Statistics

2023-12-26

Agriculture is one of the key development in human existence, farming helps to feed billions of people all over the world, this has partly been possible because of extensive research involving how crops are grown. Soil PH and and the factors involving soil management are very important. The dataset used in this report contains observations about soil PH, the type of crops being grown on them, the mineral content Potassium, Nitrogen and Phosphorous and also humidity and temperature. I will attempt to carry out statistical analysis on the dataset, to show relationships between the variables and how they relate with one another.

I will be using a confidence level of 95% indicating a p-value of 0.05.

Loading and viewing the structure of the dataset.

crop\_data <- read.csv("Crop\_recommendation.csv")  
  
str(crop\_data)

## 'data.frame': 2200 obs. of 8 variables:  
## $ N : int 90 85 60 74 78 69 69 94 89 68 ...  
## $ P : int 42 58 55 35 42 37 55 53 54 58 ...  
## $ K : int 43 41 44 40 42 42 38 40 38 38 ...  
## $ temperature : num 20.9 21.8 23 26.5 20.1 ...  
## $ humidity : num 82 80.3 82.3 80.2 81.6 ...  
## $ ph : num 6.5 7.04 7.84 6.98 7.63 ...  
## $ rainfall : num 203 227 264 243 263 ...  
## $ crop\_suitable\_to\_grow: chr "rice" "rice" "rice" "rice" ...

The dataset contains 2200 obsevations and 8 columns N - This is the ratio of Nitrogen content in the soil P - This is the ratio of Phosphorous in the soil K - This is the ration of Potassium in the soil temperature - Weather temperature in degree celcius Humidity - This is the relative humidity in % PH - This is the ph value of the soil Rainfall - is the average rainfall measured in mm Crop suitable to grow - This is the ideal crop which would grow under certain conditions. (Kaggle, 2021)

Running the head command to see the first 6 rows of the dataset.

head(crop\_data)

## N P K temperature humidity ph rainfall crop\_suitable\_to\_grow  
## 1 90 42 43 20.87974 82.00274 6.502985 202.9355 rice  
## 2 85 58 41 21.77046 80.31964 7.038096 226.6555 rice  
## 3 60 55 44 23.00446 82.32076 7.840207 263.9642 rice  
## 4 74 35 40 26.49110 80.15836 6.980401 242.8640 rice  
## 5 78 42 42 20.13017 81.60487 7.628473 262.7173 rice  
## 6 69 37 42 23.05805 83.37012 7.073454 251.0550 rice

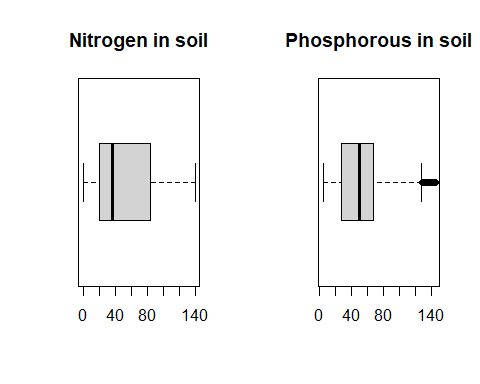
The summary command will give me an idea of how spread out the numerical columns are. I can infer from the summary that temperature and ph are normally distributed, because the mean and the medium do not differ so much from each other. I will however need to confirm this graphically by plotting both a boxplot and also a qqnorm plot.

summary\_stats <- summary(crop\_data)  
print(summary\_stats)

## N P K temperature   
## Min. : 0.00 Min. : 5.00 Min. : 5.00 Min. : 8.826   
## 1st Qu.: 21.00 1st Qu.: 28.00 1st Qu.: 20.00 1st Qu.:22.769   
## Median : 37.00 Median : 51.00 Median : 32.00 Median :25.599   
## Mean : 50.55 Mean : 53.36 Mean : 48.15 Mean :25.616   
## 3rd Qu.: 84.25 3rd Qu.: 68.00 3rd Qu.: 49.00 3rd Qu.:28.562   
## Max. :140.00 Max. :145.00 Max. :205.00 Max. :43.675   
## humidity ph rainfall crop\_suitable\_to\_grow  
## Min. :14.26 Min. :3.505 Min. : 20.21 Length:2200   
## 1st Qu.:60.26 1st Qu.:5.972 1st Qu.: 64.55 Class :character   
## Median :80.47 Median :6.425 Median : 94.87 Mode :character   
## Mean :71.48 Mean :6.469 Mean :103.46   
## 3rd Qu.:89.95 3rd Qu.:6.924 3rd Qu.:124.27   
## Max. :99.98 Max. :9.935 Max. :298.56

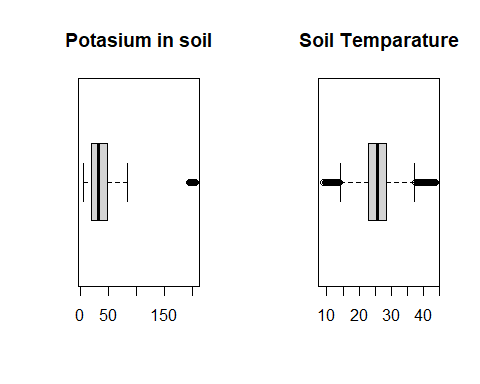
Ploting a boxplot of Nitrogen and Phosphorous using the par(mfrow) to split the screen and plot it side by side. We can see from the Nitrogen boxplot the data is skewed to the right and the same with Phosphorous is also skewed to the right but also having outliers to the right of the data.

par(mfrow = c(1, 2))  
boxplot(crop\_data$N, main = "Nitrogen in soil", horizontal = TRUE)  
boxplot(crop\_data$P, main = "Phosphorous in soil", horizontal = TRUE)



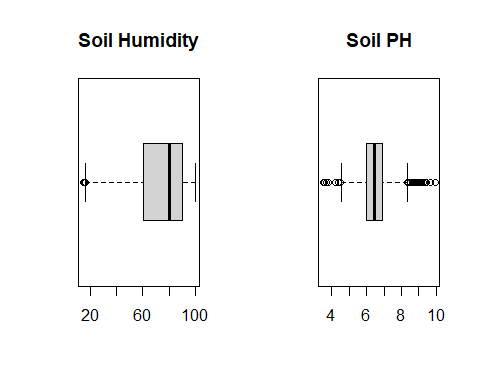
The Potasium observations has data concentrated to the left and outliers to the right, the temperature observation has its median in the middle as reported earlier but also has outliers to both left and right of the observation, this can be assumed to be a normal distribution. I will further confirm this assumption by plotting both a histogram and a qqplot and examining the distribution.

par(mfrow = c(1, 2))  
boxplot(crop\_data$K, main = "Potasium in soil", horizontal = TRUE)  
boxplot(crop\_data$temperature, main = "Soil Temparature", horizontal = TRUE)



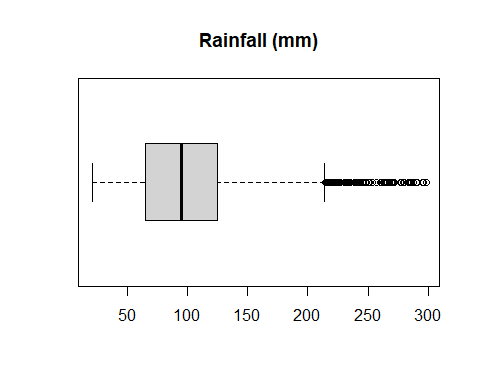
The soil observation is left skewed with outliers as well on the left of the data, The PH observations has its median in the middle of the data with outliers on both the left and right.

par(mfrow = c(1, 2))  
boxplot(crop\_data$humidity, main = "Soil Humidity", horizontal = TRUE)  
boxplot(crop\_data$ph, main = "Soil PH", horizontal = TRUE)



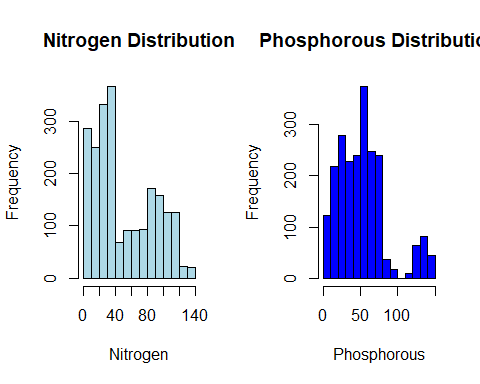
The rainfall observation is skewed to the right and has outliers also on the right.

boxplot(crop\_data$rainfall, main = "Rainfall (mm)", horizontal = TRUE)



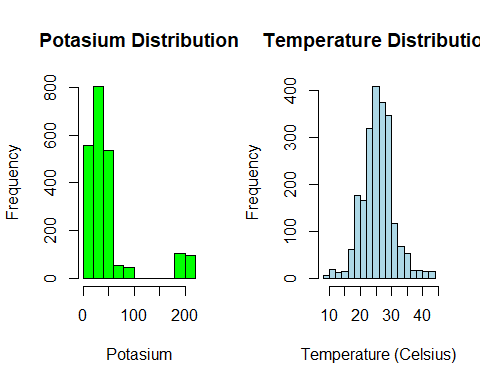
Plotting a histogram of both Nitrogen and Phosphorous, as was observed in the box plot, the data is right skewed.

par(mfrow = c(1, 2))  
hist(crop\_data$N, main = "Nitrogen Distribution", xlab = "Nitrogen", col = "lightblue", border = "black")  
hist(crop\_data$P, main = "Phosphorous Distribution", xlab = "Phosphorous", col = "blue", border = "black")



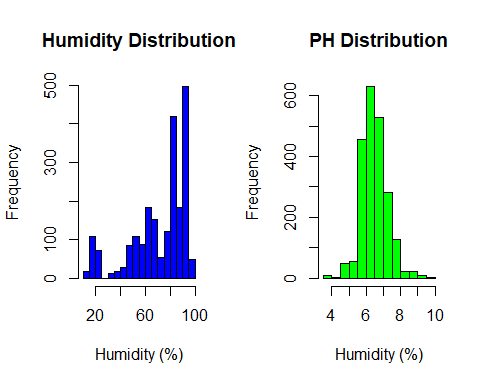
Potasium as well plotting the histogram shows that it is right skewed while the temperature follows a normal distribution curve with most of the data clustered in the middle

par(mfrow = c(1, 2))  
hist(crop\_data$K, main = "Potasium Distribution", xlab = "Potasium", col = "green", border = "black")  
hist(crop\_data$temperature, main = "Temperature Distribution", xlab = "Temperature (Celsius)", col = "lightblue", border = "black")



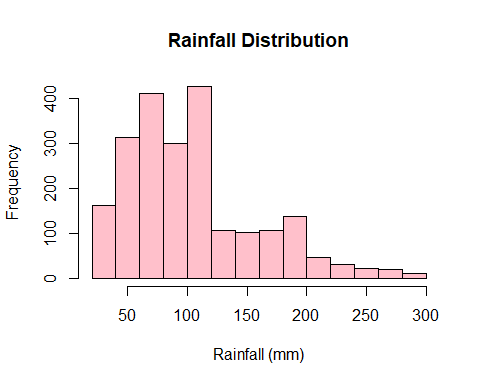
The humidity histogram shows that it is left skewed while PH follows a normal distribution.

par(mfrow = c(1, 2))  
hist(crop\_data$humidity, main = "Humidity Distribution", xlab = "Humidity (%)", col = "blue", border = "black")  
hist(crop\_data$ph, main = "PH Distribution", xlab = "Humidity (%)", col = "green", border = "black")



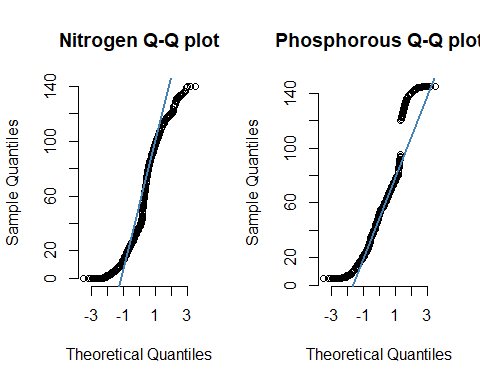
The rainfall data is skewed to the right as can be seen with the histogram.

hist(crop\_data$rainfall, main = "Rainfall Distribution", xlab = "Rainfall (mm)", col = "Pink", border = "black")



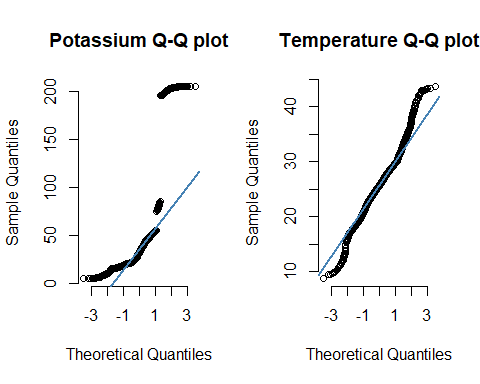
Plot both Nitrogen and Phosphorous on a QQ Plot and fitting a line through the data point we can conclude graphically that the line does not fit most of the data point well and can say that the data is not normally distributed

par(mfrow = c(1, 2))  
qqnorm(crop\_data$N, main = "Nitrogen Q-Q plot", pch = 1, frame = FALSE)  
qqline(crop\_data$N, col = "steelblue", lwd = 2)  
  
qqnorm(crop\_data$P, main = "Phosphorous Q-Q plot", pch = 1, frame = FALSE)  
qqline(crop\_data$P, col = "steelblue", lwd = 2)



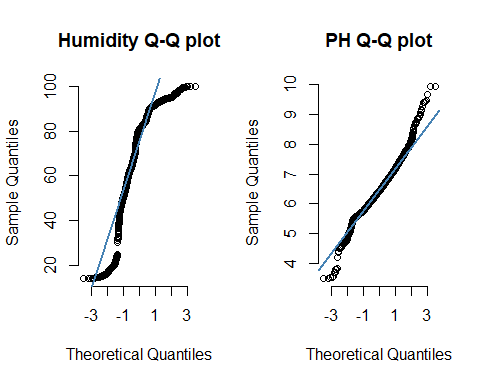
We can also conclude that the Potassium data does not follow a normal distribution as well, however for the temperature most of the data points fit on the line with the exception of outliers, we can conclude that it is approximately normal

par(mfrow = c(1, 2))  
qqnorm(crop\_data$K, main = "Potassium Q-Q plot", pch = 1, frame = FALSE)  
qqline(crop\_data$K, col = "steelblue", lwd = 2)  
  
qqnorm(crop\_data$temperature,main = "Temperature Q-Q plot", pch = 1, frame = FALSE)  
qqline(crop\_data$temperature, col = "steelblue", lwd = 2)

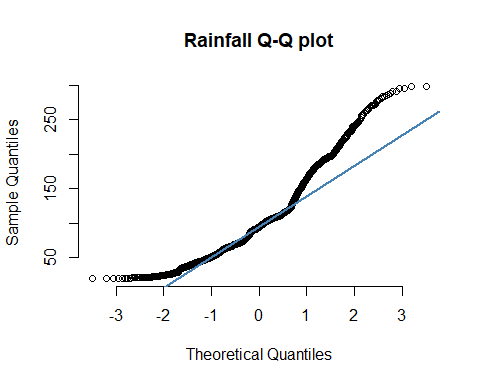


We can also conclude that the Humidity data does not follow a normal distribution as well, however for the PH most of the data points fit on the line with the exception of outliers, we can conclude that it is approximately normal.

par(mfrow = c(1, 2))  
qqnorm(crop\_data$humidity, main = "Humidity Q-Q plot", pch = 1, frame = FALSE)  
qqline(crop\_data$humidity, col = "steelblue", lwd = 2)  
  
qqnorm(crop\_data$ph, main = "PH Q-Q plot", pch = 1, frame = FALSE)  
qqline(crop\_data$ph, col = "steelblue", lwd = 2)

 Lastly we can also conclude from the QQ plot of rainfall that the data does not follow a normal distribution

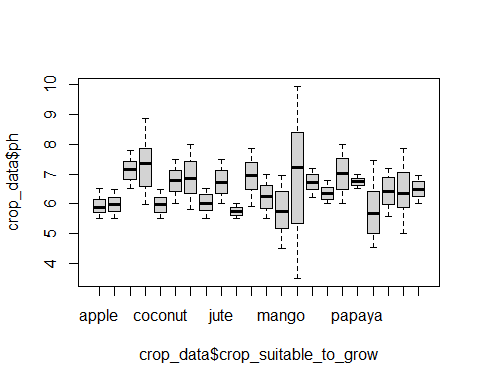
qqnorm(crop\_data$rainfall, main = "Rainfall Q-Q plot", pch = 1, frame = FALSE)  
qqline(crop\_data$rainfall, col = "steelblue", lwd = 2)



`

The PH of the soil is very important for plant growth, and when we plot the PH against the crop suitable to grow, we see the medians of crops spread across 5 and 7. According to Frontier Agriculture a PH of 6.5 is beneficial for most crops to thrive on. (Campbell, 2018) We will carry out statistical test confirm if our data aligns with this.

boxplot(crop\_data$ph ~ crop\_data$crop\_suitable\_to\_grow)



**Single sample t-test**

Hypothesis testing: H0: the average PH for crops to do well in is 6.5 H1: the average PH for crops to do well is not 6.5

The observation from our sample data is that the average PH level for crops is 6.46. The t.test will prove if this is statistically number significant.

ph\_soil.ttest <- t.test(x=crop\_data$ph, mu=6.5)  
  
ph\_soil.ttest

##   
## One Sample t-test  
##   
## data: crop\_data$ph  
## t = -1.8496, df = 2199, p-value = 0.0645  
## alternative hypothesis: true mean is not equal to 6.5  
## 95 percent confidence interval:  
## 6.437122 6.501838  
## sample estimates:  
## mean of x   
## 6.46948

We fail to reject the Null hypothesis as the p-value is greater than 0.05 and conclude that the PH for crops to grow and thrive is 6.5. We can confirm this with a 95% confidence level.

**Two-sided test for difference of means**

The Food and Agricultural Organization of The United Nations classifies Maize and Rice under the Poaceae family. (ICC, 2010) In this two sided test, I want to check if the avearge PH levels needed to grow rice is the same PH levels needed to grow maize. I will carry out a two sided t.test to check the mean of the two PH levels.

We will be using the library tidy verse and I have attached the dataset so as to make writing the code in R easier when referencing the columns in the data

attach(crop\_data)

Here I am piping the crop\_data and filtering for only Maize and Rice as those are the two category I want to test for. Earlier, we confirmed graphically that the PH column follows a normal distribution so using a parametric t.test is appropriate.

H0: The mean of maize and rice are the same H1: The mean of maize is less than rice

crop\_data %>%   
 filter(crop\_suitable\_to\_grow %in% c("rice", "maize")) %>%   
 t.test(ph ~ crop\_suitable\_to\_grow, data = .,  
 alternative = "two.sided")

##   
## Welch Two Sample t-test  
##   
## data: ph by crop\_suitable\_to\_grow  
## t = -2.0663, df = 151.69, p-value = 0.0405  
## alternative hypothesis: true difference in means between group maize and group rice is not equal to 0  
## 95 percent confidence interval:  
## -0.352656182 -0.007906218  
## sample estimates:  
## mean in group maize mean in group rice   
## 6.245190 6.425471

Given a p-value = 0.0405, we reject the null hypothesis that there is no difference in the mean PH for rice and maize.

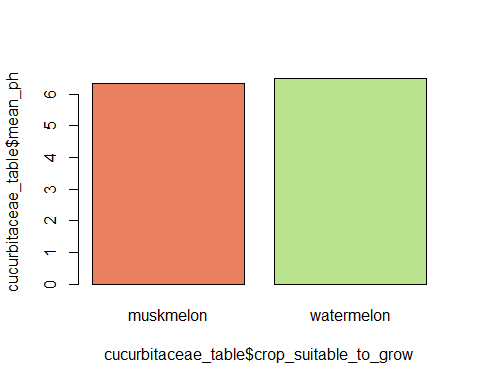
**One sided T.test for difference of means**

According to the Food and Agriculture Organization of the United Nations, Watermelon and muskmelon belong to the Cucurbitaceae family. (ICC, 2010)

As a farmer suppose we wants to plant both muskmelon and Watermelon and assume that the PH levels for both crop are the same, and wants to confirm if the PH level of muskmelon is less than that of Watermelon. We will use the One sided test to check if this difference is statistically significant with a confidence level of 95%.

Plotting the means on a bar plot, graphically it is quite difficult to say if there is a difference in the mean and if even the difference we observe is statistically significant.

cucurbitaceae\_table <- crop\_data %>%   
 filter(crop\_suitable\_to\_grow %in% c("watermelon", "muskmelon") ) %>%   
 group\_by(crop\_suitable\_to\_grow) %>%   
 summarise(mean\_ph = mean(ph)) %>%  
 select(mean\_ph,crop\_suitable\_to\_grow)  
  
  
 barplot(cucurbitaceae\_table$mean\_ph~cucurbitaceae\_table$crop\_suitable\_to\_grow, col = c("#eb8060", "#b9e38d"))



#arrange(mean\_ph)

crop\_data %>%   
 filter(crop\_suitable\_to\_grow %in% c("watermelon","muskmelon")) %>%   
 mutate(csg = factor(crop\_suitable\_to\_grow, levels = c("watermelon","muskmelon"))) %>%   
 t.test(ph ~ crop\_suitable\_to\_grow, data =.,  
 alternative = "less",  
 conf.level = 0.95)

##   
## Welch Two Sample t-test  
##   
## data: ph by crop\_suitable\_to\_grow  
## t = -3.7587, df = 190.91, p-value = 0.0001135  
## alternative hypothesis: true difference in means between group muskmelon and group watermelon is less than 0  
## 95 percent confidence interval:  
## -Inf -0.07673926  
## sample estimates:  
## mean in group muskmelon mean in group watermelon   
## 6.358805 6.495778

Having a p-value = 0.0001135 we have sufficient evidence to reject the null hypothesis that the mean PH levels are the same.

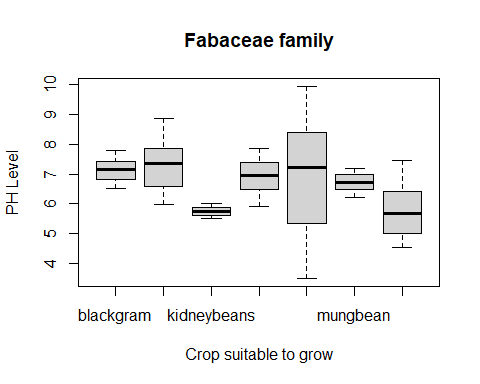
**Analysis of Variance**

The Fabaceae family which includes Chickpea, kidney beans, pigeon peas, moth beans, mung bean, black gram, and lentil has been filtered from the dataset. The aim here is to use ANOVA to test the difference in the soil mean of PH levels of these crop family.

fabaceae <- crop\_data %>%   
 filter(crop\_suitable\_to\_grow %in% c("pigeonpeas", "mothbeans","mungbean","blackgram","pigeonpeas","lentil", "chickpea","kidneybeans")) %>%   
 select(crop\_suitable\_to\_grow, ph)

The box plot of the Fabaceae family filter from the data shows the median clustered around PH soil value range of 6 -7.

boxplot(fabaceae$ph ~ as.factor(fabaceae$crop\_suitable\_to\_grow), xlab = "Crop suitable to grow", ylab = "PH Level", main = "Fabaceae family")

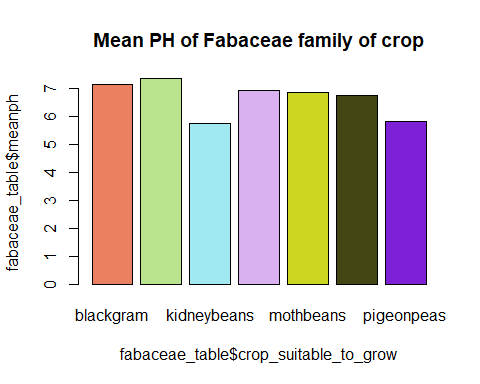


We can observe a difference of means from the Fabaceae family when we plot both a box plot and bar plot, however, is this difference we see statiscally significant and can we say with 95% confidence that the difference is significant. An ANOVA test will confirm this. We assume equal variane as the sample are coming from the same PH distribution

fabaceae %>%   
 group\_by(crop\_suitable\_to\_grow) %>%   
 summarise(meanph = mean(ph)) %>%   
 arrange(meanph)

## # A tibble: 7 × 2  
## crop\_suitable\_to\_grow meanph  
## <chr> <dbl>  
## 1 kidneybeans 5.75  
## 2 pigeonpeas 5.79  
## 3 mungbean 6.72  
## 4 mothbeans 6.83  
## 5 lentil 6.93  
## 6 blackgram 7.13  
## 7 chickpea 7.34

fabaceae\_table <- fabaceae %>%   
 group\_by(crop\_suitable\_to\_grow) %>%   
 summarise(meanph = mean(ph)) %>%   
 select(crop\_suitable\_to\_grow,meanph)   
   
 barplot(fabaceae\_table$meanph ~ fabaceae\_table$crop\_suitable\_to\_grow, main= "Mean PH of Fabaceae family of crop", col = c("#eb8060", "#b9e38d", "#a1e9f0", "#d9b1f0","#cad620","#444713","#7e20d6"))

 H0: The mean across the group of Fabaceae family are the same H1: The mean across the group of Fabaceae family are not the same

fabaceae %>%   
 aov(ph ~ crop\_suitable\_to\_grow, data =.) %>%   
 summary()

## Df Sum Sq Mean Sq F value Pr(>F)   
## crop\_suitable\_to\_grow 6 236.5 39.41 51.64 <2e-16 \*\*\*  
## Residuals 693 528.9 0.76   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Having a p-value = 2e-16 we can reject the null hypothesis that the mean are the same

Although we have concluded that there is at least a difference in the mean, we however do not know which crop differs in mean to the other. Using the TukeyHSD we can show where the difference occur

fabaceae %>%   
 aov(ph ~ crop\_suitable\_to\_grow, data =.) %>%   
 TukeyHSD()

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = ph ~ crop\_suitable\_to\_grow, data = .)  
##   
## $crop\_suitable\_to\_grow  
## diff lwr upr p adj  
## chickpea-blackgram 0.20300499 -0.1623314 0.56834138 0.6542447  
## kidneybeans-blackgram -1.38454104 -1.7498774 -1.01920466 0.0000000  
## lentil-blackgram -0.20602006 -0.5713564 0.15931633 0.6380536  
## mothbeans-blackgram -0.30277755 -0.6681139 0.06255884 0.1793794  
## mungbean-blackgram -0.40999469 -0.7753311 -0.04465830 0.0164927  
## pigeonpeas-blackgram -1.33977675 -1.7051131 -0.97444036 0.0000000  
## kidneybeans-chickpea -1.58754604 -1.9528824 -1.22220965 0.0000000  
## lentil-chickpea -0.40902505 -0.7743614 -0.04368867 0.0169225  
## mothbeans-chickpea -0.50578254 -0.8711189 -0.14044616 0.0009307  
## mungbean-chickpea -0.61299968 -0.9783361 -0.24766330 0.0000182  
## pigeonpeas-chickpea -1.54278174 -1.9081181 -1.17744536 0.0000000  
## lentil-kidneybeans 1.17852099 0.8131846 1.54385737 0.0000000  
## mothbeans-kidneybeans 1.08176350 0.7164271 1.44709988 0.0000000  
## mungbean-kidneybeans 0.97454635 0.6092100 1.33988274 0.0000000  
## pigeonpeas-kidneybeans 0.04476429 -0.3205721 0.41010068 0.9998211  
## mothbeans-lentil -0.09675749 -0.4620939 0.26857890 0.9865446  
## mungbean-lentil -0.20397463 -0.5693110 0.16136175 0.6490546  
## pigeonpeas-lentil -1.13375669 -1.4990931 -0.76842031 0.0000000  
## mungbean-mothbeans -0.10721714 -0.4725535 0.25811924 0.9771505  
## pigeonpeas-mothbeans -1.03699920 -1.4023356 -0.67166282 0.0000000  
## pigeonpeas-mungbean -0.92978206 -1.2951184 -0.56444567 0.0000000

We can see the difference in mean using the tukeyHSD, which takes each combination of crop\_suitable\_to\_grop and pairs them. So far it has considered the difference between the each pair and stated the adjusted p-value.

Just to point out: chickpea-blackgram, there is a difference of 0.20300499, with a lower limit of -0.1623314 and an upper limit of 0.56834138, this confidence interval includes the value of 0, this would mean there is no difference in the mean between chickpea and blackgram. Having an adjusted p-value of 0.6542447 suggest that there is no difference in the mean and we go with the null hypothesis

The Opposite can be observed between kidneybeans-blackgram when the p-value is practically 0.0, and we will have to go with the alternate hypothesis.

**Chi squared goodness of fit test**

According to the Royal Horticultural Society we can classify soil types with the level of acidity or alkalinity in the soil and this can be measured using the PH level. When you measure your soils PH level we can group the soil type into “Very acid soil” when the PH value is between 3.0 - 5.0 “Acid soil” PH level between 5.1 - 6.0 “Moderately acid soil” PH level between 6.1 - 7.0 “Alkaline soil” PH level between 7.1 - 8.0

Question: is there a statistically significant difference in the proportion of soil that are ‘Very acid soil’, ‘Acid soil’, ‘Moderately acid soil’ and ‘Alkaline soil’ using a alpha = 0.05

H0: The proportion of soil that are ‘Very acid soil’, ‘Acid soil’, ‘Moderately acid soil’, ‘Alkaline soil’ are equal H1: The proportion of soil that are ‘Very acid soil’, ‘Acid soil’, ‘Moderately acid soil’, ‘Alkaline soil’ are not equal

Looking at the graph below we see the proportion of soil type to be different, the question now is to confirm if this difference statistically significant

df <- data.frame(crop\_data$ph, crop\_data$crop\_suitable\_to\_grow)

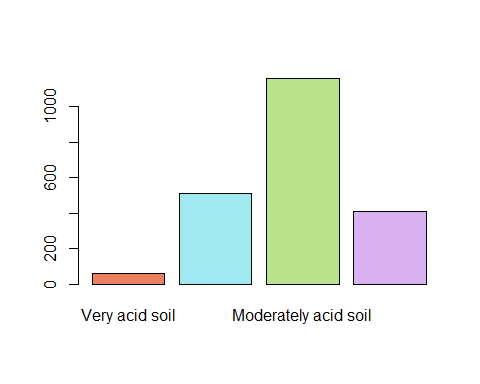
df$phlevels <- cut(crop\_data$ph,  
 breaks=c(0, 5, 6, 7, 8),  
 labels=c('Very acid soil', 'Acid soil', 'Moderately acid soil', 'Alkaline soil'))  
head(df)

## crop\_data.ph crop\_data.crop\_suitable\_to\_grow phlevels  
## 1 6.502985 rice Moderately acid soil  
## 2 7.038096 rice Alkaline soil  
## 3 7.840207 rice Alkaline soil  
## 4 6.980401 rice Moderately acid soil  
## 5 7.628473 rice Alkaline soil  
## 6 7.073454 rice Alkaline soil

table(df$phlevels)

##   
## Very acid soil Acid soil Moderately acid soil   
## 62 511 1159   
## Alkaline soil   
## 412

barplot(table(df$phlevels), col = c("#eb8060","#a1e9f0" , "#b9e38d", "#d9b1f0"))



chisq.test(table(df$phlevels))

##   
## Chi-squared test for given probabilities  
##   
## data: table(df$phlevels)  
## X-squared = 1173.1, df = 3, p-value < 2.2e-16

p value = 2.2e-16 if the proportions were equal, the probability of the sample data providing a difference of this magnitude or more is less than 0.05 therefore we reject the Null H0 hypothesis that the PH of soil types are equal.

**Chi Squared Test of Independence**

H0: The variables are independent. There is no relationship between the variables Knowing the value of one variable does not help to predict the value of the other variable

H1: The variable are dependent. Knowing the value of one variable does help to predict the value of the other variable

table(df$crop\_data.crop\_suitable\_to\_grow,df$phlevels)

##   
## Very acid soil Acid soil Moderately acid soil Alkaline soil  
## apple 0 60 40 0  
## banana 0 50 50 0  
## blackgram 0 0 36 64  
## chickpea 0 2 34 44  
## coconut 0 50 50 0  
## coffee 0 0 68 32  
## cotton 0 5 50 45  
## grapes 0 49 51 0  
## jute 0 0 64 36  
## kidneybeans 0 100 0 0  
## lentil 0 4 48 48  
## maize 0 34 66 0  
## mango 18 42 40 0  
## mothbeans 20 19 9 16  
## mungbean 0 0 78 22  
## muskmelon 0 0 100 0  
## orange 0 0 49 51  
## papaya 0 0 100 0  
## pigeonpeas 24 37 28 11  
## pomegranate 0 26 58 16  
## rice 0 33 40 27  
## watermelon 0 0 100 0

chisq.test(table(df$crop\_data.crop\_suitable\_to\_grow,df$phlevels))

## Warning in chisq.test(table(df$crop\_data.crop\_suitable\_to\_grow, df$phlevels)):  
## Chi-squared approximation may be incorrect

##   
## Pearson's Chi-squared test  
##   
## data: table(df$crop\_data.crop\_suitable\_to\_grow, df$phlevels)  
## X-squared = 1870, df = 63, p-value < 2.2e-16

One of the criteria for conducting the Chi-squared test of independence is that the cell value of the frequency table should not be less than 5 and thus I get a warning when I proceed with my Ch-squared test of independence. (Haberman, 1988)

The null hypothesis is being rejected as the p-value is less than 0.05 however, this result could be biased due to the missing cell values in my frequency table.

In an attempt to ensure that the frequency in each cell is at least 5, I have grouped the temperature data into 3 category, low, optimal and high. Referencing the article written by the EOS Data Analytics about soil temperature as a factor of crop development, where the average soil temperature for bioactivity range from 10 - 24 degrees.

Using the Chi-Squared Test of independence to see if there is a relationship between temperature and soil PH

df1 <- data.frame(crop\_data$temperature, crop\_data$ph,crop\_data$crop\_suitable\_to\_grow )  
df1$templevels <- cut(crop\_data$temperature,  
 3,  
 labels=c('Low', 'Optimal', 'High'))  
df1$phlevels <- cut(crop\_data$ph,  
 breaks=c(0, 5, 6, 7, 8),  
 labels=c('Very acid soil', 'Acid soil', 'Moderately acid soil', 'Alkaline soil'))  
head(df1)

## crop\_data.temperature crop\_data.ph crop\_data.crop\_suitable\_to\_grow templevels  
## 1 20.87974 6.502985 rice Optimal  
## 2 21.77046 7.038096 rice Optimal  
## 3 23.00446 7.840207 rice Optimal  
## 4 26.49110 6.980401 rice Optimal  
## 5 20.13017 7.628473 rice Low  
## 6 23.05805 7.073454 rice Optimal  
## phlevels  
## 1 Moderately acid soil  
## 2 Alkaline soil  
## 3 Alkaline soil  
## 4 Moderately acid soil  
## 5 Alkaline soil  
## 6 Alkaline soil

table(df1$templevels,df1$phlevels)

##   
## Very acid soil Acid soil Moderately acid soil Alkaline soil  
## Low 8 110 126 68  
## Optimal 41 365 927 318  
## High 13 36 106 26

H0: There is no relationship between the temperature and ph of soil. the variable are independent. H1: The variables are dependent

chisq.test(table(df1$templevels,df1$phlevels))

##   
## Pearson's Chi-squared test  
##   
## data: table(df1$templevels, df1$phlevels)  
## X-squared = 49.018, df = 6, p-value = 7.394e-09

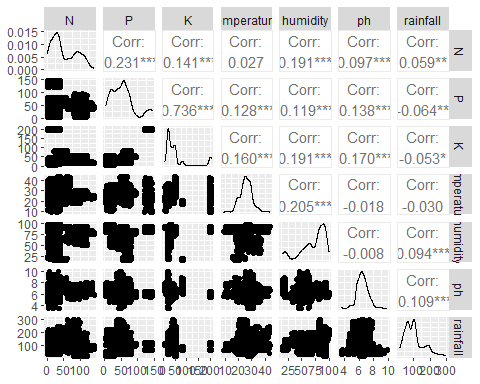
After running the code we see that there is no warning message which occured earlier. Also having a p-value of 7.394e-09 indicate that there is in fact a relationship between temperature and soil ph the null hypothesis is rejected.

**Correlation Analysis**

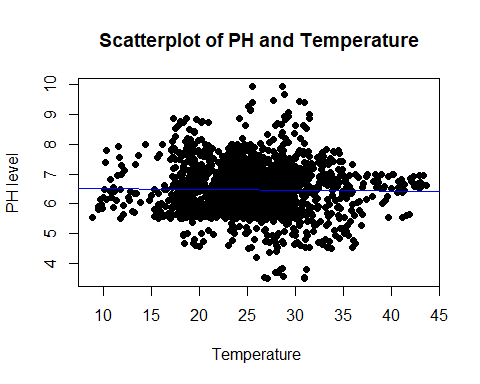
In this section we want to check the effect of temperature on the soil PH level. I will be using scatter plot, the correlation function and linear regression, the response variable here will be the PH level.

Using ggpairs to plot the correlation between the numeric varioables in the dataset

crop\_data %>%   
 select(N,P,K,temperature,humidity,ph,rainfall) %>%   
 #pairs()  
 ggpairs()



plot(crop\_data$temperature, crop\_data$ph, main="Scatterplot of PH and Temperature", xlab="Temperature ", ylab="PH level", pch=19)  
abline(lm(ph ~ temperature, data=crop\_data), col= "blue")

 The scatter plot does not show any relationship between the temperature and soil ph. However we see data points concentrated at 20 degrees to 30 degrees intervals.

cor(crop\_data$ph,crop\_data$temperature)

## [1] -0.01779502

Calculating the correlation coefficient we see that it is small and negative, which is not surprising as we know that an increase in temperature has a negative effect on soil PH. This effect is however not a strong one.

Although the correlation coefficient is small, I would like to see how this variable performs in a linear model.

crop\_data %>%   
 lm(ph ~ temperature, data=.) %>%   
 summary()

##   
## Call:  
## lm(formula = ph ~ temperature, data = .)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.9551 -0.5055 -0.0420 0.4503 3.4740   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.53915 0.08511 76.831 <2e-16 \*\*\*  
## temperature -0.00272 0.00326 -0.834 0.404   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.774 on 2198 degrees of freedom  
## Multiple R-squared: 0.0003167, Adjusted R-squared: -0.0001382   
## F-statistic: 0.6962 on 1 and 2198 DF, p-value: 0.4041

Looking at the summary of the simple linear model, we can see a very small F-statitics and a p-value of 0.4041, this confirms that the variable temperature is a poor predictor of soil PH. The R-squared which tells us how much our model can predict soil PH is approximately zero.

**Non-Parametric Test**

Are you having trouble speaking or understanding? Are you suddenly confused? There is a numbness or weakness in your face, arm or leg? These questions are being asked because you might be having a stroke. (cdc.gov, 2022) Stroke has been a life-threatening medical condition which affect millions of people around the world. In 2013 approximately 6.9 people had ischemic stroke and 3.4 million people had hermorrhagic stroke. (Wikipedia, 2017)

In this second part of the report, I will be using data about stroke patients to carry out statistical techniques, attempting to create a logistic model that predicts the likelihood a patient will have stroke. The techniques in this section will be non parametric and I will be using an alpha value of 0.05.

The data set contains 4908 observation and 12 columns. Id: This is a unique identifier to each patient Gender: Sex of the patient Age: Age of the patient Hypertension: This column indicates if the patient has hypertension or not heart\_disease: This column indicates if the patient has hearty disease Ever\_married: Contains information about the marital status of the patient work\_type: The occupation which the patient does Residence\_type: This contains the area where the patient stays avg\_glucose\_level: The average glucose level measured BMI: Body mass index of the patient Smoking\_status: Indicates the smoking status of the patient Stroke: Indicates if the patient had stroke. (Kaggle, 2021)

stroke\_data <- read.csv("healthcare\_dataset\_stroke.csv")  
head(stroke\_data)

## id gender age hypertension heart\_disease ever\_married work\_type  
## 1 9046 Male 67 0 1 Yes Private  
## 2 31112 Male 80 0 1 Yes Private  
## 3 60182 Female 49 0 0 Yes Private  
## 4 1665 Female 79 1 0 Yes Self-employed  
## 5 56669 Male 81 0 0 Yes Private  
## 6 53882 Male 74 1 1 Yes Private  
## Residence\_type avg\_glucose\_level bmi smoking\_status stroke  
## 1 Urban 228.69 36.6 formerly smoked 1  
## 2 Rural 105.92 32.5 never smoked 1  
## 3 Urban 171.23 34.4 smokes 1  
## 4 Rural 174.12 24.0 never smoked 1  
## 5 Urban 186.21 29.0 formerly smoked 1  
## 6 Rural 70.09 27.4 never smoked 1

dim(stroke\_data)

## [1] 4908 12

str(stroke\_data)

## 'data.frame': 4908 obs. of 12 variables:  
## $ id : int 9046 31112 60182 1665 56669 53882 10434 60491 12109 12095 ...  
## $ gender : chr "Male" "Male" "Female" "Female" ...  
## $ age : num 67 80 49 79 81 74 69 78 81 61 ...  
## $ hypertension : int 0 0 0 1 0 1 0 0 1 0 ...  
## $ heart\_disease : int 1 1 0 0 0 1 0 0 0 1 ...  
## $ ever\_married : chr "Yes" "Yes" "Yes" "Yes" ...  
## $ work\_type : chr "Private" "Private" "Private" "Self-employed" ...  
## $ Residence\_type : chr "Urban" "Rural" "Urban" "Rural" ...  
## $ avg\_glucose\_level: num 229 106 171 174 186 ...  
## $ bmi : num 36.6 32.5 34.4 24 29 27.4 22.8 24.2 29.7 36.8 ...  
## $ smoking\_status : chr "formerly smoked" "never smoked" "smokes" "never smoked" ...  
## $ stroke : int 1 1 1 1 1 1 1 1 1 1 ...

summary(stroke\_data)

## id gender age hypertension   
## Min. : 77 Length:4908 Min. : 0.08 Min. :0.00000   
## 1st Qu.:18603 Class :character 1st Qu.:25.00 1st Qu.:0.00000   
## Median :37581 Mode :character Median :44.00 Median :0.00000   
## Mean :37060 Mean :42.87 Mean :0.09189   
## 3rd Qu.:55182 3rd Qu.:60.00 3rd Qu.:0.00000   
## Max. :72940 Max. :82.00 Max. :1.00000   
## heart\_disease ever\_married work\_type Residence\_type   
## Min. :0.00000 Length:4908 Length:4908 Length:4908   
## 1st Qu.:0.00000 Class :character Class :character Class :character   
## Median :0.00000 Mode :character Mode :character Mode :character   
## Mean :0.04951   
## 3rd Qu.:0.00000   
## Max. :1.00000   
## avg\_glucose\_level bmi smoking\_status stroke   
## Min. : 55.12 Min. :10.30 Length:4908 Min. :0.00000   
## 1st Qu.: 77.07 1st Qu.:23.50 Class :character 1st Qu.:0.00000   
## Median : 91.68 Median :28.10 Mode :character Median :0.00000   
## Mean :105.30 Mean :28.89 Mean :0.04258   
## 3rd Qu.:113.50 3rd Qu.:33.10 3rd Qu.:0.00000   
## Max. :271.74 Max. :97.60 Max. :1.00000

Looking at the data set we see that hypertension contains either 1 or 0, representing having hypertension or not having hypertension, I will clean this column so the new values will represent 1 for Yes and 0 for No and also store it as a factor, so that R knows to treat the column as a categorical variable rather than character.

stroke\_data[stroke\_data$hypertension==0,]$hypertension <- "No"  
stroke\_data[stroke\_data$hypertension==1,]$hypertension <- "Yes"  
stroke\_data$hypertension <- as.factor(stroke\_data$hypertension)

I will do the same with heart\_disease and every other column that has the same format

stroke\_data[stroke\_data$heart\_disease==0,]$heart\_disease <- "No"  
stroke\_data[stroke\_data$heart\_disease==1,]$heart\_disease <- "Yes"  
stroke\_data$heart\_disease <- as.factor(stroke\_data$heart\_disease)

stroke\_data[stroke\_data$stroke==0,]$stroke <- "No"  
stroke\_data[stroke\_data$stroke==1,]$stroke <- "Yes"  
stroke\_data$stroke <- as.factor(stroke\_data$stroke)

stroke\_data$ever\_married <- as.factor(stroke\_data$ever\_married)  
stroke\_data$work\_type <- as.factor(stroke\_data$work\_type)  
stroke\_data$Residence\_type <- as.factor(stroke\_data$Residence\_type)  
stroke\_data$smoking\_status <- as.factor(stroke\_data$smoking\_status)  
stroke\_data$gender <- as.factor(stroke\_data$gender)

After cleaning the data and transforming column we can take a glimpse at the data set to see that all changes have been effected

glimpse(stroke\_data)

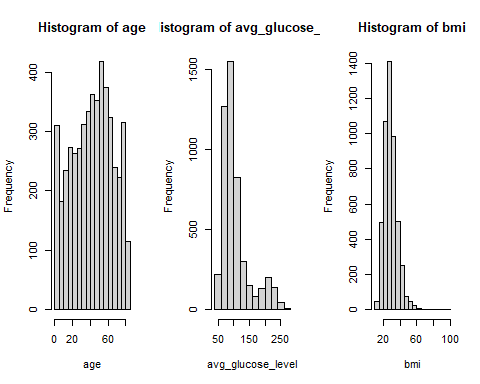
## Rows: 4,908  
## Columns: 12  
## $ id <int> 9046, 31112, 60182, 1665, 56669, 53882, 10434, 60491…  
## $ gender <fct> Male, Male, Female, Female, Male, Male, Female, Fema…  
## $ age <dbl> 67, 80, 49, 79, 81, 74, 69, 78, 81, 61, 54, 79, 50, …  
## $ hypertension <fct> No, No, No, Yes, No, Yes, No, No, Yes, No, No, No, Y…  
## $ heart\_disease <fct> Yes, Yes, No, No, No, Yes, No, No, No, Yes, No, Yes,…  
## $ ever\_married <fct> Yes, Yes, Yes, Yes, Yes, Yes, No, Yes, Yes, Yes, Yes…  
## $ work\_type <fct> Private, Private, Private, Self-employed, Private, P…  
## $ Residence\_type <fct> Urban, Rural, Urban, Rural, Urban, Rural, Urban, Urb…  
## $ avg\_glucose\_level <dbl> 228.69, 105.92, 171.23, 174.12, 186.21, 70.09, 94.39…  
## $ bmi <dbl> 36.6, 32.5, 34.4, 24.0, 29.0, 27.4, 22.8, 24.2, 29.7…  
## $ smoking\_status <fct> formerly smoked, never smoked, smokes, never smoked,…  
## $ stroke <fct> Yes, Yes, Yes, Yes, Yes, Yes, Yes, Yes, Yes, Yes, Ye…

attach(stroke\_data)

Plotting the histogram of the numeric columns so we can have an idea of what the spread and shape of the data is.

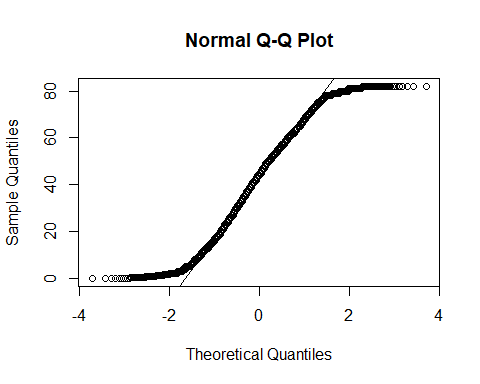
Since we will be carrying out non parametric test with this data, we make no assumptions of normality.

par(mfrow = c(1, 3))  
hist(age)  
hist(avg\_glucose\_level)  
hist(bmi)

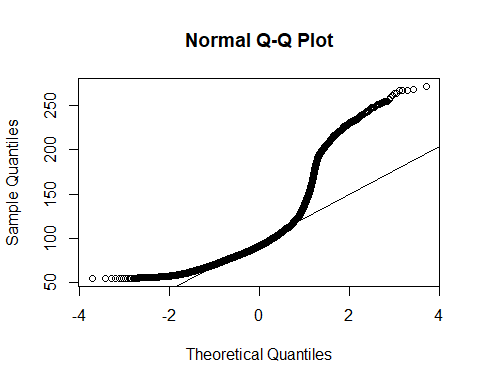


Plotting QQ Plot of the dataset

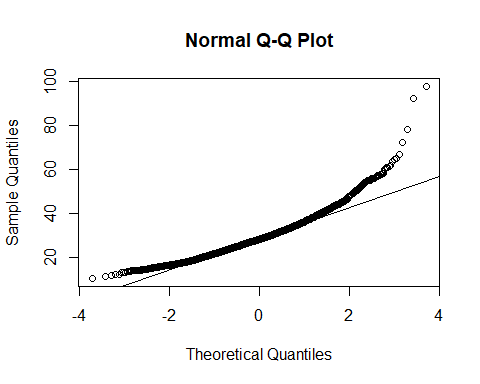
qqnorm(age)  
qqline(age)



qqnorm(avg\_glucose\_level)  
qqline(avg\_glucose\_level)



qqnorm(bmi)  
qqline(bmi)



The table below shows that stroke patients and patients without stroke are both represented in the female and male samples

xtabs(~ stroke + gender, data=stroke\_data)

## gender  
## stroke Female Male  
## No 2777 1922  
## Yes 120 89

There is also a good representation of patient who had stroke and hypertension

xtabs(~ stroke + hypertension, data=stroke\_data)

## hypertension  
## stroke No Yes  
## No 4308 391  
## Yes 149 60

There is a good representation of patient with heart disease in relations to stroke

xtabs(~ stroke + heart\_disease, data=stroke\_data)

## heart\_disease  
## stroke No Yes  
## No 4496 203  
## Yes 169 40

The marital status of patient is also well represented

xtabs(~ stroke + ever\_married, data=stroke\_data)

## ever\_married  
## stroke No Yes  
## No 1681 3018  
## Yes 23 186

There is little to no representation when it concerns children and patients who never worked having stroke, this might potentially get in the way of finding the best fitting line. I will however still utilize this coulumn and see what effects it has.

xtabs(~ stroke + work\_type, data=stroke\_data)

## work\_type  
## stroke children Govt\_job Never\_worked Private Self-employed  
## No 670 602 22 2683 722  
## Yes 1 28 0 127 53

xtabs(~ stroke + Residence\_type, data=stroke\_data)

## Residence\_type  
## stroke Rural Urban  
## No 2318 2381  
## Yes 100 109

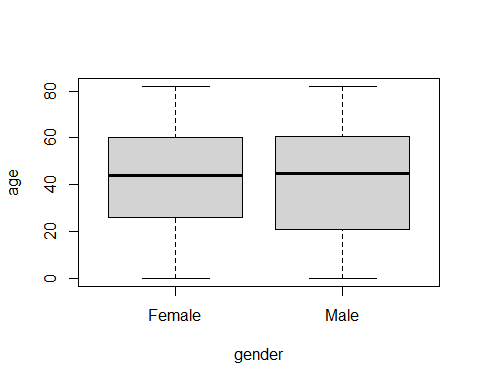
xtabs(~ stroke + smoking\_status, data=stroke\_data)

## smoking\_status  
## stroke formerly smoked never smoked smokes Unknown  
## No 779 1768 698 1454  
## Yes 57 84 39 29

**Wilcox Rank Sum Test**

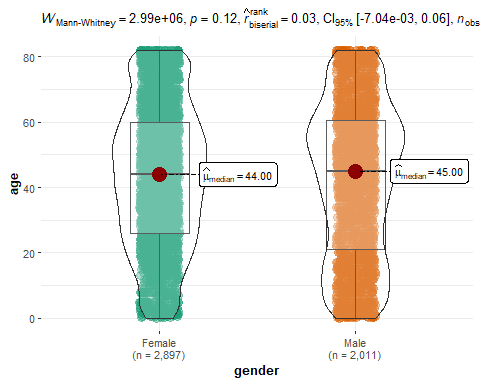
The aim of this test is to confirm if there is a difference in the age of gender in the data. The Box plot

boxplot(age ~ gender)



H0: The median difference between the ages is zero H1: The median difference between the ages is not zero Alpha = 0.05

ggbetweenstats(  
 data = stroke\_data,  
 x = gender,  
 y = age,  
 type = "nonparametric"  
)



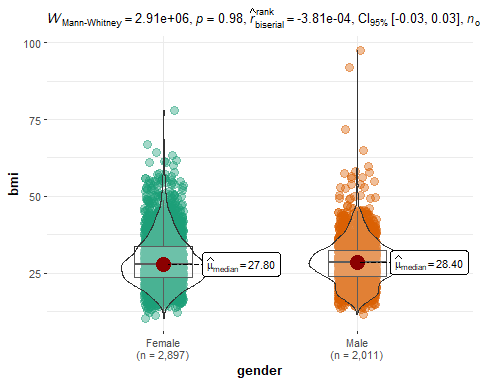
wilcox.test(stroke\_data$age ~ stroke\_data$gender, exact=F)

##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: stroke\_data$age by stroke\_data$gender  
## W = 2988082, p-value = 0.1237  
## alternative hypothesis: true location shift is not equal to 0

Given the p-value = 0.1237 we can conclude that the difference in median age between genders in our dataset is zero

H0: The median difference between the bmi is zero H1: The median difference between the bmi is not zero Alpha = 0.05

ggbetweenstats(  
 data = stroke\_data,  
 x = gender,  
 y = bmi,  
 type = "nonparametric"  
)



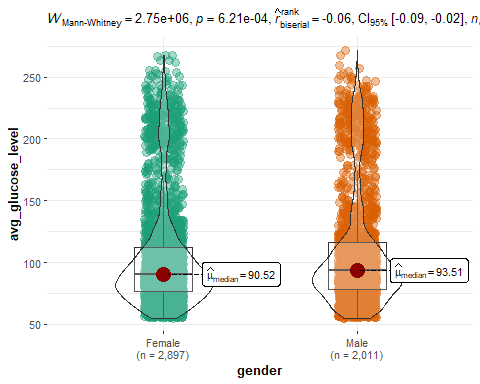
wilcox.test(stroke\_data$bmi ~ stroke\_data$gender, exact=F)

##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: stroke\_data$bmi by stroke\_data$gender  
## W = 2911825, p-value = 0.9819  
## alternative hypothesis: true location shift is not equal to 0

With a P-value=0.9819 we fail to reject the null hypothesis

H0: The median difference between the average glucose level is zero H1: The median difference between the average glucose level is not zero Alpha = 0.05

ggbetweenstats(  
 data = stroke\_data,  
 x = gender,  
 y = avg\_glucose\_level,  
 type = "nonparametric"  
)



wilcox.test(stroke\_data$avg\_glucose\_level ~ stroke\_data$gender, exact=F)

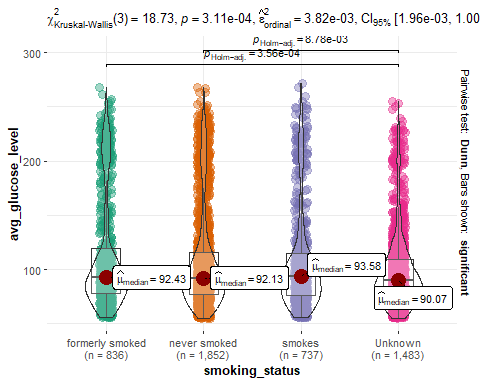
##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: stroke\_data$avg\_glucose\_level by stroke\_data$gender  
## W = 2745859, p-value = 0.0006209  
## alternative hypothesis: true location shift is not equal to 0

We reject the null hypothesis and have to agree with the alternate H1 that there is a difference in the average glucose level

**Kruskal Wallis Test**

To test if the average glucose level across smoking status of patients H0 There is no difference across the group H1 There is a difference across the group

ggbetweenstats(  
 data = stroke\_data,  
 x = smoking\_status,  
 y = avg\_glucose\_level,  
 type = "nonparametric"  
)



kruskal.test(avg\_glucose\_level ~ smoking\_status)

##   
## Kruskal-Wallis rank sum test  
##   
## data: avg\_glucose\_level by smoking\_status  
## Kruskal-Wallis chi-squared = 18.729, df = 3, p-value = 0.000311

We reject the null hypothesis that there is no observable difference across smoking statuses of patients as the p-value = 3.11e-4 this is less than 0.05

**Logistic Regression**

**Model1**: In this simple logistic regression, we would like to predict the likelihood of stroke using the age of patients.

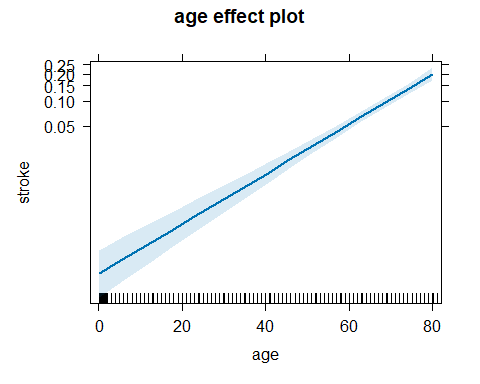
Model1 <- glm(stroke ~ age, data=stroke\_data, family = "binomial")  
summary(Model1)

##   
## Call:  
## glm(formula = stroke ~ age, family = "binomial", data = stroke\_data)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -7.376871 0.362399 -20.36 <2e-16 \*\*\*  
## age 0.074964 0.005318 14.10 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1728.3 on 4907 degrees of freedom  
## Residual deviance: 1407.7 on 4906 degrees of freedom  
## AIC: 1411.7  
##   
## Number of Fisher Scoring iterations: 7

We can see that the coefficient of age is 0.074964 and the y intercept is -7.376871. what this means is that for every one increase in age, the predicted probability of stroke increases by 7.49%. Our p-value is less than 0.05 so therefore the coefficients are statistically significant and not due to chance. The residual deviance is significantly lower than the null deviance which will suggest a good fit.

The current AIC of this simple model is 1411.7, I will compare this to another model to see which is better.

plot(allEffects(Model1))

 The graph shows the effect of age and the probability of having stroke.

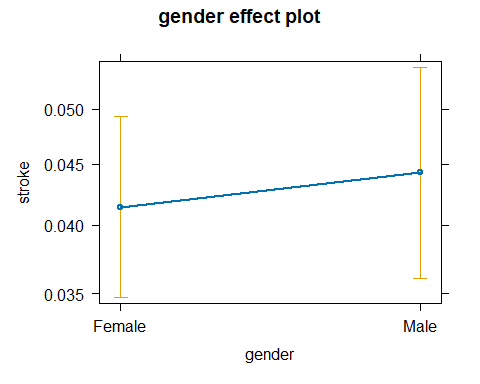
**Model2**: I will be using gender to predict the likelihood of stroke

Model2 <- glm(stroke ~ gender, data=stroke\_data, family = "binomial")  
summary(Model2)

##   
## Call:  
## glm(formula = stroke ~ gender, family = "binomial", data = stroke\_data)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -3.14163 0.09324 -33.695 <2e-16 \*\*\*  
## genderMale 0.06915 0.14300 0.484 0.629   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1728.3 on 4907 degrees of freedom  
## Residual deviance: 1728.1 on 4906 degrees of freedom  
## AIC: 1732.1  
##   
## Number of Fisher Scoring iterations: 6

We see from the summary that gender is not a good predictor of having stroke as the p-value is 0.629, implying that gender is not significantly impacting the likelihood of stroke.

plot(allEffects(Model2))

 The graph shows the probability of having stroke vs your gender. This difference is minimal just as we have seen by the model.

**Model3**: This is a multi logistic model where I will be using both the categorical and numerical variables in the dataset to predict the likelihood of stroke

Model3 <- glm(stroke ~ ., data=stroke\_data, family = "binomial")  
summary(Model3)

##   
## Call:  
## glm(formula = stroke ~ ., family = "binomial", data = stroke\_data)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -7.374e+00 1.075e+00 -6.858 6.99e-12 \*\*\*  
## id 3.983e-07 3.520e-06 0.113 0.909915   
## genderMale -1.481e-02 1.544e-01 -0.096 0.923561   
## age 7.347e-02 6.346e-03 11.577 < 2e-16 \*\*\*  
## hypertensionYes 5.244e-01 1.751e-01 2.995 0.002740 \*\*   
## heart\_diseaseYes 3.481e-01 2.073e-01 1.679 0.093174 .   
## ever\_marriedYes -1.155e-01 2.473e-01 -0.467 0.640438   
## work\_typeGovt\_job -6.815e-01 1.114e+00 -0.612 0.540755   
## work\_typeNever\_worked -9.824e+00 3.087e+02 -0.032 0.974612   
## work\_typePrivate -5.202e-01 1.100e+00 -0.473 0.636329   
## work\_typeSelf-employed -9.444e-01 1.119e+00 -0.844 0.398686   
## Residence\_typeUrban 4.790e-03 1.500e-01 0.032 0.974528   
## avg\_glucose\_level 4.655e-03 1.294e-03 3.597 0.000322 \*\*\*  
## bmi 4.055e-03 1.188e-02 0.341 0.732914   
## smoking\_statusnever smoked -6.731e-02 1.886e-01 -0.357 0.721203   
## smoking\_statussmokes 3.131e-01 2.296e-01 1.364 0.172683   
## smoking\_statusUnknown -2.752e-01 2.471e-01 -1.114 0.265482   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1728.3 on 4907 degrees of freedom  
## Residual deviance: 1363.2 on 4891 degrees of freedom  
## AIC: 1397.2  
##   
## Number of Fisher Scoring iterations: 14

This model has incorporated all the variables in the dataset to predict stroke. We can see that age, hypertension and the average glucose level have significant positive coefficients, their p-values are less than 0.05. This factors indicate a higher possibility of stroke. Heart disease might also be associated with higher possibility of stroke however the p-value 0.093 is not statistically significant. The non significant predictors are gender, marital status, work type, residence type, BMI and smoking status, these variables have p-value > than 0.05 and are not statistically significant to the model.

The residual deviance is lower than the null deviance which suggests a reasonable fit and the AIC 1397.2 is lower than both model1 and model2

Lastly Model4: In the previous model we saw variables which did not contribute to the prediction likelihood of stroke. What will be done in this model is to take out those variable and see if the new model makes a better prediction of stroke.

Model4 <- glm(stroke ~ age + hypertension + avg\_glucose\_level, data=stroke\_data, family = "binomial")  
summary(Model4)

##   
## Call:  
## glm(formula = stroke ~ age + hypertension + avg\_glucose\_level,   
## family = "binomial", data = stroke\_data)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -7.766457 0.385221 -20.161 < 2e-16 \*\*\*  
## age 0.069565 0.005490 12.671 < 2e-16 \*\*\*  
## hypertensionYes 0.547010 0.172629 3.169 0.00153 \*\*   
## avg\_glucose\_level 0.005047 0.001245 4.052 5.07e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1728.3 on 4907 degrees of freedom  
## Residual deviance: 1378.4 on 4904 degrees of freedom  
## AIC: 1386.4  
##   
## Number of Fisher Scoring iterations: 7

This model focuses on the variables that is statistical significance, and we can see all three predictors having a high positive coefficient with p-values less than 0.05, reinforcing their association with stroke. The model is a good fit with the residual deviance being lower than the null deviance. The AIC is also lower than the 3 previous models, which suggests a better predictor of stroke.

**Conclusion**

We have been able to carry out statistical analysis on two datasets. Soil PH is quite interesting as we see that the type of soil in which crop grows can be quite selective/sensitive and we cannot assume that a PH level which is optimal for a crop in the same family or classification will also be ideal for another crop in the same family. A logistic regression model to predict the likelihood of stroke was developed, with age, hypertension and the average glucose level being strong indicators to the likelihood that a patient has stroke.

# References

Campbell, G., 2018. *Correcting your soil pH.* [Online]   
Available at: https://www.frontierag.co.uk/blog/correcting-your-soil-ph

cdc.gov, 2022. *Centers for Disease Control and Prevention.* [Online]   
Available at: https://www.cdc.gov/stroke/signs\_symptoms.htm#:~:text=Sudden%20confusion%2C%20trouble%20speaking%2C%20or,headache%20with%20no%20known%20cause.

Haberman, S. J., 1988. In: *A Warning on the Use of Chi-Squared Statistics With Frequency Tables With Small Expected Cell Counts.* s.l.:Taylor & Francis, Ltd, p. 555.

ICC, 2010. *Food and Agriculture Organization of the United Nations.* [Online]   
Available at: https://www.fao.org/fileadmin/templates/ess/documents/world\_census\_of\_agriculture/appendix3\_r7.pdf

Kaggle, 2021. *Crop Recommendation Dataset.* [Online]   
Available at: https://www.kaggle.com/datasets/siddharthss/crop-recommendation-dataset/data

Kaggle, 2021. *Stroke Prediction Dataset.* [Online]   
Available at: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

Wikipedia, 2017. *Stroke.* [Online]   
Available at: https://en.wikipedia.org/wiki/Stroke