Statistical analyses and visualization in R:I Assingment 1

Yuka Tatsumi

2021-10-15

Result

Part 1. Correlation between GDP and access to personal computers in the year of 2005

Question 1. Is there a relationship between GDP/capita and the number of personal computers per 100 individuals measured in 2005?

Median, mean and SD

The median, mean and SD for GDP per Capita are as follows according to the calculation with R.

GDP Capita
Median :2196.2, Mean:7963.5, SD:12328.80814

Regarding the number of personal computers per 100(PC_PER.100), the median, mean and SD are as follows.

Number of personal computers per 100(PC_PER.100)
Median:5.900, Mean:15.781, SD:22.11974

Summaries: Active data set

```
summary(Dataset1)
```

```
##                 Country     GDPperCAPITA       PC_PER.100     
##  Albania            :  1   Min.   :  128.3   Min.   : 0.070  
##  Algeria            :  1   1st Qu.:  618.6   1st Qu.: 1.755  
##  Angola             :  1   Median : 2196.2   Median : 5.900  
##  Antigua and Barbuda:  1   Mean   : 7963.5   Mean   :15.781  
##  Argentina          :  1   3rd Qu.: 9271.1   3rd Qu.:16.480  
##  Armenia            :  1   Max.   :81828.0   Max.   :88.660  
##  (Other)            :149
```

Summaries: Numerical summaries

```
numSummary(Dataset1[,
  c("GDPperCAPITA",
  "PC_PER.100"), drop=FALSE],
  statistics=c("mean", "sd",
  "IQR", "quantiles"),
  quantiles=c(0,.25,.5,.75,1))
```

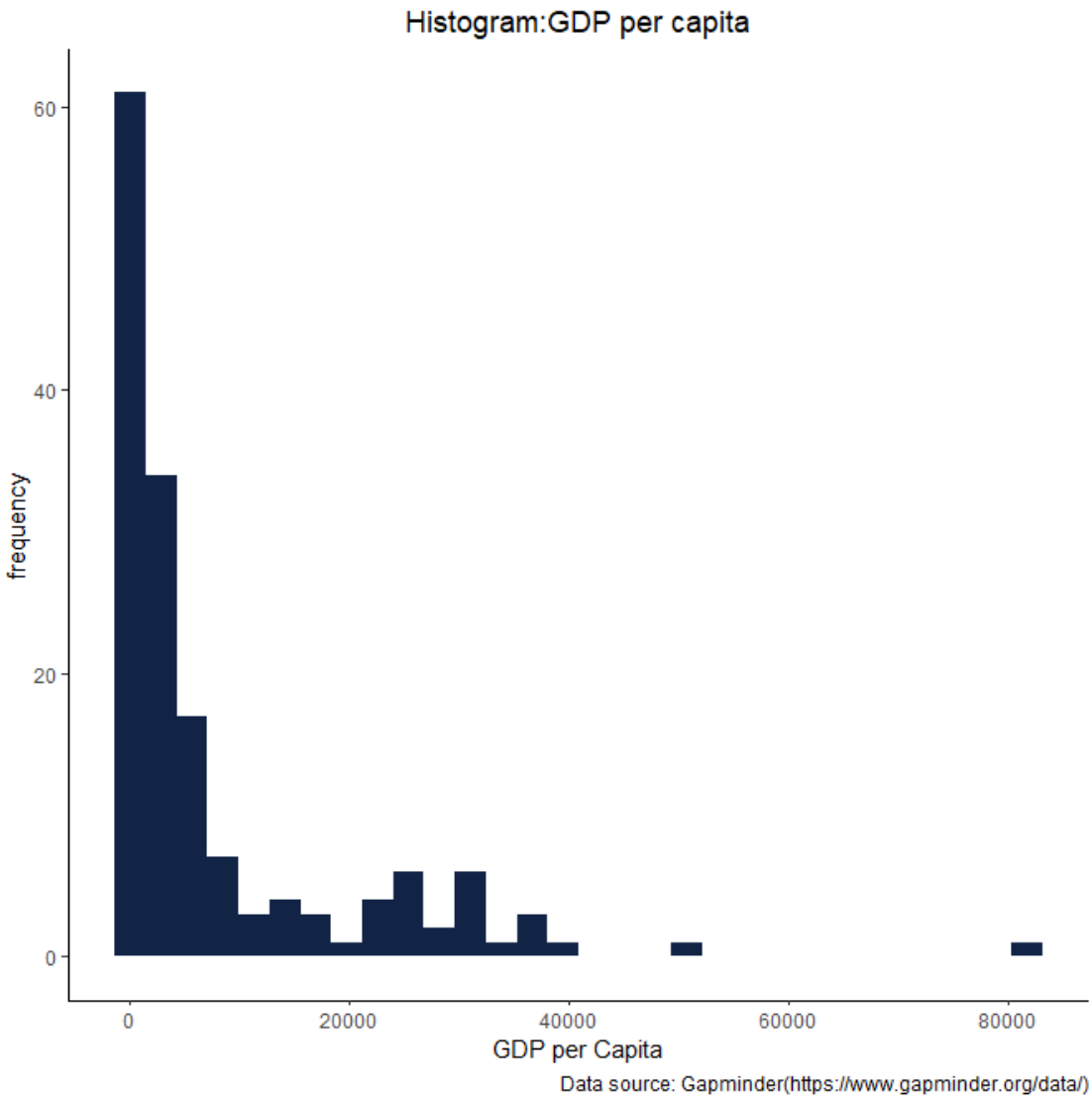
```
##                    mean           sd       IQR        0%       25%       50%
75%
## GDPperCAPITA 7963.5139 12328.80814 8652.498 128.2961 618.6427 2196.247
9271.14
## PC_PER.100     15.7809    22.11974   14.725   0.0700   1.7550    5.900
16.48
##                   100%   n
## GDPperCAPITA 81827.96 155
## PC_PER.100      88.66 155
```

A plot to check normal distribution for variables and Choose the appropriate test.

The histograms for both GDP per capita and for the number of personal computers per 100 individuals below show that both variables are not normally distributed.

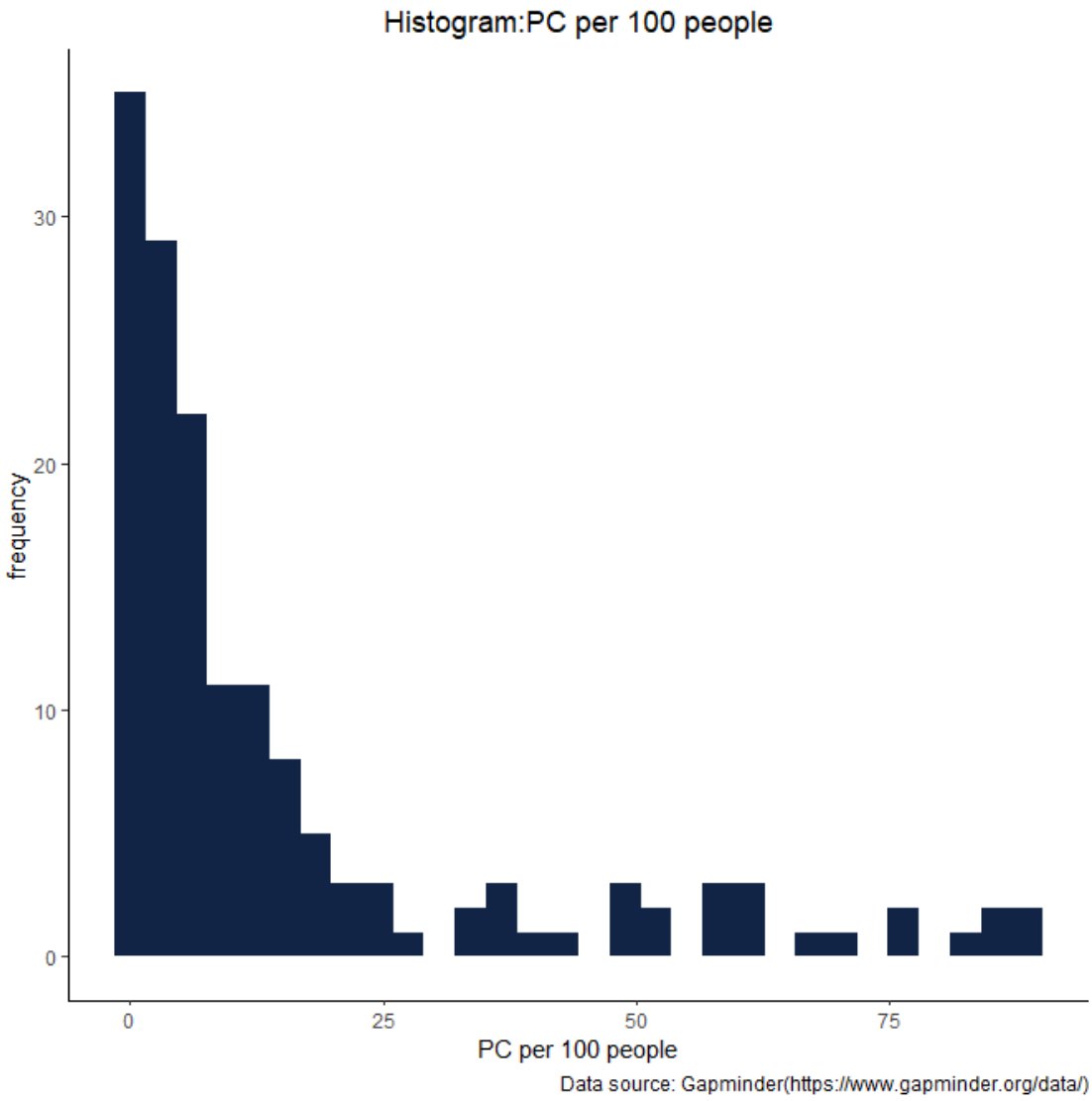
Histogram to check normal distribution for GDP per Capita

```
ggplot(Dataset1) +
  aes(x = GDPperCAPITA) +
  geom_histogram(bins = 30L, fill = "#112446") +
  labs(x = "GDP per Capita", y="frequency", title ="Histogram:GDP per
capita", caption = "Data source: Gapminder(https://www.gapminder.org/data/)"
) +
  theme_classic()+
  theme(plot.title = element_text(hjust = 0.5))
```



** Histogram to check normal distribution for PC per 100 people **

```
ggplot(Dataset1) +
  aes(x = PC_PER.100) +
  geom_histogram(bins = 30L, fill = "#112446") +
  labs(x = "PC per 100 people", y="frequency", title="Histogram:PC per 100
people", caption="Data source: Gapminder(https://www.gapminder.org/data/)" )
+
  theme_classic()+
  theme(plot.title = element_text(hjust = 0.5))
```



Choice of appropriate test and motivation

According to the histograms above, both of the variables are not normally distributed. Therefore, Spearman Rank-Order correlation is a test to be used because Pearson product-moment correlation requires normal distribution of the variables.

Result and analysis of the statistic values and a scatter plot

Result and analysis of the statistic

According the Spearman Rank-Order correlation's test result below, correlation coefficient is 0.8547637, degree of freedom= 153 and P value is 2.2e-16. The p value suggests that correlation coefficient 0.8547637 is statistically significant. It implies that there is a correlation between GDP per capita and number of personal computers per 100 individuals in 2005. (95% confidence)

```
with(Dataset1,
  cor.test(GDPperCAPITA, PC_PER.100,
  alternative="two.sided",
  method="spearman"))
```

```
## Warning in cor.test.default(GDPperCAPITA, PC_PER.100, alternative =
## "two.sided", : Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  GDPperCAPITA and PC_PER.100
## S = 90137, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.8547637
```

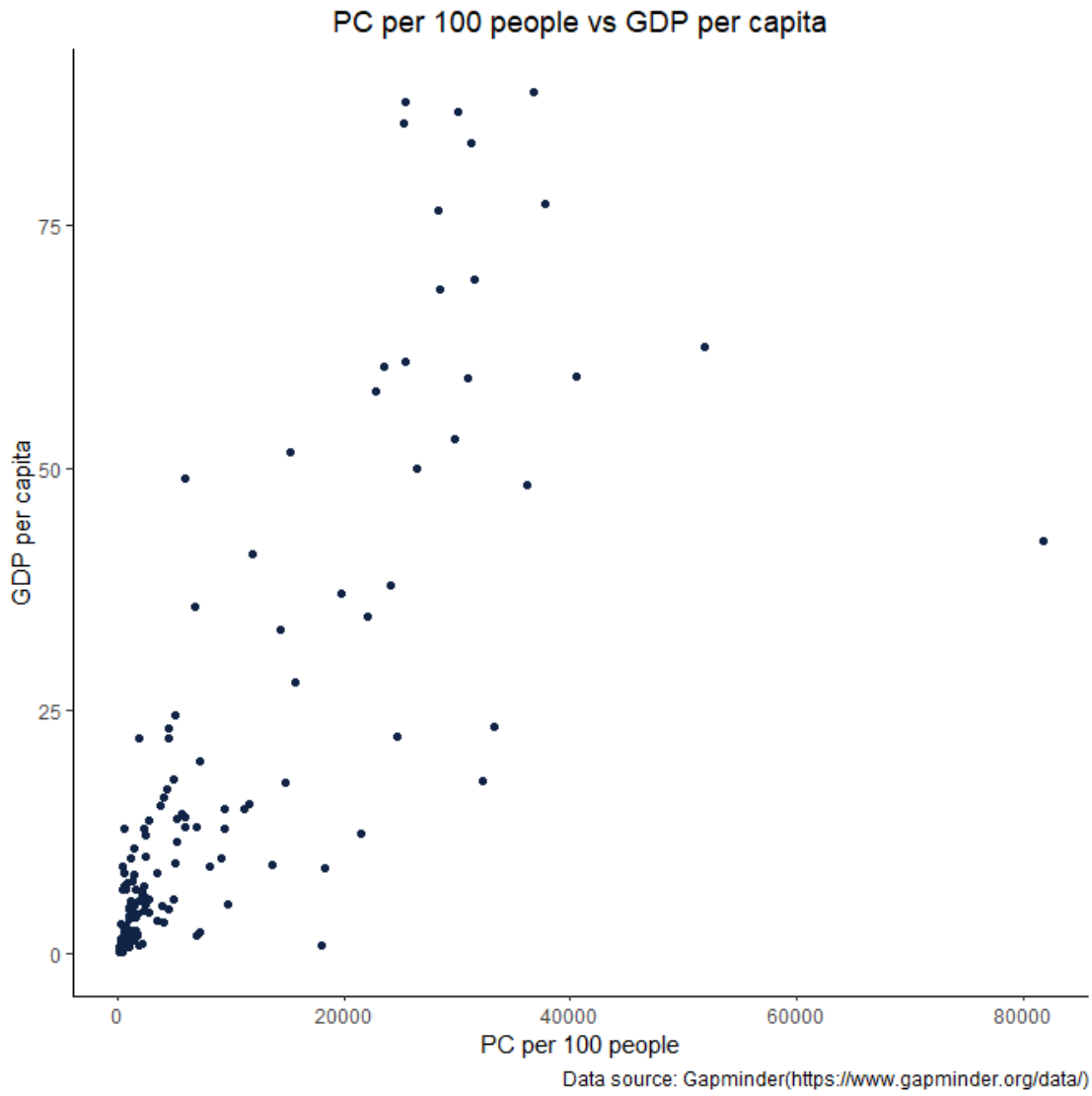
A scatter plot

The scatter plot below shows the positive linear relationship between between GDP per capita and number of personal computers per 100 individuals in 2005. It indicates that the higher the GDP per capita in the country is, the higher the percentage of PC per 100 individuals is.

However, there are several outliers especially around higher GDP capita such as Luxembourg and Monaco.

Luxembourg GDP: 51927,36096, PC per 100:62,38
Monaco GDP: 81827,95612, PC per 100: 42,55

```
ggplot(Dataset1) +
 aes(x = GDPperCAPITA, y = PC_PER.100) +
 geom_point(shape = "circle", size = 1.5, colour = "#112446") +
 labs(x = "PC per 100 people", y = "GDP per capita",title="PC per 100 people
vs GDP per capita",caption="Data source:
Gapminder(https://www.gapminder.org/data/)") +
theme_classic()+
theme(plot.title = element_text(hjust = 0.5))
```



Part 2 Regression analysis

Question 2 Has the electricity generation per capita in China increased from 1990 to 2005?

Regression Analysis

Making a linear regression

The following linear regression model is created, using R.

```
RegModel.1 <- lm(EL_China~Year, data=Dataset2)
RegModel.1 %>% as_flextable()
```

	Estimate	Standard Error	t value	Pr(>|t|)	
(Intercept)	-211,961.557	19,293.924	-10.986	0.0000	***

	Estimate	Standard Error	t value	Pr(>|t|)	
Year	106.657	9.652	11.051	0.0000	***

*Signif. codes: 0 <= '***' < 0.001 < '**' < 0.01 < '*' < 0.05 < '.' < 0.1 < '' < 1*

Residual standard error: 230.4 on 17 degrees of freedom

Multiple R-squared: 0.8778, Adjusted R-squared: 0.8706

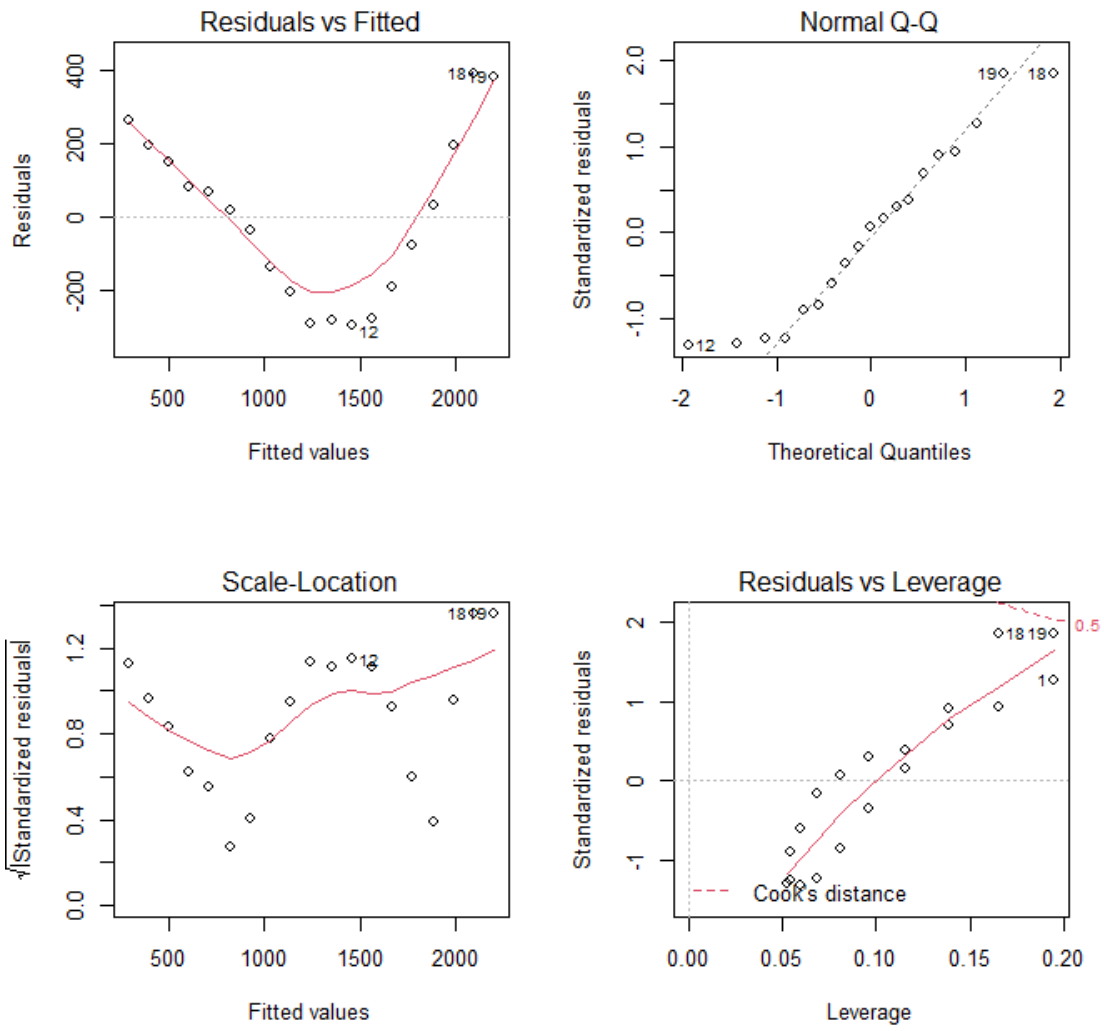
F-statistic: 122.1 on 17 and 1 DF, p-value: 0.0000

Checking if the model is well adapted to the data set.

Graph for "Residuals vs Fitted" in the basic diagnostic plots below shows U shape patterns and does not show horizontal line. This means weak linear relationships in this model. "Normal Q-Q shows plots" follows roughly reference line, which means the residual are roughly normally distributed. "Scale-location plots" do not show a horizontal line and plots spread unequally, which indicates heteroscedasticity. "Resiudals vs Leverage" chart shows that there are no outliers that may affect the model as all plots are within Cook's distance line.

Basic Diagnostic plots

```
par(mfrow=c(2,2))
plot(RegModel.1)
```

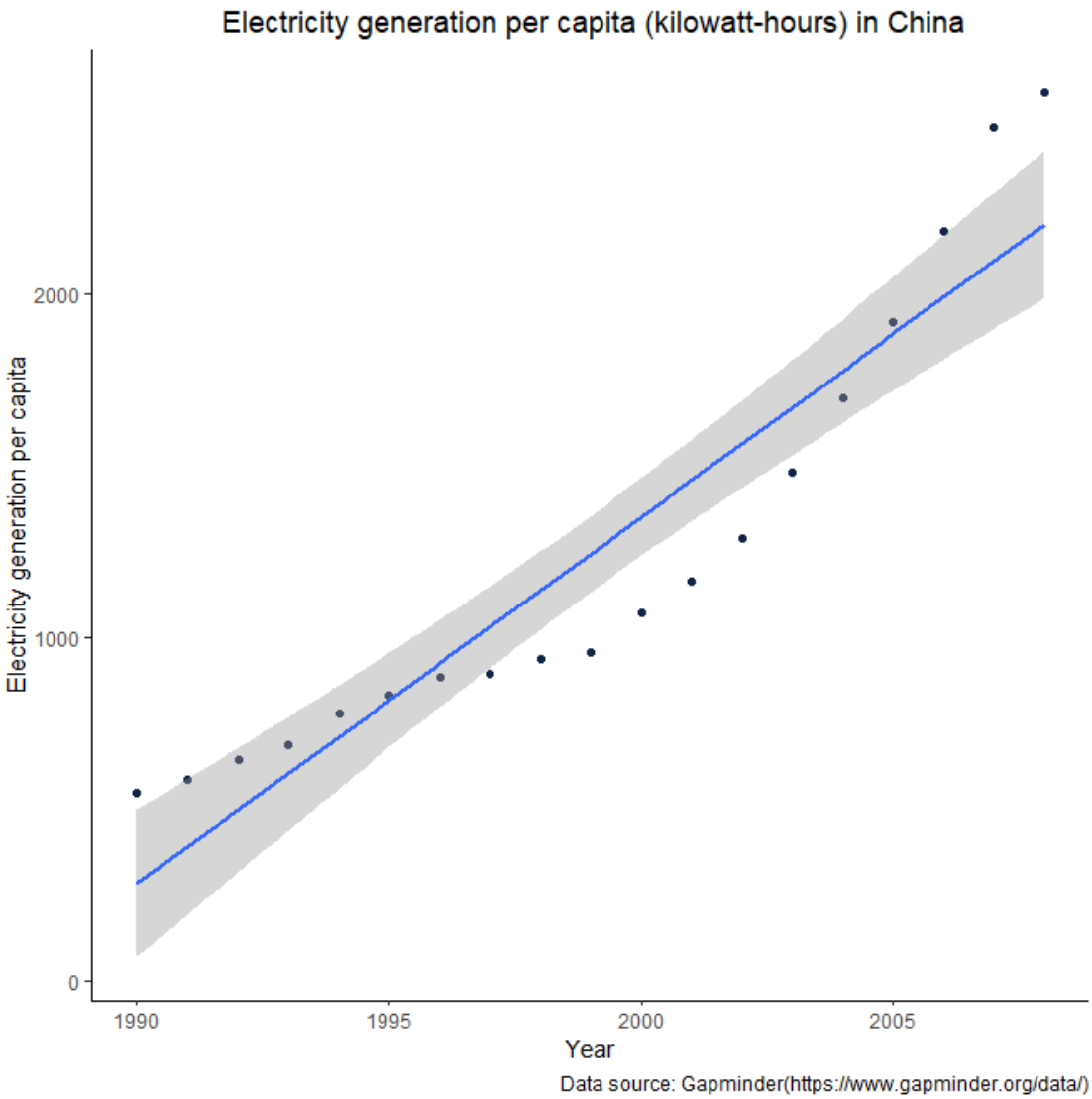



```
par(mfrow=c(1,1))
```

Scatterplot

The scatterplot (year and electricity generation per capita in china) below does not show complete linear. There is a slight curve around year 2000, which shows higher growth rate of electricity generation after 2000.

```
ggplot(Dataset2) +
  aes(x = Year, y = EL_China) +
  geom_point(shape = "circle", size = 1.5, colour = "#112446") +
  labs(x = "Year", y = "Electricity generation per capita", title = "
Electricity generation per capita (kilowatt-hours) in China", caption="Data
source: Gapminder(https://www.gapminder.org/data/)") +
  theme_classic()+
  stat_smooth(method="lm", formula=y~x, geom="smooth")+
  theme(plot.title = element_text(hjust = 0.5))
```



Transform the response variable, using log 10.

The evaluation of basic diagnostic plots and the scatterplot (year and electricity generation per capita in China) above shows that the regression model(RegModel.1) that is created above needs to be improved. In order to improve the regression model, log 10 transformation is used to create the new model as below.

```
RegModel.2 <-
  lm(log10(EL_China)~Year,
  data=Dataset2)
```

According to the test result of the model after log 10 transformation(RegModel.2)below, Adjusted R-squared has improved from 0.8706 to 0.9689. It indicates that this model is better fit than the first model.

```
RegModel.2 %>% as_flextable()
```

	Estimate	Standard Error	t value	Pr(>|t|)	
(Intercept)	-70.312	3.093	-22.731	0.0000	***
Year	0.037	0.002	23.715	0.0000	***

*Signif. codes: 0 <= '***' < 0.001 < '**' < 0.01 < '*' < 0.05 < '.' < 0.1 < ' ' < 1*

Residual standard error: 0.03694 on 17 degrees of freedom

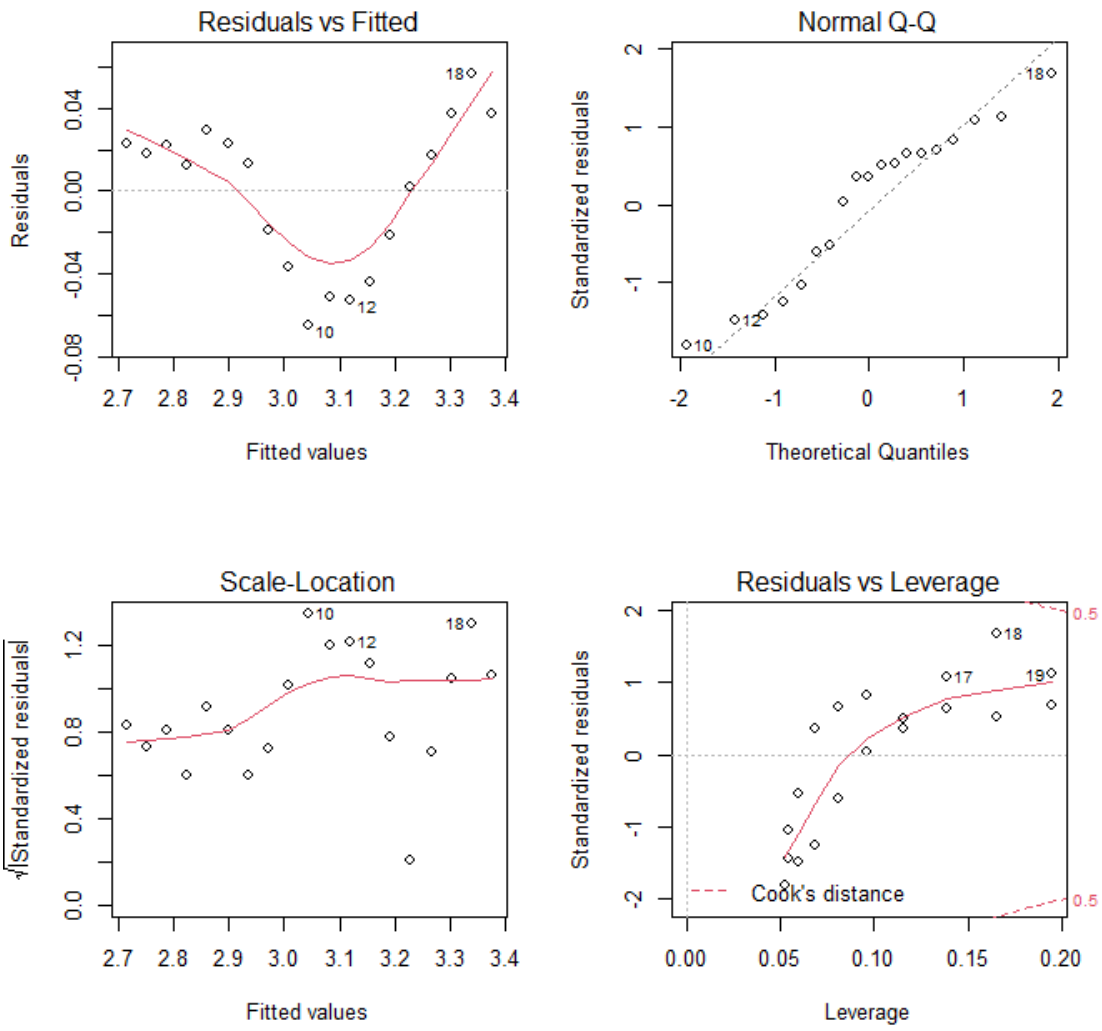
Multiple R-squared: 0.9707, Adjusted R-squared: 0.9689

F-statistic: 562.4 on 17 and 1 DF, p-value: 0.0000

Diagnostic Diagram for RegModel.2

Graph for "Residuals vs Fitted" in the basic diagnostic plots still shows U shape patterns but less than the previous model (RegModel.1). "Normal Q-Q shows plots" follows roughly reference line, which means the residual are roughly normally distributed. In "Scale-location" chart, plots do not show a complete horizontal lines that indicates homoscedasticity but closer to horizontal than the previous model. "Residuals vs Leverage" chart shows that there are no outliers that may affect the model as all plots are within Cook's distance line.

```
par(mfrow=c(2,2))
plot(RegModel.2)
```




```
par(mfrow=c(1,1))
```

Transforming the response variable, using square root.

In order to search for a better fit regression model, the response model is transformed, using square root.

According to the test result below, Adjuststed R-squared (0.9278) in the model (RegModel.3) has not improved from the previous model(RegModel.2, Adjusted R-squared (0.9689)). It suggests that the previous model(RegModel.2) is better fit than the model(Regmodel.3).

```
RegModel.3 <-
  lm(sqrt(EL_China)~Year,
  data=Dataset2)

RegModel.3 %>% as_flextable()
```

	Estimate	Standard Error	t value	Pr(>|t|)	
(Intercept)	-2,914.440	193.462	-15.065	0.0000	***
Year	1.475	0.097	15.242	0.0000	***

*Signif. codes: 0 <= '***' < 0.001 < '**' < 0.01 < '*' < 0.05 < '.' < 0.1 < ' ' < 1*

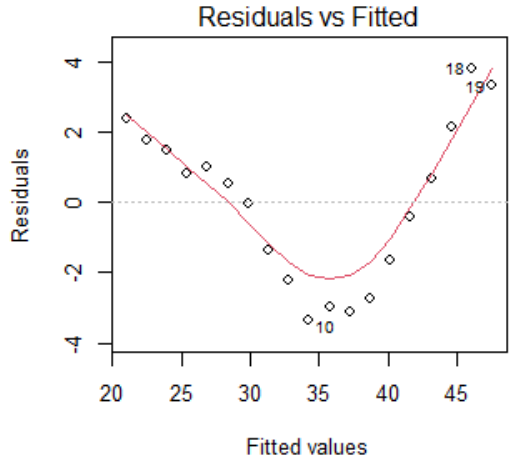
Residual standard error: 2.311 on 17 degrees of freedom

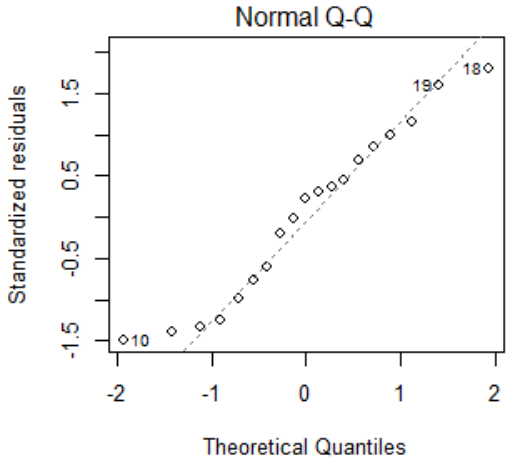
Multiple R-squared: 0.9318, Adjusted R-squared: 0.9278

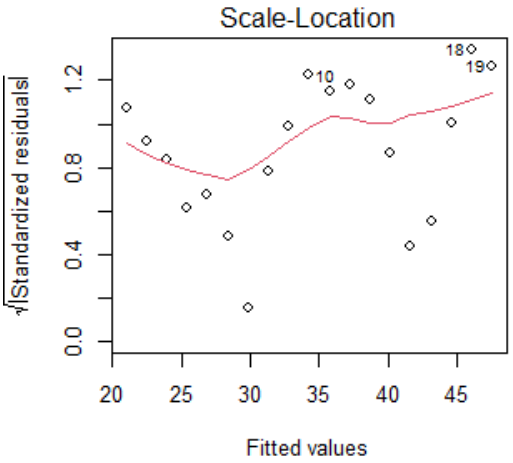
F-statistic: 232.3 on 17 and 1 DF, p-value: 0.0000

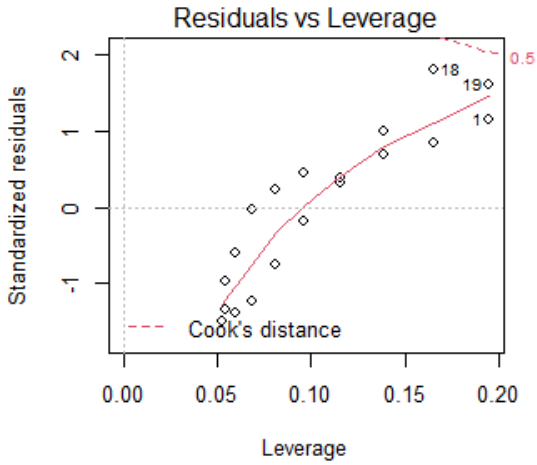
Graph for "Residuals vs Fitted" in the basic diagnostic plots below shows U shape patterns. "Normal Q-Q", plots follows reference line. In "Scale-location" chart,plots do not show complete horizontal lines. "Residuals vs Leverage" chart shows that there are no outliers that may affect the model as all plots are within Cook's distance line.

```
par(mfrow=c(2,2))
plot(RegModel.3)
```

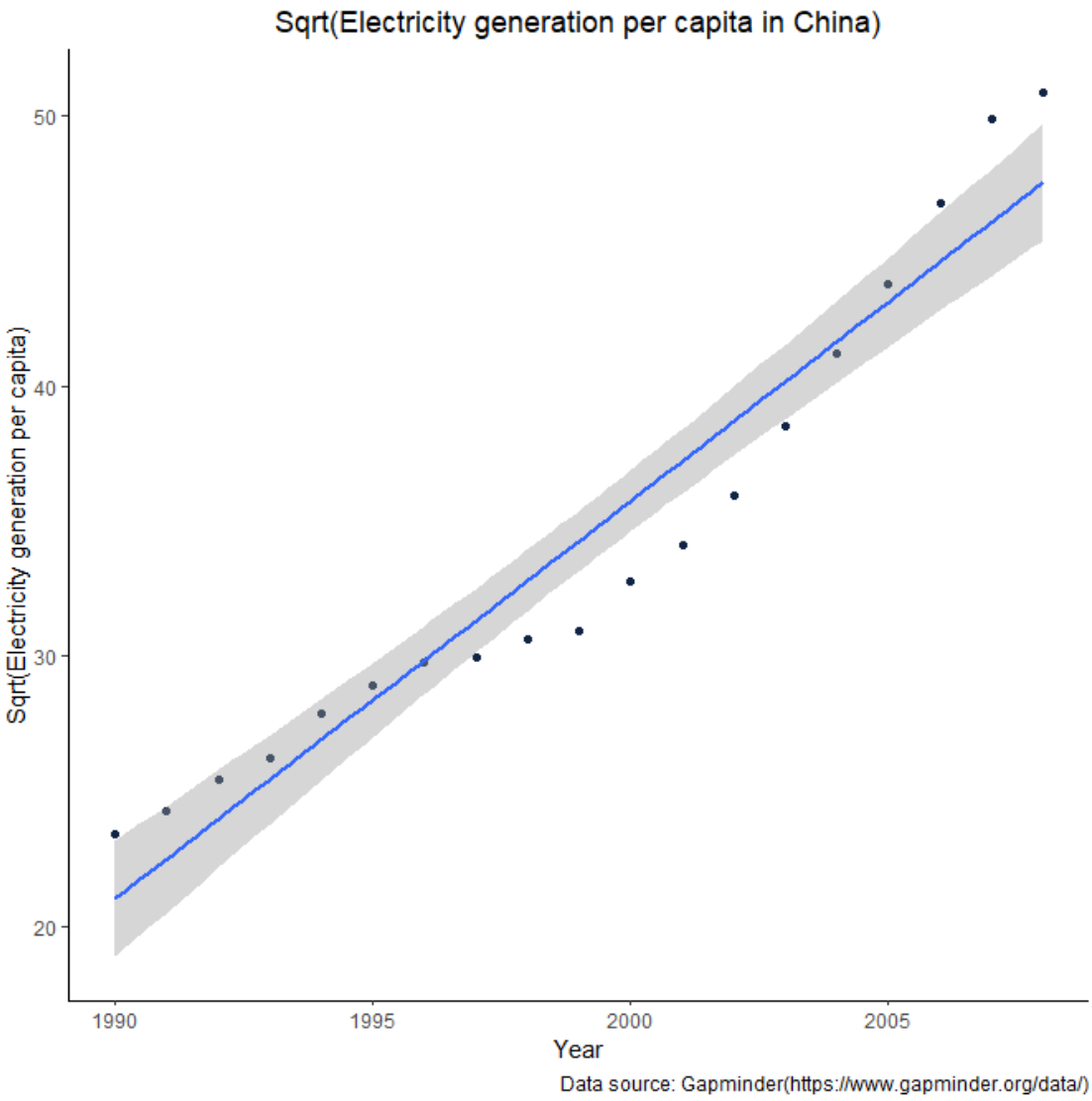









```
par(mfrow=c(1,1))

ggplot(Dataset2) +
  aes(x = Year, y = sqrt(EL_China)) +
  geom_point(shape = "circle", size = 1.5, colour = "#112446") +
  labs(x = "Year", y = "Sqrt(Electricity generation per capita)",title =
"Sqrt(Electricity generation per capita in China)", caption = "Data source:
Gapminder(https://www.gapminder.org/data/)" )+
  theme_classic()+
  stat_smooth(method="lm", formula=y~x, geom="smooth")+
  theme(plot.title = element_text(hjust = 0.5))
```



Comparison of all the different models

Comparison of Adjusted R2 value of the 3 models suggested that RegModel.2 are best model to use. (R-squared: 0.9707, Adjusted R-squared: 0.9689 F-statistic: 562.4 on 1 and 17 DF, p-value<0.01)

Therefore, RegModel.2 (b(regression coefficient): 0.036687, SEb = 0.001547, t=23.71, p<0.01 df=17) below is the best fit among 3 models.

```
RegModel.2 %>% as_flextable()
```

	Estimate	Standard Error	t value	Pr(>|t|)	
(Intercept)	-70.312	3.093	-22.731	0.0000	***
Year	0.037	0.002	23.715	0.0000	***

	Estimate	Standard Error	t value	Pr(>|t|)

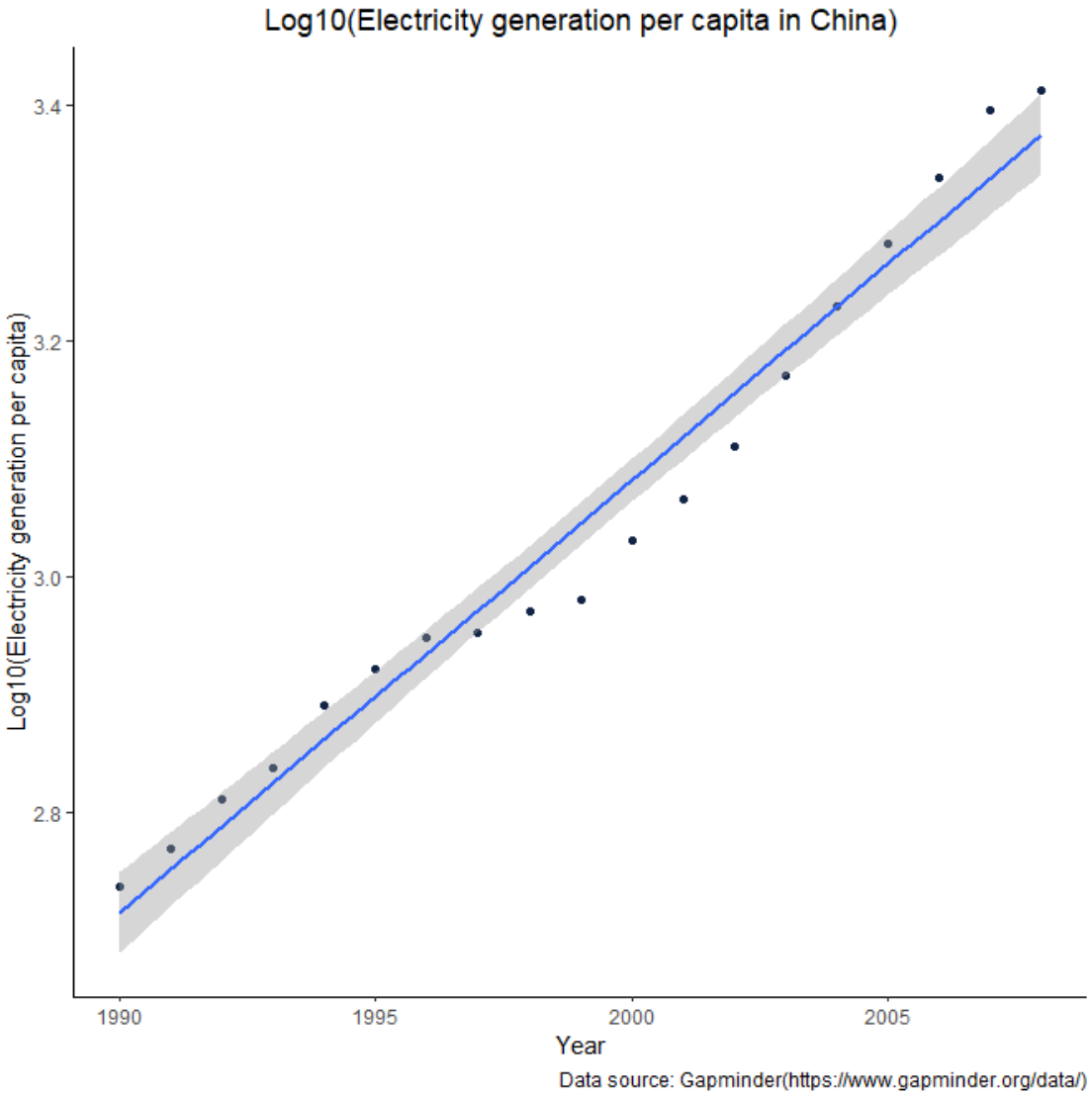
*Signif. codes: 0 <= '***' < 0.001 < '**' < 0.01 < '*' < 0.05 < '.' < 0.1 < ' ' < 1*

Residual standard error: 0.03694 on 17 degrees of freedom

Multiple R-squared: 0.9707, Adjusted R-squared: 0.9689

F-statistic: 562.4 on 17 and 1 DF, p-value: 0.0000

```
ggplot(Dataset2) +
  aes(x = Year, y = log10(EL_China)) +
  geom_point(shape = "circle", size = 1.5, colour = "#112446") +
  labs(x = "Year", y = "Log10(Electricity generation per capita)", title =
"Log10(Electricity generation per capita in China)", caption = "Data source:
Gapminder(https://www.gapminder.org/data/)")+
  theme_classic()+
  stat_smooth(method="lm", formula=y~x, geom="smooth")+
  theme(plot.title = element_text(hjust = 0.5))
```



Part 3. Testing differences between groups

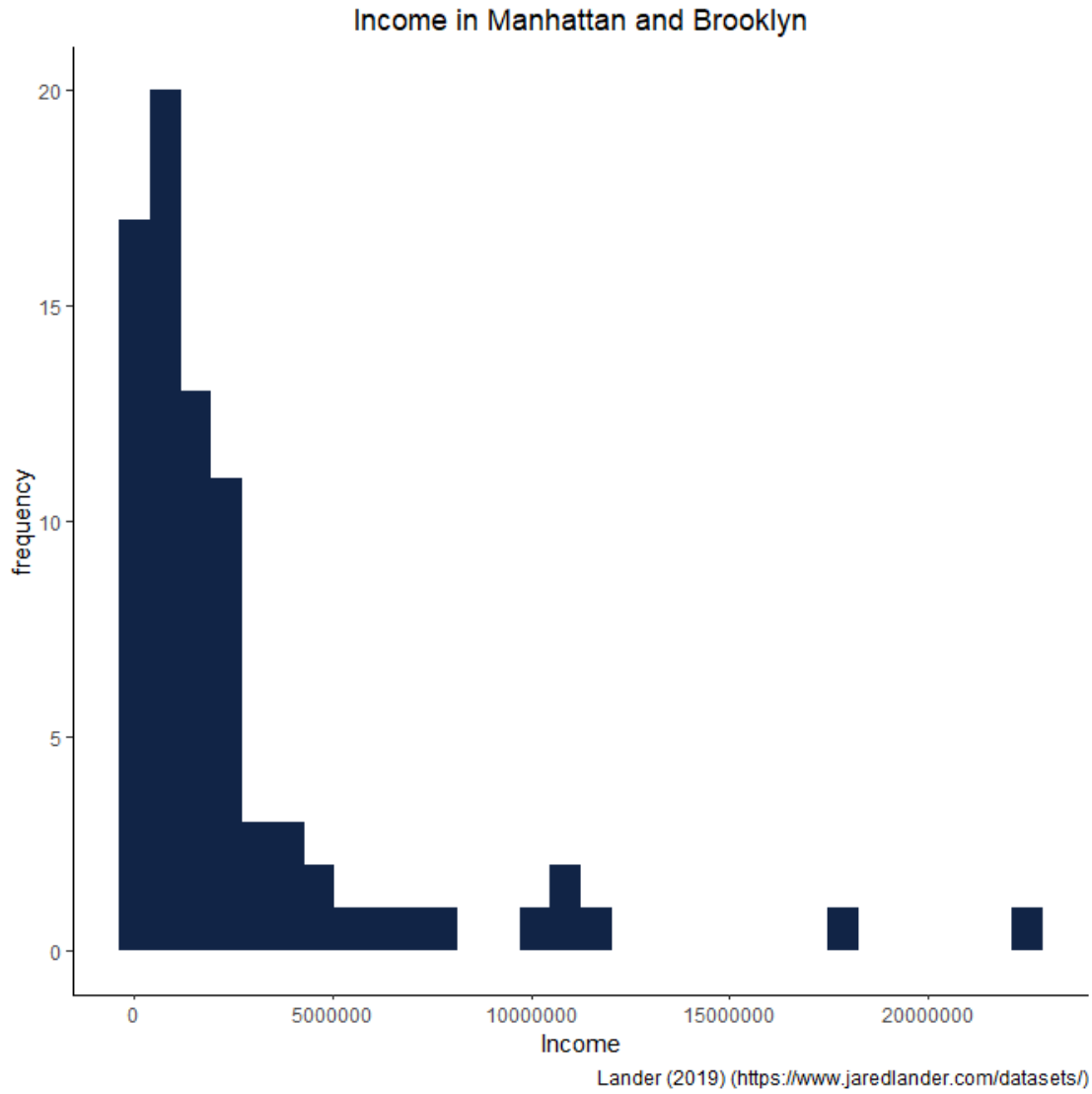
Question 3. Is there a difference in income between the New York districts, Manhattan and Brooklyn?

Histogram and appropriate test to use

According to the histogram below, variable income is not normally distributed. Therefore, Wilcoxon's rank-sum test is more appropriate to use.

```
options(scipen=999)
ggplot(Dataset3) +
  aes(x = Income) +
  geom_histogram(bins = 30L, fill = "#112446") +
  labs(x = "Income", y="frequency", title="Income in Manhattan and Brooklyn"
,caption="Lander (2019) (https://www.jaredlander.com/datasets/)") +
```

```
  theme_classic()+
  theme(plot.title = element_text(hjust = 0.5))
```



2. Run Wilcoxon's rank-sum test

The result of Wilcoxon's rank-sum test is as below (W = 223, p-value = 0.00001019, df= 21). The p-value(0.00001019) indicates that there are statistically significant differences in income between Manhattan and Brooklyn.

```
wilcox.test(Income ~ Boro, alternative="two.sided", data=Dataset3)

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Income by Boro
## W = 223, p-value = 0.00001019
## alternative hypothesis: true location shift is not equal to 0
```

Calculation of median, mean, SD

Median

There is a big difference in median of income between in Manhattan and Brooklyn in the dataset. The median in Manhattan is much higher than the one in Brooklyn.

```
Tapply(Income ~ Boro, median, na.action=na.omit, data=Dataset3)

##  Brooklyn Manhattan
##  443850.5 1742515.0

with(Dataset3, tapply(Income, Boro,numSummary))

## $Brooklyn
##      mean       sd      IQR     0%    25%      50%      75%    100%  n
##  639548.1 598728.4 482661.5 147206 257152 443850.5 739813.5 2338189 22
##
## $Manhattan
##     mean      sd     IQR     0%    25%     50%     75%     100%  n
##  3344961 4345457 2695419 147229 941958 1742515 3637377 22673513 57
```

MEAN

There is a big difference in mean of income between Manhattan and Brooklyn in the dataset.The mean in Manhattan is much higher than the one in Brooklyn.

```
with(Dataset3,tapply(Income,Boro, mean))

##  Brooklyn Manhattan
##  639548.1 3344961.0
```

SD

There is a big difference in standard deviation of income between Manhattan and Brooklyn in the dataset. The standard deviation in Manhattan is much bigger than the one in Brooklyn.

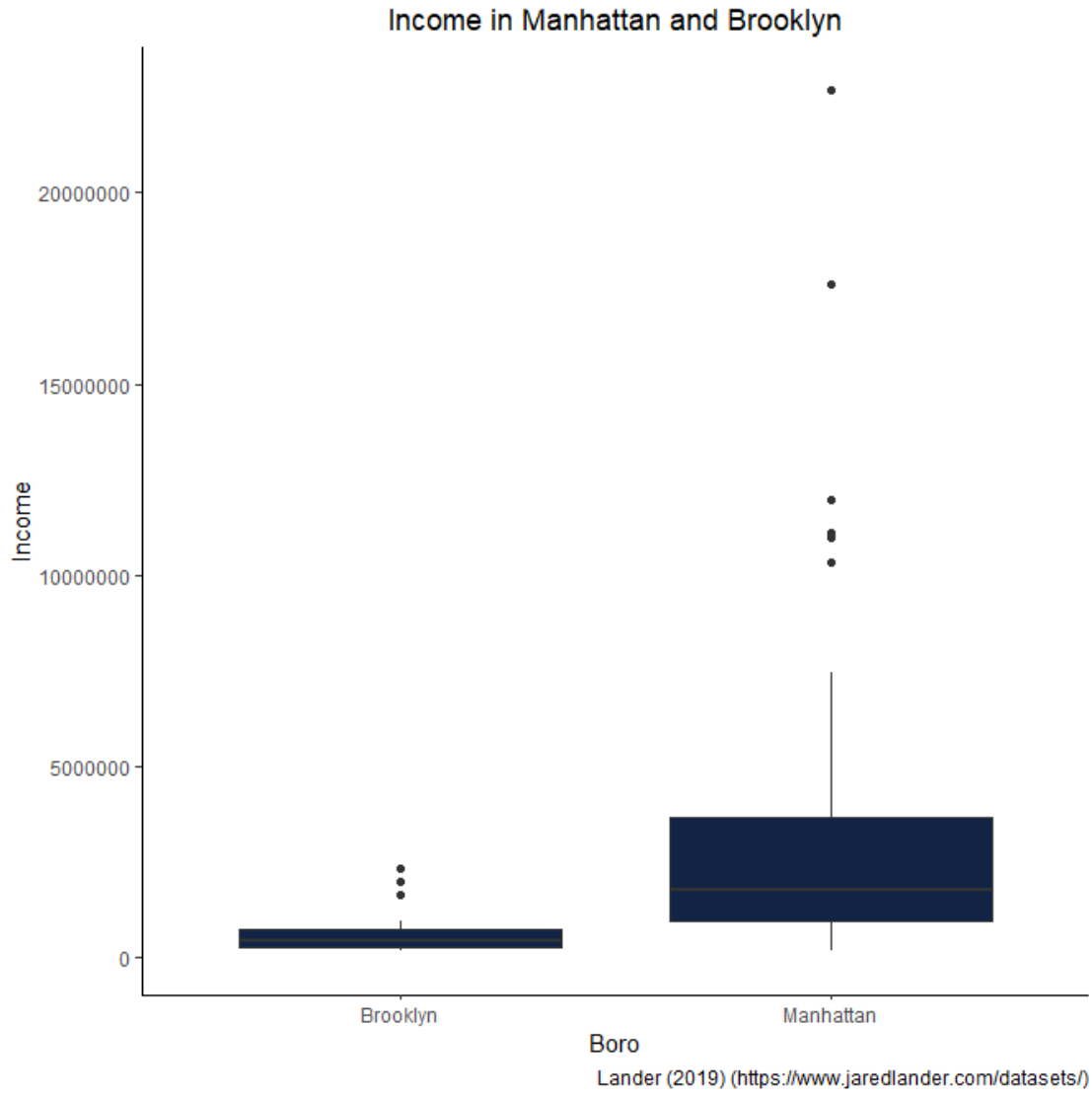
```
with(Dataset3,tapply(Income,Boro, sd))

##  Brooklyn Manhattan
##  598728.4 4345456.9
```

Graphs to visualize findings

The boxplot below shows that the median of income in Manhattan is higher than in Brooklyn and there is a big difference. However, the difference in upper quartiles as well as the difference in maximum between Brooklyn and Manhattan is much bigger than the difference of median. Additionaly, the dataset for Manhattan is right skewed with a large variance/spread and several extreme outliers exist around higher income in Manhattan.

```
options(scipen=999)
ggplot(Dataset3) +
```

```
  aes(x = Boro, y = Income) +
  geom_boxplot(shape = "circle", fill = "#112446") +
  labs(x = "Boro", y = "Income", title="Income in Manhattan and Brooklyn"
,caption="Lander (2019) (https://www.jaredlander.com/datasets/)") +
  theme_classic()+
  theme(plot.title = element_text(hjust = 0.5))
```



Interpretation of the results

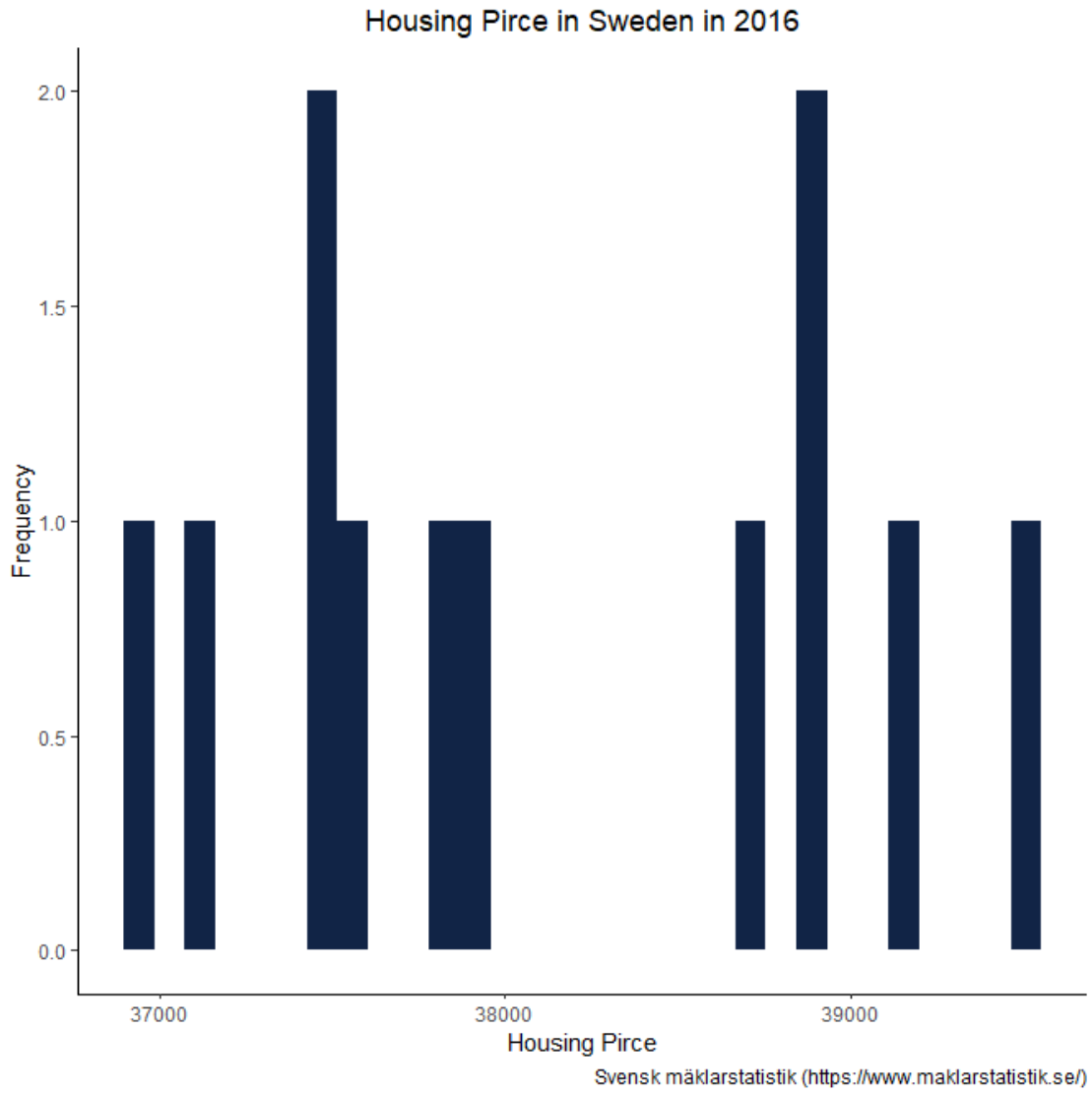
According to the the result of Wilcoxon's rank-sum test mentioned above(W = 223, p-value = 0.00001019 and df= 21), there are significant difference in income between Manhattan and Brooklyn.

Question 4 Are there differences in house pricing (SEK/m2) in Sweden between 2016 and 2017?

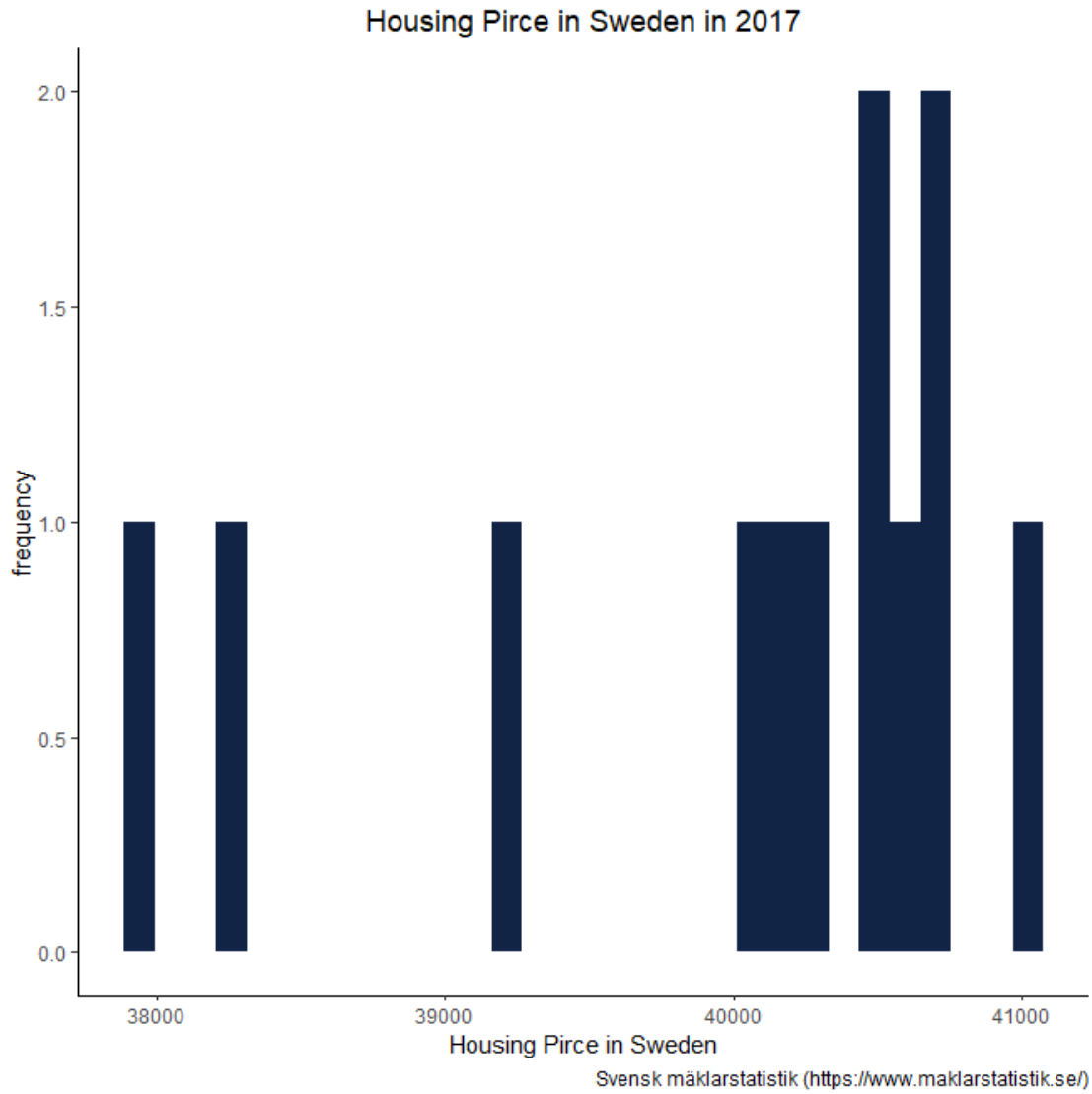
Histograms and appropriate test to use

According to the histograms below, both variable(housing price in Sweden in 2016 and 2017) are not normally distributed. Therefore, wilcoxon singed rank exact test is appropriate to use.

```
ggplot(Dataset4) +
  aes(x = X2016_sek_sqrm) +
  geom_histogram(bins = 30L, fill = "#112446") +
  labs(x = "Housing Pirce", y = "Frequency", title="Housing Pirce in Sweden
in 2016", caption="Svensk mäklarstatistik (https://www.maklarstatistik.se/)")
+
  theme_classic()+
 theme(plot.title = element_text(hjust = 0.5))
```



```
ggplot(Dataset4) +
  aes(x = X2017_sek_sqrm) +
  geom_histogram(bins = 30L, fill = "#112446") +
  labs(x = "Housing Pirce in Sweden", y = "frequency", title="Housing Pirce
in Sweden in 2017", caption="Svensk mäklarstatistik
(https://www.maklarstatistik.se/)") +
  theme_classic()+
  theme(plot.title = element_text(hjust = 0.5))
```



Perform Wilcoxon signed rank exact test

The paired samples test is appropriate because both variables(X2016_sek_sqrm, x2017_sek_sqrm) come from the same sample(same month in different year) and they are not totally independent from each other.

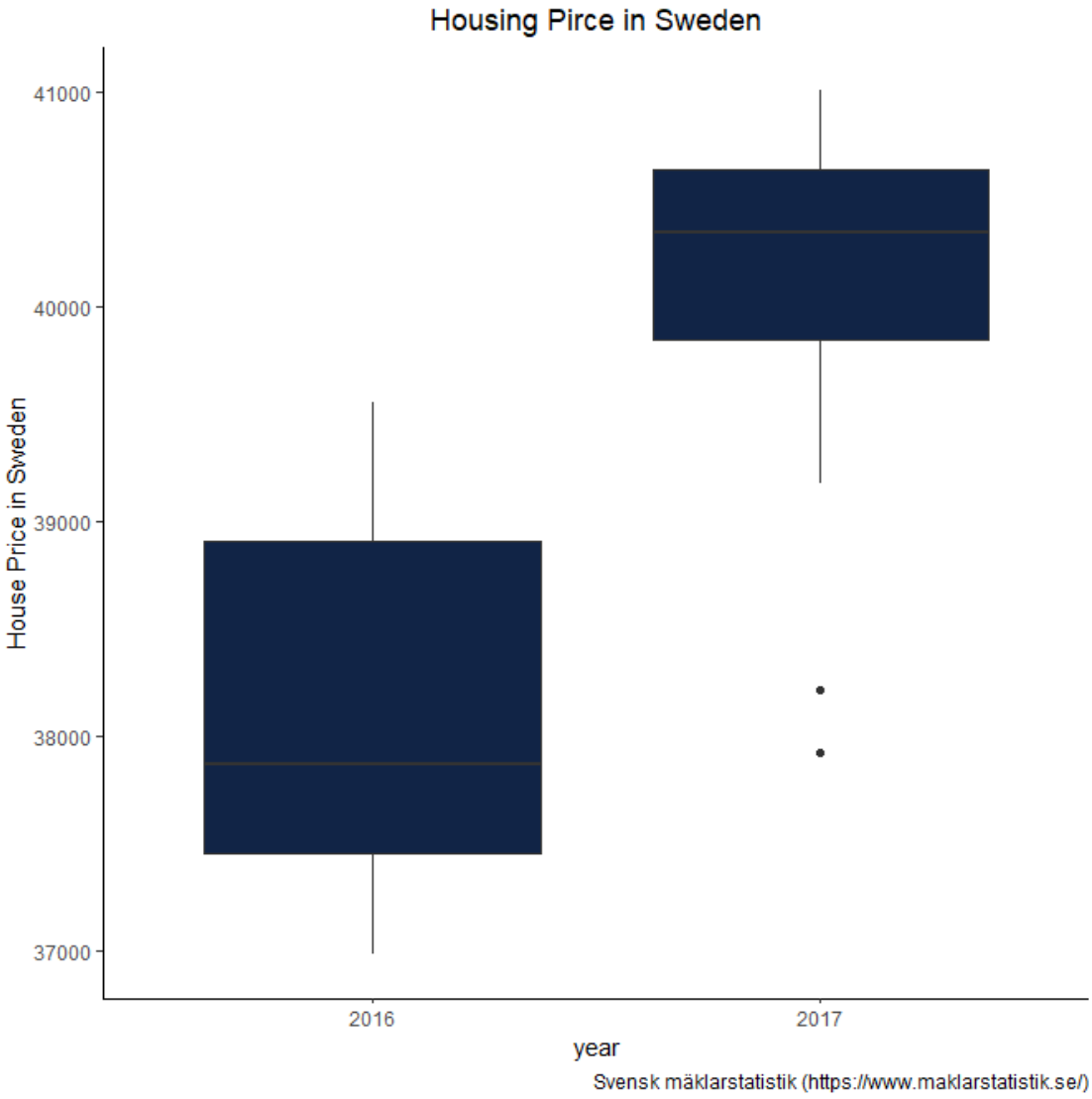
P-value from the test is 0.006836 with V=6 and degree of freedom 11. This suggests that the difference between 2016 and 2017 is statistically significant.

```
with(Dataset4, median(X2016_sek_sqrm - X2017_sek_sqrm, na.rm=TRUE))

## [1] -2759

  # median difference
with(Dataset4, wilcox.test(X2016_sek_sqrm, X2017_sek_sqrm,
  alternative='two.sided', paired=TRUE))

##
##  Wilcoxon signed rank exact test
##
## data:  X2016_sek_sqrm and X2017_sek_sqrm
## V = 6, p-value = 0.006836
## alternative hypothesis: true location shift is not equal to 0
```

A Plot to dispalay the diffrence

The plot below also shows that there is a difference between housing price in Sweden in 2016 and 2017 and that housing price in 2017 is higher than 2016. But it also shows that 2 outliers (november and december) in 2017 exists.

```
sqrm<-append(Dataset4$X2016_sek_sqrm,Dataset4$X2017_sek_sqrm)
label<-factor(c(rep("2016",12),rep("2017",12)))

Dataset4a<-data.frame(label, sqrm)
ggplot(Dataset4a) +
  aes(x = label, y = sqrm) +
  geom_boxplot(shape = "circle", fill = "#112446") +
  labs(x = "year", y = "House Price in Sweden", title="Housing Pirce in
Sweden", caption="Svensk mäklarstatistik (https://www.maklarstatistik.se/)")
+
  theme_classic()+
  theme(plot.title = element_text(hjust = 0.5))
```



Discussion

Question 1 Is there a relationship between GDP/capita and the number of personal computers per 100 individuals measured in 2005?

The Spearman Rank-Order correlation's test result shows that, there is correlation between GDP per capita and number of personal computers per 100 individuals in 2005. (95% confidence) It indicates that the country who has higher GDP per capita tend to have a higher percentage of PC per 100 individuals. However, there are some exceptions especially around higher GDP per capita such as Luxembourg and Monaco, which suggests that there might be another factor that can influence the percentage of PC per 100. This can be investigated further.

Question 2 Has the electricity generation per capita in China increased from 1990 to 2005?

According to comparison between the 3 linear regression models based on Adjusted R-squared, it can be concluded that Regmodel2 with log 10 transformation of dependent variable(electricity generation) are best fit. This implies that electricity generation per capita in China increased exponentially from 1990 to 2005. It increased more rapidly after around 2000.

Question 3 Is there a difference in income between the New York districts, Manhattan and Brooklyn?

The results of Wilcoxon's rank-sum test indicates that there are statistically significant difference in income between Manhattan and Brooklyn and Income in Manhattan is more likley to be higher than Brooklyn. However, several extreme outliers (high income) and a large variance with right skewed are found in data of Manhattan, which means that several very high incomes in Manhattan increased the average income in Manhattan.

Question 4. Are there differences in house pricing (SEK/m2) in Sweden between 2016 and 2017?

The result suggests that the difference in house pricing between 2016 and 2017 is statistically significant. It indicates that there are differences in house pricing (SEK/m2) in Sweden between 2016 and 2017. It can be said that the housing price in Sweden has increased between 2016 and 2017 in total(Jan-Oct). However, it is good to notice that the dataset shows that the housing price decreased in Oct, Nov and December in 2017 and the housing price in November and December in 2017 was lower than the same months in 2016. It means that there was a relatively large decline in housing price from October to December 2017 in contrast to the constant increase that was seen before.