# Part 3 Assignment 1

Sanjiv Kumar

10/14/2021

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

### Question 1

### Introduction

We are analyzing the relationship between GDP per capita and access to the personal computers per 100 measured in 2005 (data # Ref 1). The data set from two data (from GDP_per_capita_2005.xlxs, Personalcomputer_2005.xlxs) were combined in the form of `gdp_pc` (data import). Data for the countries missing either of the data, i.e. GDP per capita or access to personal computer were removed from the dataset for this analysis.

### Method

Initially, we see the general summary statistics using `summary(gdp_pc)` as shown in result section. Then we calculate the mean, median and standard deviation (sd) for both the columns in the dataset (i.e. `gdp_pc`) and we report Report the Mean, SD and Median.

### Result

The mean GDPperCPITA is 7963.5138784 and mean PC_per_100 is 15.7809032. The median GDPperCPITA and PC_per_100 is 2196.247196 and 5.9 respectively. The standard deviation for GDPperCPITA and PC_per_100 is $1.2328808 \times 10^4$ and 22.1197352, respectively.

```
summary(gdp_pc)
```

```
##                    Data      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
```

```
sd(gdp_pc$GDPperCAPITA)
```

```
## [1] 12328.81
```

```
sd(gdp_pc$PC_per_100)
```

```
## [1] 22.11974
```

#### Histogram

From the histogram below, both the GDPperCAPITA and PC__per__100 does not appear to be normally distributed. Since the data was collected over same countries, and to see the relationship between the continuous variables GDPperCAPITA and PC__per__100 across the data set, we need to perform correlation analysis. To check the relationship between GDPperCAPITA and PC__per__100, scatter plot was made which shows linear correlation between the chosen data. In this case, Pearson product-moment correlation was used for the analysis, as Spearman Rank-Order correlation is for the non-parametric analysis, requiring at least one ordinal variable.

```
with(gdp_pc, Hist(GDPperCAPITA, scale="frequency", breaks="Sturges",
                  col="darkgray", xlab = "GDP Per Capita", ylab = "Frequency"))
```

```
with(gdp_pc, Hist(PC_per_100, scale="frequency", breaks="Sturges",
                  col="darkgray", xlab = "PC Per 100", ylab = "Frequency"))
```

#### Analysis

Pearson product-moment correlation:

- Correlation coefficient: 0.7882171
- df: 153
- p-value: $< 2.2e\text{-}16$
- name of test: Pearson's product-moment correlation

```
with(gdp_pc, cor.test(GDPperCAPITA, PC_per_100, alternative="two.sided",
  method="pearson"))
```

```
##
##  Pearson's product-moment correlation
##
## data:  GDPperCAPITA and PC_per_100
## t = 15.843, df = 153, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7200420 0.8413218
## sample estimates:
##       cor
## 0.7882171
```

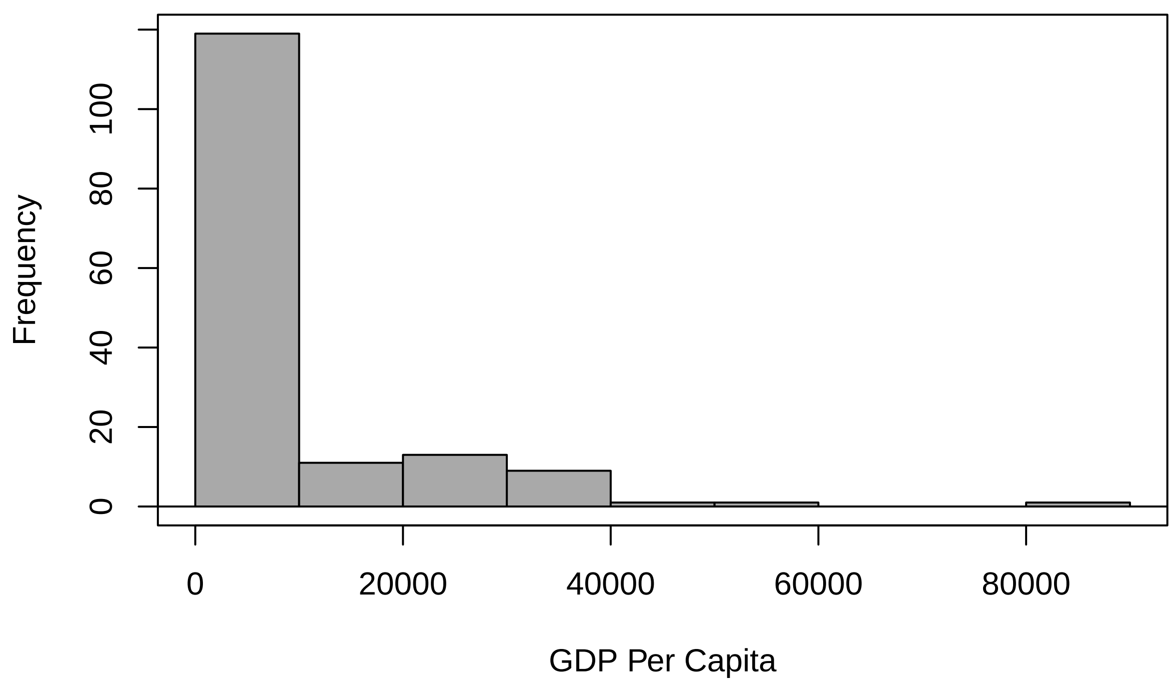



Figure 1: Histogram of GDPperCAPITA

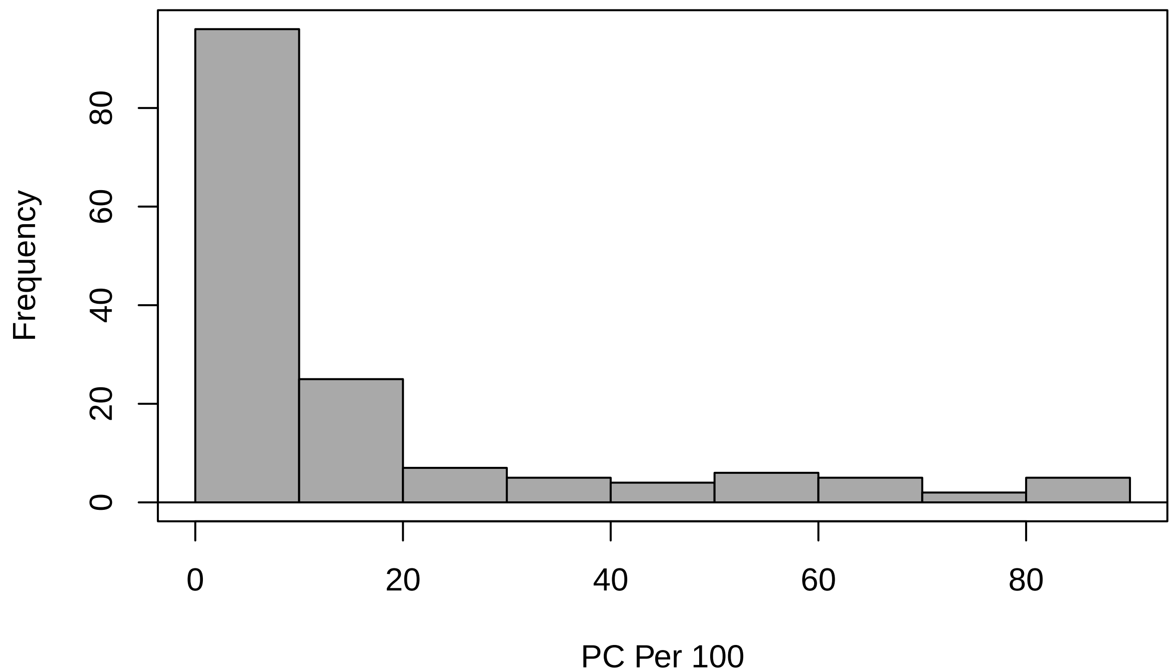



Figure 2: Histogram of PC per 100

**Scatterplot**

```
scatterplot(PC_per_100~GDPperCAPITA, regLine=FALSE, smooth=FALSE,
            boxplots=FALSE, data=gdp_pc, xlab = "GDP Per Capita",
            ylab = "PC Per 1000")
```

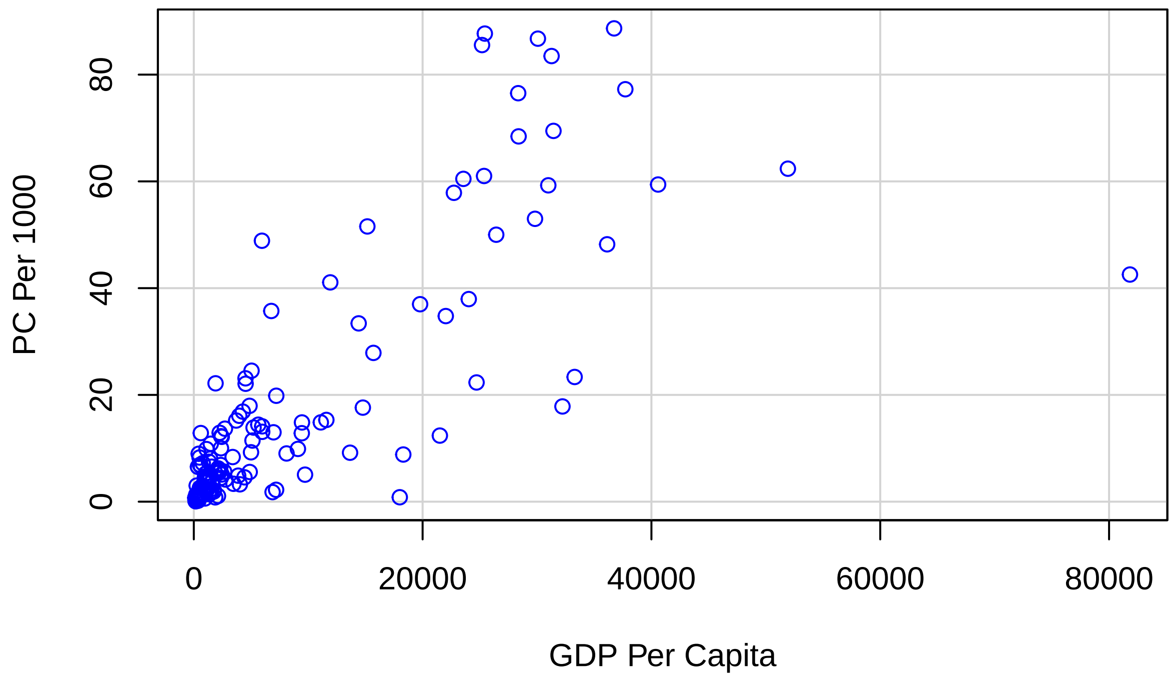



Figure 3: Scatterplot of GDPperCAPITA vs PC_per_100

## Discussion

The data shows significant high linear correlation between the GDP Per Capita and PC per 1000. However, from the dataset, the data points are not uniformly distributed but are concentrated near x, y intercept (as apparent from scatterplot). Also shapiro.test, shows p $<$ 0.005 (and therefore, hypothesis of normally distributed is rejected) for both GDPperCAPITA and PC_per_100. Therefor, for this data we may also choose Spearman Rank-Order correlation, which also shows similar trend and conclusion. In this case:

- Correlation rho: 0.8547637
- p-value: $<$ 2.2e-16
- name of test: Spearman's rank correlation rho

Both the correlation methods give positive correlation between GDPperCAPITA and PC_per_100.

```r
shapiro.test(gdp_pc$GDPperCAPITA)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  gdp_pc$GDPperCAPITA
## W = 0.66031, p-value < 2.2e-16
```

```r
shapiro.test(gdp_pc$PC_per_100)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  gdp_pc$PC_per_100
## W = 0.69647, p-value < 2.2e-16
```

```r
with(gdp_pc, cor.test(GDPperCAPITA, PC_per_100, alternative="two.sided",
  method="spearman", exact=FALSE))
```

```
##
##  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
```

### References


1. GDP_per_capita_2005.xlxs, Personalcomputer_2005.xlxs from Gapminder (https://www.gapminder.org/data/).


# Part 2. Regression analysis

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

### Introduction

Here we analyse a linear regression model to study the electricity generation per capita from 1990 to 2005 (data # Ref 1).

### Method

Here we make three different regression models i.e. normal data, log10 transformed and square root transformed data to make these models.

## Result

Three different regression models were made as follows.

```
Reg.Model.1 <- lm(China~Year, data=chinaElectricity)
summary(Reg.Model.1)
```

```
##
## Call:
## lm(formula = China ~ Year, data = chinaElectricity)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -187.50  -98.50   14.09   62.55  300.70
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.569e+05  1.446e+04  -10.85 3.36e-08 ***
## Year         7.908e+01  7.239e+00   10.92 3.09e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 133.5 on 14 degrees of freedom
## Multiple R-squared:  0.895,  Adjusted R-squared:  0.8875
## F-statistic: 119.3 on 1 and 14 DF,  p-value: 3.092e-08
```

```
#plot(Reg.Model.1)
```

- r-squared: 0.887505
- b (regression coefficient): $-1.5694468 \times 10^5$
- SEb: 7.2392692
- t: 10.9242484
- df: 2, 14, 2
- p: $3.0922238 \times 10^{-8}$

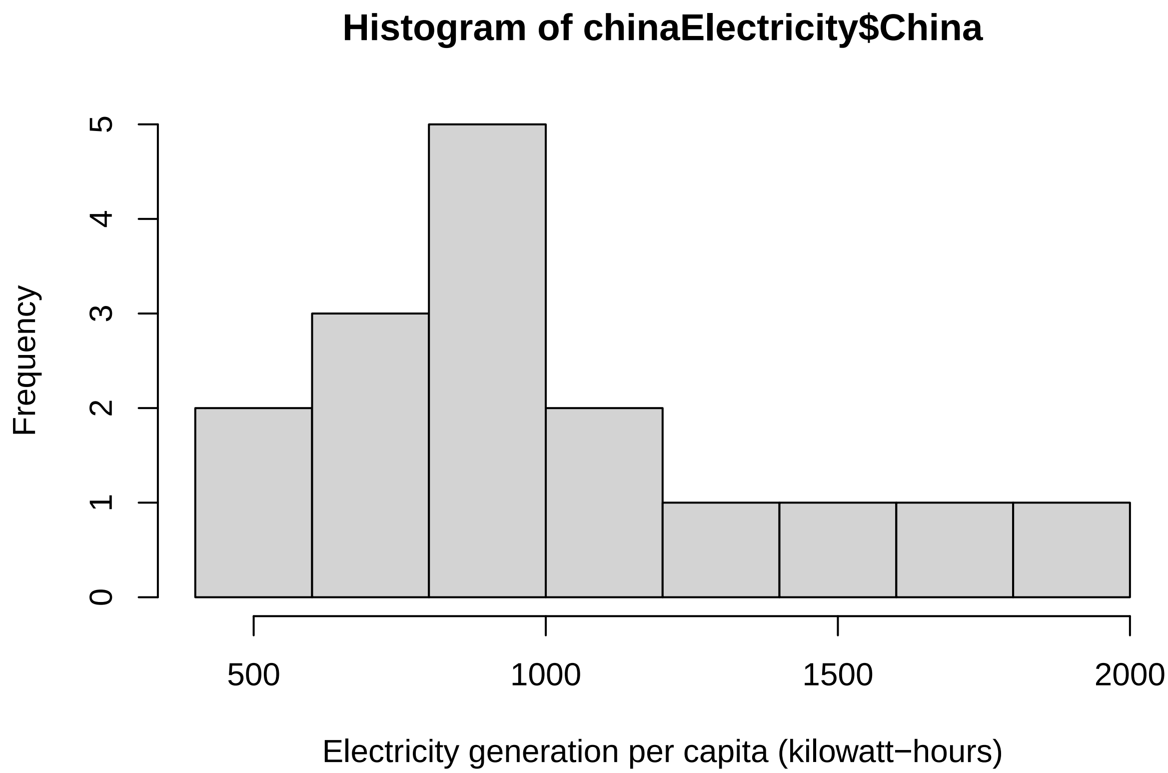
## Histogram

Distribution of electricty production looks like normal distribution, skewed towards left.

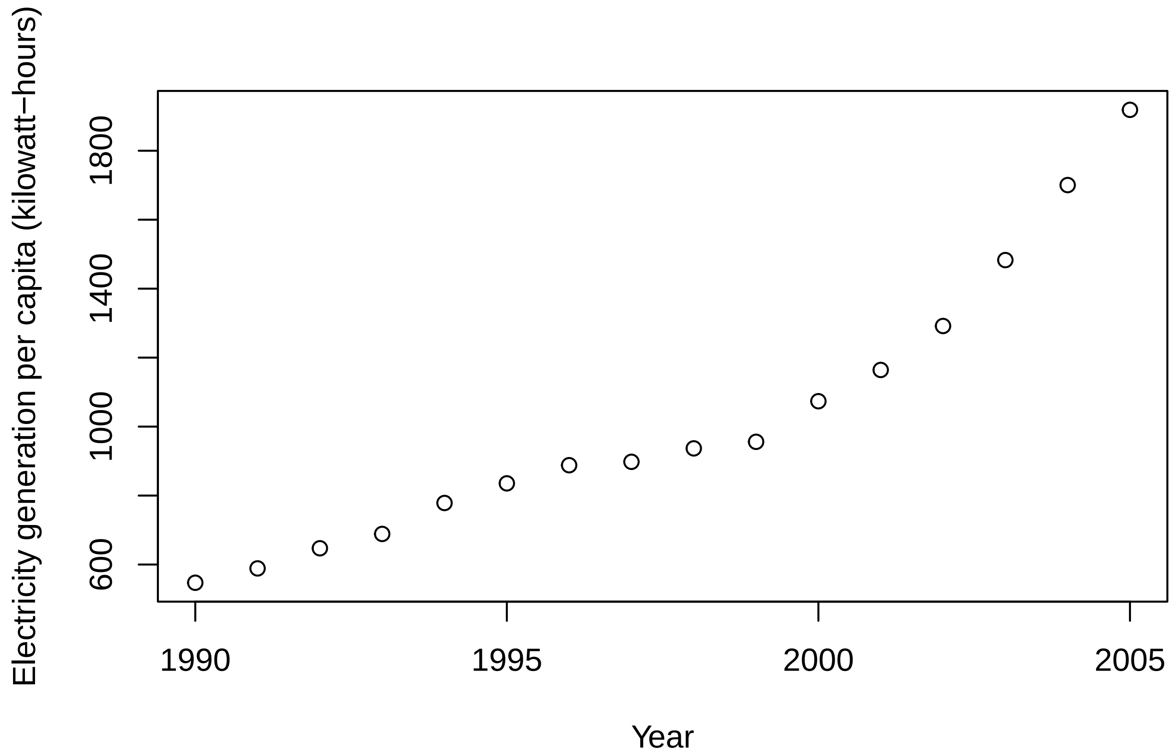
```
summary(chinaElectricity$China)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   547.2   756.0   917.4  1024.8  1196.2  1918.6
```

```
hist(chinaElectricity$China,
     xlab = "Electricity generation per capita (kilowatt-hours)",
     ylab = "Frequency")
```

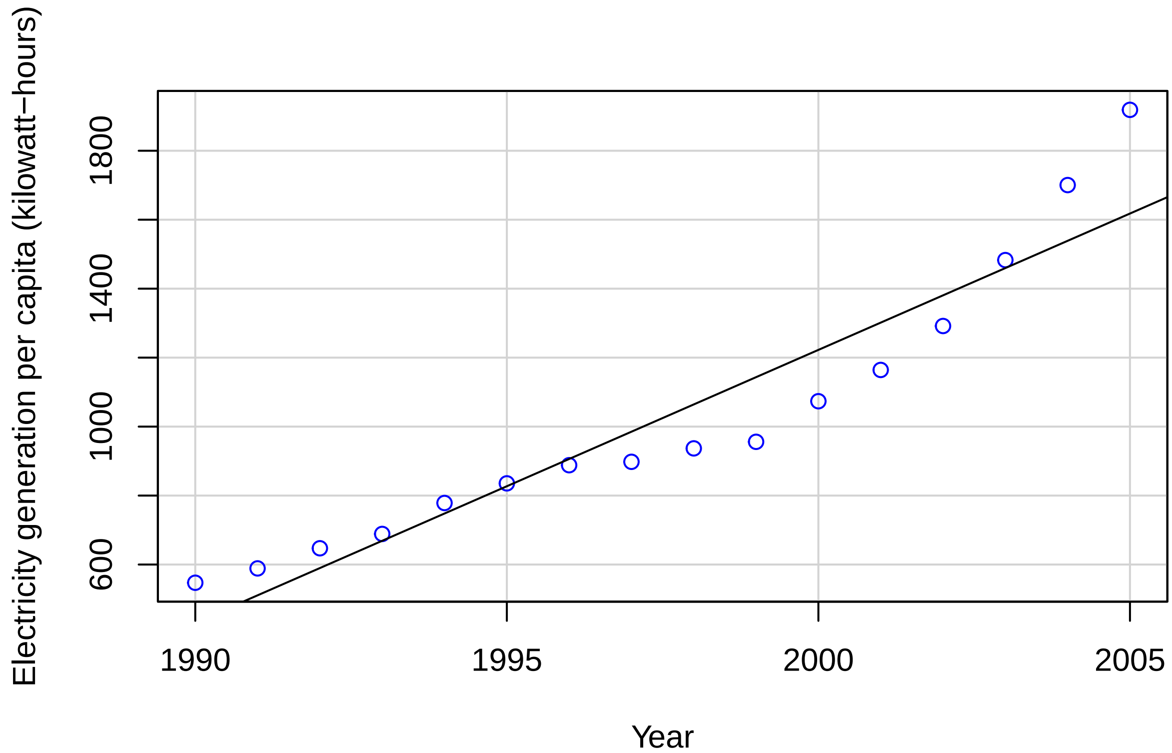



```
plot(chinaElectricity$Year, chinaElectricity$China, xlab = "Year",
     ylab = "Electricity generation per capita (kilowatt-hours)")
```




### Scatterplot

```
scatterplot(chinaElectricity$China~chinaElectricity$Year, regLine=FALSE,
            smooth=FALSE, boxplots=FALSE, xlab = "Year",
            ylab = "Electricity generation per capita (kilowatt-hours)")
abline(Reg.Model.1)
```




Does the relationship look linear? Yes, there is linear increase in the electricity generation per capita from 1990 to 2005 in China.

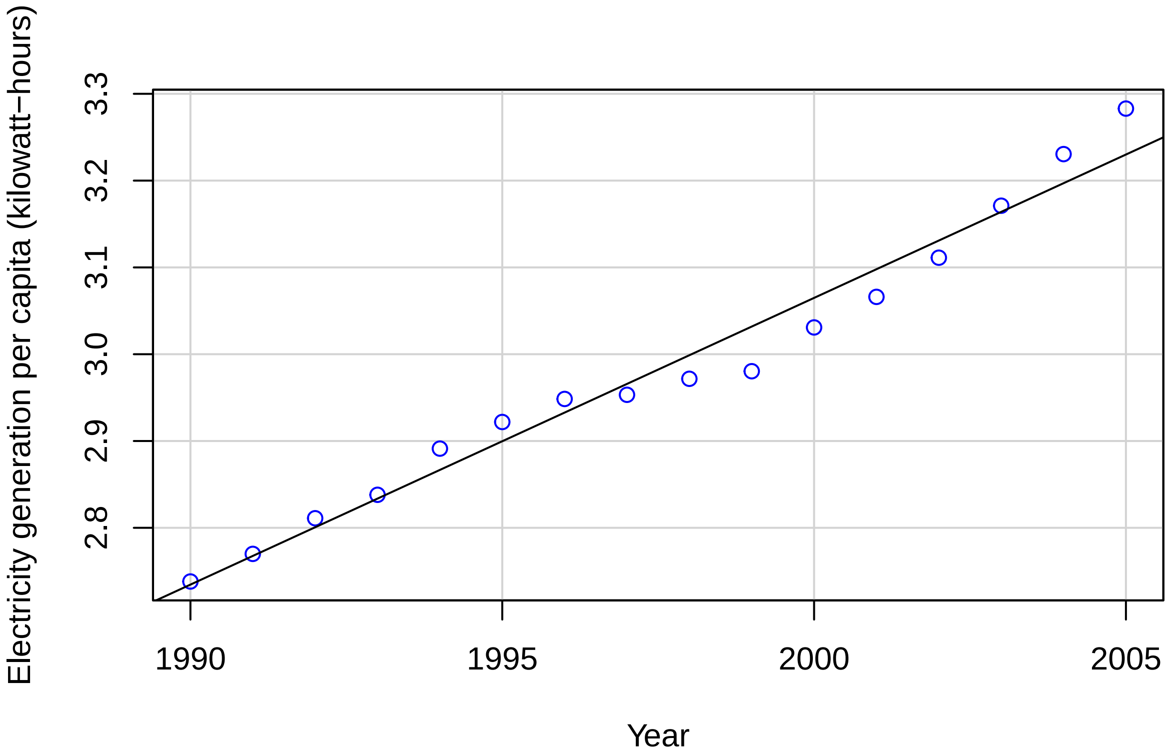
### log10 transformation

```
Reg.Model.2 <-lm(log10(China)~Year, data=chinaElectricity)
summary(Reg.Model.2)
```

```
##
## Call:
## lm(formula = log10(China) ~ Year, data = chinaElectricity)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.051445 -0.021607  0.003993  0.017326  0.052931
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -63.000976   3.114238  -20.23 9.21e-12 ***
## Year          0.033033   0.001559   21.19 4.91e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02875 on 14 degrees of freedom
```

```
## Multiple R-squared:  0.9698, Adjusted R-squared:  0.9676
## F-statistic: 448.9 on 1 and 14 DF,  p-value: 4.913e-12
```

```
scatterplot(chinaElectricity$Year, log10(chinaElectricity$China), regLine=FALSE,
            smooth=FALSE, boxplots=FALSE, xlab = "Year",
            ylab = "Electricity generation per capita (kilowatt-hours)")
abline(Reg.Model.2)
```




```
#plot(Reg.Model.2)
```

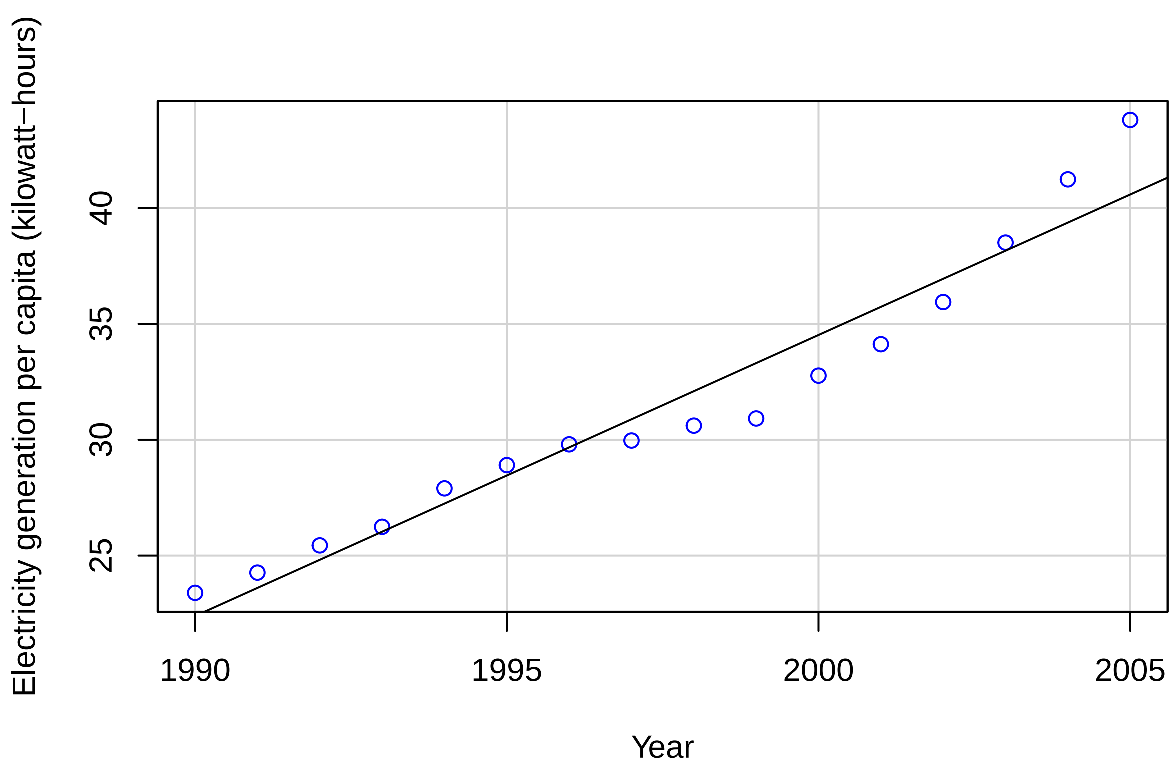
- r-squared: 0.9675968
- b (regression coefficient): -63.0009764
- SEb: 0.0015591
- t: 21.1876694
- df: 2, 14, 2
- p: $4.9127545 \times 10^{-12}$

### Square root transformation

```
Reg.Model.3 <-lm(sqrt(China)~Year, data=chinaElectricity)
summary(Reg.Model.3)
```

```
##
## Call:
## lm(formula = sqrt(China) ~ Year, data = chinaElectricity)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3899 -1.1247  0.2812  0.6577  3.2179
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.391e+03  1.636e+02  -14.62 7.16e-10 ***
## Year         1.213e+00  8.188e-02   14.81 6.02e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.51 on 14 degrees of freedom
## Multiple R-squared:   0.94,  Adjusted R-squared:  0.9357
## F-statistic: 219.4 on 1 and 14 DF,  p-value: 6.02e-10
```

```r
scatterplot(chinaElectricity$Year, sqrt(chinaElectricity$China), regLine=FALSE,
            smooth=FALSE, boxplots=FALSE, xlab = "Year",
            ylab = "Electricity generation per capita (kilowatt-hours)")
abline(Reg.Model.3)
```



```r
#plot(Reg.Model.3)
```

- r-squared: 0.9357204
- b (regression coefficient): -2390.9385343
- SEb: 0.0818822
- t: 14.8106541
- df: 2, 14, 2
- p: $6.0202529 \times 10^{-10}$

## Comparision of different models

- r-squared(linear): 0.887505
- r-squared(log10): 0.9675968
- r-squared (sqrt): 0.9357204

From above value, the r-squared(log10) is the highest (i.e. 0.9675968) and therefore this regression model explains the data optimally. Also see below (mtable):

```r
mtable(Reg.Model.1, Reg.Model.2, Reg.Model.3) # Ref 2
```

```
##
## Calls:
## Reg.Model.1: lm(formula = China ~ Year, data = chinaElectricity)
## Reg.Model.2: lm(formula = log10(China) ~ Year, data = chinaElectricity)
## Reg.Model.3: lm(formula = sqrt(China) ~ Year, data = chinaElectricity)
##
## ============================================================
##                 Reg.Model.1    Reg.Model.2   Reg.Model.3
##               --------------  ------------  ------------
##                    China       log10(China)  sqrt(China)
## ------------------------------------------------------------
##   (Intercept)  -156944.684***   -63.001***   -2390.939***
##                 (14460.479)      (3.114)      (163.560)
##   Year              79.084***     0.033***       1.213***
##                     (7.239)      (0.002)        (0.082)
## ------------------------------------------------------------
##   R-squared          0.895        0.970          0.940
##   N                 16           16             16
## ============================================================
##   Significance: *** = p < 0.001; ** = p < 0.01;
##                 * = p < 0.05
```

## References


1. Electricity Generation per capita.xlxs from Gapminder (https://www.gapminder.org/data/).
2. mtable from https://bookdown.org/josiesmith/labbook/bivariate-linear-regression.html

## Part 3. Testing differences between groups

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

### Introduction

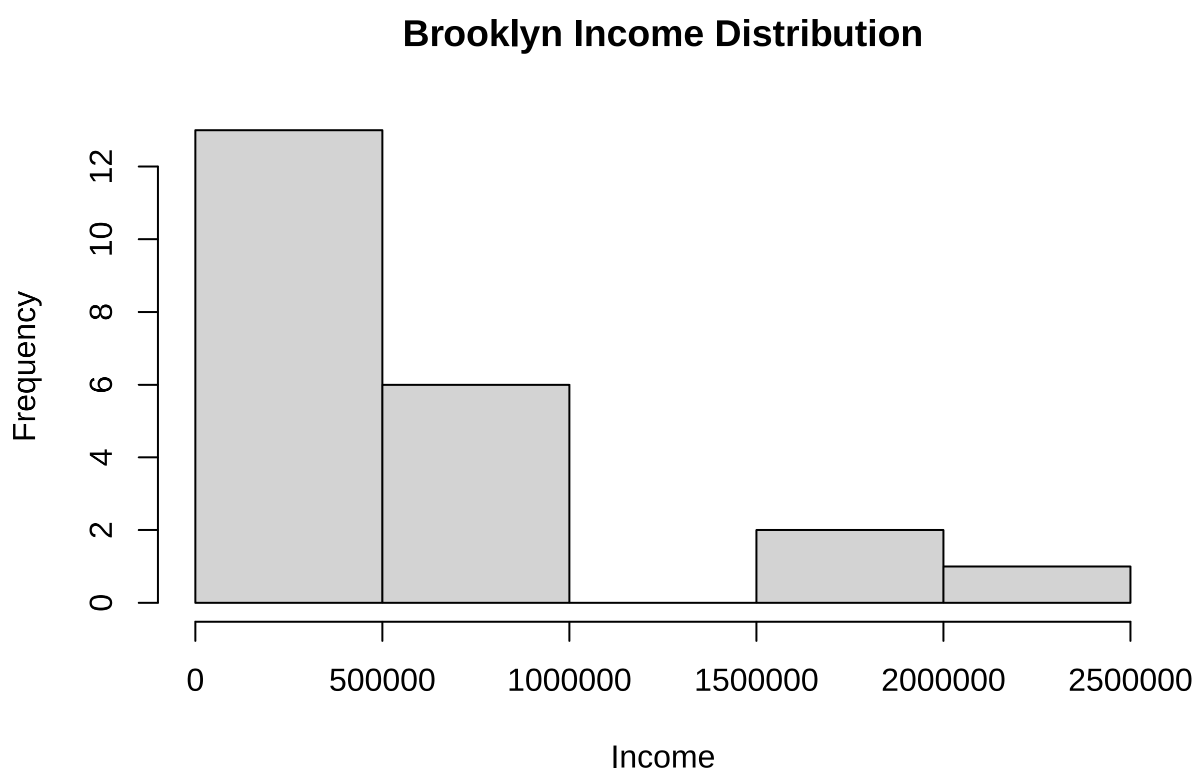
Here we analyse the data from Lander (2019) (Ref 1) to study the difference in the income between two districts in New York i.e. Manhattan and Brooklyn.

### Histogram

### Histogram by cities

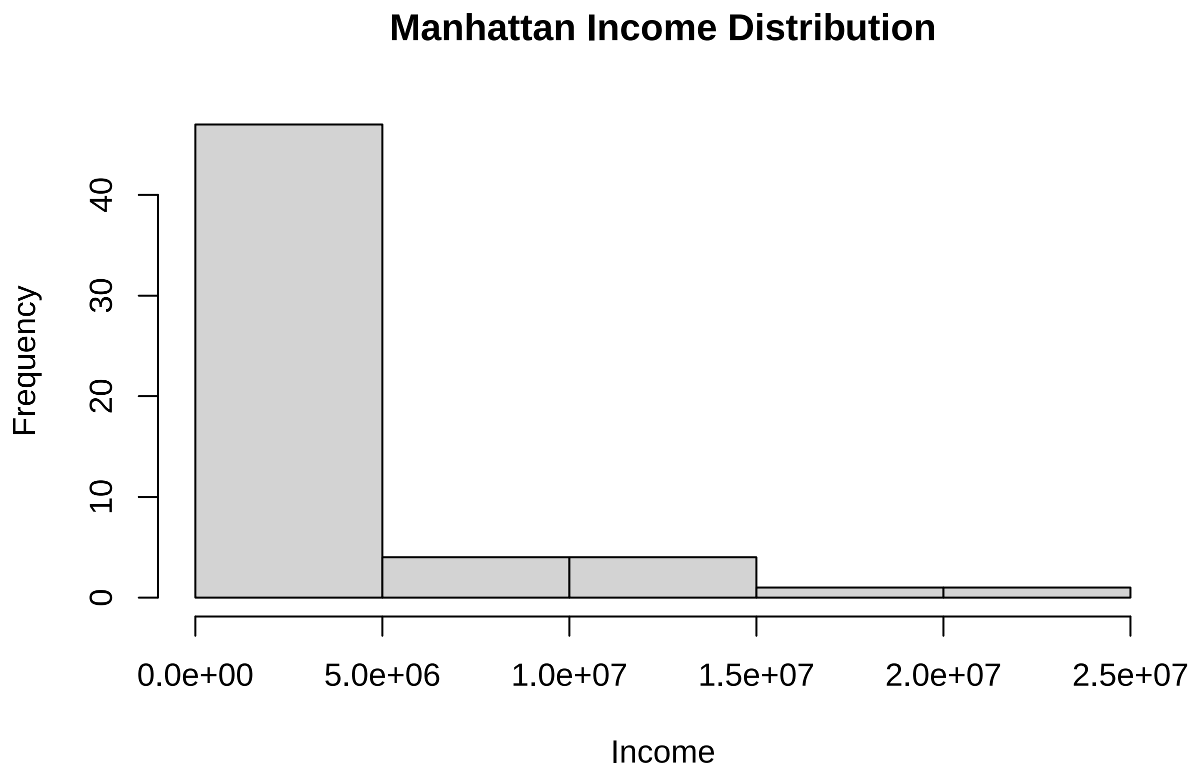
Brooklyn

```
hist(subset(LanderHousingNew, Boro == "Brooklyn")$Income,
  main = "Brooklyn Income Distribution",
  xlab = "Income")
```




Manhattan

```
hist(subset(LanderHousingNew, Boro == "Manhattan")$Income,
  main = "Manhattan Income Distribution",
  xlab = "Income")
```




Which test is most appropriate to use? The LanderHousingNew Income data is not normally distributed data, it is left skewed and therefore, non-parametric test Wilcoxon's rank-sum test would be used in this case.

This is also supprted by shapiro.test.

For Brooklyn

```
shapiro.test(subset(LanderHousingNew, Boro == "Brooklyn")$Income)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  subset(LanderHousingNew, Boro == "Brooklyn")$Income
## W = 0.7567, p-value = 0.0001131
```

For Manhattan

```
shapiro.test(subset(LanderHousingNew, Boro == "Manhattan")$Income)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  subset(LanderHousingNew, Boro == "Manhattan")$Income
## W = 0.67163, p-value = 5.125e-10
```

Summary of the data

```
summary(LanderHousingNew)
```

```
##                    Neighborhood             Class        Units
##  LOWER EAST SIDE        : 6    R2-CONDOMINIUM: 9   Min.   :  5.00
##  WILLIAMSBURG-CENTRAL   : 6    R4-CONDOMINIUM:57   1st Qu.: 13.00
##  MIDTOWN EAST           : 5    R9-CONDOMINIUM:11   Median : 28.00
##  HARLEM-CENTRAL         : 4    RR-CONDOMINIUM: 2   Mean   : 62.61
##  TRIBECA                : 4                        3rd Qu.: 81.00
##  UPPER EAST SIDE (79-96): 4                        Max.   :372.00
##  (Other)                :50
##    YearBuilt         SqFt            Income          IncomePerSqFt
##  Min.   :1900   Min.   :  5700   Min.   :  147206   Min.   :12.51
##  1st Qu.:1917   1st Qu.: 21587   1st Qu.:  445324   1st Qu.:21.84
##  Median :1986   Median : 43065   Median : 1393233   Median :32.86
##  Mean   :1965   Mean   : 73678   Mean   : 2591555   Mean   :30.76
##  3rd Qu.:2004   3rd Qu.: 93001   3rd Qu.: 2496518   3rd Qu.:38.30
##  Max.   :2009   Max.   :512280   Max.   :22673513   Max.   :61.11
##  NA's   :1
##     Expense        ExpensePerSqFt    NetIncome           Value
##  Min.   :  51987   Min.   : 6.03   Min.   :   76001   Min.   :   536802
##  1st Qu.: 167086   1st Qu.: 8.16   1st Qu.:  274948   1st Qu.:  2002000
##  Median : 443851   Median :10.45   Median :  907440   Median :  6328000
##  Mean   : 801372   Mean   :10.20   Mean   : 1790183   Mean   : 13122118
##  3rd Qu.:1023117   3rd Qu.:11.39   3rd Qu.: 1652922   3rd Qu.: 12163499
##  Max.   :5609466   Max.   :17.45   Max.   :17064047   Max.   :124320032
##
##   ValuePerSqFt           Boro
##  Min.   : 37.40   Brooklyn :22
##  1st Qu.: 92.01   Manhattan:57
##  Median :160.47
##  Mean   :151.13
##  3rd Qu.:207.76
##  Max.   :366.64
##
```

## Wilcoxon rank-sum test

```
wilcox.test(LanderHousingNew$Income~LanderHousingNew$Boro)
```

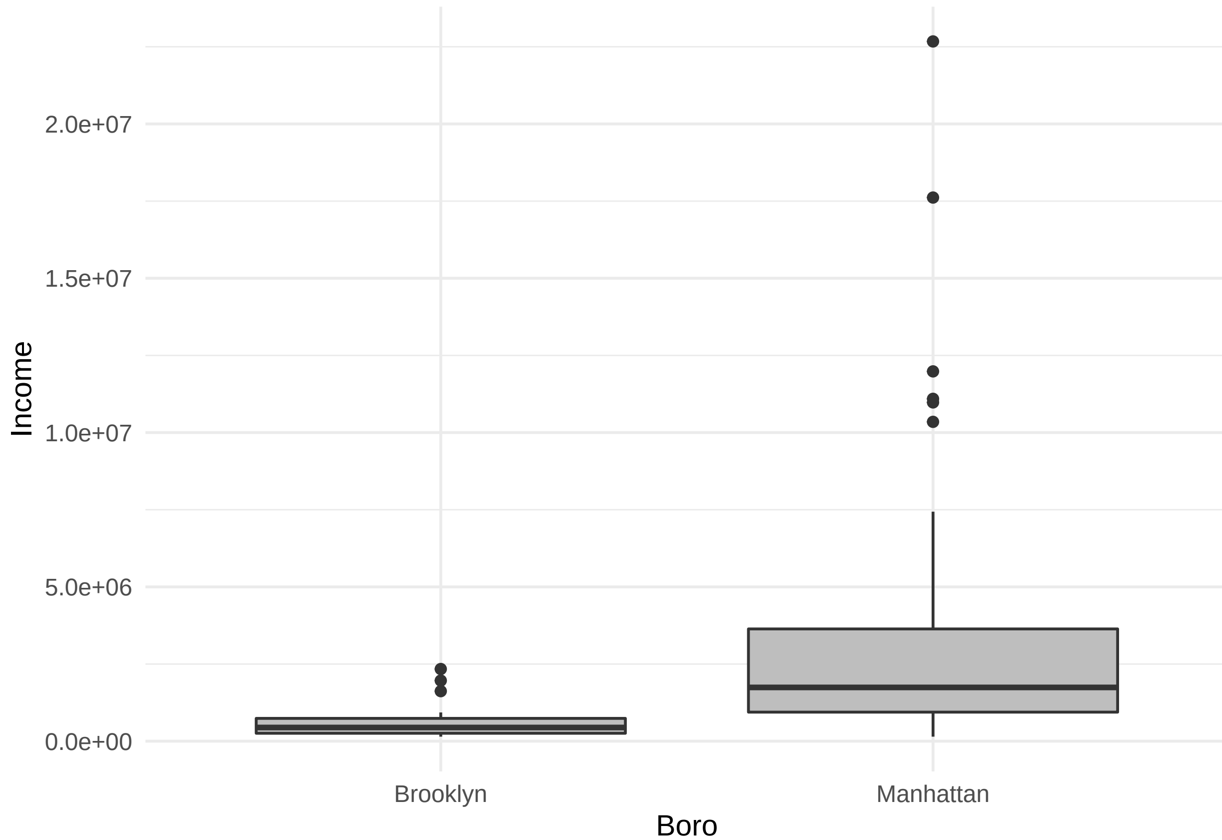
```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  LanderHousingNew$Income by LanderHousingNew$Boro
## W = 223, p-value = 1.019e-05
## alternative hypothesis: true location shift is not equal to 0
```

Since, the p-value $1.0191702 \times 10^{-5}$ is less then than the significance level 0.05, it is therefore concluded that the difference in the Income between the two group i.e. Brooklyn and Manhattan is significant. W = 223.

## Distribution by cities

Graph representing differences among the districts.

```
ggplot(LanderHousingNew) +
  aes(x = Boro, y = Income) +
  geom_boxplot(fill = "grey") +
  theme_minimal()
```




Above plot shows that income at Manhattan is higher than the income at Brooklyn.

What are the measured central tendencies for income in the two districts?

```
LanderHousingNew %>%
  group_by(Boro) %>%
  summarise("n" = length(Income), "Mean" = mean(Income), "Median" = median(Income),
            "Mode" = Mode(Income), "SD" = sd(Income)) %>%
  kbl(caption = "Measured central tendencies for income in the two districts") %>%
  kable_minimal()
```

## References


1. Lander (2019) (https://www.jaredlander.com/datasets/)

Table 1: Measured central tendencies for income in the two districts

| Boro | n | Mean | Median | Mode | SD |
|---|---|---|---|---|---|
| Brooklyn | 22 | 639548.1 | 443850.5 | 147206 | 598728.4 |
| Manhattan | 57 | 3344961.0 | 1742515.0 | 147229 | 4345456.9 |

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

## Introduction

Here we analyse the data from "Svensk mäklarstatistik" (Ref 1) to study the difference in housing pricing in Sweden between 2016 and 2017.

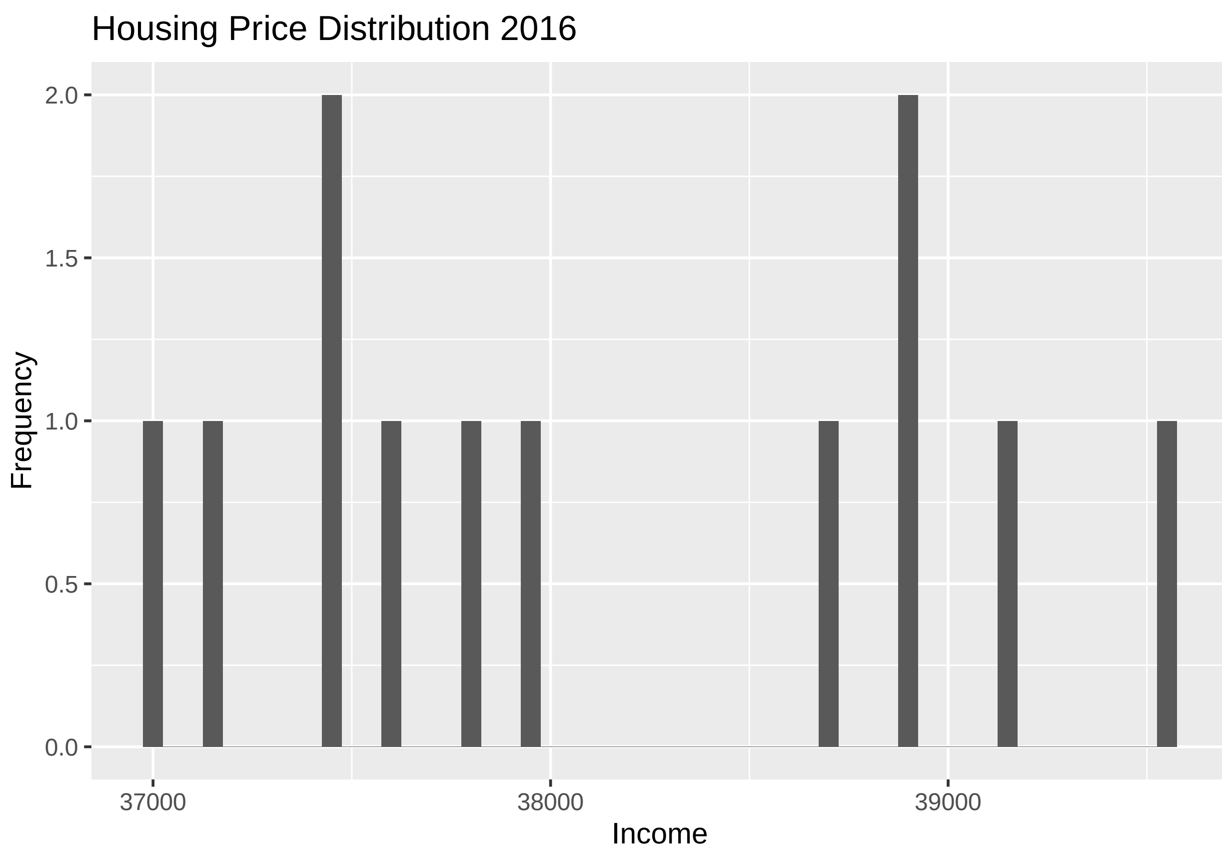
## Methods and Results

## Histogram

## Histogram by year

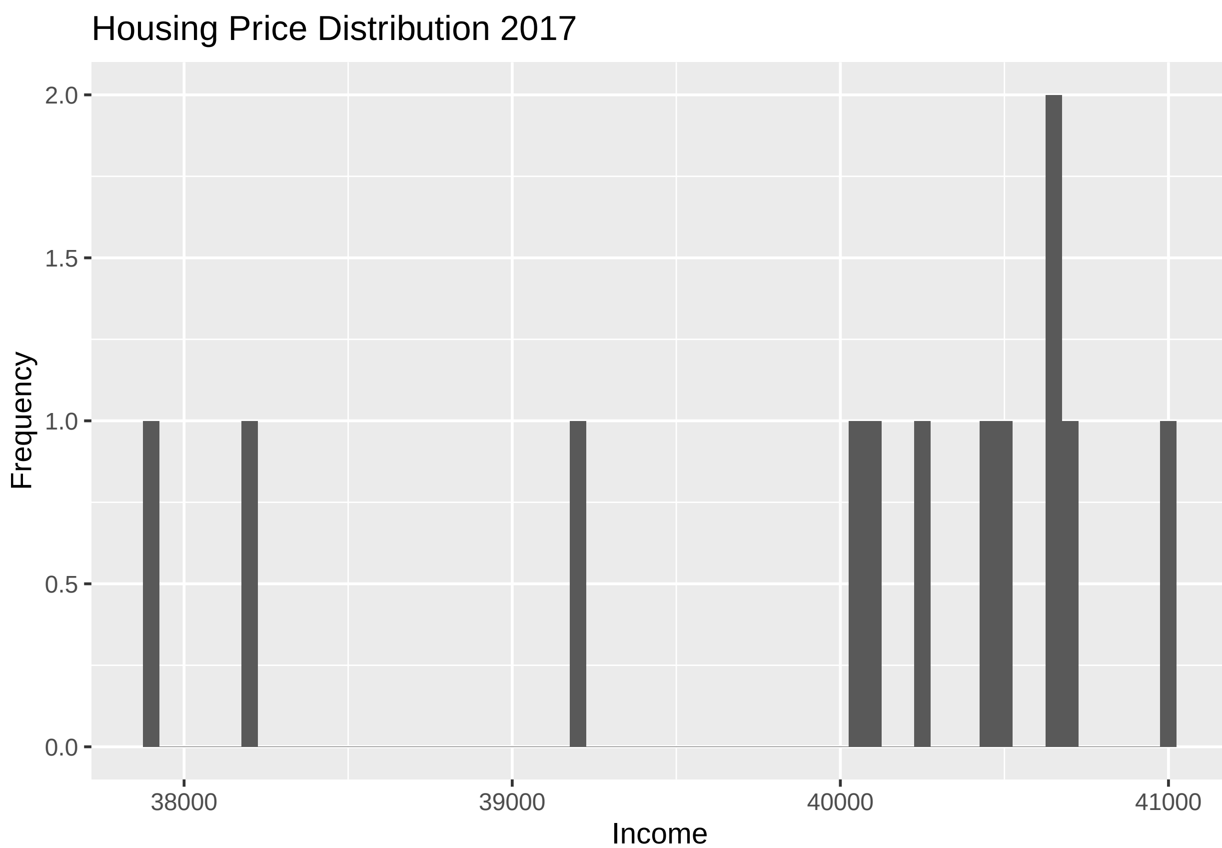
Year 2016: The data is normally distributed as apparent from the histogram below.

```
ggplot(Housepricing_sweden, aes(x=X2016_sek_sqrm)) +
  geom_histogram(binwidth = 50) +
  labs(title="Housing Price Distribution 2016", x="Income", y="Frequency")
```




Year 2017: The data is not-normally distributed (skewed).

```
ggplot(Housepricing_sweden, aes(x=X2017_sek_sqrm)) +
  geom_histogram(binwidth = 50) +
  labs(title="Housing Price Distribution 2017", x="Income", y="Frequency")
```




On an average, the there is increase in minimum, mean and maximum house pricing in 2017 as compare to 2016.

```
summary(Housepricing_sweden)
```

```
##      Month   X2016_sek_sqrm  X2017_sek_sqrm
##  apr    :1   Min.   :36982   Min.   :37924
##  aug    :1   1st Qu.:37458   1st Qu.:39850
##  dec    :1   Median :37873   Median :40346
##  feb    :1   Mean   :38129   Mean   :39977
##  jan    :1   3rd Qu.:38909   3rd Qu.:40643
##  july   :1   Max.   :39551   Max.   :41006
##  (Other):6
```

```
sd(Housepricing_sweden$X2016_sek_sqrm)
```

```
## [1] 862.8587
```

```
sd(Housepricing_sweden$X2017_sek_sqrm)
```

```
## [1] 1001.695
```

```r
shapiro.test(Housepricing_sweden$X2016_sek_sqrm)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Housepricing_sweden$X2016_sek_sqrm
## W = 0.9216, p-value = 0.2994
```

```r
shapiro.test(Housepricing_sweden$X2017_sek_sqrm)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Housepricing_sweden$X2017_sek_sqrm
## W = 0.81025, p-value = 0.01228
```

The two data set from 2016 and 2017 are not identical with (V = 6, p-value 0.006836)

```r
wilcox.test(Housepricing_sweden$X2016_sek_sqrm, Housepricing_sweden$X2017_sek_sqrm,
            paired=TRUE)
```

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

Paired samples test

```r
t.test(Housepricing_sweden$X2017_sek_sqrm, Housepricing_sweden$X2016_sek_sqrm,
       paired = TRUE, alternative = "two.sided")
```

```
##
##  Paired t-test
##
## data:  Housepricing_sweden$X2017_sek_sqrm and Housepricing_sweden$X2016_sek_sqrm
## t = 3.8116, df = 11, p-value = 0.002885
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   780.8864 2915.1136
## sample estimates:
## mean of the differences
##                    1848
```

## Discussion

The p-value (0.002885) of the test is less than the p-value (0.05). Therefore, we can reject the null hypothesis. There is a significant increase in the housing prices between 2016 and 2017 in Sweden ($t = 3.8116$, $df = 11$, p-value = 0.002885).

## References


1. "Svensk mäklarstatistik" https://www.maklarstatistik.se/