INDIAN STATISTICAL INSTITUTE

POST GRADUATE DIPLOMA IN BUSINESS ANALYTICS

(PGDBA)

Statistical Structures in Data

Numerical Assignment

Instructor

Professor Subhajit Datta

Name: Ankamreddi Praveen

ROLL NO. 24BM6JP07

Wine quality

1. Data Overview

```
'data.frame':   6497 obs. of  13 variables:
 $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
 $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
 $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
 $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
 $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071
...
 $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
 $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
 $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
 $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...
 $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...
 $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
 $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...
 $ type                : chr  "red" "red" "red" "red" ...
```

Figure 1.Overview of the Data

The Wine Quality dataset consists of 6497 samples of wine with 13 attributes, including chemical properties like acidity, sugar, alcohol content, pH, and others, as well as a target variable, quality, which indicates the quality of the wine (ranging from 0 to 10). The dataset is split between **red** and **white wines**

2. Summary Statistics(residual.sugar)

The statistical properties of the variable – "residual.sugar" are tabulated in (Table.1).

1	Mean	5.443235
2	Median	3
3	Standard Deviation	4.757804
4	Minimum	0.6
5	Maximum	65.8

3.Distribution Visualization

From the Histogram, it can be inferred that the data is **Right Skewed**, which explains the median being close to the min.value than the max.value. The data is suggesting that most wines have low sugar content, but there are a few wines with higher sugar content.

Also from the box plot it is evident that the data is right skewed as the **median** value is **closer to Q1**. And there are **outliers** present in the data which is evident from the Boxplot.

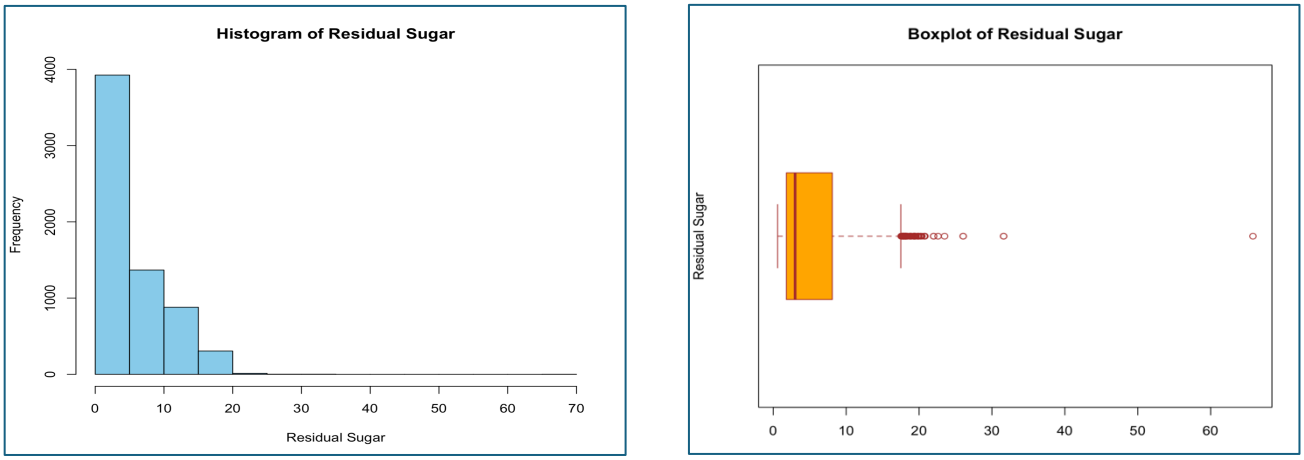

Figure 2.Histogram and Box plot

4.Categorical Analysis

The dataset contains a **categorical variable wine type** (white vs red) and a **target variable quality**.

From **Wine Type Distribution**, White wine is more frequent in the dataset, accounting for approximately **75%** of the samples.

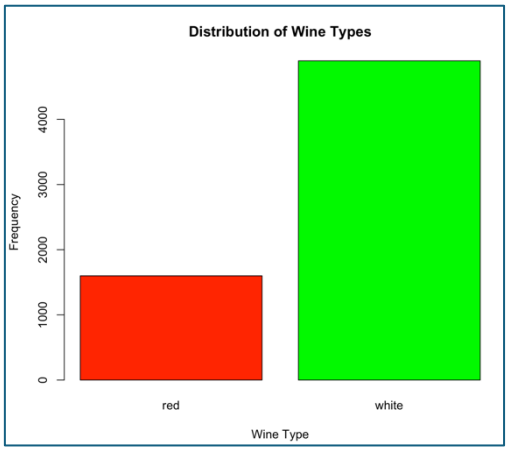

Figure 3.Distribution of Wine Types

5.Correlation Analysis

We are analyzing the linear dependency between the variables fixed.acidity and residual.sugar. The Pearson correlation coefficient came out to be -0.1119813 , indicating a weak negative correlation.

6.Scatter Plot Visualization

The data is right skewed and it is evident from the scatter plot too. The two variables are weakly negatively correlated visually too which can be inferred from the slope of the trend line.

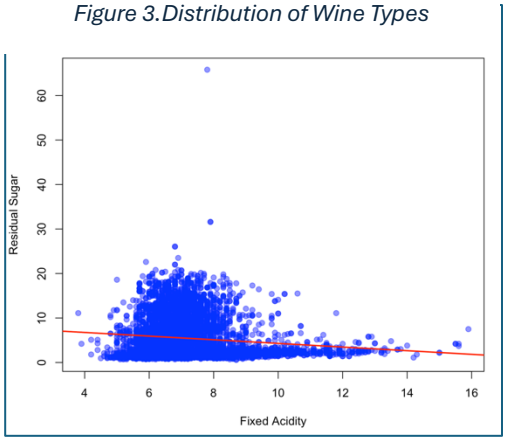

Figure 4.Scatter Plot

7.Multiple Linear Regression

We are now trying to fit a multi variable linear regression model for predicting the ‘density’ of wine using the variables ‘fixed.acidity’ and ‘residual.sugar’.

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    9.838e-01  1.430e-04 6881.55   <2e-16 ***
fixed.acidity  1.220e-03  1.872e-05   65.16   <2e-16 ***
residual.sugar 3.855e-04  5.101e-06   75.56   <2e-16 ***
---
```

Figure 5.Regression Parameters

After fitting the regression model , the p-values for the coefficients of both the features came out to be less than 0.05 , indicating both the **coefficients are significant** in the regression model. Both the coefficients of ‘fixed.acidity’ and ‘residual.sugar’ are positive indicating a positive relationship between density and the two features.

Density=0.9838+0.00122(Fixed Acidity)+0.0003855(Residual Sugar)

8.Model Diagnostics

From the Q-Q plot, it can be seen that the residuals follow a normal distribution. But it can also be seen that the variance of residuals is decreasing with the fitted values indicating heterosadasticity.

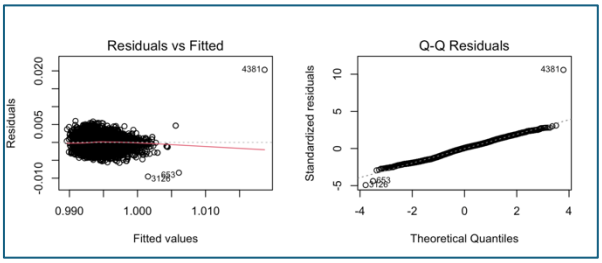

Figure 6.Residual Plots

The **R-squared** value of approximately **0.5799** indicates that the model explains **57.99%** of the variance in the target variable, suggesting a moderate fit.

```
F-statistic:  4483 on 2 and 6494 DF,  p-value: < 2.2e-16
```

9.PCA

From the Scree plot it can be seen that 80% of the variance can be explained by the first 4 principal components.

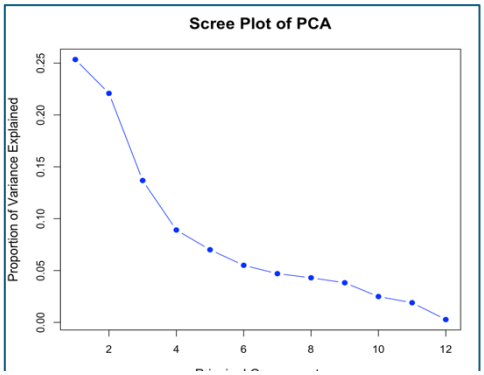

Figure 7.Scree Plot

From the **biplot**, it is evident that **fixed acidity** and **residual sugar** are positioned **far apart** from each other, which suggests that these two variables are **not highly correlated**.

when visualizing the data using the first **two principal components (PCs)**, we observe the presence of a **single cluster**. This indicates that the data points are concentrated in one area of the plot, suggesting that the dataset is relatively homogenous with respect to the principal components.

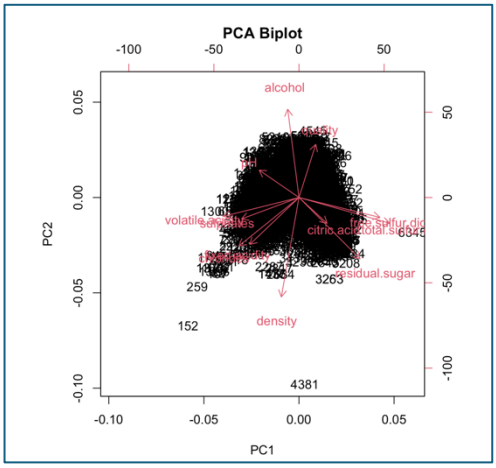

Figure 8.Biplot

Seoul Bike Rental Data:

1.Data Overview

The dataset consists of various factors affecting bike rental counts in Seoul. The primary variables influencing bike rental counts are weather-related (such as temperature, humidity, wind speed), time-based (such as hour of the day), and other factors like season and holiday status. The dataset has a total Observations of 600 described with 14 variables.

```
$ Date                     : chr  "1/12/17" "1/12/17" "1/12/17" "1/12/17" ...
$ Rented_Bike_Count        : int  254 204 173 107 78 100 181 460 930 490 ...
$ Hour                     : int  0 1 2 3 4 5 6 7 8 9 ...
$ Temperature_c            : num  -5.2 -5.5 -6 -6.2 -6 -6.4 -6.6 -7.4 -7.6 -6.5 ...
$ Humidity_percent         : int  37 38 39 40 36 37 35 38 37 27 ...
$ Wind.speed_m_s           : num  2.2 0.8 1 0.9 2.3 1.5 1.3 0.9 1.1 0.5 ...
$ Visibility_10m           : int  2000 2000 2000 2000 2000 2000 2000 2000 2000 1928 ...
$ Dew.point.temperature_k: num  256 256 255 256 255 ...
$ Solar.Radiation..MJ.m2.: num  0 0 0 0 0 0 0 0 0.01 0.23 ...
$ Rainfall.mm.             : num  0 0 0 0 0 0 0 0 0 0 ...
$ Snowfall..cm.            : num  0 0 0 0 0 0 0 0 0 0 ...
$ Seasons                  : chr  "Winter" "Winter" "Winter" "Winter" ...
$ Holiday                  : chr  "No Holiday" "No Holiday" "No Holiday" "No Holiday"
$ Functioning.Day          : chr  "Yes" "Yes" "Yes" "Yes" ...
```

Figure 9.Overview of the Data

2.Summary Statistics

1	Mean	254.445
2	Median	237
3	Standard Deviation	163.436
4	Minimum	3
5	Maximum	937

This indicates that bike rentals are heavily skewed, with a few hours having a very high number of bike rentals, while others have significantly lower counts.

3.Distribution Visualization

The distribution of **Rented_Bike_Count** is **right-skewed**, which suggests a higher frequency of low rental counts, with occasional spikes in the number of bikes rented. The presence of outliers in the **Rented_Bike_Count** is clear. These outliers

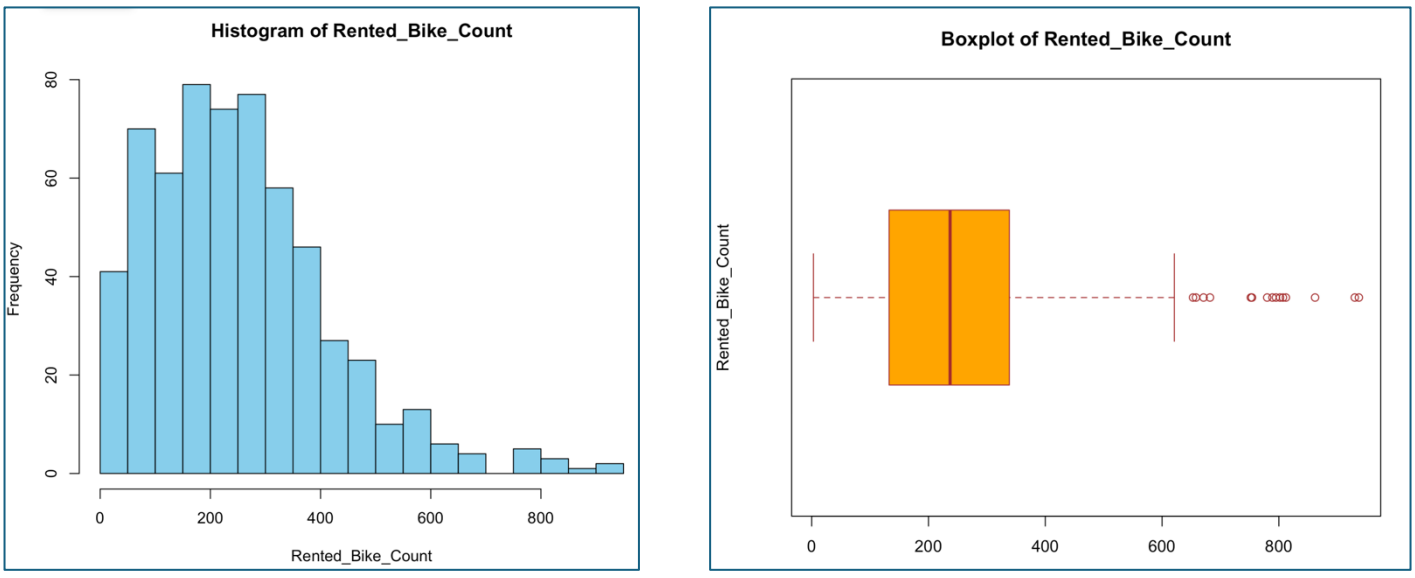

Figure 10.Histogram and Box plot

could represent special circumstances, such as peak times when bike rentals are high.

4.Categorical Analysis

This shows that the majority of the data points correspond to days without holidays, which could influence the bike rental demand .

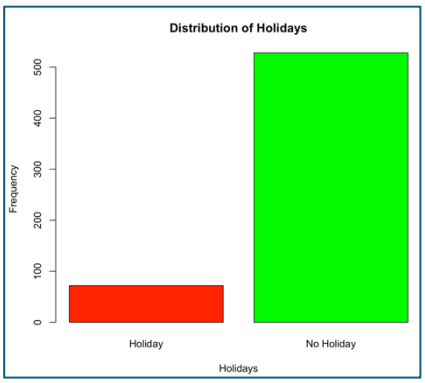

Figure 11.Histogram for Holidays

5.Correlation Analysis

Dew.point.temperature_k and **Humidity_percent** is **0.847**, indicating a strong positive correlation between dew point temperature and humidity.

6.Scatter Plot

The scatter plot shows a positive linear relationship between Dew.point.temperature_k and Humidity_percent, with higher dew point temperatures corresponding to higher humidity levels

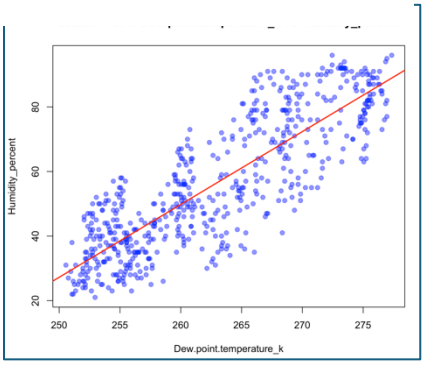

Figure 12.Scatter Plot

7.Multiple Linear Regression

We are now trying to fit a multi variable linear regression model for predicting the 'Rented_Bike_Count' using the variables 'Dew.point.temperature_k' and 'Humidity_percent'.

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         -3233.2167   357.5125  -9.044   <2e-16 ***
Dew.point.temperature_k   14.7752     1.4576  10.136   <2e-16 ***
Humidity_percent          -7.0407     0.5473 -12.865   <2e-16 ***
---
```

Figure 13.Regression Parameters

After fitting the regression model , the p-values for the coefficients of both the features came out to be less than 0.05 , indicating both the **coefficients are significant** in the regression model. The coefficient of 'Dew.point.temperature_k' is positive indicating a positive relationship between Rented_Bike_Count and the Dew.point.temperature_k.

The coefficient of 'Humidity_percent' is negative indicating an inverse relationship between Rented_Bike_Count and Humidity_percent.

Rented_Bike_Count = -3233.2167 + 14.7752 x(Dew.point.temperature_k) 7.04x(Humidity_percent)

The R-squared value of 0.2192 suggesting the model can be improved. We tried fitting a polynomial regression model which is giving the cofficients of 2nd order variables to be non-significant.

Also the R squared(polynomial fit) came out as 0.2232, that need to be improvised.

```
Residual standard error: 144.7 on 597 degrees of freedom
Multiple R-squared:  0.2192,    Adjusted R-squared:  0.2166
F-statistic: 83.79 on 2 and 597 DF,  p-value: < 2.2e-16
```

```
Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                       254.445      5.910  43.052   <2e-16 ***
poly(Dew.point.temperature_k, 3)1 2788.130    279.599   9.972   <2e-16 ***
poly(Dew.point.temperature_k, 3)2 -200.840    165.138  -1.216    0.224
poly(Dew.point.temperature_k, 3)3 -151.762    153.209  -0.991    0.322
poly(Humidity_percent, 3)1       -3536.160    279.480 -12.653   <2e-16 ***
poly(Humidity_percent, 3)2         -11.278    165.770  -0.068    0.946
poly(Humidity_percent, 3)3           2.057    152.743   0.013    0.989
---
```

Figure 14.Regression Parameters

8.Model Diagnostics

The **Residuals** appear randomly scattered, confirming the linearity assumption.From **Q-Q plot** the residuals deviates from normal distribution and the presence of **heteroscedasticity** is suggested by the variance of residuals decreasing with increasing fitted values.

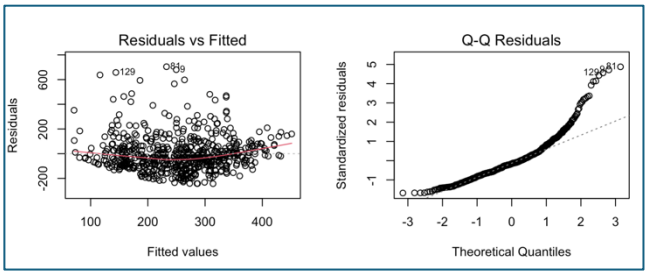

Figure 15.Residual Plot

9.PCA

The **first principal component (PC1)** explained **35.84%** of the variance, and together with **PC2**, explained over **54%** of the variance. The **Scree plot** suggests that the most important components are **PC1** and **PC2** together explaining more than 80% of the variance of data. There are **no visible clusters** in the biplot, suggesting that the a clear separation based on the first two principal components

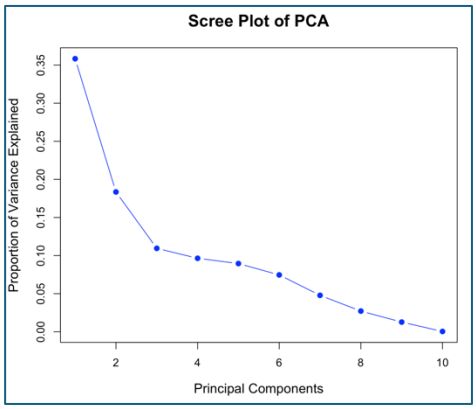

Figure 16.Scree Plot

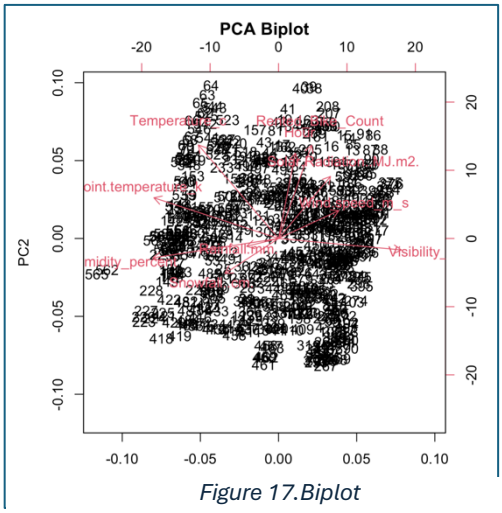

Figure 17.Biplot

Abalone Dataset:

1.Data Overview

The Abalone dataset contains 4177 observations and 9 variables related to the physical measurements of abalones. The primary goal is to predict the age of an abalone based on these measurements, which serve as easier alternatives to the traditional method of cutting the shell and counting the rings.
Summary Statistics

```
'data.frame':   4177 obs. of  9 variables:
$ Sex           : chr  "M" "M" "F" "M" ...
$ Length        : num  0.455 0.35 0.53 0.44 0.33 0.425 0.53 0.545 0.475 0.55 ...
$ Diameter      : num  0.365 0.265 0.42 0.365 0.255 0.3 0.415 0.425 0.37 0.44 ...
$ Height        : num  0.095 0.09 0.135 0.125 0.08 0.095 0.15 0.125 0.125 0.15 ...
$ Whole.weight  : num  0.514 0.226 0.677 0.516 0.205 ...
$ Shucked.weight: num  0.2245 0.0995 0.2565 0.2155 0.0895 ...
$ Viscera.weight: num  0.101 0.0485 0.1415 0.114 0.0395 ...
$ Shell.weight  : num  0.15 0.07 0.21 0.155 0.055 0.12 0.33 0.26 0.165 0.32 ...
$ Rings         : int  15 7 9 10 7 8 20 16 9 19 ...
```

Figure 18.Overview of Data

2.Summary Statistics (Diameter)

Summary statistics provide an overview of the central tendency and spread of each variable.
The mean and median are close, suggesting a roughly symmetric distribution.

1	Mean	0.055
2	Median	0.65
3	Standard Deviation	0.4078
4	Minimum	0.425
5	Maximum	0.0992

3.Distribution Visualization

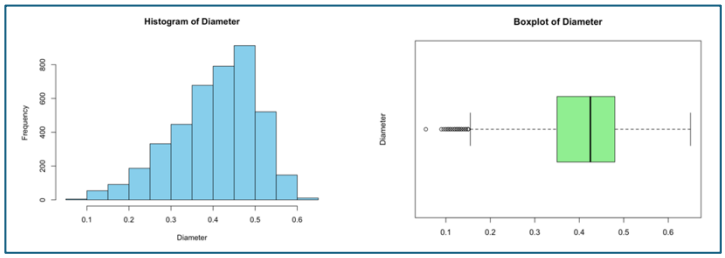

Figure 19.Histogram and BoxPlot

The histogram of Diameter exhibits a unimodal, right-skewed distribution, with most observations concentrated between 0.4 and 0.5. The boxplot confirms the skewness and reveals potential outliers on the lower end of the range, below approximately 0.2. The central tendency is evident around the interquartile range (0.4 to 0.5), with a median close to 0.45. These outliers suggest variability among smaller diameter measurements.

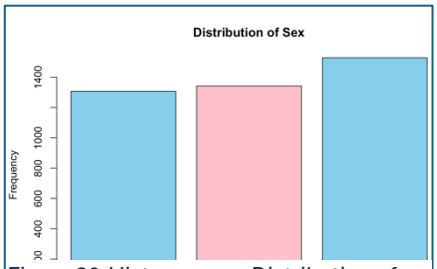

Figure 20.Histogram on Distribution of sex

4.Categorical Variable(Sex) Analysis:

Categories F and M have a similar distribution, with a small difference in frequency. Category I(Infant) is relatively less frequent compared to F(Female) and M(Male). The distribution is not perfectly balanced but shows that all three categories are relatively comparable in frequency.

5.Correlation Analysis:

Choosen variables for the correlation analysis are Height and Viscera.Weight and the pearson correlation coefficient is 0.7983193 which indicates that the two variables have strong positive correlation

6.Scatterplot Visualisation:

The scatter plot of Height versus Viscera.Weight shows a clustered distribution with some outlying points. While there is a general upward trend, the relationship is not strictly linear, as evidenced by the uneven spread of points and deviations from the trendline. The concentration of points at lower values indicates a possible non-linear or heteroscedastic relationship, where Height may not predict Viscera.Weight effectively using a simple linear model.

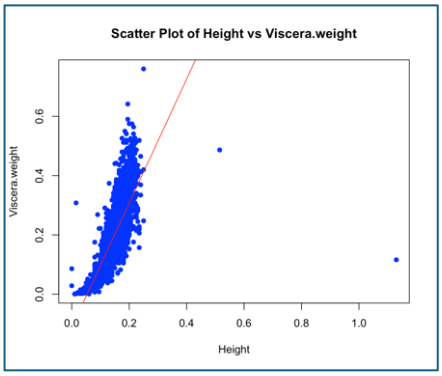

Figure 21.Scatter Plot

7.Multiple Regression:

Four models were used to predict Rings. Model 1, using Height and Viscera.weight, explained 32% of variability (R^2=0.3203). Model 2, with all predictors, improved R^2 to 0.5275 but showed heteroscedasticity. Model 3 added polynomial terms, raising R^2 to 0.5634. Model 4, a log-transformed polynomial regression, achieved the best performance with R^2=0.6402 and stable variance.

```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    2.24568    0.00297 756.202  < 2e-16
poly(Diameter, 2)1             4.59796    0.84630   5.433 5.86e-08
poly(Diameter, 2)2            -4.70629    0.30923 -15.220  < 2e-16
poly(Height, 2)1               3.26418    0.49521   6.592 4.90e-11
poly(Height, 2)2              -1.33149    0.27623  -4.820 1.48e-06
poly(Whole.weight, 2)1        31.35541    2.35857  13.294  < 2e-16
poly(Whole.weight, 2)2        -5.74583    1.08532  -5.294 1.26e-07
poly(Shucked.weight, 2)1 -28.82723    1.16484 -24.748  < 2e-16
poly(Shucked.weight, 2)2       6.98124    0.61970  11.266  < 2e-16
poly(Viscera.weight, 2)1      -5.66554    0.94843  -5.974 2.52e-09
poly(Viscera.weight, 2)2       0.95825    0.52252   1.834  0.06674
poly(Shell.weight, 2)1         9.07709    1.09028   8.325  < 2e-16
poly(Shell.weight, 2)2        -1.50321    0.50025  -3.005  0.00267
```

Figure 22.Regression Parameters

The model shows strong linear and quadratic effects for most variables. Whole.weight and Shell.weight have the strongest positive linear impacts (**$p < 2e\text{-}16$**), while Shucked.weight has a significant negative relationship. Non-linear effects are evident for Diameter, Height, and

Viscera.weight, with quadratic terms like poly(Whole.weight, 2)_2 and poly(Shucked.weight, 2)_2 also highly significant, reflecting complex relationships with log(Rings).

8.Model Diagnostics:

The residuals vs. fitted plot shows random scatter, indicating homoscedasticity, with minor non-linearity or outliers. The Q-Q plot confirms approximate normality, suggesting the model fits well while leaving room for slight improvements.

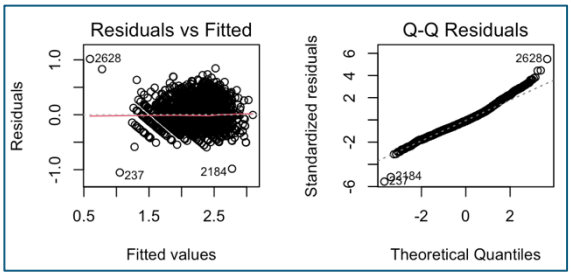

Figure 23.Residual Plots

9.Principal Component Analysis(PCA):

Based on the scree plot, I would select the first principal component, which explains over 80% of the variance. This significant variance indicates that the first component effectively captures the dataset's structure for dimensionality reduction.

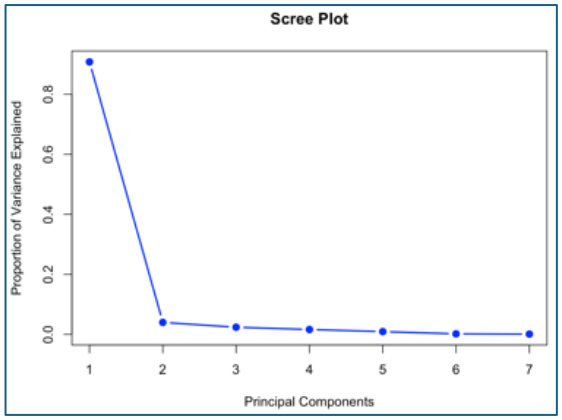

Figure 24.Scree Plot

The biplot reveals that **PC1** is heavily influenced by weight-related variables (Shell Weight, Viscera Weight, Shucked Weight), indicating their strong correlation. **PC2** is driven primarily by Height, which separates from the weight variables. This pattern highlights that weight characteristics dominate overall variance, while height provides unique structural variance in the data.

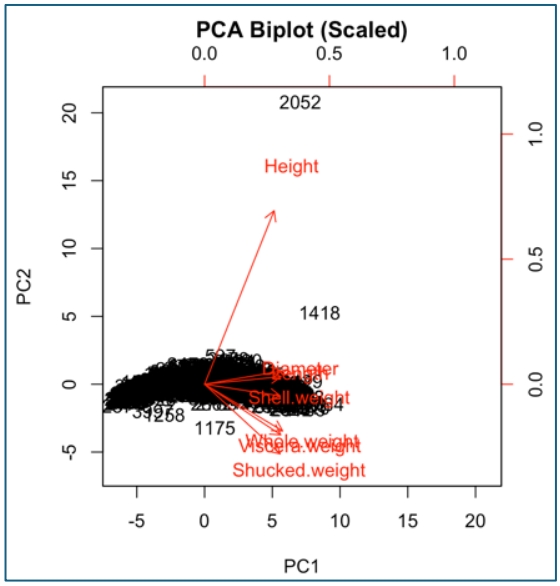

Figure 25.Biplot

Students Performance Dataset

1.Data Overview:

```
$ gender                     : chr  "female" "female" "female" "male" ...
$ race.ethnicity             : chr  "group B" "group C" "group B" "group A" ...
$ parental.level.of.education: chr  "bachelor's degree" "some college" "master's degree"
"associate's degree" ...
$ lunch                      : chr  "standard" "standard" "standard" "free/reduced" ...
$ test.preparation.course    : chr  "none" "completed" "none" "none" ...
$ math.score                 : int  72 69 90 47 76 71 88 40 64 38 ...
$ reading.score              : int  72 90 95 57 78 83 95 43 64 60 ...
$ writing.score              : int  74 88 93 44 75 78 92 39 67 50 ...
```

Figure 26,Overview of the Data

The dataset contains information about students' performance in three subjects: Math, Reading, and Writing. Along with these, there are several other features such as gender, parental education, and whether the students completed a test preparation course. There are 6497 observations with 13 variables.

2.Summary Statistics

1	Mean	66.089
2	Median	66
3	Standard Deviation	15.16308
4	Minimum	0
5	Maximum	100

The mean math score is 66.089, indicating the average performance of the students. The median score is 66, showing that half the students scored above and half below this value. The standard deviation of 15.16 reflects a moderate spread of scores around the mean, with scores ranging from a minimum of 0 to a maximum of 100.

3.Distribution Visualization

The histogram indicated that the Math scores are slightly **right-skewed**, with a concentration of scores in the lower range. This suggests that a majority of the students are scoring on the lower end, with fewer students achieving higher scores.

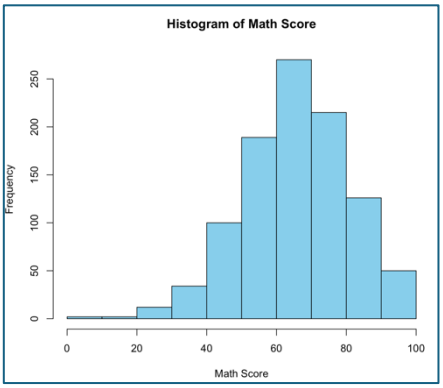

Figure 27.Histogram of Mathscore

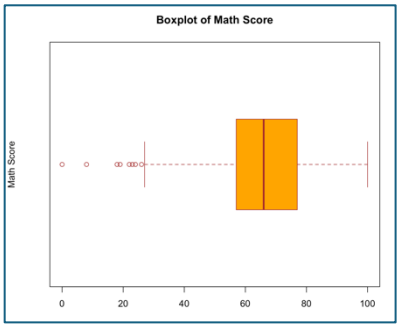

Figure 28.Box Plot

The **boxplot** further confirmed this right-skewness, with the median close to the lower quartile and outliers observed on the higher side of the scores. The distribution is fairly spread out, with the presence of a few extreme values, as indicated by the range from the minimum to the maximum score.

4.Categorical Analysis :

The **gender** distribution in the dataset is nearly balanced, with 518 female students and 482 male students. This indicates an almost equal representation of genders.

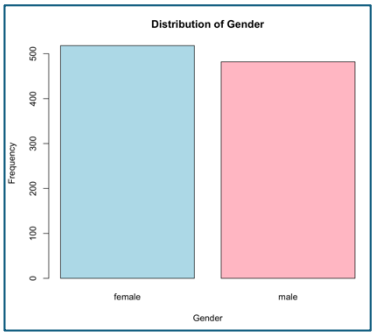

Figure 29.Distribution of Gender

5.Correlation Analysis:

The Pearson correlation between **Math Score** and **Reading Score** was found to be **0.82**, indicating a strong positive correlation.

6.Scatter plot:

The Scatter Plot confirmed the positive linear association between the Math and Reading scores.

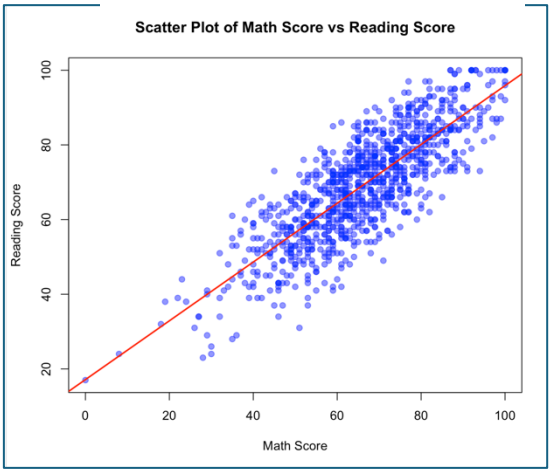

Figure 30.Scatter Plot

7.Multiple Linear Regression

We are now trying to fit a multi variable linear regression model for predicting the 'Writing_Score' using the variables 'math_Score' and 'reading_Score'

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.16084    0.69866  -1.662   0.0969 .
math.score    0.06705    0.01628   4.118 4.14e-05 ***
reading.score 0.93660    0.01691  55.389  < 2e-16 ***
```

Figure 31.Regression Parameters

After fitting the regression model , the p-values for the coefficients of both the features came out to be less than 0.05 , indicating both the **coefficients are significant** in the regression model. Both the coefficients of 'math_Score' and 'reading_Score' are positive indicating a positive relationship between 'Writing_Score' and the two features.

Writing_Score = 0.06705 x (math_Score')+0.9366 x (reading_Score)

The **Multiple R-squared** value of 0.9127 indicates that approximately 91% of the variance in Writing scores can be explained by Math and Reading scores. This high R-squared value suggests that the model fits the data very well.

```
Residual standard error: 4.493 on 997 degrees of freedom
Multiple R-squared:  0.9127,    Adjusted R-squared:  0.9126
F-statistic:  5214 on 2 and 997 DF,  p-value: < 2.2e-16
```

8.Model Diagnostics

The residuals were randomly scattered in the **Residuals vs Fitted Plot**, indicating that there is no obvious pattern and the assumption of homoscedasticity (constant variance of errors) holds. The **Normal Q-Q Plot** shows that the residuals are approximately normally distributed, confirming the assumption of normality.

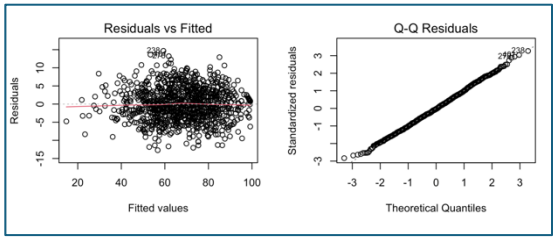

Figure 32.Residual Plots

9.PCA

The first two principal components (PC1 and PC2) together explain **approximately 90% of the variance** in the data, as indicated by the **Scree Plot**.

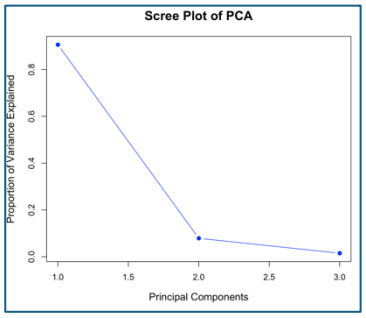

Figure 33.Scree Plot

The **PCA Biplot** showed that the **Math Score**, **Reading Score**, and **Writing Score** are highly loaded on the first two principal components. This indicates that these three variables contribute significantly to the variance in the dataset, which is consistent with the findings of the correlation analysis and the regression model.

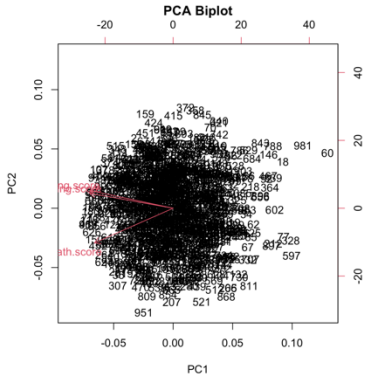

Figure 34.Biplot