# Analysis in R

**How will you read the data into R for analysis?**

**Dataset Sources:** Google Trend Data is information gleaned from a Wikipedia page. The data in Trends is a random sample of our Google search data. It's anonymized (no one's identity is revealed), classified (the subject of a search query is determined), and aggregated (grouped together). Google Trends is a Google website that analyzes the popularity of top Google Search queries in different regions. We connect gtrends_TinderPremium.csv data with another source of data that is Tinder_IAPrM.csv.

**Read library**

```
library(dplyr)
```

**Read datasets**

```
df <- read.csv("gtrends_TinderPremium.csv", skip = 2)
df2 <- read.csv("Tinder_IAPrM.csv", skip = 2)
```

**Pre-Precess** datasets for merge it into one dataset.

```
df$Months <- format(as.Date(df$Week, format="%Y-%m-%d"), "%Y-%m")

df <- df %>%
  group_by(Months) %>%
  summarize(`Tinder Premium` = sum(Tinder.Premium...Worldwide.))
df2 <- df2 %>% rename(Months = X)
```

**Merge dataset** using merge function and then remove null value from our dataset.

```
df3 <- merge(x=df, y=df2, by="Months")
data <- na.omit(df3)
head(data)
```

```
##    Months Tinder Premium    APAC      EMEA      NALA       SUM
## 1 2015-11             69 664,444 1,894,438 3,967,859 6,526,741
## 2 2015-12             28 729,850 1,912,092 3,676,738 6,318,680
## 3 2016-01             49 741,240 1,997,469 3,659,085 6,397,794
## 4 2016-02             16 798,023 2,207,649 3,950,965 6,956,637
## 5 2016-03             50 922,720 2,390,147 4,831,493 8,144,360
## 6 2016-04             65 970,083 2,494,149 5,253,195 8,717,427
```

Select dependent and independent **variables** from final dataset.

```
x<-as.numeric(as.factor(data$`Tinder Premium`))
x

##  [1] 14  3  9  1 10 13  2  7 11  8  8 19  5  6 12 12  4 22 18 20 26 17 23
30 16
## [26] 25 15 29 20 28 31 33 48 21 38 24 32 41 27 36 37 39 40 52 57 34 42 35
45 50
## [51] 44 43 56 52 53 51 51 49 46 54 55 47

y<-as.numeric(as.factor(data$SUM))
y

##  [1] 49 47 48 50 60 61 62  1  2  3  4  5  6  7  8  9 11 10 12 13 14 15 16
18 17
## [26] 19 21 20 27 22 24 26 30 39 35 34 32 43 59 44 58 56 46 28 31 37 33 45
41 38
## [51] 52 40 42 23 25 29 36 55 53 57 54 51
```

**Methods and Results:**

**What statical method will you use to analyses the data?**

I will use correlation and regression to analyse the data and seek the correlation between variables. To conduct a regression analysis, the following assumptions will be applied:

X: The number of searches for “Tinder Premium”

Y: SUM from dataset.

1. There are signification correlation between X and Y
2. The underlying data is normally distributed.

**Which R functions will you use to carry out the statistical analysis?**

Calculate Pearson's correlation coefficient to test for significant correlation between two variables.

```
cor.test(x, y)
```

```
##
##  Pearson's product-moment correlation
##
## data:  x and y
## t = 3.0991, df = 60, p-value = 0.002954
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1341382 0.5684857
## sample estimates:
##       cor
## 0.3714611
```

**Interpreting result:**

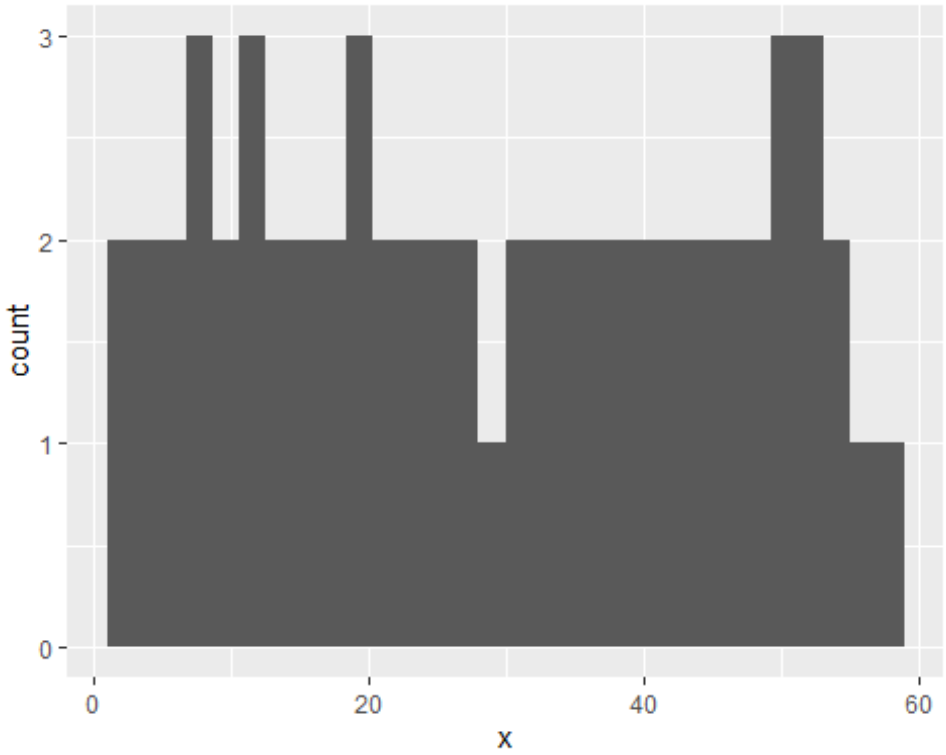
The correlation coefficient, abbreviated as r, is a measure of the strength of a linear or straight-line relationship between two variables. A positive linear relationship between variables is indicated by values between 0.37. Where sample size (degrees of freedom) is 60.

Testing for data **normality:**

```
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.0.4

ggplot(data=data, aes(x=x),xlab='Tinder Premium') + stat_bin(bins = 30)
```




**Approximately normal**

**Interpreting result**:
A perfectly smooth normal curve will be unlikely to emerge from a histogram of sample data, particularly if the sample size is small. A parametric test can be used if the data is roughly normally distributed, with a peak in the middle and is fairly symmetrical.

Check if the data is **normally distributed?**

```
shapiro.test(x)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  x
## W = 0.94745, p-value = 0.01008
```

**Result:** p-value is less then .05 that means its statistically significant. For normality tests, the Shapiro-method is commonly recommended because it has more power. It is based on a correlation between the data and the normal scores that correspond to it.

Calculate Kendall's tau

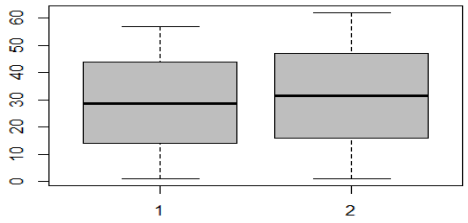
```
cor.test(x, y, method="kendall")
```

```
##
## Kendall's rank correlation tau
##
## data:  x and y
## z = 3.3046, p-value = 0.0009511
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##       tau
## 0.2880596
```

**Interpretation Result:**
tau is the Kendall correlation coefficient. The correlation coefficient between x and y are 0.2880596and the p-value is 0.0009511. Which indicates significant result.

**Boxplot for two variables:**

```
boxplot(x,y,col = "gray")
```




**Boxplot**

**Interpreting Result:** The line that separates the box into two sections shows the median (middle quartile), which is the mid-point of the data. Half of the scores are higher than or equal to this value, while the other half are lower. The middle "box" reflects the group's middle 50 percent of scores. Build a regression model.

Build a **regression model.**

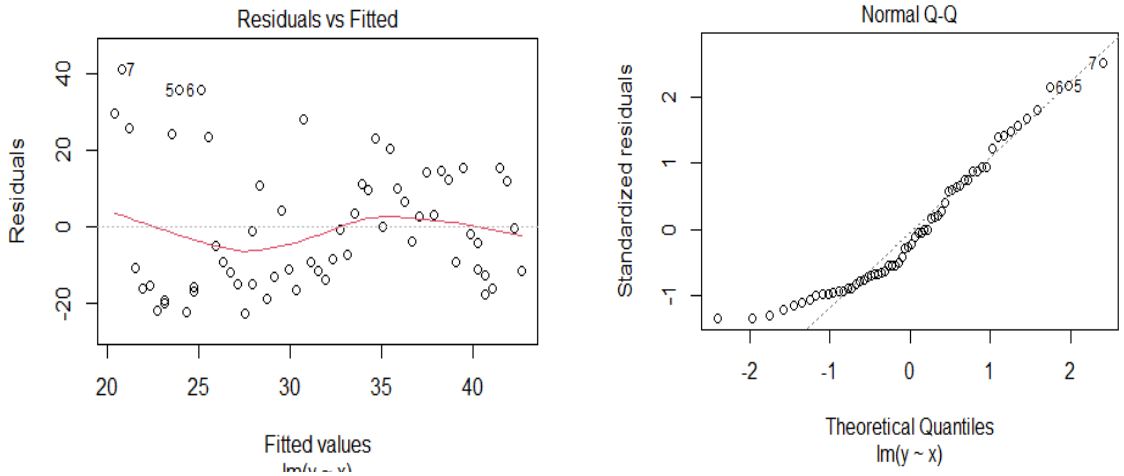
```
Model <- lm(y ~ x, data=data)
summary(Model)
```

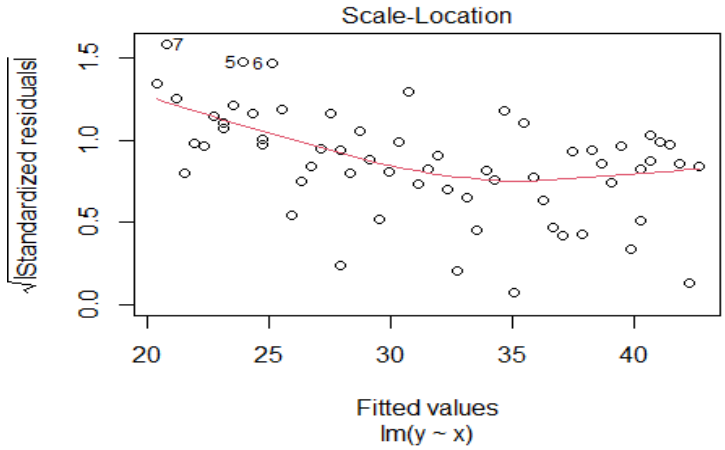
```
##
## Call:
## lm(formula = y ~ x, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.528 -13.714  -3.987  11.869  41.247
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  19.9562     4.2984   4.643 1.92e-05 ***
## x             0.3985     0.1286   3.099  0.00295 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.89 on 60 degrees of freedom
## Multiple R-squared:  0.138,  Adjusted R-squared:  0.1236
## F-statistic: 9.604 on 1 and 60 DF,  p-value: 0.002954
```

**INTERPRETING MODEL:** The lm() function in R was used to construct the model above, and the summary() function on the model was used to call the output. The intercept and slope terms in the model are expressed by the coefficients, which are two unknown constants. We would estimate the coefficients to use in the model formula if we wanted to predict the Y needed for X. In other words, 0.3985 is needed. The slope, or in our case, the second row in the Coefficients, is the second row. A small p-value suggests that a relationship between the predictor and response variables is unlikely to be observed. The p-values in our model example are very close to zero, indicating that the relationship between variables is important.

**Regression model plot:**

```
plot (Model)
```




**Interpretation of model Plot:** The most common plot generated during a residual analysis is a "**residuals versus fits plot.**" On the y axis are residuals, and on the x, axis is fitted values

(estimated responses). Non-linearity, unequal error variances, and outliers are all detected using this plot.

A scatterplot generated by plotting two sets of quantiles against each other is known as a Q-Q plot. If both sets of quantiles came from the same distribution, the points should form an approximately straight line. When both sets of quantiles actually come from Normal distributions, this is an example of a **Normal Q-Q plot.**

And then **scale-location diagram** shows the fitted values of a regression model on the x-axis and the square root of the standardized residuals on the y-axis.

**The End**