NUMERICAL ASSIGNMENT

STATISTICAL STRUCTURES IN DATA

Instructor: Prof. (Dr.) Subhajit Dutta

P Prabhanjan
24BM6JP40
BATCH – 10, PGDBA

CONTENTS

DATASET 1: AIR QUALITY

INTRODUCTION

This dataset provides information on the daily air quality measurements in New York from May to September 1973. The dataset is available in the 'datasets' package.

UNIVARIATE ANALYSIS

DATA OVERVIEW

The detail of the dataset is given below:

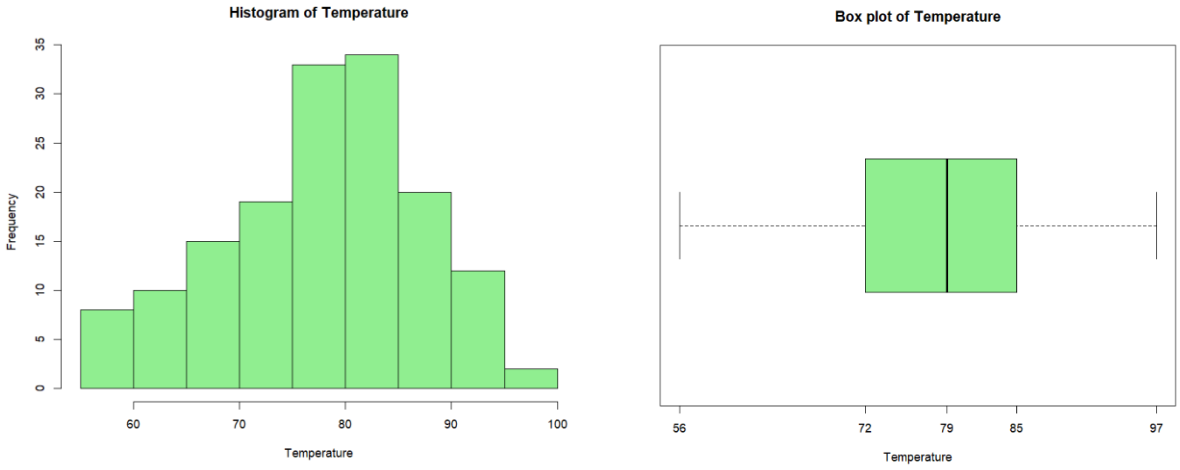
Dataset	Package	Variables	Observations
airquality	datasets	6	153

SUMMARY STATISTICS

The temperature (degree F) numerical variable is chosen for further analysis. The summary is shown below:

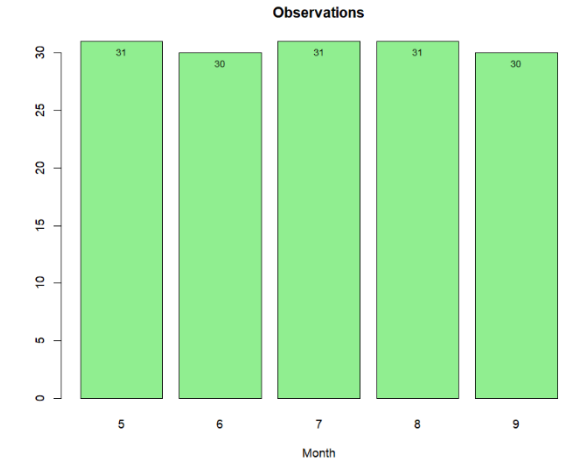
Variable	Mean	Median	Standard Deviation	Minimum	Maximum	Missing Values
Temperature (F)	77.88	79	9.46	56	97	0

DISTRIBUTION VISUALIZATION



The temperature variable is visualized using the Histogram and Boxplot. The most frequent temperature range in New York is between **75°F and 85°F**. The distribution is **left-skewed**, as visible from the boxplot. Skewness of **-0.37** proves the same. There are no outliers in the dataset.

CATEGORICAL VARIABLE ANALYSIS

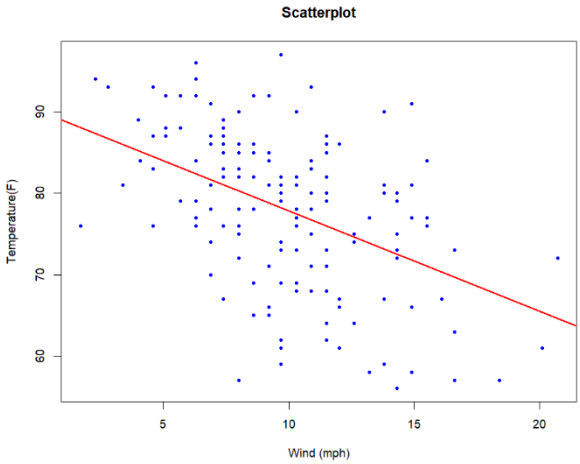


The analysis is done on the **Month variable**. The Bar plot suggests the number of Air quality recordings taken from May to September.

The bar plot shows that air quality observations were made daily for five months.

MULTIVARIATE ANALYSIS

CORRELATION ANALYSIS AND SCATTER PLOT VISUALIZATION



The correlation between Temperature and the wind variable is **-0.45**, which indicates a **negative linear relationship**.

The trend line fit to the model supports this, but the linear relationship is not strong as the observations are not linearly spread out.

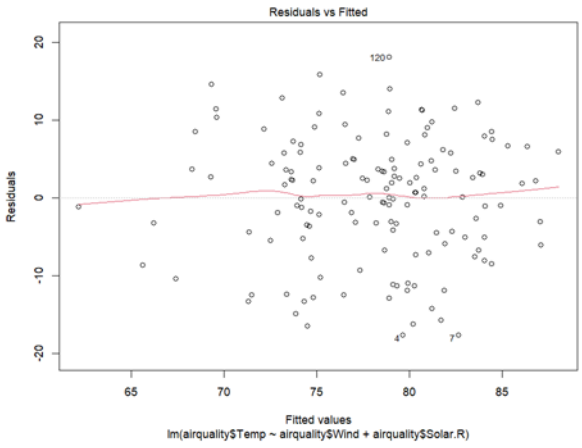
Higher-order relations may give a better relationship between the two.

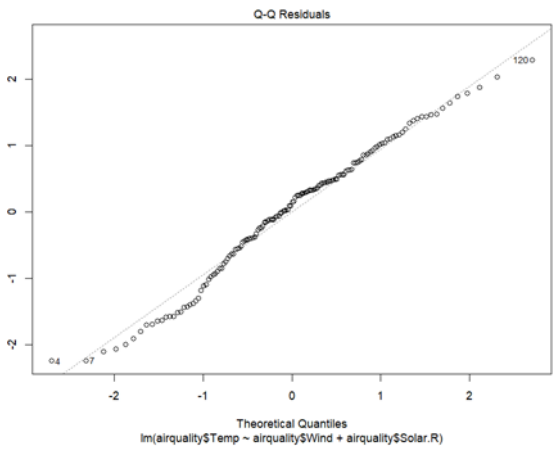
MULTIPLE LINEAR REGRESSION AND DIAGNOSTICS

Multiple Linear Regression was done to interpret the effect of **Wind and Solar Radiation on the Temperature**. Both the variables significantly affect the Temperature as the p-value of both is less than the significant level of 0.05. **Wind** affects it **negatively**, and **Solar Radiation** affects it **positively**.

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     84.899708   2.476595  34.281  < 2e-16 ***
airquality$Wind -1.155726   0.188311  -6.137 7.82e-09 ***
airquality$Solar.R  0.025697   0.007337   3.503 0.000615 ***
```

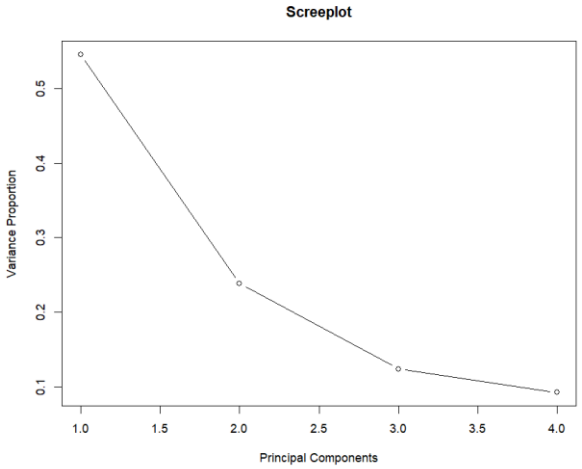
The Residual plot suggests that the residuals are **homoscedastic**. The QQ Plot suggests that the residuals are following the normal distribution.

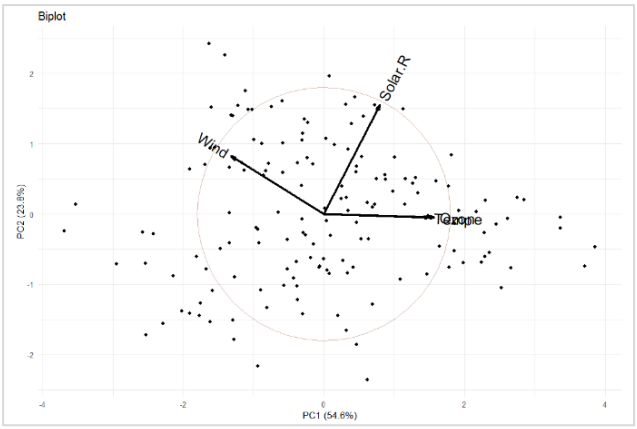




ADVANCED ANALYSIS

PRINCIPAL COMPONENT ANALYSIS (PCA)





PCA was done on the four numerical variables to reduce the dimensions of the dataset. Based on the Scree plot for the Air quality dataset, we can preserve around **90%** of the **variance** by considering the **first three principal component**

This biplot shows the correlation between variables Wind, Solar Radiation, Temp, and Ozone and their contributions to the first two principal components (PC 1 and PC 2).

PC 1 captures the most variance of the data (**54.6%).** Temperature and Ozone contribute most to PC 1, while Solar Radiation contributes the most to PC 2. Solar Radiation is positively correlated to the Ozone and Temperature variables, while Wind is negatively correlated with the temperature and Ozone. Ozone is closely aligned with Temperature, suggesting a potential association.

DATASET 2: MT CARS

INTRODUCTION

The data from the 1974 Motor Trend US magazine comprised fuel consumption and ten aspects of automobile design and performance for 32 automobiles (1973–74 models). The dataset is available in the 'datasets' package.

UNIVARIATE ANALYSIS

DATA OVERVIEW

The detail of the dataset is given below:

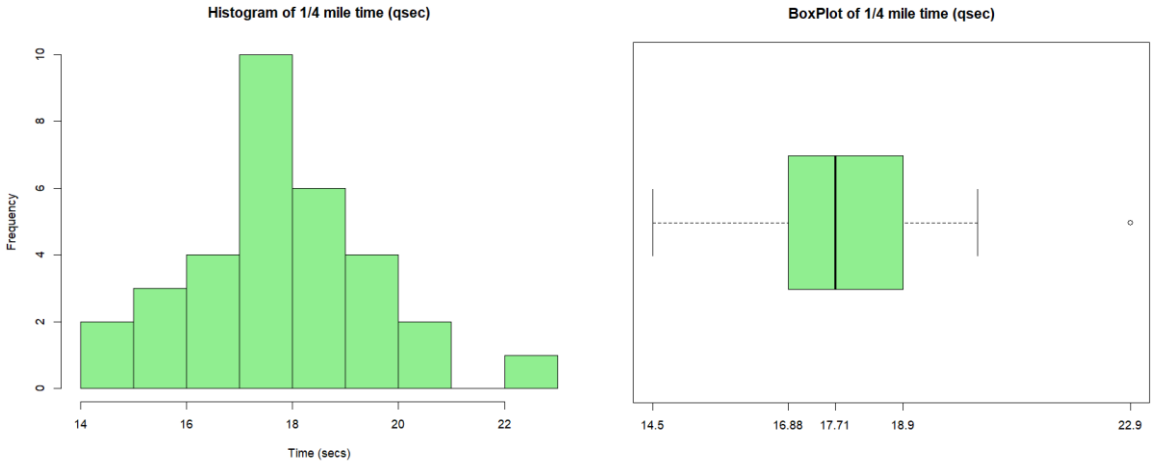
Dataset	Package	Variables	Observations
mtcars	datasets	11	32

SUMMARY STATISTICS

Qsec variable is considered for further analysis. Q sec is the time the car takes to a quarter-mile distance. It is measured in seconds. Following is the summary:

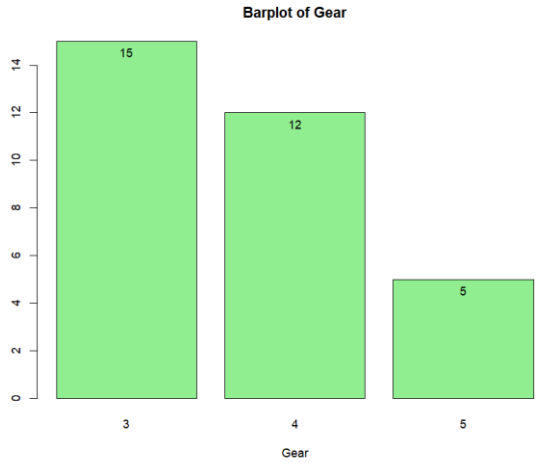
Variable	Mean	Median	Standard Deviation	Minimum	Maximum	Missing Values
Qsec (s)	17.85	17.71	1.79	14.5	22.9	0

DISTRIBUTION VISUALIZATION



The distribution of the variable is right skewed. This is due to the outlier, as evidenced by the box plot. **Skewness** of the qsec variable is **0.39. 50%** of the **cars** observed take **17 to 19 seconds** to cover a quarter of a mile.

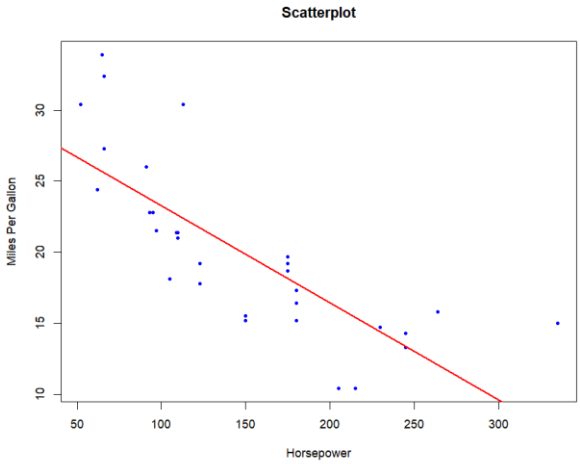
CATEGORICAL VARIABLE ANALYSIS



Gear is considered for the categorical variable analysis. Around **50%** of the **cars** observed have **three gears**. The number of cars decreases as the number of gears increases.

MULTIVARIATE ANALYSIS

CORRELATION ANALYSIS AND SCATTER PLOT VISUALIZATION



Linear Relationship between mileage (Miles per Gallon) and the horsepower is analysed.

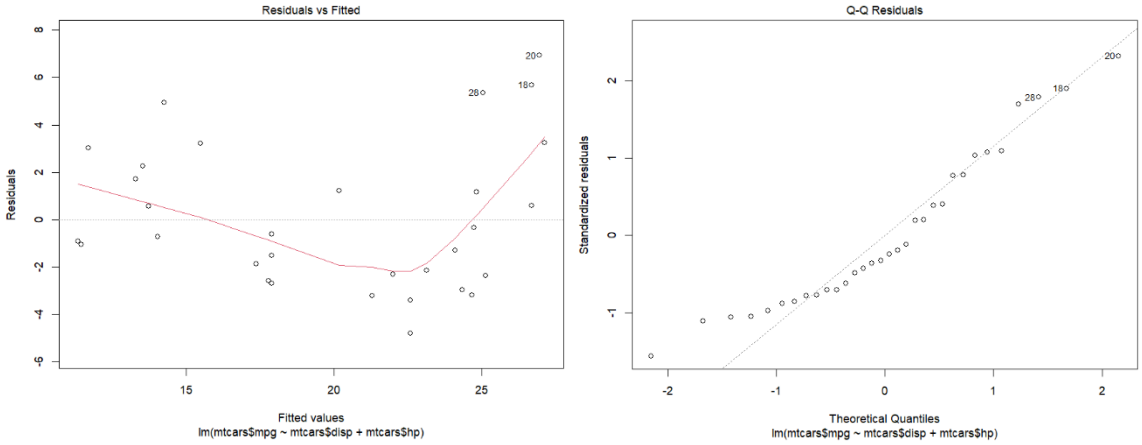
A strong correlation of **-0.77** between the variable x and y which indicating the horsepower inversely affect the Mileage of the car.

However, from the scatterplot, it can be inferred that a **nonlinear relationship** can better describe the relationship.

MULTIPLE LINEAR REGRESSION AND DIAGNOSTICS

Multiple Linear Regression was performed to understand the effect of Displacement and the Horsepower of the car on the Mileage (Miles Per Gallon). The **displacement** of the car significantly contributes to the **Mileage** in an **inverse** way. This is visible by the p-value of 0.003 which is much lesser than 0.05. However, **Horsepower** of the car does **not** play an **important** role on the mileage of the car.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 30.735904   1.331566  23.083  < 2e-16 ***
mtcars$disp -0.030346   0.007405  -4.098 0.000306 ***
mtcars$hp   -0.024840   0.013385  -1.856 0.073679 .
```

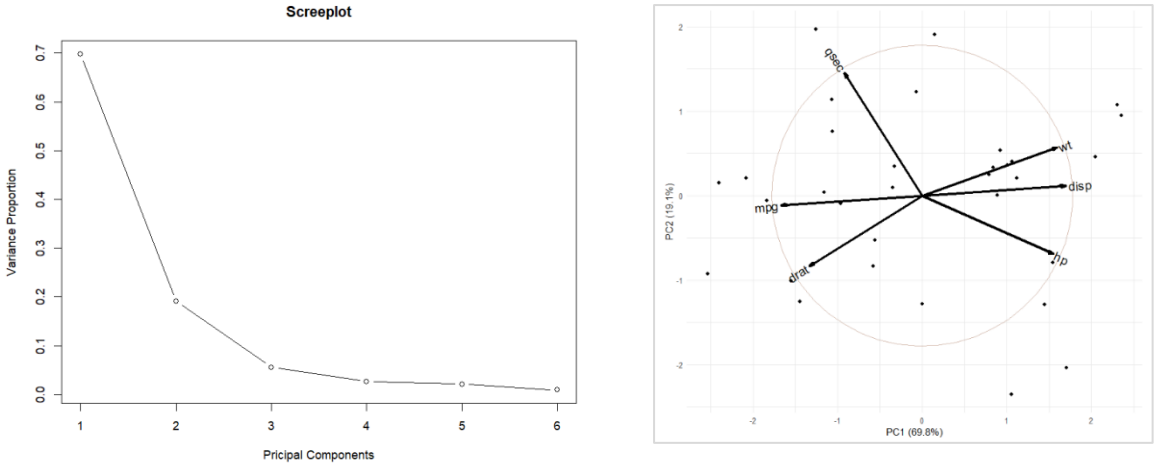


The residual vs fitted plot shows that the residuals are **Heteroskedastic**. The QQ-plot that the residuals do not follow the normal distribution. This implies that the multiple linear regression **cannot be relied upon**.

ADVANCED ANALYSIS

PRINCIPAL COMPONENT ANALYSIS (PCA)

A total of six numerical values are considered for the analysis. Observing the Scree plot, we can conclude that considering the **first two principal** components will capture roughly **90%** of the variance present in the data. PCA will help us reduce the dimensions to four from Six.



The Biplot shows the correlation among six different variables and the contributions towards PC1 and PC2.

PC 1 captures 70% of the variance. Qsec, mpg and drat are positively correlated to each other. It is similar with wt, disp and hp. All the variables have similar contributions to the variance as they all are closer to the circle. Disp and mpg contribute highest to the PC1 and qsec contributes highest to the PC2.

DATASET 3: PENGUINS

INTRODUCTION

The dataset provides data on three species of penguins collected from the Palmer Archipelago in Antarctica. There are a total of 344 observations on 8 variables. This dataset is available in the 'palmerpenguins' Package

UNIVARIATE ANALYSIS

DATA OVERVIEW

The detail of the dataset is given below:

Dataset	Package	Variables	Observations
penguins	palmerpenguins	8	344

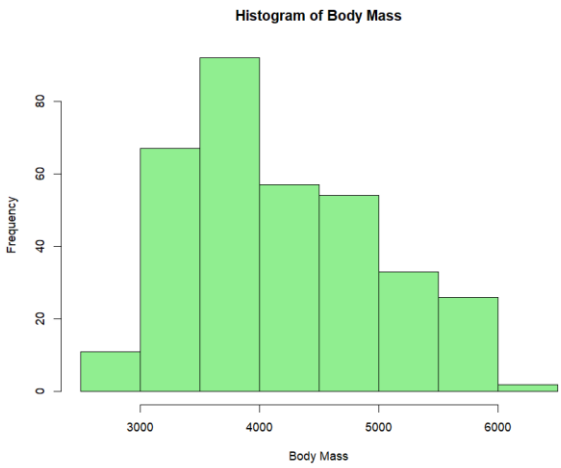
Two observations had **null values** in 6 of the 8 variables. These two observations are removed from the dataset. Also, 9 observations have the gender value missing. These were not treated as the gender variable in not being used for the analysis.

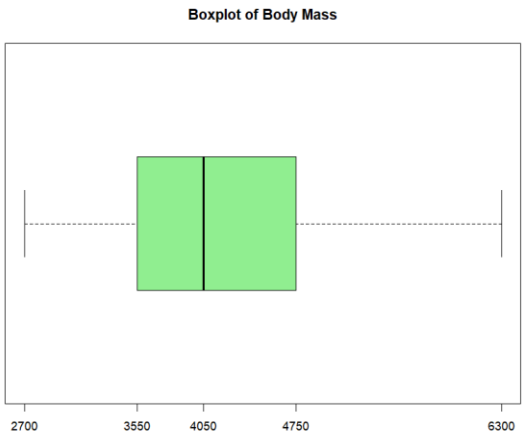
SUMMARY STATISTICS

Body mass of the penguins used for the numerical variable analysis.

Variable	Mean	Median	Standard Deviation	Minimum	Maximum	Missing Values
Body Mass (g)	4202	4050	801.95	2700	6300	0

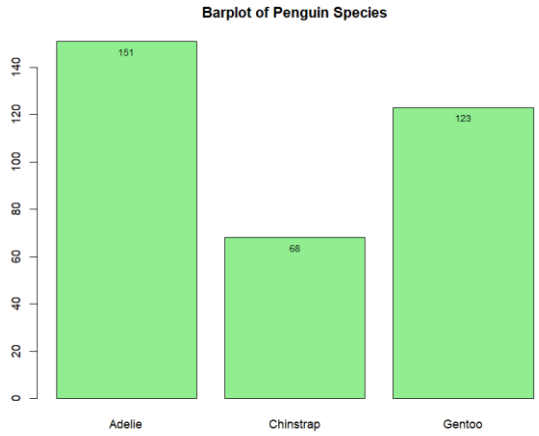
DISTRIBUTION VISUALIZATION





The body mass variable doesn't follow the normal distribution. As evident from the box plot, the data is right skewed (**Skewness = 0.47**). There are no outliers in the dataset**. 50%** of the observed species have their **body mass** in between **3.55kg and 4.75kg**

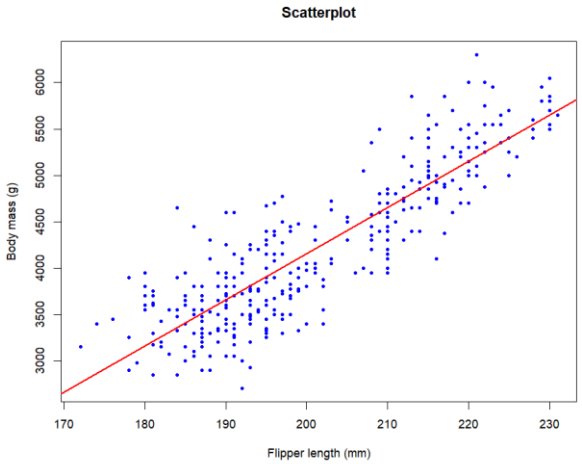
CATEGORICAL VARIABLE ANALYSIS



Penguin Species variable is considered for the analysis. Out of the 342 observations, **44%** (151) penguins belong to **Adelle**, **20%** (68) belong to **Chinstrap** and **36%** (123) belong to the **Gentoo** species.

MULTIVARIATE ANALYSIS

CORRELATION ANALYSIS AND SCATTER PLOT VISUALIZATION

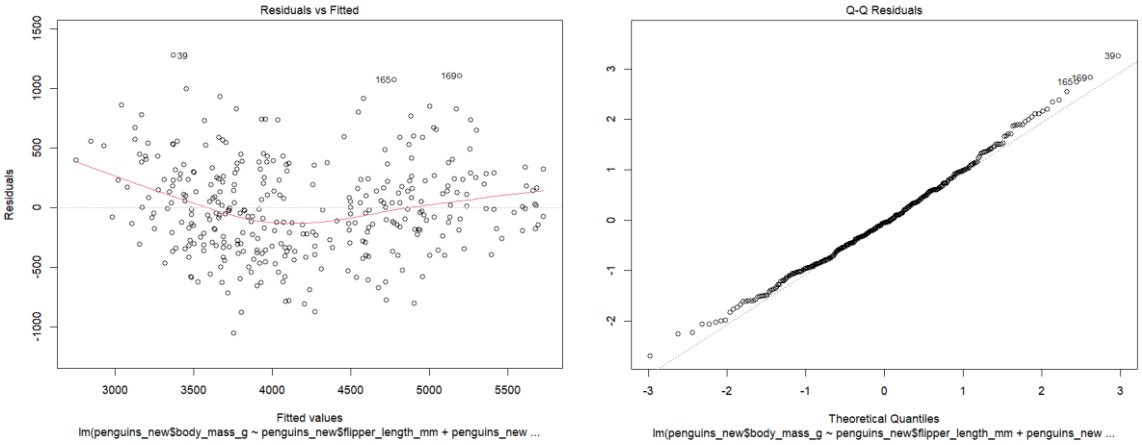


Correlation of **0.87** between the variable flipper length and the body mass variable suggests a **strong linear relationship** between the two.

MULTIPLE LINEAR REGRESSION AND DIAGNOSTICS

Understand the dependency of the body mass of the penguin on other body features, a Multiple Linear regression was performed. The **flipper length** of the penguin significantly **affects** the **body mass** (p value is much lesser than 0.05) **unlike** other features **bill length** and **bill depth**.

```
Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                -6424.765    561.469 -11.443   <2e-16 ***
penguins_new$flipper_length_mm    50.269      2.477  20.293   <2e-16 ***
penguins_new$bill_length_mm        4.162      5.329   0.781    0.435
penguins_new$bill_depth_mm        20.050     13.694   1.464    0.144
```

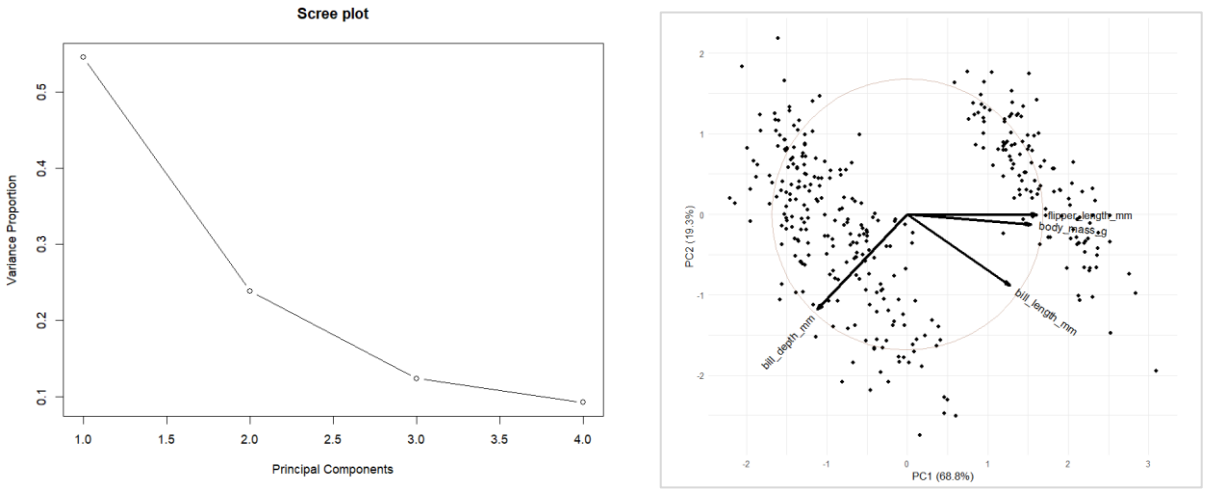


The plot shows that the residuals are **homoscedastic**. QQ plot shows that the residuals seem to follow the normal distribution.

ADVANCED ANALYSIS

PRINCIPAL COMPONENT ANALYSIS (PCA) AND INTERPRETATION

Four variables body mass, bill_length, bill_depth and flipper_length are considered for the principal component Analysis. From the Scree plot we can deduce that 90% of the variance can be captured using three principal components. Although, there is a single dimension reduction in the data, the data can now be visualized in a three-dimension.



The biplot shows the correlation between flipper length, bill length, bill depth and the body mass of the penguins. **All variable have similar contributions** to the variance as they all are closer to the circle. **PC 1 captures 69%** of the variance present in the data.

Flipper length, bill length and the body mass are positively correlated to each other. Bill length and bill depth are not related with each other.

Flipper Length and body mass contribute highest to the PC1. Bill_depth contributes highest to the PC1

DATASET 4: WINE

INTRODUCTION

The dataset contains the results of a chemical analysis of wines grown in the same region in Italy, derived from three different cultivars. The analysis determined the quantities of 13 chemical constituents found in each of the three types of wines. The dataset is available in the 'rattle' package

UNIVARIATE ANALYSIS

DATA OVERVIEW

The detail of the dataset is given below:

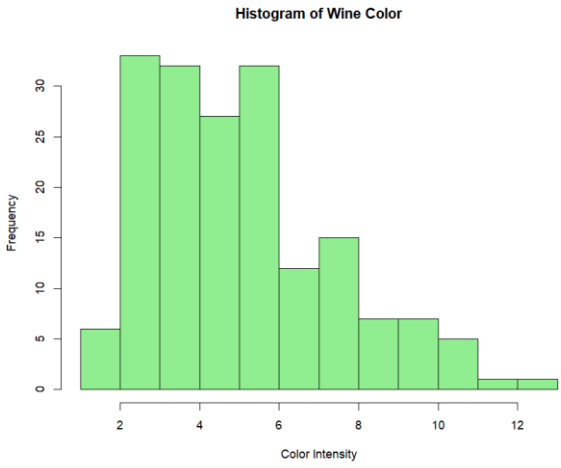
Dataset	Package	Variables	Observations
wine	rattle	13	178

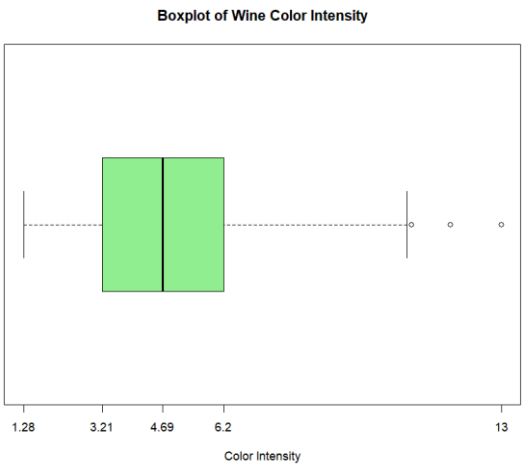
SUMMARY STATISTICS

Color numerical considered for the analysis:

Variable	Mean	Median	Standard Deviation	Minimum	Maximum	Missing Values
Color	5.05	4.69	1.79	1.28	13	0

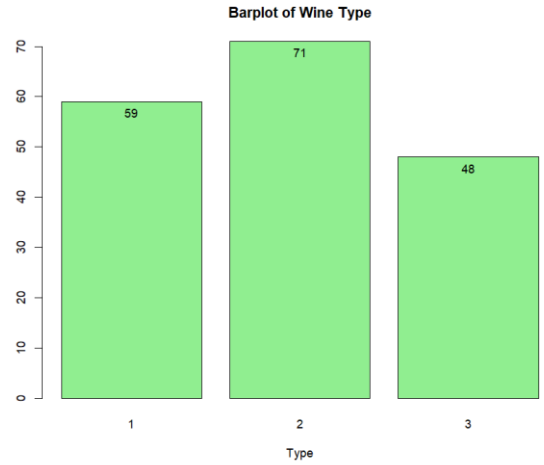
DISTRIBUTION VISUALIZATION





From the histogram we can infer that the color intensity of the wine is not following the normal distribution. This is further explained from the Box Plot which shows the positive skewness (**skewness = 0.86**). **Three outliers** are present in the color variable. 50% of the wine have color intensity between 3.2 and 6.2.

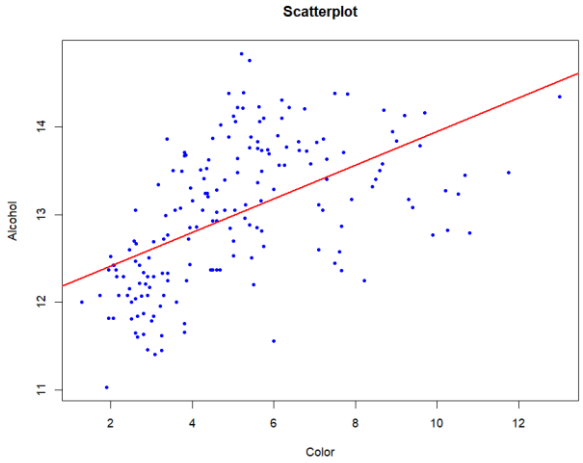
CATEGORICAL VARIABLE ANALYSIS



Three types of wine observed in the dataset. Nearly **40%** of the observations are of **type 2** and **33%** of the observations are of **type 1**

MULTIVARIATE ANALYSIS

CORRELATION ANALYSIS AND SCATTER PLOT VISUALIZATION

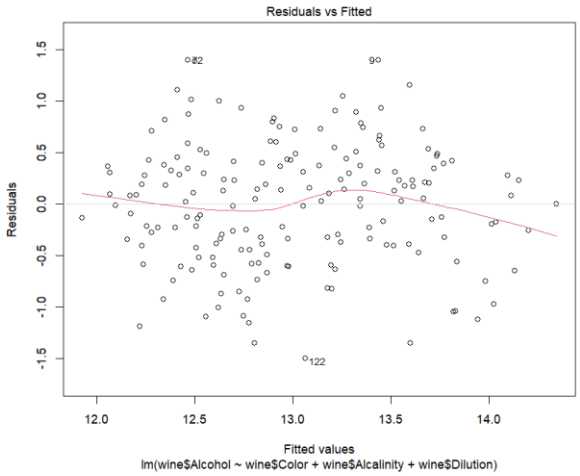


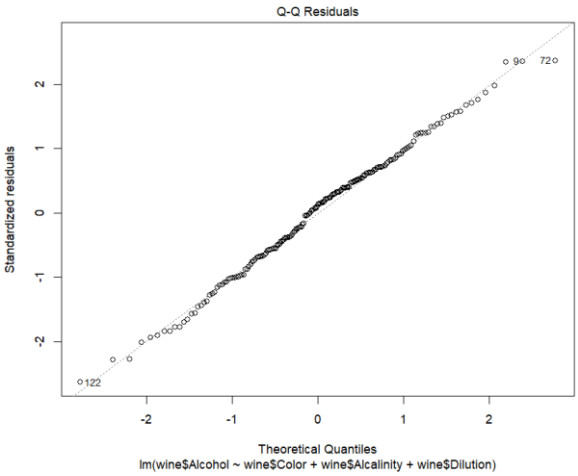
The correlation between color intensity of the wine and the alcohol level is **0.55,** which suggests a **weak linear relationship** between them. However, from the scatterplot, a **higher order relationship** might describe the relationship better.

MULTIPLE LINEAR REGRESSION AND DIAGNOSTICS

To understand the dependency of alcohol level on other features like color, dilution and Alkalinity, multiple linear regression is performed. From the results, we can observe that all the three features **significantly** contribute to the alcohol level. However, alkalinity affects inversely.

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    12.05628    0.42780  28.182  < 2e-16 ***
wine$Color      0.23742    0.02161  10.989  < 2e-16 ***
wine$Alcalinity -0.05856    0.01410  -4.153 5.13e-05 ***
wine$Dilution   0.33891    0.07340   4.617 7.54e-06 ***
```



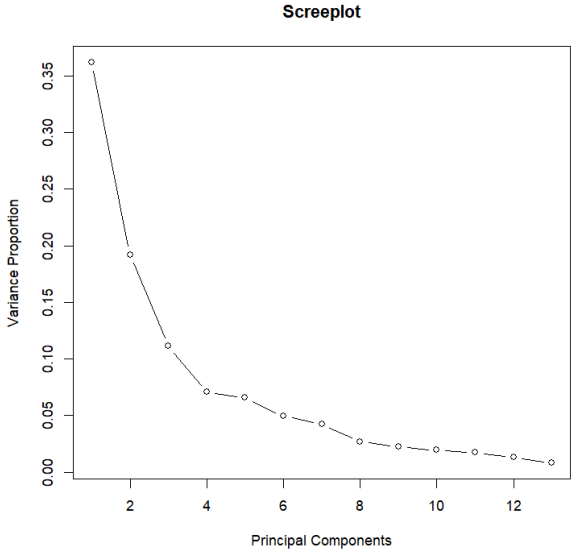


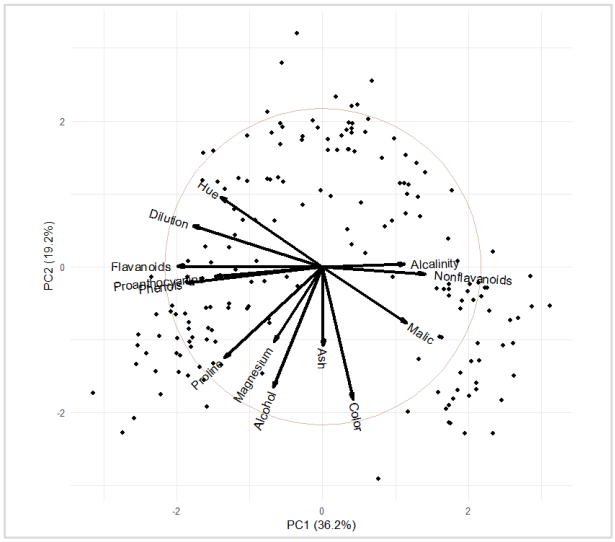
From the residual plot, the residuals don't seem to follow any patten suggesting that the residuals are **Homoscedastic**. The residuals follow the Normal distribution as evident from the QQ Plot.

ADVANCED ANALYSIS

PRINCIPAL COMPONENT ANALYSIS (PCA) AND INTERPRETATION

A total of 13 Wine attributes are considered for analysis to reduce the data to a lower dimension. As evident from the Screeplot, by considering 7 principal components, 90% of the variance can be captured.





The biplot illustrates relationships between chemical components in wine and their contributions to the first two principal components. **PC1 explains 36.2%** and PC2 explains 19.2% of the variability in the data.

Flavonoids and Phenols are strongly correlated and have significant and similar contributions to PC1. Nonflavonoids and Alkalinity show an inverse relationship with Flavonoids and Phenols.

Variables like Alcohol, Magnesium, and Colour are **grouped closely** suggesting some shared influence on variance