STATISTICAL STRUCTURES IN DATA

NUMERICAL ASSIGNMENT

Under the Guidance of

Prof. Subhajit Dutta

Name- **Neeraj Santosh Ahire**

Roll No- **24BM6JP36**

Date- **08/12/2024**

The following Datasets have been selected for analysis-

1. **Mtcars-** The data set was extracted from the 1974 *Motor Trend* US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models).

2. **Iris**- The data set gives the measurements in centimeters of the variable's sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are *Iris setosa*, *versicolor*, and *virginica*.

3. **Wholesale Customers-** The data set refers to clients of a wholesale distributor. It includes the annual spending in monetary units (m.u.) on diverse product categories.

4. **Concrete Compressive Strength**- The data set refers to clients of a wholesale distributor. It includes the annual spending in monetary units (m.u.) on diverse product categories.

Dataset 1 - MTCARS

Univariate Analysis-

1. Data overview-

```
Following is the result of the dataframe mtcars :
'data.frame':   32 obs. of  11 variables:
 $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
 $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
 $ disp: num  160 160 108 258 360 ...
 $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
 $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
 $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
 $ qsec: num  16.5 17 18.6 19.4 17 ...
 $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
 $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
 $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
 $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
Number of observations: 32
Number of variables: 11
```

Following is the structure of the dataset

Number of observations: 32
Number of variables: 11

2. Summary Statistics-

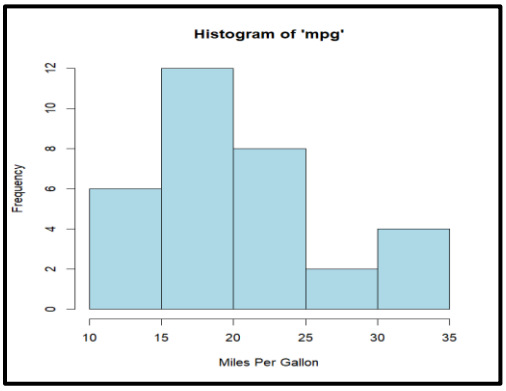
Following is the summary statistics of the variable Miles per Gallon (mpg)

```
Following are the summmary statistics of the variable mpg :
Mean: 20.09062
Median: 19.2
Standard Deviation: 6.026948
Minimum: 10.4
Maximum: 33.9
```

The mpg values in this dataset range widely from **10.4 to 33.9**, with a moderate spread. The slightly **higher mean compared to the median** suggests the presence of some higher mpg outliers pulling the average upwards and suggests a **slight right skew** in the data.

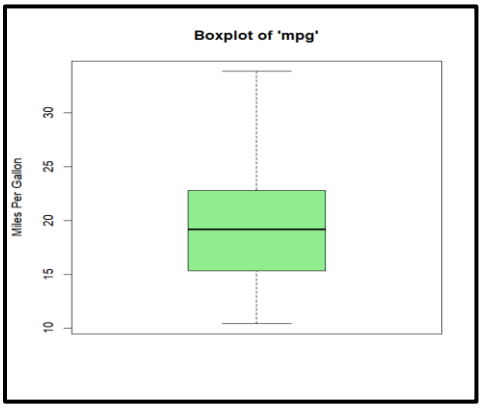
3. Distribution Visualization-

Histogram of the variable Miles per Gallon (mpg)



The histogram of 'mpg' shows that the data is **slightly right skewed** with presence of **potential outliers** which is in agreement with the summary statistics for the variable.

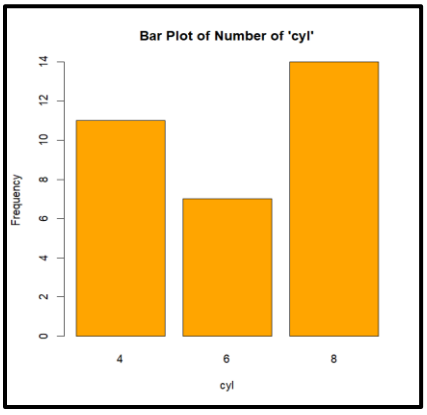
Boxplot of the variable Miles per Gallon (mpg)



The box plot further shows that the data is **slightly right skewed**, however, we don't observe any particular outliers as such.

4. Categorial Variable Analysis-

Bar plot of the variable Number of cylinders (cyl)



The dataset mainly features cars with 4 or 8 cylinders, reflecting a **mix of fuel-efficient designs and high-power** preferences common in the sampled vehicles.

Multivariate Analysis-

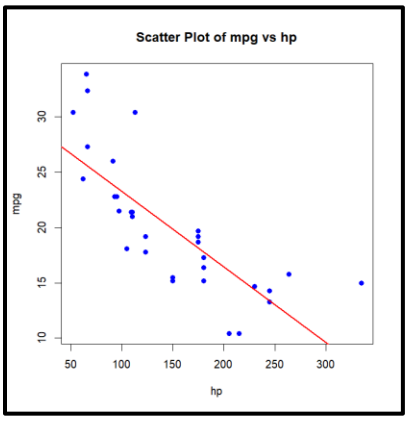
5. Correlation Analysis-

The selected variables were **Miles per Gallon (mpg) and Horsepower (hp)**.

Pearson Correlation Coefficient between mpg and hp: **-0.7761684**

This suggests that the mpg and hp are **negatively correlated** which is expected because high-performance engines (with higher horsepower) often prioritize power and speed over fuel efficiency.

6. **Scatter Plot visualization-**



The scatter plot indicates the mpg and hp are **negatively correlated** with a negatively sloped trend line. This is in agreement with the calculated Pearson Correlation Coefficient.

7. **Multiple Linear Regression-**

Selected variables-

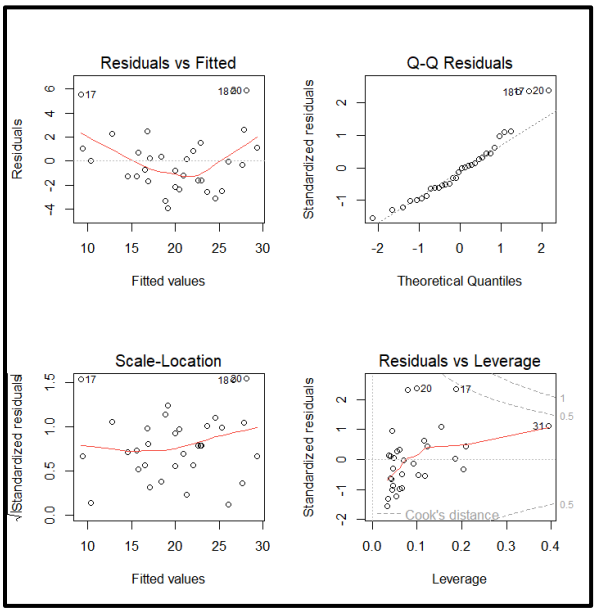
Y= 'mpg', X1= 'hp', X2 = 'wt'

```
Call:
lm(formula = df_y ~ df_x1 + df_x2)

Residuals:
   Min     1Q Median     3Q    Max
-3.941 -1.600 -0.182  1.050  5.854

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 37.22727    1.59879  23.285  < 2e-16 ***
df_x1       -0.03177    0.00903  -3.519  0.00145 **
df_x2       -3.87783    0.63273  -6.129 1.12e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.593 on 29 degrees of freedom
Multiple R-squared:  0.8268,    Adjusted R-squared:  0.8148
F-statistic: 69.21 on 2 and 29 DF,  p-value: 9.109e-12
```

1. The model explains a large portion of the variation in fuel efficiency (R-squared = 0.8268) and shows that **both weight (wt) and horsepower (hp) significantly affect miles per gallon (mpg).**

2. Both weight and horsepower reduce fuel efficiency, but **weight has a stronger negative impact than horsepower**, as seen in the larger coefficient value.

8. **Model Diagnostics-**



1. Residuals vs Fitted-

The residuals show a **slight curve** and increasing spread at higher fitted values, suggesting the model might miss some non-linear patterns and could have issues with uneven variance.

2. Q-Q Residuals-

The residuals mostly follow the 45-degree line, showing **approximate normality**, though slight deviations at the tails hint at potential outliers.

3. Scale-Location-

The red line shows a slight upward trend, suggesting that the variance of residuals may increase as the fitted values increase. This indicates **mild heteroscedasticity**.

4. Residuals vs Leverage-

Some points, like 17, 18, 20, and 30, have high leverage, but none are overly influential based on Cook's distance. Observation 30 might need closer attention.

Overall Assessment-

Residuals are approximately normally distributed, but there may be **slight issues with non-linearity and heteroscedasticity**.

Advanced Analysis-

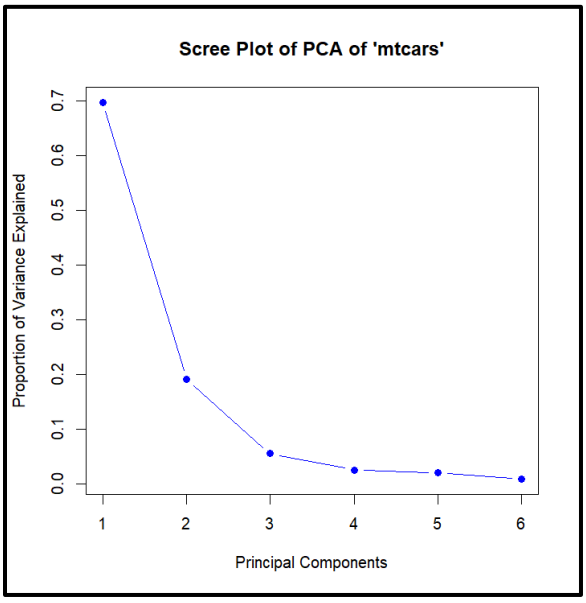
9. **Principal Component Analysis (PCA)-**

Variables selected for PCA-

"mpg", "disp", "hp", "drat", "wt", "qsec"

Following are the results of PCA-

```
Importance of components:
                          PC1    PC2     PC3     PC4    PC5     PC6
Standard deviation     2.0463 1.0715 0.57737 0.39289 0.3533 0.22799
Proportion of Variance 0.6979 0.1913 0.05556 0.02573 0.0208 0.00866
Cumulative Proportion  0.6979 0.8892 0.94481 0.97054 0.9913 1.00000
```



Since the Scree Plot has an elbow at 3 and the Cumulative Proportion of Variance crosses 0.9 as well, **the number of Principal components chosen should be 3**.

10. **PCA Interpretation-**

PCA loadings-

	PC1	PC2
mpg	-0.4586835	-0.05867609
disp	0.4660354	0.06065296
hp	0.4258534	-0.36147576
drat	-0.3670963	-0.43652537
wt	0.4386179	0.29953457
qsec	-0.2528320	0.76284877

Biplot in Appendix

PC-1:
PC1 captures the "power and size" dimension, with variables like engine size (disp), weight (wt), and horsepower (hp) contributing positively, while mpg and drat show a **trade-off** with power.

PC-2:
PC2 highlights acceleration (qsec) with a strong positive impact, contrasting with rear axle ratio (drat) and horsepower (hp), which contribute negatively.

Biplot observations-

i) MPG (fuel efficiency) and qsec (acceleration) are negatively correlated with weight and horsepower, showing a **trade-off between fuel economy and vehicle power**.
ii) Fuel-efficient, lightweight cars like Honda Civic and Toyota Corolla are grouped on the left, while heavy, powerful cars like Cadillac Fleetwood and Lincoln Continental are on the right.
iii) **Acceleration (qsec) stands apart from other variables, capturing a distinct aspect of car performance**. PC1 and PC2 together highlight key trade-offs between size, power (PC1) and acceleration (PC2).

Conclusion-

1. From the **univariate analysis**, the mpg variable shows a slightly **right-skewed** distribution with no significant outliers, indicating the presence of more cars with lower fuel efficiency.
2. The bar plot for cyl highlights the **dominance of cars with 4 and 8 cylinders**, reflecting a preference for either fuel efficiency or high power.
3. **Negative correlation** between mpg and hp suggests that higher horsepower results in reduced fuel efficiency, as expected in performance-oriented vehicles.
4. The multiple linear regression model confirms that **both weight (wt) and horsepower (hp) negatively affect mpg**, with weight having a stronger influence.
5. PCA highlights two key dimensions: **PC1 (size and power)** and **PC2 (acceleration)**. Cars are clustered by characteristics like fuel efficiency, engine power, and weight, with **clear trade-offs between performance and efficiency.**

Dataset 2- Iris

Univariate Analysis-

1. Data overview-

Following is the structure of the dataset

```
Following is the result of the dataframe iris :
'data.frame':   150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
Number of observations: 150
Number of variables: 5
```

Number of observations: 150
Number of variables: 5

2. Summary Statistics-

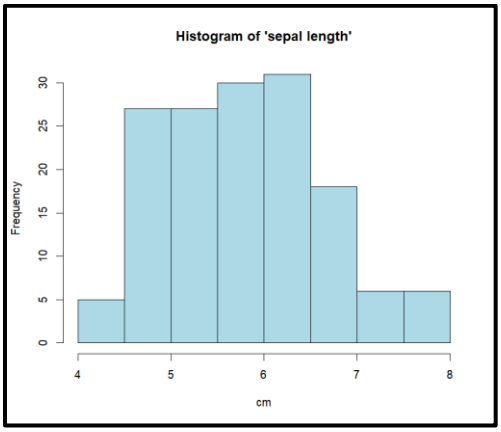
Following is the summary statistics of the variable Sepal length

```
Following are the summmary statistics of the variable sepal length (cm) :
Mean: 5.843333
Median: 5.8
Standard Deviation: 0.8280661
Minimum: 4.3
Maximum: 7.9
```

The average sepal length of the flowers is about **5.84 cm**, with most values ranging between **4.3 cm and 7.9 cm**, showing some variation but generally staying close to the mean, reflecting moderate differences in size across the flowers.

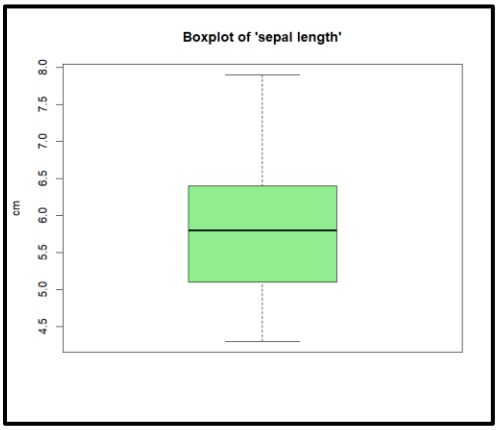
3. Distribution Visualization-

Histogram of the variable Sepal length



The histogram of sepal length shows that most flowers have sepal lengths between 5.5 cm and 6.5 cm, with a few stretching towards 6-7 cm. There's a bit more variation towards the larger side, but **overall, the sepal lengths are fairly consistent** with fewer flowers having very short or very long sepals.

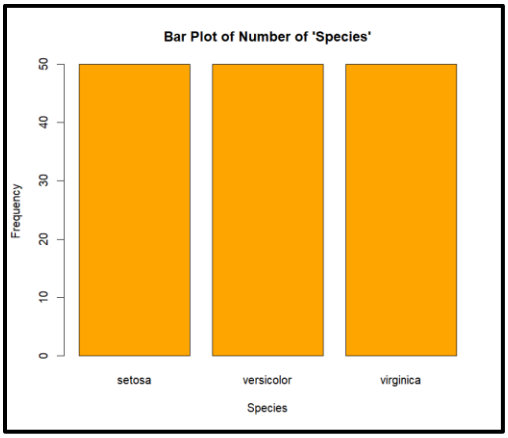
Boxplot of the variable Sepal length



The boxplot shows that most sepal lengths are centered around 6 cm, with a **tight range between 5.5 cm and 6.5 cm**. There are no significant outliers, and the **distribution is fairly even**, indicating consistent sepal lengths across the data.

4. Categorial Variable Analysis-

Bar plot of the variable Species



The bar plot indicates the data is **equally divided** into the 3 categories for the categorical variable ‘Species’.

Multivariate Analysis-

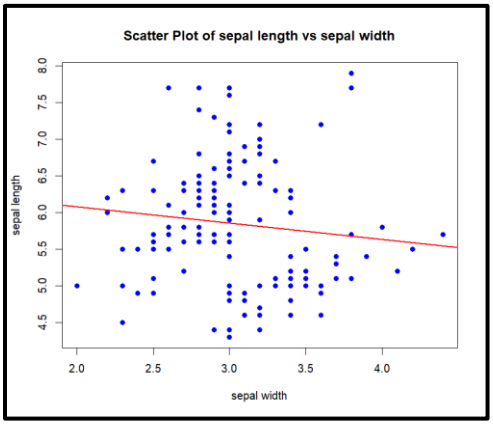
5. Correlation Analysis-

The selected variables were **Sepal length** and **Sepal width**.

Pearson Correlation Coefficient between Sepal length and Sepal width: **-0.1175698**

There is a **very slight negative relationship** between Sepal length and Sepal width. However, this relationship is weak, so the **two variables are mostly independent of each other**.

6. Scatter Plot visualization-



The scatter plot shows the **at best very mild negative correlation between** Sepal length and Sepal width with the mild negatively sloped linear fit line.

7. Multiple Linear Regression-

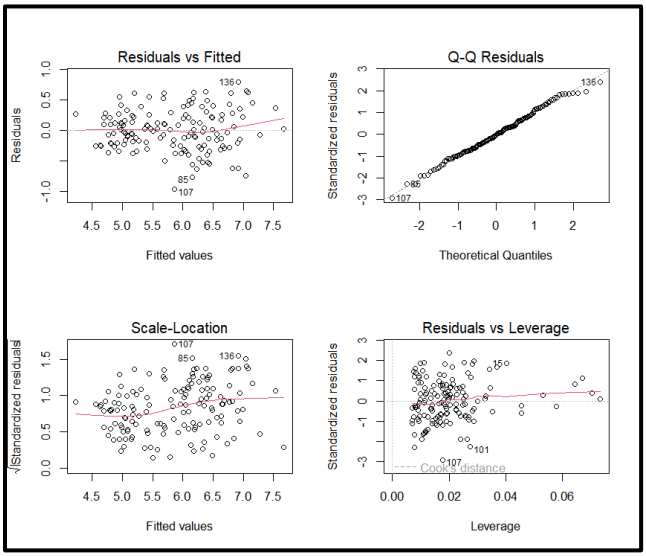
Selected variables-

Y= 'Sepal length', X1= 'Sepal width', X2 = 'Petal length'

```
Call:
lm(formula = df_y ~ df_x1 + df_x2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.96159 -0.23489  0.00077  0.21453  0.78557

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.24914    0.24797    9.07 7.04e-16 ***
df_x1        0.59552    0.06933    8.59 1.16e-14 ***
df_x2        0.47192    0.01712   27.57  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3333 on 147 degrees of freedom
Multiple R-squared:  0.8402,    Adjusted R-squared:  0.838
F-statistic: 386.4 on 2 and 147 DF,  p-value: < 2.2e-16
```

1. The model shows that Sepal width and Petal length are **strong predictors** of Sepal length. Both relationships are **highly significant**, indicating they strongly influence Sepal length.
2. The model explains **84%** of the variation in Sepal length, which is a **strong fit**. The predictions are quite accurate, with an average prediction error of only 0.33 cm. The model as a whole is very reliable, with a high F-statistic showing that both predictors together are highly useful for estimating Sepal length.

8. Model Diagnostics-



1. Residuals vs Fitted-

The residuals appear to be randomly scattered, suggesting that the **assumption of linearity is reasonable**. However, **there is a slight curvature, indicating potential non-linearity**.

2. Q-Q Residuals-

The points mostly lie along the reference line, indicating that the residuals are approximately normally distributed. There are a few deviations at the tails, but they are not severe.

3. Scale-Location-

the residuals appear to be spread equally, suggesting that the assumption of homoscedasticity is reasonable. There is a slight increase in spread at higher fitted values, but it is not very pronounced.

4. Residuals vs Leverage-

Most points have low leverage, and there are no points with both high leverage and large residuals. The Cook's distance lines indicate that there are no highly influential points.

Overall Assessment-

Diagnostic plots suggest that the **regression model assumptions are reasonably met**, although with some very minor deviations

Advanced Analysis-

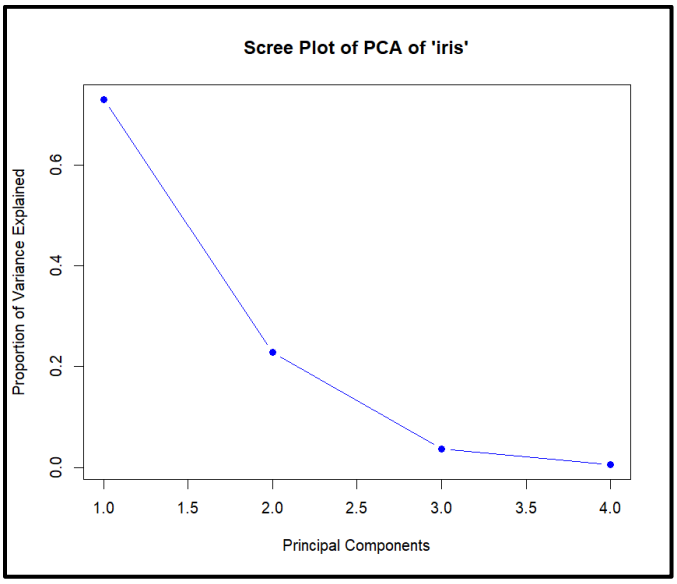
9. Principal Component Analysis (PCA)-

Variables selected for PCA-

"Sepal length", "Sepal width", "Petal length", "Petal width"

Following are the results of PCA-

```
Importance of components:
                          PC1    PC2     PC3     PC4
Standard deviation     1.7084 0.9560 0.38309 0.14393
Proportion of Variance 0.7296 0.2285 0.03669 0.00518
Cumulative Proportion  0.7296 0.9581 0.99482 1.00000
```



The Scree Plot has an elbow at 3 but the Cumulative Proportion of Variance crosses 0.9 at PC 2 itself, so, the **number of Principal components chosen should be 2**.

10. **PCA Interpretation-**

PCA loadings-

	PC1	PC2
Sepal.Length	0.5210659	-0.37741762
Sepal.Width	-0.2693474	-0.92329566
Petal.Length	0.5804131	-0.02449161
Petal.Width	0.5648565	-0.06694199

Biplot in Appendix

PC-1:
PC1 seems to capture the **"overall size and flower dimension"** with variables like **Sepal length**, **Petal length**, and **Petal width** contributing positively, while **Sepal width** shows a trade-off with size.

PC-2:
PC2 highlights the **"shape and spread dimension"** with **Sepal width** contributing strongly and negatively, contrasting with **Sepal length**, **Petal length**, and **Petal width**, which contribute positively.

Biplot observations-

i) **Sepal width** is **negatively correlated** with **Sepal length**, **Petal length**, and **Petal width**, showing a trade-off between Sepal width and the other dimensions of the flowers.

ii) **Petal length** and **Petal width** stand apart from the other variables, capturing distinct aspects of the flower's morphology. **PC1 highlights trade-offs between size attributes** like petal length and width, while P**C2 focuses more on the variation in sepal width**.

Conclusion-

1. From the univariate analysis, the variable Sepal length shows **moderate variation** at best, with most flowers having sepal lengths around 6 cm. The distribution is fairly consistent with no significant outliers.
2. The bar plot for Species shows an **equal distribution** across the three species, indicating balanced representation in the dataset.
3. A **very weak negative correlation between Sepal length and Sepal width suggests that these variables are mostly independent of each other**, as confirmed by the scatter plot.
4. The multiple linear regression model highlights that both **Sepal width and Petal length are strong predictors of Sepal length**, explaining 84% of its variation with a good model fit and minimal prediction error.
5. PCA identifies two key dimensions: **PC1 (overall size and flower dimensions)** and **PC2 (shape and spread)**, highlighting trade-offs between sepal width and other dimensions like petal length and width.

Dataset 3- Wholesale customers

Univariate Analysis-

1. Data overview-

Following is the structure of the dataset

```
Following is the result of the dataframe Wholesale customers :
'data.frame':   440 obs. of  8 variables:
 $ Channel         : int  2 2 2 1 2 2 2 2 1 2 ...
 $ Region          : int  3 3 3 3 3 3 3 3 3 3 ...
 $ Fresh           : int  12669 7057 6353 13265 22615 9413 12126 7579 5963 6006 ...
 $ Milk            : int  9656 9810 8808 1196 5410 8259 3199 4956 3648 11093 ...
 $ Grocery         : int  7561 9568 7684 4221 7198 5126 6975 9426 6192 18881 ...
 $ Frozen          : int  214 1762 2405 6404 3915 666 480 1669 425 1159 ...
 $ Detergents_Paper: int  2674 3293 3516 507 1777 1795 3140 3321 1716 7425 ...
 $ Delicassen      : int  1338 1776 7844 1788 5185 1451 545 2566 750 2098 ...
Number of observations: 440
Number of variables: 8
```

Number of observations: 440
Number of variables: 8

2. Summary Statistics-

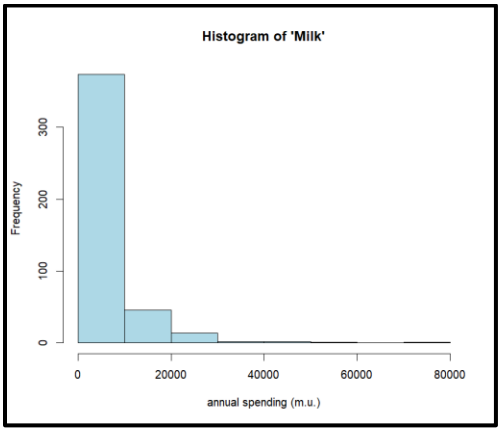
Following is the summary statistics of the variable Milk

```
Following are the summmary statistics of the variable Milk (annual spending (m.u.)) :
Mean: 5796.266
Median: 3627
Standard Deviation: 7380.377
Minimum: 55
Maximum: 73498
```

Households spend an average of **5796.27 m.u.** annually on milk, but most spend closer to **3627 m.u.**, showing that a few households with very high spending pull the average up. The wide range (**55 to 73,498 m.u.**) and large variation in spending suggest **big differences** in milk consumption across households.

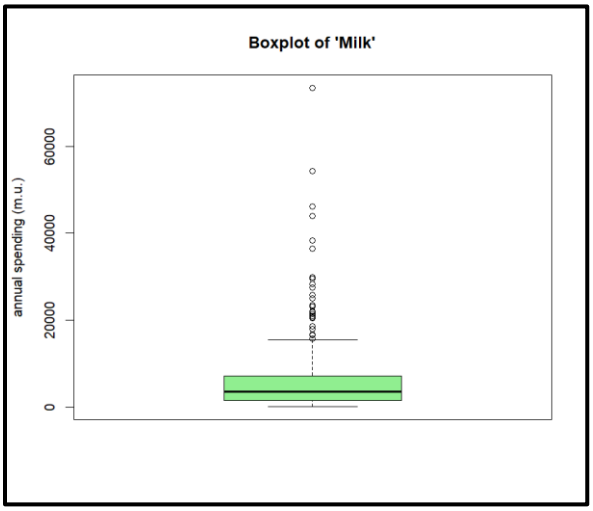
3. Distribution Visualization-

Histogram of the variable Milk



Most people spend between **0 and 20,000 m.u.** on milk annually. This range has the highest frequency, with over 300 occurrences.
The distribution is **right-skewed**, with a long tail extending towards higher spending values, up to 80,000 m.u. This suggests that while most people spend relatively little on milk, **there are a few outliers who spend significantly more**.

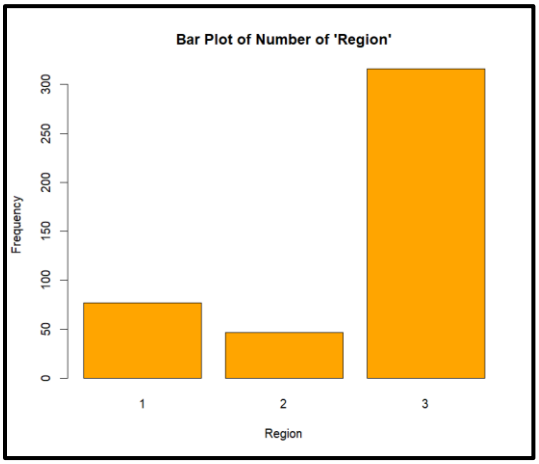
Boxplot of the variable Sepal length



The **majority of spending falls between 10,000 and 30,000 m.u., but there are some notable outliers**. These outliers indicate that a few individuals spend significantly more than the rest, suggesting a wide range of milk purchasing behaviors. The overall trend is right-skewed, highlighting that while most people spend moderately, **there are occasional big spenders**. The observations are consistent with those before.

4. Categorial Variable Analysis-

Bar plot of the variable Region



The data is **heavily skewed** and majority of customers belong to **Region-3** hence, Region 3 is largest and most dominant customer base.

Multivariate Analysis-

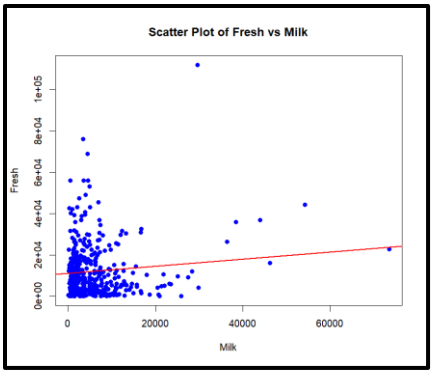
5. Correlation Analysis-

The selected variables were **Fresh** and **Milk**.

Pearson Correlation Coefficient between Fresh and Milk: **0.1005098**

There is a **very slight positive relationship** between Sepal length and Sepal width. However, this **relationship is weak, so the two variables are mostly independent** of each other.

6. Scatter Plot visualization-



The scatter plot shows the **at best very mild positive correlation** between Sepal length and Sepal width with the mild positively sloped linear fit line.

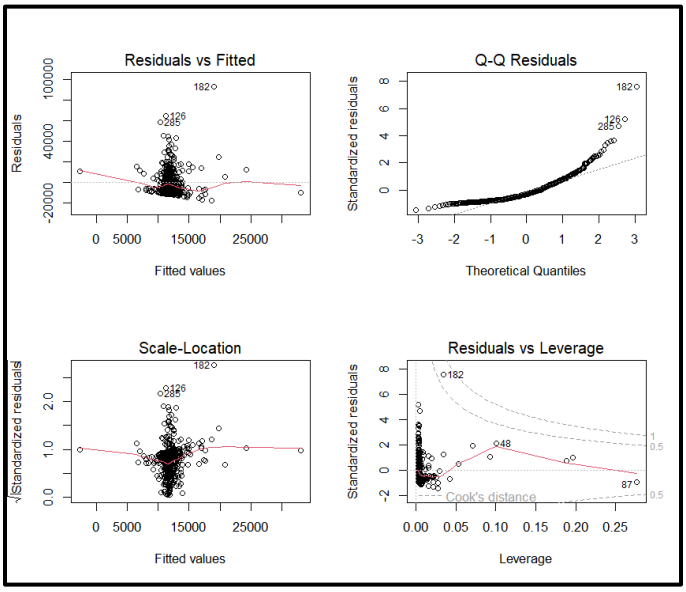
7. Multiple Linear Regression-

Selected variables-
Y= 'Fresh', X1= 'Milk', X2 = 'Grocery'

```
Call:
lm(formula = df_y ~ df_x1 + df_x2)

Residuals:
   Min     1Q Median     3Q    Max
-17884  -8636  -3749   5251  93116

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.161e+04  7.932e+02  14.636  < 2e-16 ***
df_x1        3.983e-01  1.181e-01   3.373  0.00081 ***
df_x2       -2.411e-01  9.172e-02  -2.629  0.00887 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12510 on 437 degrees of freedom
Multiple R-squared:  0.02551,   Adjusted R-squared:  0.02105
F-statistic:  5.72 on 2 and 437 DF,  p-value: 0.00353
```

1. The model shows that spending on **Milk has a positive impact on Fresh spending, while Grocery spending has a small negative impact**. Both are statistically significant, meaning their influence on Fresh spending is not by chance. However, their combined effect is quite small, explaining only about 2.5% of the variation in Fresh spending.
2. While Milk and Grocery spending do affect Fresh spending slightly, t**here are likely other factors that play a much bigger role**. The model suggests more exploration is needed to understand what really drives Fresh spending.

8. Model Diagnostics-



1. Residuals vs Fitted-
Ideally, the points should be all over the place with no obvious pattern. Here, though, **there's a bit of a curve**, hinting that the relationship between spending on Milk, Grocery, and Fresh isn't perfectly straightforward.
2. Q-Q Residuals-
Most of the points line up nicely along the straight line, which means our model's assumptions are mostly right. **But there are a few points that stray off course**, especially at the edges.
3. Scale-Location-
The spread of residuals is relatively even, indicating that the **assumption of homoscedasticity (constant variance) is mostly valid**.
4. Residuals vs Leverage-
Most points have low leverage, and there aren't any points with both high leverage and large residuals. The Cook's distance lines show there are no highly influential points.
Overall Assessment-
The diagnostic plots suggest that the regression model assumptions are reasonably met, though there are some minor deviations. Finally, Milk and Grocery spending have a significant but very small combined effect on Fresh spending.

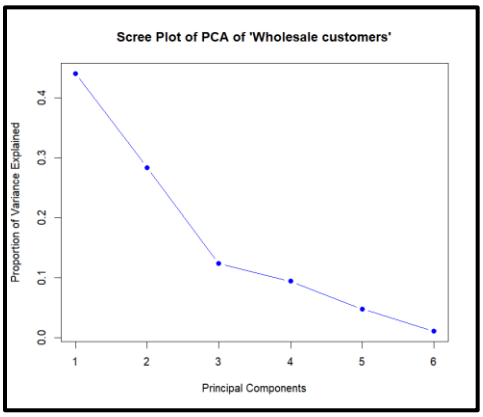
Advanced Analysis-

9. Principal Component Analysis (PCA)-

Variables selected for PCA-
"Fresh", "Milk", "Grocery", "Frozen", "Detergents_Paper", "Delicassen"

Following are the results of PCA-

```
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6
Standard deviation     1.6263 1.3048 0.8603 0.75082 0.53449 0.2509
Proportion of Variance 0.4408 0.2838 0.1233 0.09396 0.04761 0.0105
Cumulative Proportion  0.4408 0.7246 0.8479 0.94189 0.98950 1.0000
```



There appears to be an elbow at number of Principal Components- 3 but the cumulative variance explained is about 84.8 % so, **the number of principal components selected should be 4**.

10. **PCA Interpretation-**

PCA loadings-

	PC1	PC2
Fresh	-0.04288396	-0.52793212
Milk	-0.54511832	-0.08316765
Grocery	-0.57925635	0.14608818
Frozen	-0.05118859	-0.61127764
Detergents_Paper	-0.54864020	0.25523316
Delicassen	-0.24868198	-0.50420705

Biplot in Appendix

PC-1:
PC1 seems to capture the "overall size and quantity dimension" with variables like **Milk, Grocery, and Detergents_Paper contributing negatively**. This indicates that higher values in these variables are associated with lower PC1 scores, suggesting a trade-off between these product categories and the overall size of purchases.

PC-2:
PC2 probably highlights the "freshness and preservation dimension" with variables like **Fresh, Frozen, and Delicassen contributing negatively**. This suggests that higher values in these variables are associated with lower PC2 scores, indicating a focus on fresh and preserved products.

Biplot observations-

i) PC1 highlights trade-offs between size attributes like Milk, Grocery, and Detergents_Paper, while PC2 focuses more on the variation in fresh and preserved products.

Conclusion-

1. From the univariate analysis, annual **Milk spending shows significant variation**, with most households spending around 3627 m.u., but a few big spenders skew the average up to 5796.27 m.u. The right-skewed distribution and notable outliers highlight diverse spending habits.
2. The bar plot for 'Region' indicates that **Region-3 dominates the customer base**, reflecting a highly skewed distribution of regional representation.
3. A **very weak positive correlation between Fresh and Milk spending suggests these variables are largely independent**, as seen in both the correlation coefficient and scatter plot.
4. The multiple linear regression model shows that Milk spending positively impacts Fresh spending, while Grocery spending has a small negative effect. However, these variables collectively explain only 2.5% of the variation, **indicating other factors play a larger role**.
5. PCA identifies two key dimensions: **PC1 (size and quantity of purchases)** and **PC2 (freshness and preservation)**, highlighting trade-offs between categories like Milk, Grocery, and Detergents_Paper, and variables like Fresh, Frozen, and Delicassen.

Dataset 4- Concrete Compressive Strength

Univariate Analysis-

1. Data overview-

Following is the structure of the dataset

```
Following is the result of the dataframe Concrete compressive strength :
tibble [1,030 × 9] (S3: tbl_df/tbl/data.frame)
 $ Cement (component 1)(kg in a m^3 mixture)            : num [1:1030] 540 540 332 332 199 ...
 $ Blast Furnace Slag (component 2)(kg in a m^3 mixture): num [1:1030] 0 0 142 142 132 ...
 $ Fly Ash (component 3)(kg in a m^3 mixture)           : num [1:1030] 0 0 0 0 0 0 0 0 0 0 ...
 $ Water  (component 4)(kg in a m^3 mixture)            : num [1:1030] 162 162 228 228 192 228 228
228 228 228 ...
 $ Superplasticizer (component 5)(kg in a m^3 mixture)  : num [1:1030] 2.5 2.5 0 0 0 0 0 0 0 0 ...
 $ Coarse Aggregate  (component 6)(kg in a m^3 mixture) : num [1:1030] 1040 1055 932 932 978 ...
 $ Fine Aggregate (component 7)(kg in a m^3 mixture)    : num [1:1030] 676 676 594 594 826 ...
 $ Age (day)                                            : num [1:1030] 28 28 270 365 360 90 365 28 28
28 ...
 $ Concrete compressive strength(MPa, megapascals)      : num [1:1030] 80 61.9 40.3 41.1 44.3 ...
Number of observations: 1030
Number of variables: 9
```

Number of observations: 1030
Number of variables: 9

2. Summary Statistics-

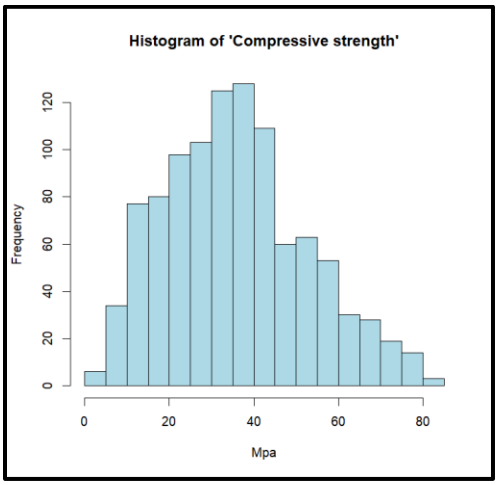
Following is the summary statistics of the variable Concrete Compressive strength

```
Following are the summmary statistics of the variable Concrete compressive strength(MPa, megapascals)
:
Mean: 35.81784
Median: 34.44277
Standard Deviation: 16.70568
Minimum: 2.331808
Maximum: 82.59922
```

The concrete samples have an average compressive strength of about **35.82 MPa**, with most values spanning from **2.33 MPa to 82.60 MPa**, showing a **wide range** of strengths. While many samples cluster around the middle, the slight difference between the mean and median suggest that a few particularly strong samples are boosting the overall average. This variation highlights the diversity in the concrete's performance.

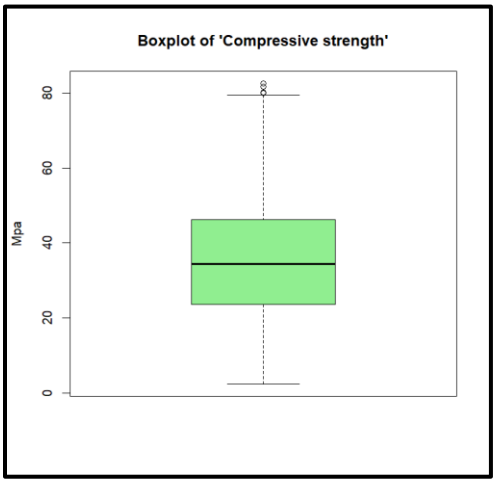
3. Distribution Visualization-

Histogram of the variable Concrete Compressive strength



Most concrete samples have compressive strengths between **30 and 50 Mpa**. The distribution of the **data is pretty balanced**, with fewer samples having very low (0-10 Mpa) or very high (70-80 Mpa) compressive strengths. This means that while most samples are in the mid-range, there are a few outliers with either much lower or much higher strengths.

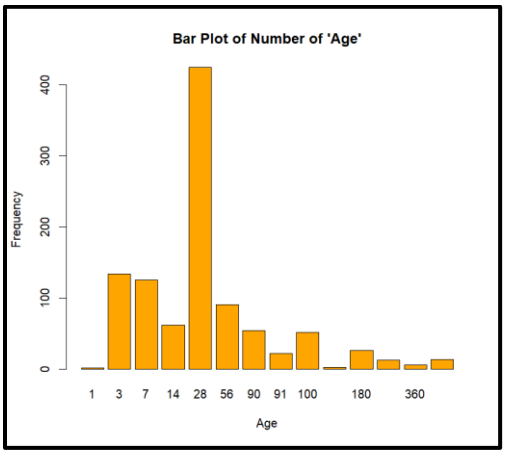
Boxplot of the variable Concrete Compressive strength



Most concrete samples have compressive strengths around ranging from **25 to 50 Mpa**. The median is close to **35 Mpa**. There are a few outliers above 75 Mpa, indicating some exceptionally strong samples. This boxplot shows that while most concrete samples are consistent, there are a few extreme values.

4. Categorial Variable Analysis-

Bar plot of the variable Age (Age can be considered a categorical variable for the data)



The bar plot shows that the **majority of concrete samples are 28 days old**, making this the most common testing age. This indicates that **28 days is the standard curing period** for concrete compressive strength tests, which is usually the case.

Multivariate Analysis-

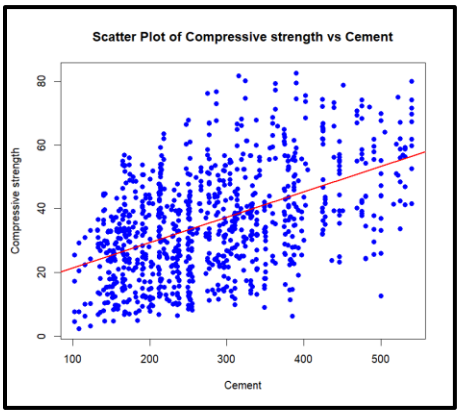
5. Correlation Analysis-

The selected variables were **Concrete Compressive strength** and **Cement content**.

Pearson Correlation Coefficient between Compressive strength and Cement content: **0.4978327**

The correlation coefficient of approximately 0.498 indicates a moderate positive relationship between cement content and concrete compressive strength.

6. Scatter Plot visualization-



The scatter plot shows the **moderate positive correlation** between Concrete compressive strength and Cement content with the moderately positively sloped linear fit line.

7. Multiple Linear Regression-

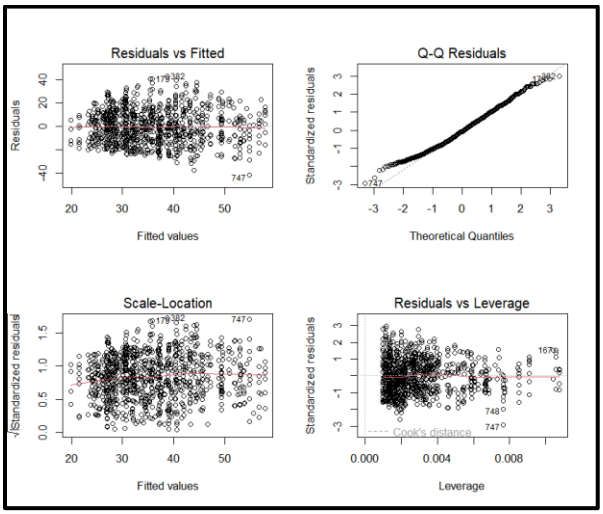
Selected variables-

Y= 'Concrete Compressive strength', X1= 'Cement content', X2 = 'Fine aggregate content'

```
Call:
lm(formula = df_y ~ df_x1 + df_x2)

Residuals:
    Min      1Q  Median      3Q     Max
-42.117 -11.145  -0.820   9.697  42.959

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.597090   4.915386   4.801 1.82e-06 ***
df_x1        0.077468   0.004428  17.497  < 2e-16 ***
df_x2       -0.012359   0.005771  -2.141   0.0325 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.47 on 1027 degrees of freedom
Multiple R-squared:  0.2512,    Adjusted R-squared:  0.2497
F-statistic: 172.2 on 2 and 1027 DF,  p-value: < 2.2e-16
```

1. The model shows that **Cement content is a strong predictor of Concrete Compressive strength**, with a highly significant positive relationship. **Fine aggregate content also influences Concrete Compressive strength, but its impact is smaller and negative**. Both relationships are statistically significant.
2. The model explains approximately 25.12% of the variation in Concrete Compressive strength, which indicates a **moderate fit**. The model as a whole is statistically significant, with a high F-statistic (172.2) indicating that both predictors together are useful for estimating Concrete Compressive strength.

8. Model Diagnostics-



1. Residuals vs Fitted-

Ideally, the points should be all over the place with no obvious pattern. Here, though, the **variance appears to change**, hinting that the relationship between Concrete Compressive Strength, Cement Content, and Fine Aggregate Content **isn't entirely linear**.

2. Q-Q Residuals-

Most of the points line up nicely along the straight line, which means our model's assumptions are mostly right. But there are a few points that stray off course, especially at the edges.

3. Scale-Location-

The spread of residuals is relatively even, indicating that the **assumption of homoscedasticity (constant variance) could be considered valid**.

4. Residuals vs Leverage-

Most points have low leverage, and there aren't any points with both high leverage and large residuals.

Overall Assessment-

The diagnostic plots suggest that the regression model **assumptions are reasonably met**, though there are some minor deviations. Cement Content and Fine Aggregate Content have a significant effect on Concrete Compressive Strength. Their combined impact is moderate.

Advanced Analysis-

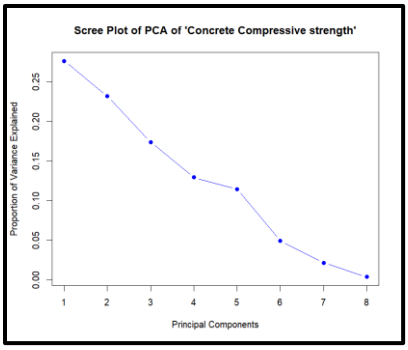
9. Principal Component Analysis (PCA)-

Variables selected for PCA-

"Cement", "Blast Furnace Slag", "Fly Ash", "Water", "Superplasticizer", "Coarse Aggregate", "Fine Aggregate", "Concrete compressive strength"

Following are the results of PCA-

```
Importance of components:
                          PC1    PC2    PC3    PC4    PC5     PC6     PC7     PC8
Standard deviation     1.4868 1.3614 1.1791 1.0170 0.9571 0.62686 0.41545 0.17227
Proportion of Variance 0.2763 0.2317 0.1738 0.1293 0.1145 0.04912 0.02157 0.00371
Cumulative Proportion  0.2763 0.5080 0.6818 0.8111 0.9256 0.97472 0.99629 1.00000
```



From the Scree plot there isn't any obvious elbow in the plot. However, we can select **the number of principal components selected to be 5**, since, the cumulative variance crosses 0.9 for PC5

10. PCA Interpretation-

PCA loadings-

	PC1	PC2
Cement (component 1)(kg in a m^3 mixture)	-0.06395444	0.55371866
Blast Furnace Slag (component 2)(kg in a m^3 mixture)	0.16132604	0.24450767
Fly Ash (component 3)(kg in a m^3 mixture)	-0.32575877	-0.39276187
Water (component 4)(kg in a m^3 mixture)	0.57775207	0.01335743
Superplasticizer (component 5)(kg in a m^3 mixture)	-0.57759184	0.12770929
Coarse Aggregate (component 6)(kg in a m^3 mixture)	0.09698076	-0.23754736
Fine Aggregate (component 7)(kg in a m^3 mixture)	-0.34692268	-0.29626312
Concrete compressive strength(MPa, megapascals)	-0.25805923	0.56450075

Biplot in Appendix

PC-1:

PC1 appears to represent a **"water and superplasticizer dimension"** with **Water** and **Superplasticizer** contributing strongly but in opposite directions. **Water has a positive contribution**, while **Superplasticizer has a negative contribution**. This could highlight a **trade-off** between the need for water and the use of chemical additives which is expected since superplasticizers are used as alternatives to reduce water content in concrete mixtures.

PC-2:

PC2 seems to emphasize a **"cement and compressive strength dimension"** with **Cement** and **Concrete compressive strength** contributing positively. This indicates that mixtures with more cement and higher compressive strength go together confirming their positive relationship. Conversely, **Fly Ash** and **Fine Aggregate** contribute negatively and implying lower compressive strength as seen before with the negative correlation of Fine Aggregate and Concrete Compressive Strength.

Biplot observations-

i) PC1 captures trade-offs between **Water** and **Superplasticizer.**

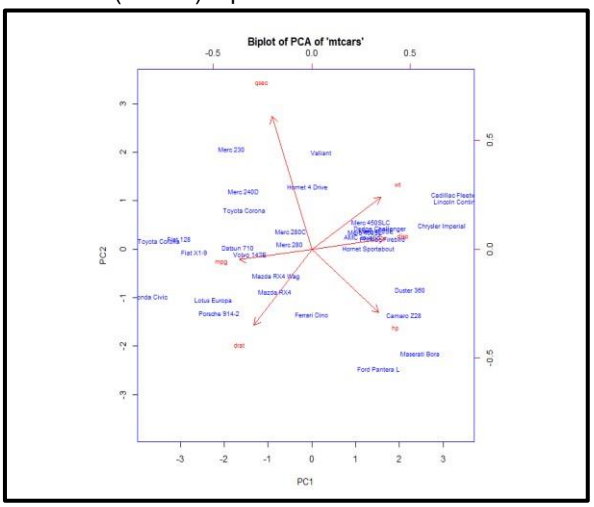
ii) PC2 highlights the influence of **Cement** and **Compressive Strength**, suggesting the importance of primary binding agents in determining the strength of the mixture.

Conclusion-

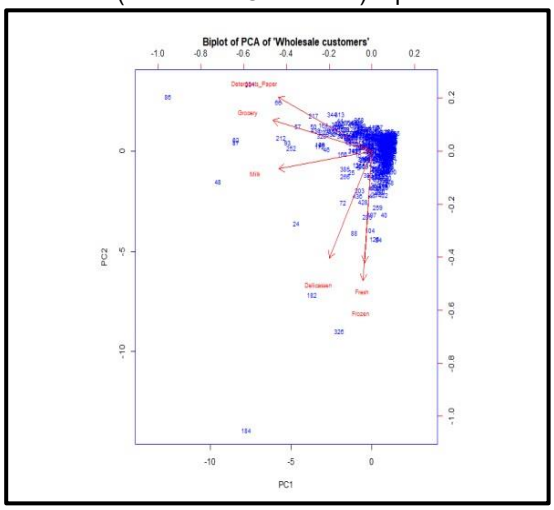
1. The univariate analysis shows that concrete compressive strength varies widely, with most samples falling between **30 MPa and 50 MPa**. While the average is **35.82 MPa**, a few very strong samples push this number higher, showing some variability in the data.
2. The bar plot for 'Age' highlights that **most concrete samples are tested after 28 days**, which is standard practice for assessing strength after curing.
3. Cement content has a **moderate positive relationship** with compressive strength (correlation: 0.498), meaning concrete tends to get stronger as more cement is used in the mix.
4. The regression model confirms that **cement content is the biggest factor influencing compressive strength**, positively contributing to it. Fine aggregate content has a smaller negative impact, and the two together explain a moderate portion of the strength variation.
5. PCA highlights two main patterns: the first focuses on the **trade-off between water and superplasticizer**, showing how they balance each other in the mix. The second emphasizes the **strong connection between cement and compressive strength**, reinforcing the importance of cement as a key ingredient in achieving stronger concrete.

Appendix

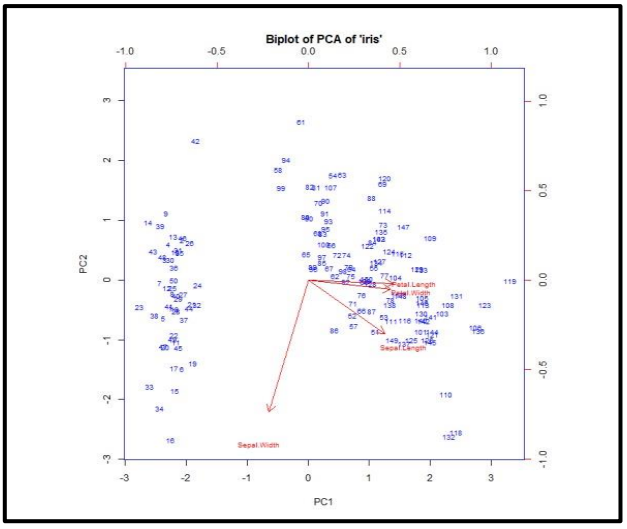
Dataset-1 (Mtcars) Biplot-



Dataset-2 (Wholesale Customers) Biplot-



Dataset-2 (Iris) Biplot-



Dataset-4 (Concrete data) Biplot-