance to the nearest MRT station is good fit for our model. Based on its p-value 2.2e-16 < 0.05 the model is statistically significant at 95%.Also, its correlation -0.6736129 which is the strongest negative linear relationship among all the other variables . Moreover, this variable, has the higher R-Squared 0.4538 meaning that this model is the best fil for our data. Furthermore, this data is normally distributed based on the Q-Q plot and the plot for the density of the residual.





III. Using a boxplot We can see that we have some outliers in all the models that influence our linear model ,it is essential to understand their impact on your predictive models, Because, it can drastically bias the fit estimates and predictions. Generally, any datapoint that lies outside the 1.5 * interquartile-range (1.5 * *IQR*) is considered an outlier, where IQR is calculated as the distance between the 25th percentile and 75th percentile values for that variable. (Taking our third model as an example)





Project 2

Achraf cherkaoui

10/8/2021

```
library(ggplot2)
library(GGally)
library(readxl)
house <- read_excel("Real Estate Price Prediction.xlsx")
colnames(house) <-c( "N","X1","X2","X3","X4","X5","X6","Y") # change the columns names
View(house)
attach(house)
ggpairs(house) # shows the correlation between all the variables
```

```
ggplot (house ,aes(X1, Y)) +
 geom_point() +
geom_smooth(method = lm ,se =F) # graph a linear model in the data
cor(Y,X1) # corrolation between Y and X1
hp1 <- lm(data = house , Y~X1) # linear regression
summary(hp1)
pE <- sigma(hp1)*100 /mean(Y)# percentage rate
pE
plot(hp1 , which = 2) #Q-QPlot to test normality
shapiro.test(resid(hp1)) # Shapiro test for testing normality
plot(density(resid(hp1)))# to test normality
```

```
ggplot (house ,aes(X2, Y)) +
 geom_point() +
stat_smooth(  method = lm ,se= F)
cor(Y,X2)
hp2 <- lm(data = house , Y~X2)
summary(hp2)
plot(hp2 , which = 2)
shapiro.test(resid(hp2))
plot(density(resid(hp2)))
```

```
library(car)
ggplot (house ,aes(X3, Y)) +
 geom_point() +
stat_smooth(method = lm , se =F)
cor(Y,X3)
hp3 <- lm(data = house , Y~X3)
summary(hp3)
plot(hp3 , which = 2)
shapiro.test(resid(hp3))
plot(density(resid(hp3)))
boxplot(Y ) # to view outliers
boxplot(X3) # to view outliers
```

```r
ggplot (house ,aes(X4, Y)) +
 geom_point() +
stat_smooth(method = lm)
cor(Y,X4)
hp4 <- lm(data = house , Y~X4)
summary(hp4)
plot(hp4 , which = 2)
shapiro.test(resid(hp4))
plot(density(resid(hp4)))
```

```r
ggplot (house ,aes(X5, Y)) +
 geom_point() +
stat_smooth(method = lm)
cor(Y,X5)
hp5 <- lm(data =house , Y~X5)
summary(hp5)
plot(hp5 , which = 2)
shapiro.test(resid(hp5))
plot(density(resid(hp5)))
```

```r
ggplot (house ,aes(X6, Y)) +
 geom_point() +
stat_smooth(method = lm)
cor(Y,X6)
hp6 <- lm(data = house , Y~X6)
summary(hp6)
plot(hp6 , which = 2)
shapiro.test(resid(hp6))
plot(density(resid(hp6)))
```