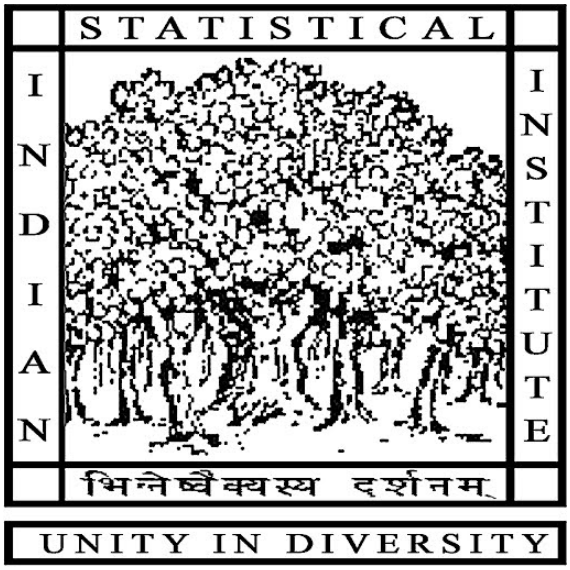
INDIAN STATISTICAL INSTITUTE

POST-GRADUATE DIPLOMA IN BUSINESS ANALYTICS (PGDBA): 2024–26

Course: Statistical Structure in Data

Assignment (08 December 2024)



Submitted by:

Harshal Bhagwat Ugalmugle

(24BM6JP20)

Submitted to:

Prof. Subhajit Dutta

Assignment Report

Dataset: *mtcars* (from R Studio)

1. Data Overview

The *mtcars* dataset contains **32 observations** and **11 variables**, providing details about different car models and their attributes, such as mileage, horsepower, weight, and cylinder count. The dataset is widely used for exploratory data analysis and statistical modelling.

2. Summary Statistics

Variable Chosen: mpg (Miles per Gallon)

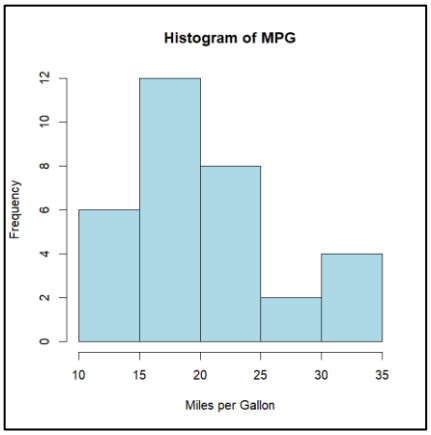
Min.	Median	Mean	Max.	Std. Dev
10.4	19.2	20.09	33.9	6.03

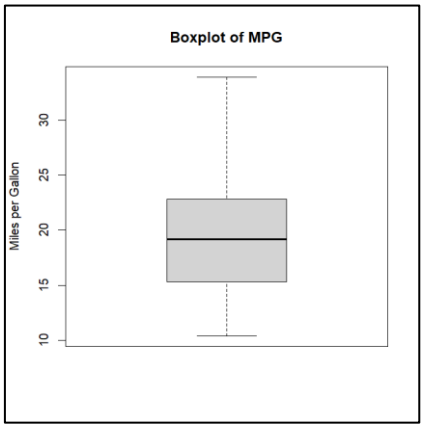
Interpretation:
The average mileage of cars is approximately 20.09 miles per gallon, with half the cars achieving less than 19.20 mpg. The data ranges from a minimum of 10.40 mpg to a maximum of 33.90 mpg, indicating considerable variation. The standard deviation of 6.03 suggests moderate dispersion around the mean.

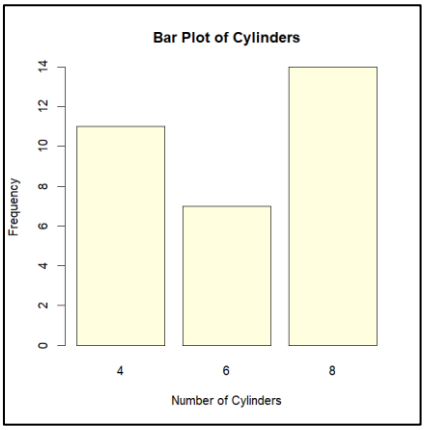
3. Distribution Visualization

Numerical Variable Chosen: mpg (Miles per Gallon)

Categorical Variable Chosen: cyl (Number of Cylinder)







Histogram: The histogram indicates a *right-skewed* distribution, meaning that most cars have mileage clustered at the lower end, with a few achieving significantly higher mileage.

Boxplot: The boxplot did not reveal any *outliers*. Most values fall within the expected range.

Bar Plot: The majority of cars in the dataset have **8 cylinders**, followed by cars with **4 cylinders** and **6 cylinders**. This reflects a prevalence of larger engine configurations, which are typical of older car models represented in this dataset.

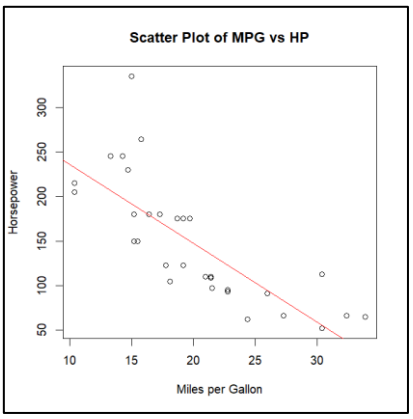
4. Correlation Analysis

Variables Chosen: mpg (Miles per Gallon) and hp (Horsepower)

Pearson Correlation Coefficient: -0.7762

Interpretation: This negative correlation indicates a strong inverse relationship between mileage (mpg) and horsepower (hp). As horsepower increases, mileage tends to decrease significantly.

5. Scatter Plot Visualization



Graph Observation: The scatter plot shows a clear negative trend, confirming the inverse relationship between mpg and hp. The trend line further emphasizes this relationship.

6. Multiple Regression

Regression Model: Predicting mpg using hp and wt (weight) as predictors.

	Coefficient	P-value	Adj.R-squared	Residual std error
Intercept	37.22	<2E-16	0.8268	2.596
hp	-0.0317	0.00145		
wt	-3.87783	1.12E-06		

Both hp and wt are statistically significant predictors of mpg based on their low p-values. Their negative coefficients indicate that increases in horsepower or weight result in reduced mileage. Identified influential points (17, 20, 31) (refer to cook's distance plot) from the Cook's distance plot, removed them, and developed a refined model.

Refined Model (After Removing Influential Points):

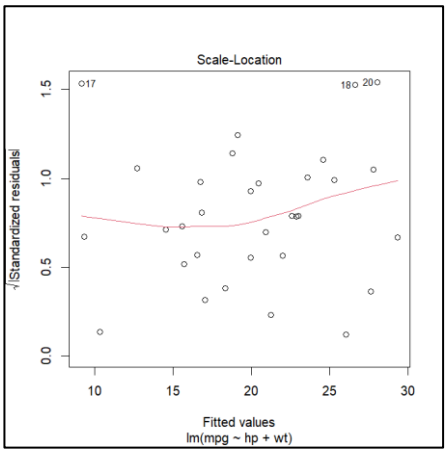
	Coefficient	P-value	Adj.R-squared	Residual std error
Intercept	37.4464	<2E-16	0.8576	2.116
hp	-0.0372	0.000476		
wt	-3.854902	8.12E-07		

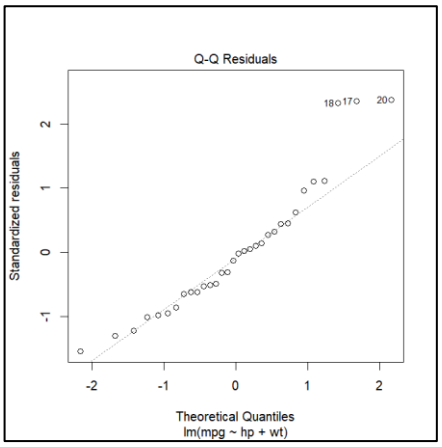
Adjusted R-squared = 0.8576, indicating improved explanatory power (85.8%). Reduced from 2.593 to 2.116, showing better model accuracy after removing influential points.

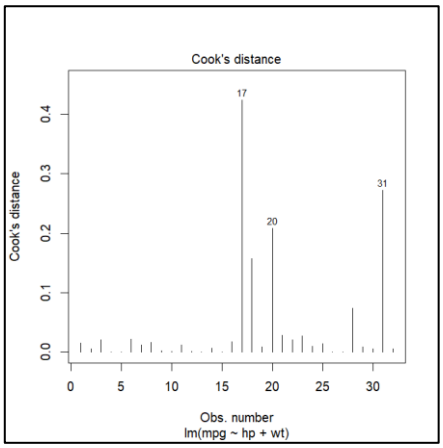
7. Model Diagnostics

Q-Q Plot: Residuals are approximately normally distributed.

Residuals vs. Fitted Plot: The initial model showed some distortion, suggesting heteroscedasticity. After removing influential points, the distortion reduced significantly, improving model validity.

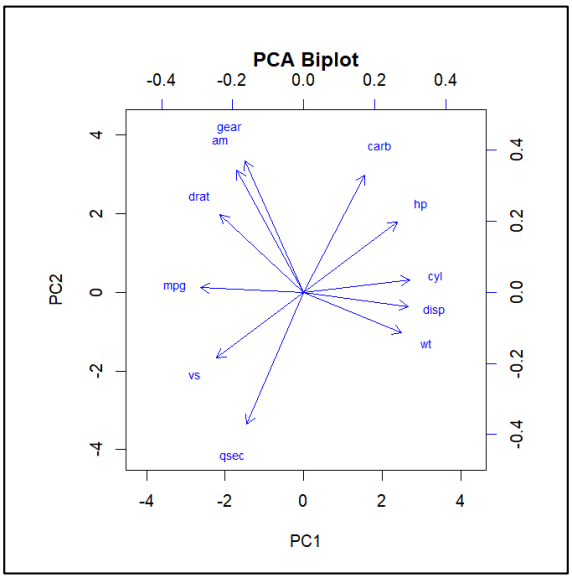


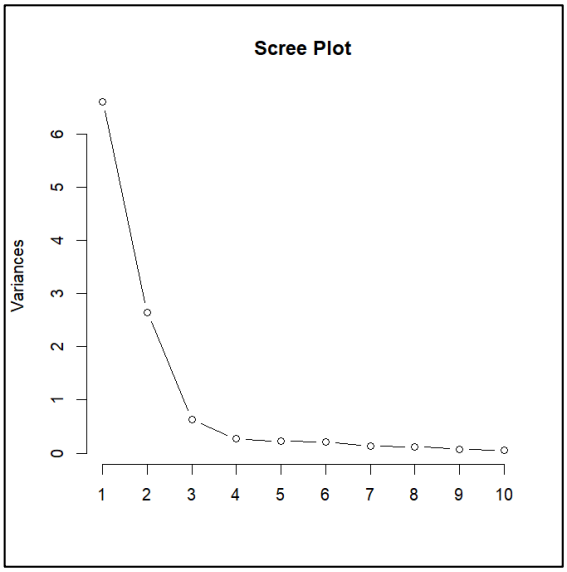




8. Principal Component Analysis (PCA

Scree Plot: The variance drops after the first three principal components, suggesting that **three components** are sufficient to explain most of the data's variance(89.87%). **PC1** is influenced heavily by variables like disp, cyl, and wt (associated with size and weight), indicating it might capture features related to vehicle size. **PC2** is dominated by am, gear, and carb, which could capture drivetrain or engine-related characteristics. Variables such as mpg (fuel efficiency) and wt (weight) are negatively correlated, as expected in vehicles. The biplot shows clear groupings: cyl, disp, wt, and hp correlate positively, representing vehicle size and power, while mpg and vs oppose them, indicating negative correlation with size (fuel efficiency). gear, am, and carb cluster, reflecting drivetrain characteristics. The plot highlights opposing trends like vehicle size vs. fuel efficiency and drivetrain vs. engine efficiency.





9. Conclusion: The univariate analysis highlighted the distribution and relationships between variables such as mpg, hp, and cyl, providing insights into the cars' fuel efficiency and engine characteristics. In the multivariate analysis, the significant negative correlation between mpg and hp, along with the regression models, revealed the impact of horsepower and weight on fuel efficiency.

The scree plot shows that the first two principal components (PCs) are sufficient to explain more than 90% of variance. The biplot shows cyl, disp, wt, and hp positively correlating, representing vehicle size and power, while mpg and vs negatively correlate, indicating fuel efficiency. Gear, am, and carb cluster around drivetrain characteristics. Key trends include vehicle size vs. fuel efficiency and drivetrain vs. engine efficiency. Overall, the *mtcars* dataset offers valuable insights into car performance characteristics, where factors like horsepower and weight significantly impact fuel efficiency.

Dataset: *USJudgeRatings* (from R Studio)

1. Data Overview

Structure: The dataset contains **43 observations** and **12 variables**, representing ratings of 43 judges across various criteria. Description of all attributes of this dataset is as follow:

CONT - Judicial integrity.
INTG - Judicial intelligence
DMNR - Judicial demeanor.
DILG - Judicial diligence.
CFMG - Case flow management.
DECI - Prompt decisions.
PREP - Preparation for trials.
FAMI - Familiarity with law.
ORAL - Clarity of oral decisions.
WRIT - Clarity of written decisions.
RTEN - Overall rating of the judge.

2. Summary Statistics

Variable Chosen: INTG (Integrity)

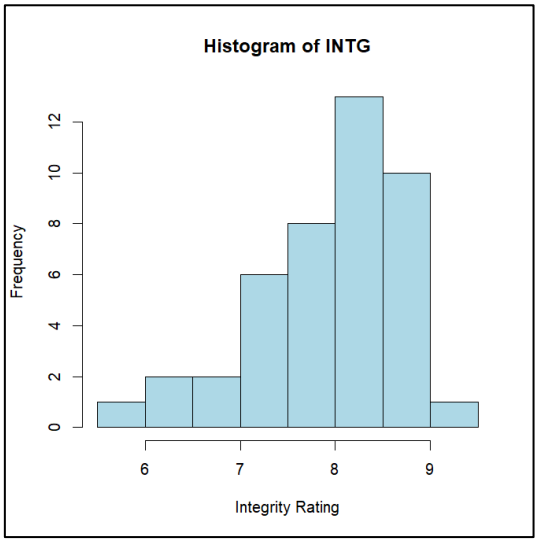
Categorical Variable Chosen: No categorical variable in dataset

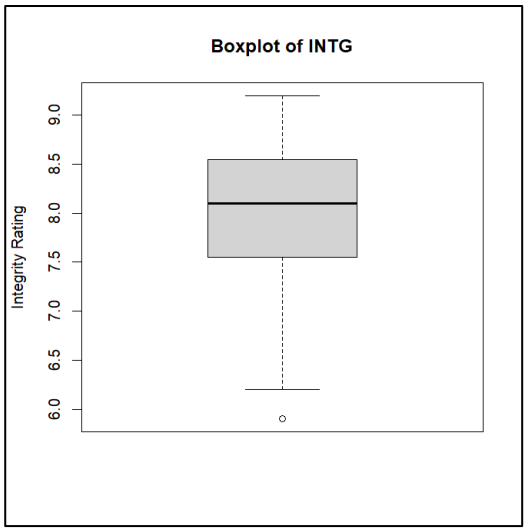
Statistics:

Min.	Median	Mean	Max.	Std. Dev
5.9	8.1	8.02	9.2	0.77

Interpretation: The INTG variable has a mean and median close to each other, indicating a fairly symmetric central tendency, but a slight left skew is present. The standard deviation of 0.77 shows moderate variability in integrity ratings.

3. Distribution Visualization





Histogram: The INTG variable shows a left-skewed distribution, indicating that higher integrity ratings are more frequent.

Boxplot: There is one outlier in the data, suggesting a judge with an unusually low integrity rating compared to others.

4. Correlation Analysis

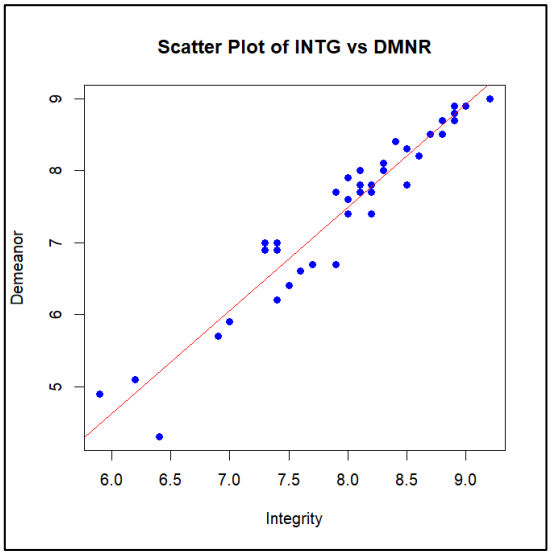
Variables Chosen: INTG (Integrity) and DMNR (Demeanor)

Result: Pearson correlation coefficient = **0.9646**.

Interpretation: A strong positive correlation exists between INTG and DMNR, indicating that judges with higher integrity tend to have better demeanor ratings.

5. Scatter Plot Visualization

Observation: The scatter plot with a trend line confirms a strong positive relationship between INTG and DMNR.



6. Multiple Regression

Model 1: After trying combination of multiple attributes to predict INTG, found DMNR and CFMG give best results in predicting INTG**.**

Key Results:

	Coefficient	P-value	R-squared(Adj)	Residual std error
Intercept	-0.9642	0.00152	0.964	0.2088
DMNR	0.5394	7.82E-14		
CFMG	0.6033	1.22E-11		

Interpretation: Both predictors significantly contribute to the model. Higher demeanor and case flow ratings strongly predict higher integrity.

Model Fit: Adjusted R-squared = 0.964, indicating the model explains 96.4% of the variance.

Model 2 (After Influential Point Removal): The refined model showed a slight improvement in fit, though not enough to warrant removing outliers. But witnessed significant improvement after removal of influential points found through cook's distance plot, results are as follow:

Key Results:

	Coefficient	P-value	R-squared(Adj)	Residual std error
Intercept	-0.6609	0.0161	0.9746	0.1709
DMNR	0.5928	6.82E-16		
CFMG	0.5102	8.71E-10		

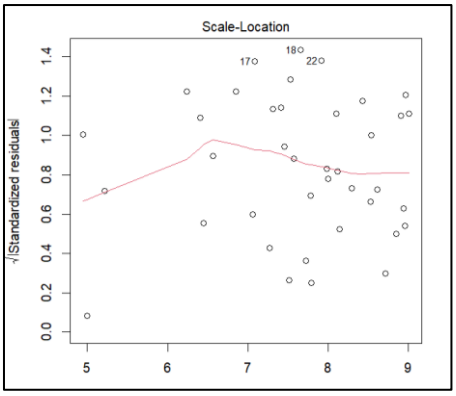
Model Fit: Adjusted R-squared = 0.9746, improved fit with fewer distortions.

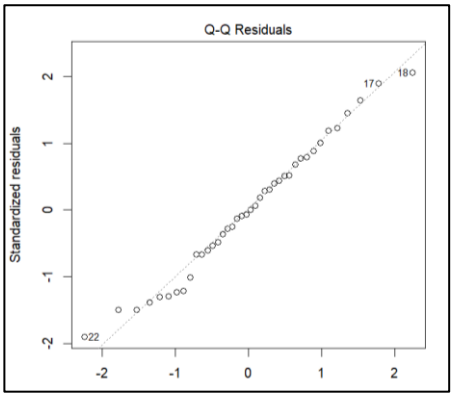
7. Model Diagnostics

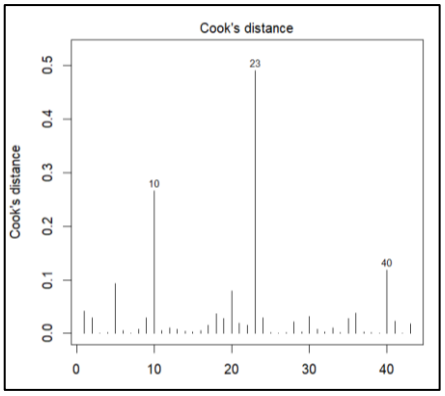
Residual Analysis:

Q-Q plot shows residuals are approximately normally distributed.

Residuals vs. fitted plot shows minor distortions in the first model, which are reduced in the second model, improving overall model reliability.



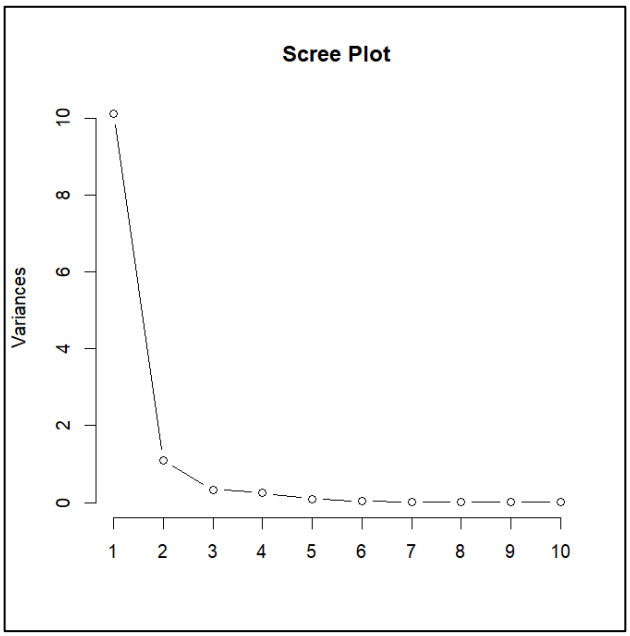


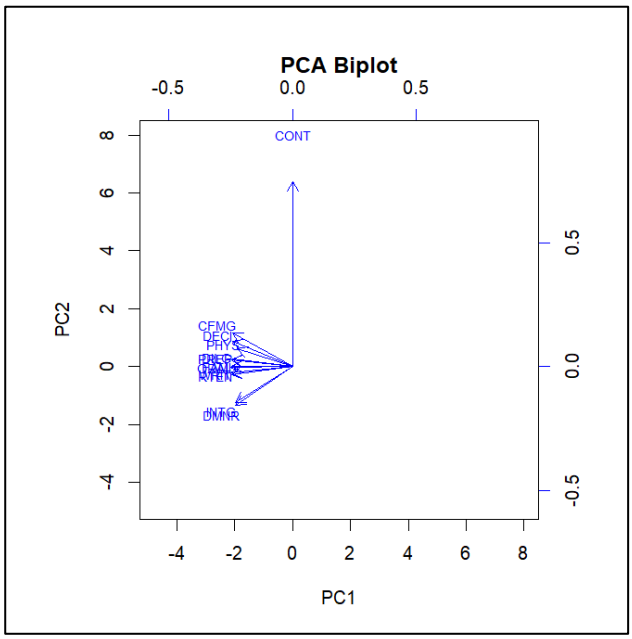


8. Principal Component Analysis (PCA)

Scree Plot Analysis: The scree plot shows a significant drop in explained variance after the first principal component (PC1). To ensure maximum variance coverage(93.64%), we select the first two components. In **PC1** Strong negative loadings for most variables (ORAL, WRIT, RTEN), indicating these contribute significantly to PC1. In **PC2** CONT has the highest positive loading, making it the dominant contributor to PC2. Variables with higher absolute loadings most strongly influence the corresponding component. The sign (positive/negative) reflects the direction of the relationship to the component's variance.

Patterns and Grouping: Most variables align in approximately the same direction, suggesting similar variance contributions, except for CONT, which stands apart. This indicates that CONT captures a unique aspect of variance compared to other variables.





9. Conclusion:

Univariate Analysis: Variables like INTG displayed left-skewed distributions with a single outlier, highlighting integrity ratings' variability. Multivariate Analysis: Strong positive correlations were found (INTG and DMNR with a coefficient of 0.9646). Regression analysis showed DMNR and CFMG as significant predictors of INTG, with robust model performance (Adjusted R-squared: 0.9746). PCA highlighted PC1 as capturing the majority of variance, with most variables contributing negatively. CONT emerged as distinct, showing unique characteristics within the data. Overall, the dataset reveals judges' ratings are highly interrelated, with demeanor and case flow management playing pivotal roles in integrity assessments. PCA further emphasized the commonality among most variables while distinguishing CONT as an outlier.

Dataset: ***trees*** (from R Studio)

1. Data Overview

The trees dataset contains 31 observations and 3 numerical variables: Girth, Height, and Volume. This dataset provides measurements of tree dimensions.

2. Summary Statistics

Variable Chosen: Height

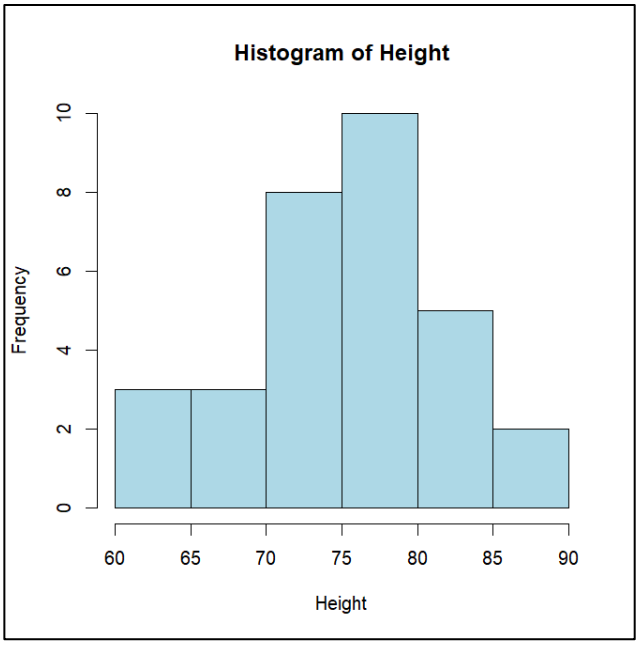
Categorical Variable Chosen: No categorical variable in dataset

Summary Statistics:

Min.	Median	Mean	Max.	Std. Dev
63	76	76	87	6.37

Interpretation:
The tree heights range from 63 to 87 units, with a mean and median of 76, indicating a fairly balanced distribution around the central value. The standard deviation of 6.37 reflects moderate variability.

3. Distribution Visualization:



Observations:

Histogram: The histogram of Height shows an approximately symmetrical distribution with most of the data concentrated around mean.

Boxplot: The boxplot indicates no outliers, suggesting the data is clean and evenly distributed.

4. Correlation Analysis

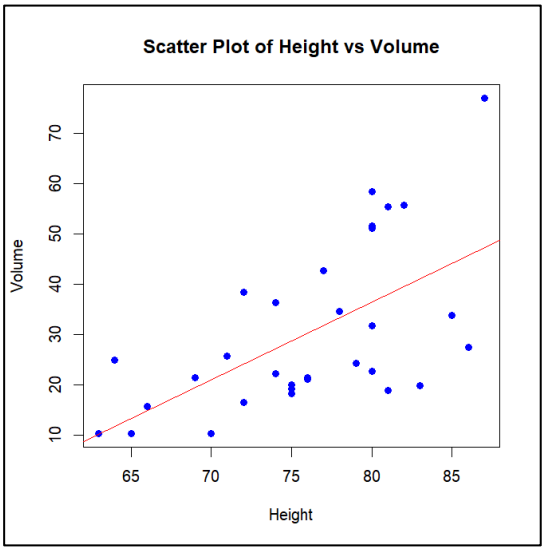
Variables Selected: Height and Volume

Pearson Correlation Coefficient: 0.598

Interpretation:
The positive correlation indicates that as Height increases, Volume tends to increase as well. However, the correlation strength is moderate, suggesting other factors, such as Girth, also significantly influence Volume.

5. Scatter Plot Visualization

Observation:
The scatter plot shows a positive trend between Height and Volume. A fitted trend line confirms this relationship, highlighting that taller trees generally have a higher volume.



6. Multiple Regression

Initial Model Summary:
Predicting Volume using Height and Girth.

	Coefficient	P-value	R-squared(Adj)	Residual std error
Intercept	-57.9877	2.75E-07	0.9331	4.252
Girth	4.7082	<2E-16		
Height	0.3393	0.0145		

The residual error is relatively high, indicating heteroscedasticity. To address this, a log transformation was applied to reduce variability.

Refined Model(After log transformation):

	Coefficient	P-value	R-squared(Adj)	Residual std error
Intercept	0.1025	6.37E-01	0.9662	0.09676
Girth	0.1452	<2E-16		
Height	0.0163	2.14E-05		

Interpretation:
The refined model, with log transformation, has a significantly improved fit and reduced residual error. Both Height and Girth significantly contribute to predicting Volume. The positive coefficients indicate that as Height or Girth increases, Volume also increases. Girth has a stronger impact compared to Height, as reflected by the larger coefficient.

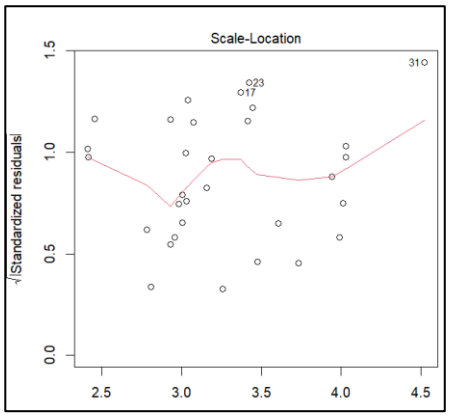
7. Model Diagnostics

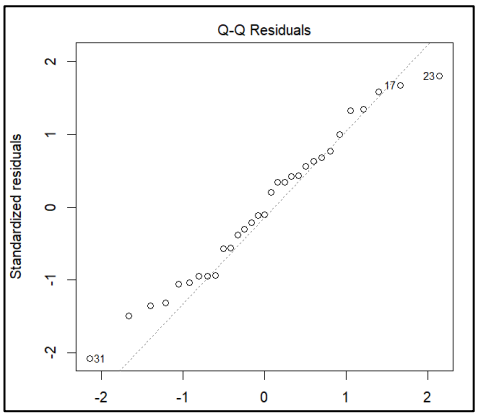
Residual Analysis:

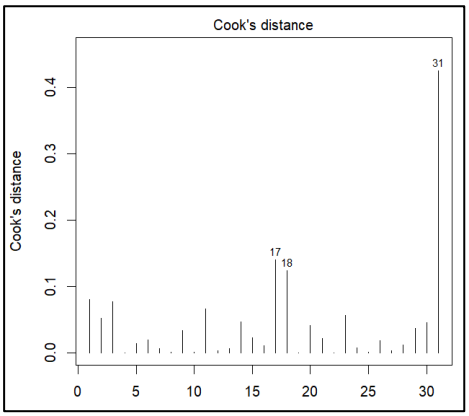
The **Q-Q plot** suggests the residuals are approximately normally distributed.

The **residuals vs. fitted plot** shows some distortion in the initial model, which is improved after the log transformation. This transformation enhances the homoscedasticity and model fit.

The model effectively captures the variability in Volume based on Height and Girth, explaining about 96.62% of the variance in the data.



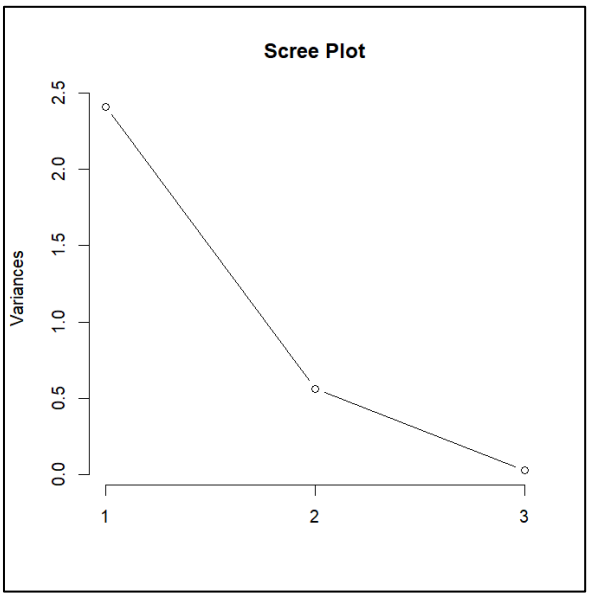


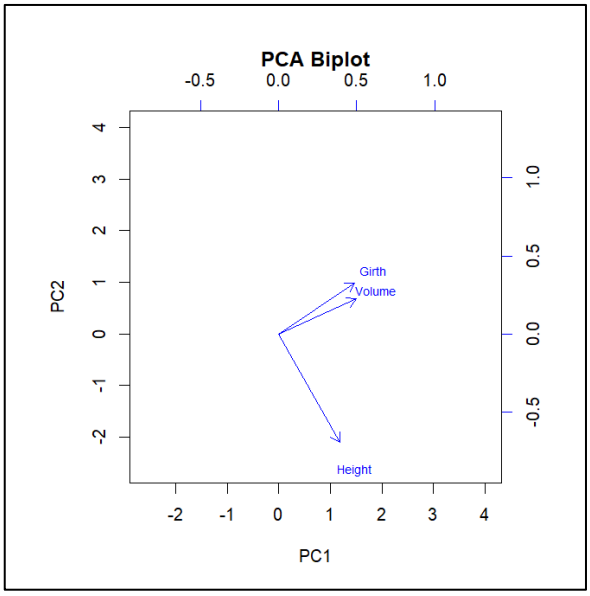


8. Principal Component Analysis (PCA):

The **scree plot** indicates a significant drop after the second principal component (PC2). Given that the dataset only has three variables, the first two components (PC1 and PC2) capture most of the variance. Therefore, we would choose two components (PC1 and PC2) as they explain 99.06% of the variance in the dataset.

Biplot Interpretation: PC1: Strongly correlates with the Volume and Girth variables, which are aligned in the same direction in the biplot. This suggests that as Volume increases, Girth tends to increase as well. **PC2:** Correlates more with Height, which is orthogonal to Volume and Girth. This suggests that Height varies independently from the other two variables. **Grouping and Patterns:** Volume and Girth are more closely related, while Height shows independent variation.





9.Conclusion

Univariate Analysis: The Height variable shows a relatively symmetrical distribution, with no significant outliers. Multivariate Analysis: The Pearson correlation coefficient between Height and Volume reveals a moderate positive relationship. The linear regression models showed good fits, with the refined model using log transformation improving homoscedasticity and reducing residual error. PCA Insights: PCA revealed that the first two components (PC1 and PC2) explain 99.06% of the variance in the data. Volume and Girth are strongly related, while Height is independent, contributing mostly to PC2. By reducing the data to just two components, we effectively capture most of the dataset's variability.

Dataset: *airquality* (from R Studio)

1. Data Overview

The airquality dataset contains air quality measurements taken in New York from May to September. It consists of **153 observations** across **6 variables**, which include numerical variables like Ozone, Solar.R, Wind, and Temp, and categorical variables representing months and days. Also there were some **missing values** in Ozone and Solar.R column which are **imputed using 'missForest'** method.

2. Summary Statistics

Variable Chosen: Temp(Temperature)

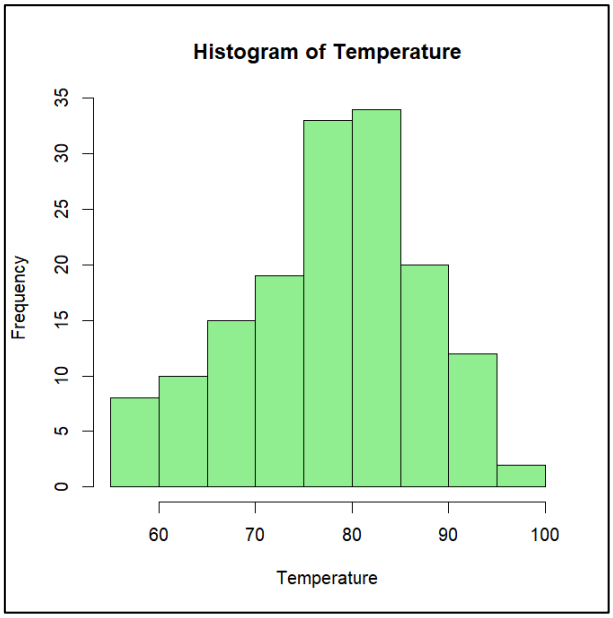
Categorical Variable Chosen: No categorical variable in dataset

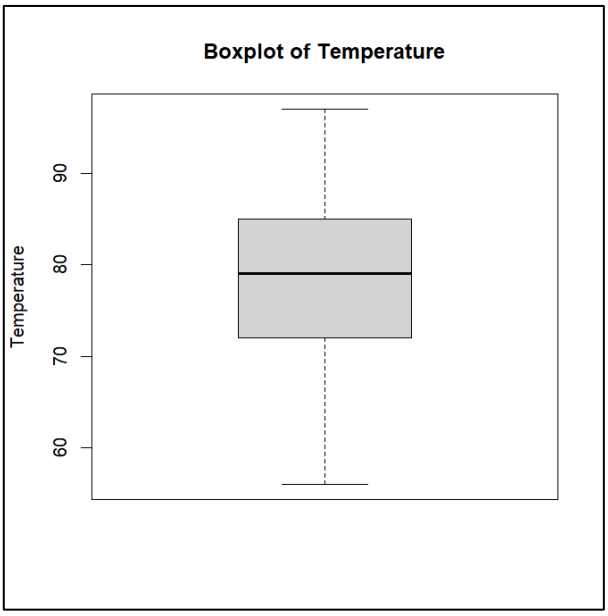
Summary Statistics:

Min.	Median	Mean	Max.	Std. Dev
56	79	77.88	97	9.4652

Interpretation:
The temperature values range from **56°F** to **97°F**, with an average temperature of approximately **77.88°F**. The distribution is moderately spread, as indicated by the standard deviation of **9.47°F**. The median value (**79°F**) being close to the mean suggests a roughly symmetrical distribution.

3. Distribution Visualization





Observations:

Histogram: The histogram of the Temp variable shows an **approximately symmetrical distribution** with a slightly thicker tail on the left side, indicating a minor skewness towards lower temperatures.

Boxplot: The boxplot reveals that there are **no significant outliers** in the data. The spread between the 1st and 3rd quartile is moderate, consistent with the histogram observations.

4. Correlation Analysis (Variables: Temp and Ozone)

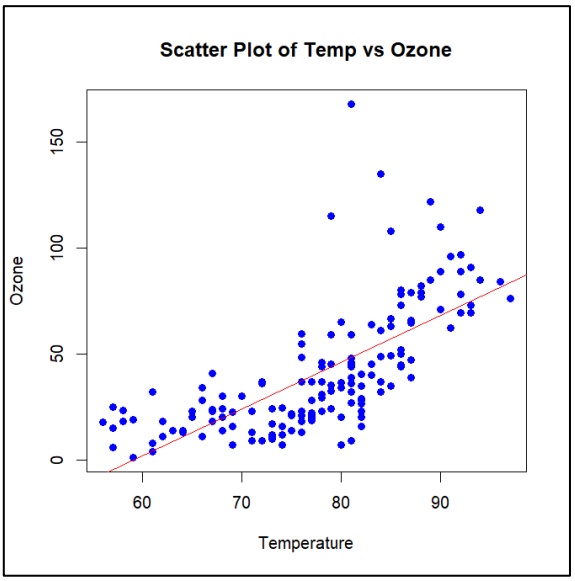
Variables Selected: Temp and Ozone

Pearson Correlation Coefficient: 0.6901.

Interpretation:
This indicates a **moderately strong positive correlation** between temperature and ozone levels. As the temperature increases, ozone levels tend to rise as well. However, this relationship is not perfect, implying other factors may also influence ozone levels.

5. Scatter Plot Visualization

Observations:
A scatter plot between Temp and Ozone reveals a **positive trend**, indicating that higher temperatures are generally associated with higher ozone levels.



6. Multiple Regression

Initial Model Summary: Predicting Ozone using Temp and Wind

	Coefficient	**P-value**	**R^2(Adj)**	**Residual std error**
Intercept	-64.8255	4.89E-04	0.5552	20.15
Temp	1.7246	1.94E-15		
Wind	-2.8082	2.81E-07		

RSE **20.15** indicates a high degree of residual variability, indicating heteroscedasticity and suggesting scope for improvement. To address this, a Box-Cox transformation was applied to reduce variability.

Refined Model (After Box-Cox Transformation):
With an optimal lambda of **0.303**, the refined model improves as follows:

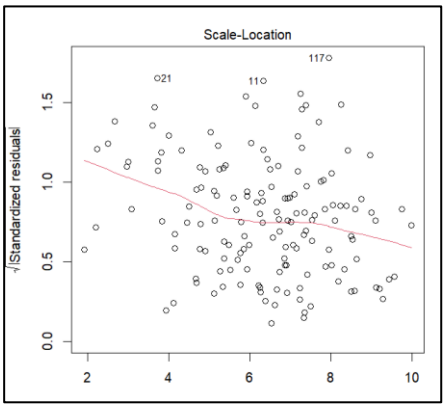
	Coefficient	P-value	R-squared(Adj)	Residual std error
Intercept	-2.9539	1.99E-02	0.6002	1.392
Temp	0.1418	<2E-16		
Wind	-0.174	3.35E-06		

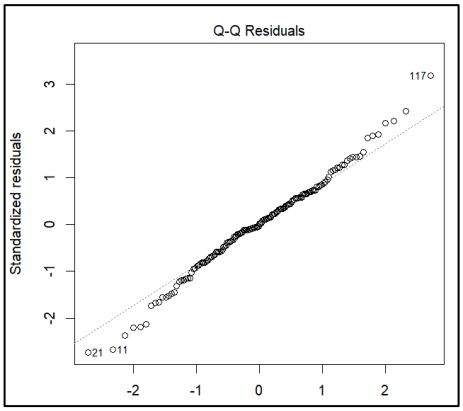
Interpretations: Residual Standard Error: **1.392**, significantly reduced. Adjusted R-squared: **0.6002**, slightly improved compared to the initial model. The Box-Cox transformation addresses heteroscedasticity, improving the normality of residuals and reducing residual variability.

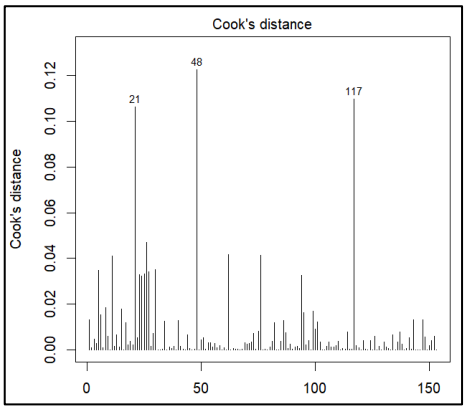
7.Model Diagnostics

Q-Q Plot: Residuals are approximately normally distributed, though slight deviations exist in the tails. The Box-Cox transformation mitigates these deviations, improving the model's robustness.
Residuals vs. Fitted Plot: The refined model shows reduced distortion, indicating improved homoscedasticity.

Cook's Distance Plot: Some influential points are identified but eliminating them provides no significant improvement, so they are retained.



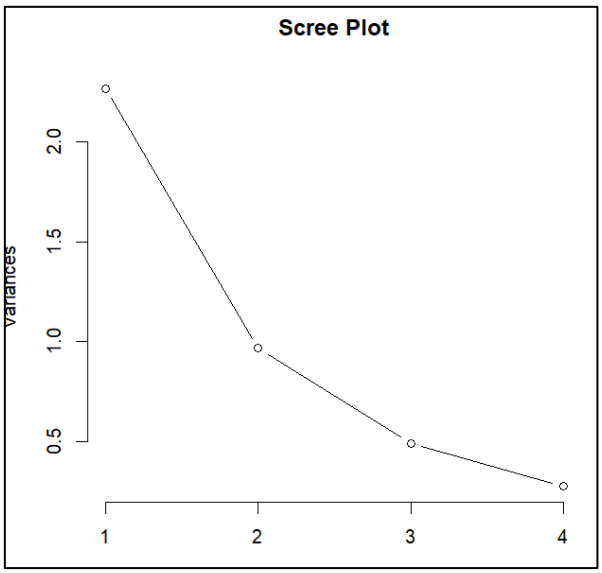


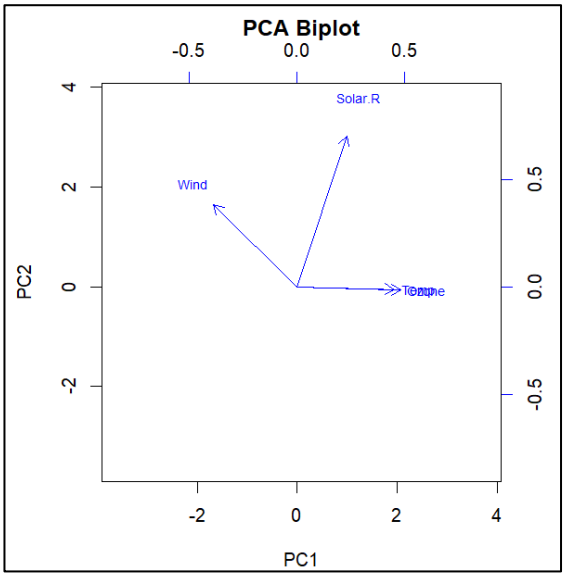


8. Principal Component Analysis (PCA):

The **scree plot** reveals that the explained variance drops significantly after the second principal component (PC2). **PC1** captures 56% of the variance. **PC2** captures 24% of the variance. Combined (PC1 + PC2) **80% of the total variance**. Adding **PC3** increases the cumulative variance explained to **95%**, which is sufficient for dimensionality reduction. **PC4** captures only 5% of the variance and can be considered negligible.

Loading Matrix Insights: PC1 Dominated by Temp and Ozone (positive direction) and inversely related to Wind. **PC2:** Primarily driven by Solar.R and Wind. **Biplot Observations:** Temp and Ozone are positively aligned, indicating a strong correlation. Wind is in the opposite direction, suggesting an inverse relationship with Temp and Ozone.





9.Conclusion:

In Univariate Analysis Temp has a symmetrical distribution with no significant outliers. Measures of central tendency and variability indicate consistent patterns across observations. **Correlation Analysis** Temp and Ozone are moderately positively correlated, with higher temperatures leading to increased ozone levels. **Scatter Plot** Visualized a clear positive trend between Temp and Ozone. **Regression Analysis:** Both Temp and Wind significantly influence Ozone levels, with Wind showing a negative impact. Box-Cox transformation improved model diagnostics, reducing heteroscedasticity and residual variability. **In PCA** PC1 and PC2 capture 80% of the variance and adequately represent the dataset. PCA identified key relationships, with Temp and Ozone aligning positively and Wind inversely related to these variables. Temperature is a critical driver of ozone levels, with an inverse relationship to wind speed.