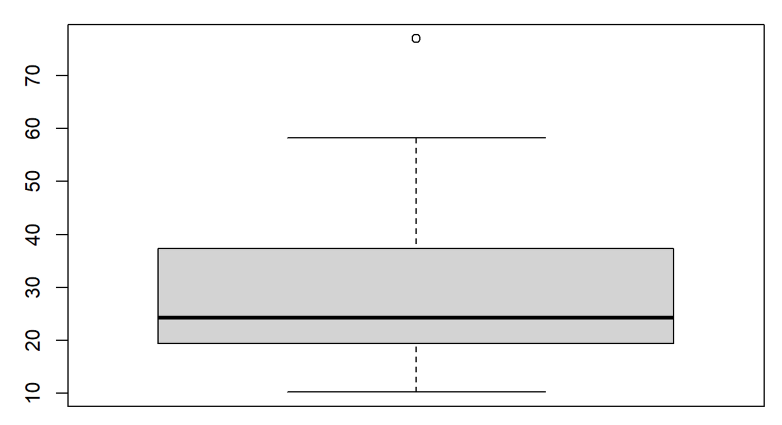
Stat 5405 Homework1 Solution

AUTHOR
Sathwik Bollepalli

PUBLISHED
September 10, 2023

Q1) a.

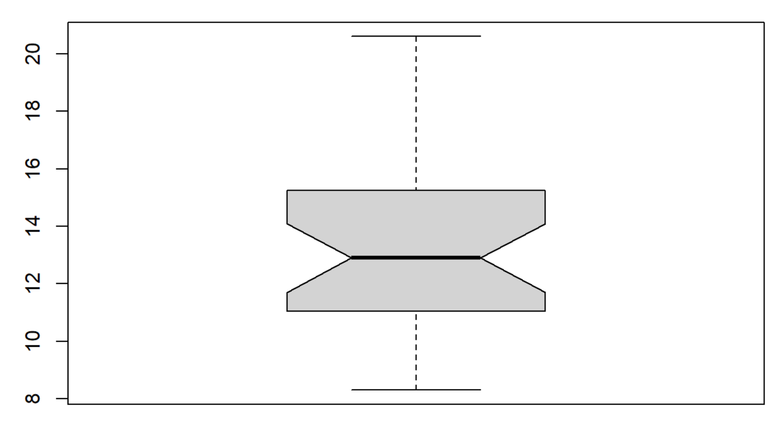
```
data(trees)
boxplot(trees$Volume,varwidth=TRUE)
```



I see a 5 number summary of the Volume of the trees present in the dataset with one outlier which has value greater than 1.5*IQR(Q3-Q1), where Q3 ~ 35 - 40 & Q1 ~ 15-20. The median of the data points lies between 20 to 25. The data seems to be right skewed as the boxplot has slight long whisker towards the right.

Q1) b.

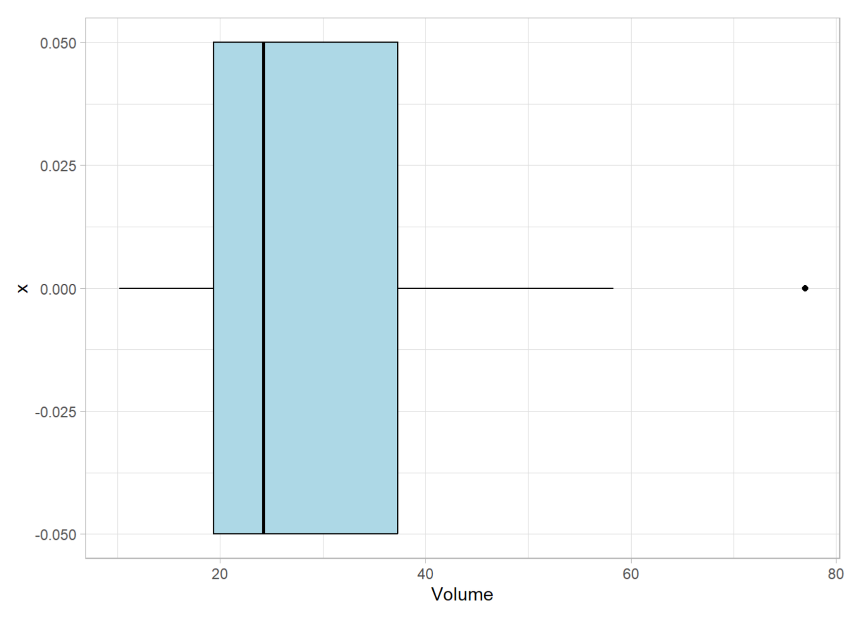
```
boxplot(trees$Girth,notch = TRUE)
```



In the above Boxplot, the notch=TRUE parameter is used to add notches to the box plot which helps in comparing the medians if multiple boxplots were there in the same visual. Notches are used to compare groups; if the notches of two boxes do not overlap, this is a strong evidence that the medians differ.

Q2) a.

```
library(ggplot2)
ggplot(data = trees,mapping = aes(y=Volume,x=0)) +
geom_boxplot(col="black",fill="lightblue",width=0.1) + coord_flip() + theme_light()
```

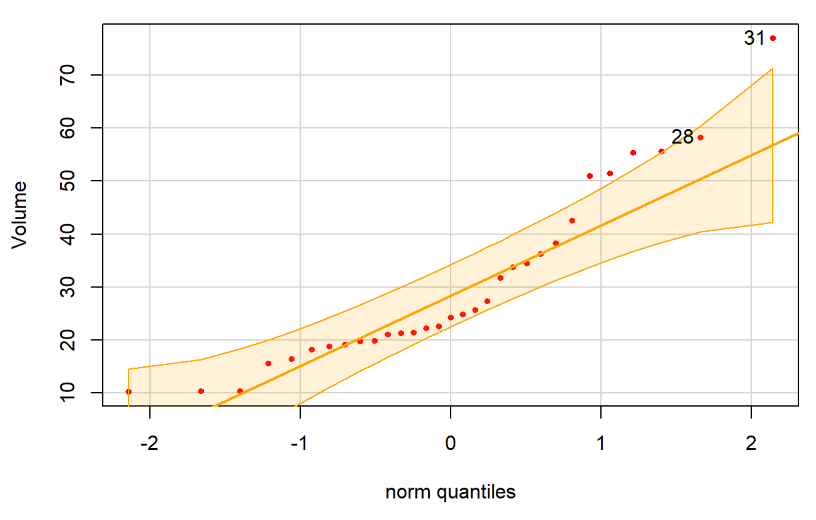


Q2) b.

```
library(car)
```

```
Loading required package: carData
```

```
qqPlot(trees$Volume, main="", ylab="Volume", cex=0.6,pch=19, col="red",
       col.lines = "orange")
```




```
[1] 31 28
```

The Volume of trees does not follow normal distribution, as we can see that the points do not lie on the straight line and some points are even going out of the confidence band.

Shapiro-Wilk Test

```
shapiro.test(trees$Volume)
```

```
    Shapiro-Wilk normality test

data:  trees$Volume
W = 0.88757, p-value = 0.003579
```

As we can see the p-value for the Shapiro test comes out to be 0.003579 which is less than 0.05(alpha), therefore we can say that trees$Volume does not follow normal distribution.

Chi-Square Test

```
library(nortest)
pearson.test(trees$Volume)
```

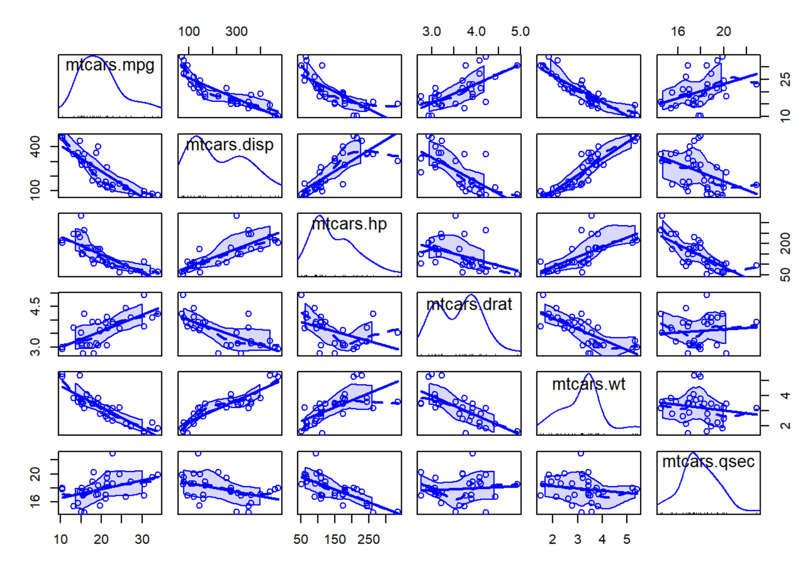
```
    Pearson chi-square normality test

data:  trees$Volume
P = 15.194, p-value = 0.009567
```

The Chi-square goodness of fit test is a statistical hypothesis test used to determine whether a variable is likely to come from a specified distribution (in this case normal distribution) or not.

As we can see, the p-value of the Pearson Chi-Square test is less than 0.05, we can say that trees$Volume does not follow normal distribution.

Q3) a.

```
scatterplotMatrix(~mtcars$mpg+mtcars$disp+mtcars$hp+mtcars$drat+mtcars$wt+mtcars$qs
```



By Looking at the first row of the scatter plot matrix, we can see that **mpg** is most **Positively assoicated** with **drat** and most **negatively associated** with **wt**.

Q3) b

```
data(mtcars)

selected_columns <- mtcars[, c("mpg", "disp", "hp", "drat", "wt", "qsec")]

correlation_matrix <- cor(selected_columns)

diag(correlation_matrix) <- NA

max_correlation <- max(correlation_matrix, na.rm = TRUE)

indices <- which(correlation_matrix == max_correlation, arr.ind = TRUE)

column1 <- rownames(correlation_matrix)[indices[1, 1]]
column2 <- colnames(correlation_matrix)[indices[1, 2]]

cat("Pair with maximum correlation:", column1, "and", column2, "with correlation ="
```

```
Pair with maximum correlation: wt and disp with correlation = 0.8879799
```

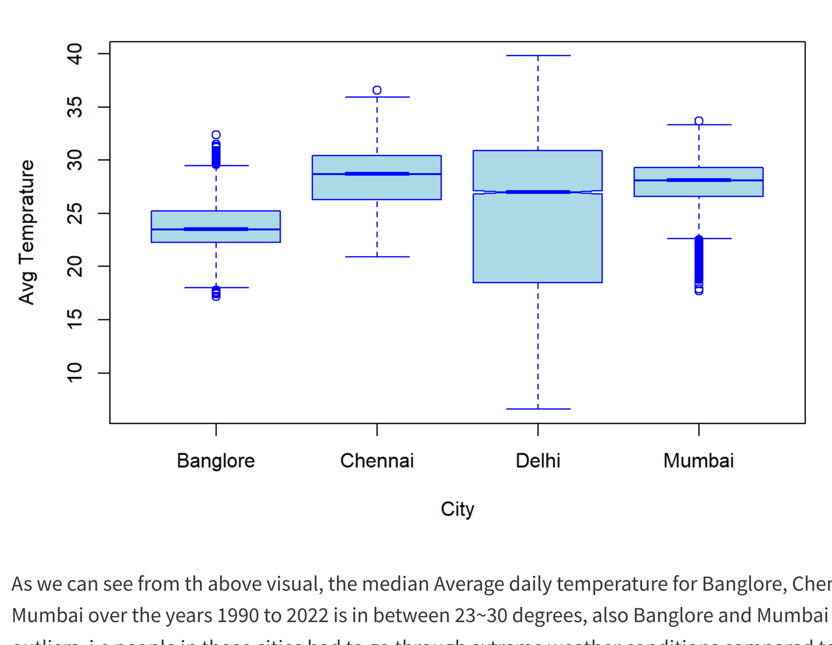
The Corelation between wt and disp is the highest with correlation = 0.8879799

Q4)

```
library(dplyr)
```

```
Attaching package: 'dplyr'

The following object is masked from 'package:car':

    recode

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
mumbai_data <- read.csv('Temperature_And_Precipitation_Cities_IN/Mumbai_1990_2022_S


chennai_data <- read.csv('Temperature_And_Precipitation_Cities_IN/Chennai_1990_2022_

delhi_data <- read.csv('Temperature_And_Precipitation_Cities_IN/Delhi_NCR_1990_2022_


banglore_data <- read.csv('Temperature_And_Precipitation_Cities_IN/Bangalore_1990_2



# Combine the datasets with a grouping variable
combined_data <- bind_rows(
  mutate(mumbai_data, Group = "Mumbai"),
  mutate(chennai_data, Group = "Chennai"),
  mutate(delhi_data, Group = "Delhi"),
  mutate(banglore_data, Group = "Banglore")
)


boxplot(combined_data$tavg ~ combined_data$Group, xlab="City",
        ylab="Avg Temprature", border = "blue", col="lightblue",notch=TRUE)
```



As we can see from th above visual, the median Average daily temperature for Banglore, Chennai, Delhi and Mumbai over the years 1990 to 2022 is in between 23~30 degrees, also Banglore and Mumbai are having a lot of outliers, i.e people in those cities had to go through extreme weather conditions compared to normal days.

Q5)

Some of the examples of unethical data visualization might be:

1) **Showing lots of variables in a pie chart**: The ideal number is in between 2 and 7, anything more than that will make the visual untidy and will over-populate the visual which make it difficut to interpret.

2) **Truncating the Y-axis** in the graphs: It will mislead or manipulate the users perception of data. Truncating the Y-axis refers to displaying a chart in a way that omits a portion of the Y-axis, making differences between data points appear more significant or less significant than they actually are.

3) **Unusual Coloring** : Good outcomes are associated with green color and bad outcomes with red, if we devaite from these then this might create misunderstandings.

4) **Lack of Labels** : If charts are lacking labels, percentages, or any specific information about the categories then viewers cannot discern the exact values or proportions represented by each category.

5) **No Data Source**: Transparency is crucial in data visualization, and viewers need to know where the data comes from and how it was collected.

6) **Misleading Labels**: Using misleading labels or titles that don't accurately represent the data, its source, or its context.

7) **Data Manipulation**: Manipulating the underlying data before visualization, such as altering data values, outliers, or summary statistics to support a particular agenda.

8) **Overlaying Charts**: Overlaying multiple datasets with different units of measurement on a single chart without proper scaling or labeling.