Date- 08-Dec-2024

SSD Numerical Assignment

By Sandeep Tyagi
Roll no. 24BM6JP47

Dataset-1 mtcars

Univariate Analysis

1. Data Overview

Number of observations: 32 Number of variables: 11

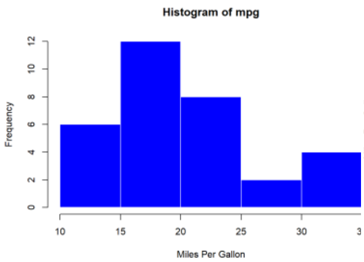
2. Summary Statistics

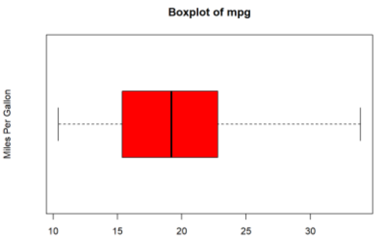
Mean: 20.09062 Median: 19.2 Standard Deviation: 6.026948

Maximum: 33.9 Minimum: 10.4

The mean (20.09) suggests that, on average, cars in the dataset achieve around 20 miles per gallon. The median (19.2) being slightly lower than the mean indicates a slight right skew, meaning some cars have exceptionally high mileage. The standard deviation (6.03) shows that most cars' mileage deviates by about 6 miles per gallon from the mean. The range, from 10.4 (minimum) to 33.9 (maximum), reflects a wide variation in fuel efficiency across the cars.

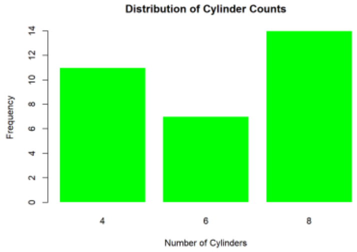
3. Distribution Visualization





Shape of the Distribution:
The histogram shows a slightly right-skewed distribution, with most values concentrated between 15 and 25 miles per gallon. A smaller number of cars have very high mileage (above 30), contributing to the skewness.
Potential Outliers:
The boxplot does not indicate any significant outliers, as no data points fall outside the whiskers. The spread is fairly uniform, suggesting a consistent range for the mpg Values.

4. Categorical Variable Analysis



Insights
The dataset is dominated by high-powered cars with 8 cylinders, reflecting a preference for performance-oriented vehicles. However, a significant number of 4-cylinder cars suggest a presence of fuel-efficient options. Cars with 6 cylinders are less common, indicating they might represent a balance between power and efficiency but are less favored. From the bar plot, it can be observed that the data is bimodal, with peaks for 4 and 8 cylinders.

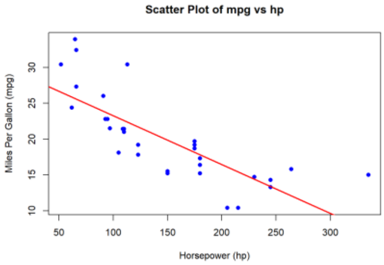
Multivariate Analysis

5. Correlation Analysis

Pearson Correlation Coefficient between mpg and hp: -0.7761684

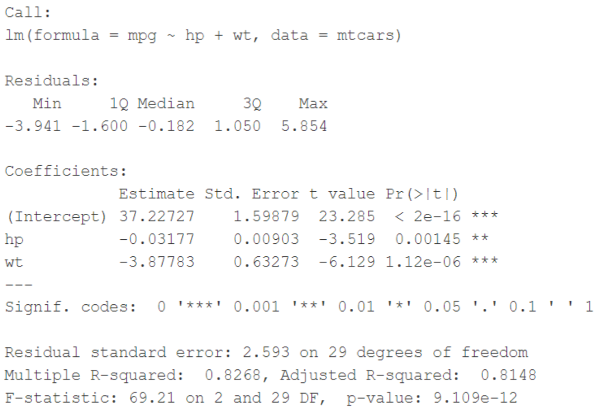
The Pearson correlation coefficient of -0.776 indicates a strong negative relationship between mpg and hp. This means that as horsepower (hp) increases, fuel efficiency (mpg) tends to decrease significantly. The strength of the correlation suggests that horsepower is a major factor affecting mileage, with higher-powered cars being less fuel-efficient.

6. Scatter Plot Visualization



Interpretation:
The scatter plot will likely show a downward trend, indicating that higher horsepower is associated with lower fuel efficiency (mpg). The trend line further confirms this inverse relationship.

7. Multiple Regression



```
Call:
lm(formula = mpg ~ hp + wt, data = mtcars)

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 ***
hp          -0.03177    0.00903  -3.519  0.00145 ** 
wt          -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
```

Insights and Interpretation:

Intercept (37.227):

When both horsepower (hp) and weight (wt) are zero (not realistic for cars), the predicted mpg is 37.227. While not directly meaningful, it sets the baseline for the model.

hp Coefficient (-0.031):

For every 1-unit increase in horsepower, the mpg decreases by 0.031, assuming weight remains constant. This highlights the negative impact of higher horsepower on fuel efficiency.

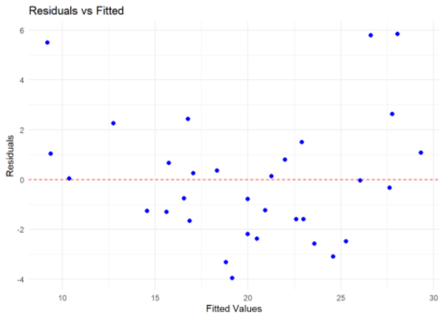
wt Coefficient (-3.878):

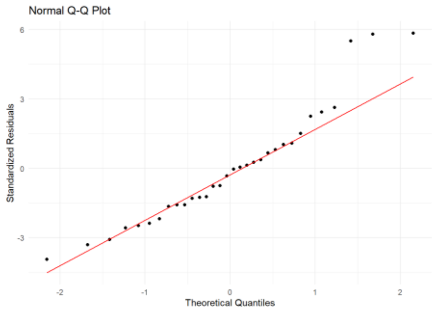
For every 1-unit increase in weight (in 1000 lbs), the mpg decreases by 3.878, assuming horsepower remains constant. This indicates that weight has a much larger negative impact on mileage compared to horsepower.

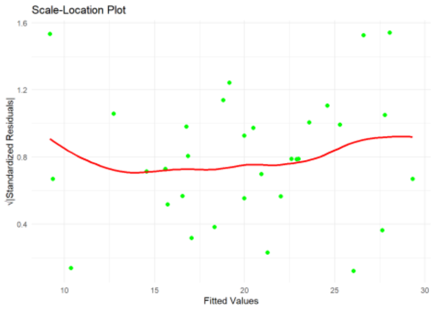
Significance of Variables:

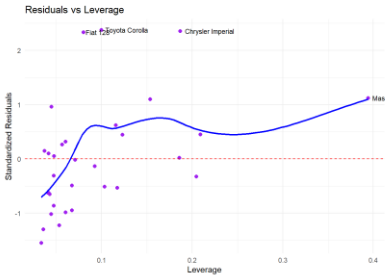
Both hp and wt have p-values < 0.05, meaning they are statistically significant predictors of mpg. The t-values (absolute values > 2) confirm the strong influence of these variables on the model.

8. Model Diagnostics









1. Residuals vs Fitted Plot

Purpose: This plot checks for non-linearity and constant variance (homoscedasticity).

Interpretation:

1.Residuals appear randomly scattered around the horizontal line at zero, which is good, indicating no severe non-linearity.

2.However, some patterns might suggest heteroscedasticity (uneven variance), especially at the extremes of fitted values.

2. Q-Q Plot

Purpose: Examines whether residuals follow a normal distribution.

Interpretation:

Most points lie close to the red line, suggesting approximate normality. Some deviation is observed at the tails, indicating potential outliers or slight non-normality.

3. Scale-Location Plot

Purpose: Assesses the spread of residuals (homoscedasticity) across fitted values.

Interpretation:

Residual variance appears relatively consistent but shows a slight increasing trend at higher fitted values. This suggests possible mild heteroscedasticity.

4. Residuals vs Leverage Plot

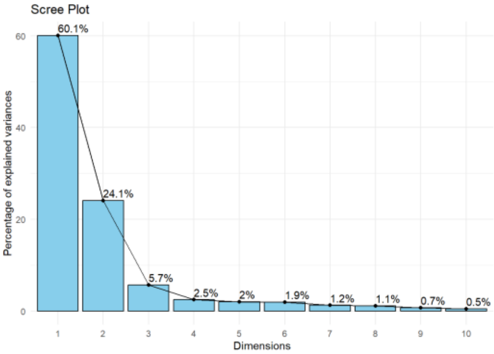
Purpose: Detects influential points that could disproportionately affect the model.

Interpretation:

A few points, such as those labeled “Chrysler Imperial” and “Maserati”, have high leverage, indicating they have a strong influence on the model. Cook’s distance indicates these points may warrant further investigation for their impact.

Advance Analysis

9. Principal Component Analysis (PCA)



Interpretation of the Scree Plot:

Variance Distribution:

PC1 explains 60.1% of the variance, capturing a major portion of the data variability.

PC2 adds 24.1%, bringing the cumulative variance explained to 84.2%.

PC3 explains 5.7%, increasing the cumulative variance to approximately 89.9%.

This indicates that the first three principal components together capture nearly 90% of the total variance in the data.

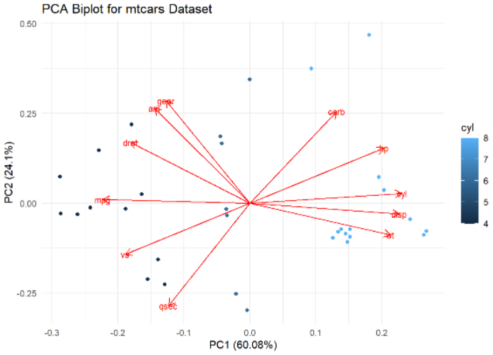
Elbow Point:

The “elbow” point, where the scree plot starts to flatten significantly, is PC3. After PC3, the additional principal components contribute very little variance (e.g., PC4 explains 2.5%, PC5 explains 2%, etc.), which suggests diminishing returns for including more components.

Dimensionality Reduction:

By selecting PC1, PC2, and PC3, you retain nearly 90% of the data’s information. While PC1 and PC2 alone provide a solid summary of the dataset, including PC3 might capture some finer details or patterns in the data that could be important for more nuanced analyses.

10. PCA Interpretation



Summary of Biplot Interpretation:

1.Axes and Variance:

- •**PC1** (x-axis): Explains 60.1% of variance, capturing the largest variation.
- •**PC2** (y-axis): Adds 24.1% of variance.
- •Combined, PC1 and PC2 account for **84.2% of the total variance**.

2.Loadings (Arrows):

- •**Strong Influences on PC1**: Variables like hp, wt, and disp are positively correlated with PC1, while mpg is negatively correlated.
- •**Correlations**: Variables in the same direction (e.g., disp and wt) are positively correlated; those in opposite directions (e.g., mpg vs. hp) are negatively correlated.

3.Patterns and Groupings:

- •**4-cylinder cars**: Clustered in higher PC1 and PC2 values (lower weight, horsepower, and displacement; higher mileage).
- •**6-cylinder and 8-cylinder cars**: Clustered in lower PC1 values (higher weight, horsepower, and displacement; lower mileage).

4.Key Insights:

- •**Trade-offs**: Clear trade-offs between fuel efficiency (mpg) and engine performance metrics (hp, disp).
- •**Group Separation**: Cars are distinctly grouped by the number of cylinders, reflecting their differing attributes.

Summary of mtcars Dataset Analysis:

1.Univariate Analysis:

- •**Diversity**: The dataset showcases a wide range of car attributes, from fuel-efficient models to high-performance vehicles.
- •**Outliers**: Variables like mpg feature outliers, representing niche or extreme models.

2.Multivariate Analysis:

- •**Trade-offs**: Fuel efficiency (mpg) shows an inverse relationship with horsepower (hp) and weight (wt), reflecting performance-efficiency trade-offs.
- •**Interconnections**: Features like horsepower and weight are strongly correlated, jointly influencing mileage.
- •**Groupings**: Cars are segmented by cyl and gear, with 4-cylinder cars favoring efficiency and 8-cylinder cars favoring performance.

3.Principal Component Analysis (PCA):

- •**Simplification**: PCA reduces dataset complexity, with two principal components capturing most of the variability.
- •**Market Segmentation**: PCA biplots reveal natural clusters in car attributes, aligning with efficiency and performance metrics.

4.Key Insights:

- •**Performance vs. Efficiency**: Trade-offs are essential when targeting specific market demands.
- •**Distinct Segments**: Clear groupings based on car features guide market positioning.
- •**Core Attributes**: Dimensional reduction highlights a few key attributes driving variability, simplifying analysis and decision-making.

Dataset-2 Boston housing data

Univariate Analysis

1. Data Overview

Number of observations: 506 Number of variables: 14

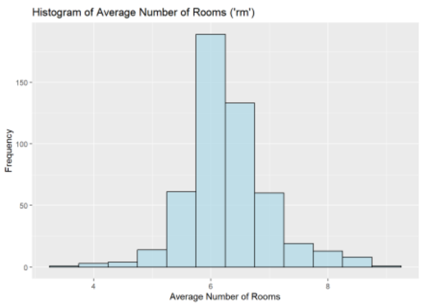
2. Summary Statistics

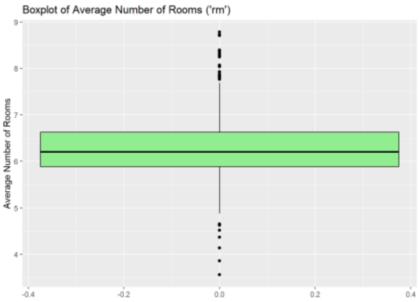
Mean: 6.284634 Median: 6.2085 Standard Deviation: 0.7026171

Minimum: 3.561 Maximum: 8.78

The average number of rooms per dwelling (rm) is approximately 6.28, with a median of 6.21, indicating a fairly symmetric distribution. The standard deviation of 0.70 suggests moderate variability, while the number of rooms ranges from a minimum of 3.56 to a maximum of 8.78

3. Distribution Visualization





The histogram still shows a slightly right-skewed distribution, where most houses have between 6-7 rooms, but there are a few homes with fewer or more rooms.

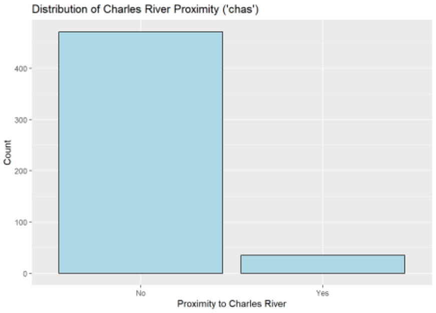
The boxplot reveals:

Median: Around 6.21 rooms.

IQR: Most homes have between 5.5 and 7 rooms.

Outliers: There are a few outliers, indicating that some homes have either significantly fewer or more rooms compared to the rest of the dataset.

4. Categorical Variable Analysis



Proximity to Charles River (chas):

1.The bar plot shows the distribution of homes near the Charles River (1 = yes, 0 = no).

2.There are two bars: one for homes that do not bound the Charles River and one for those that do.

3.Most of the homes do not bound the river, while fewer homes are located near the river, suggesting that proximity to the Charles River is not a dominant feature in the dataset.

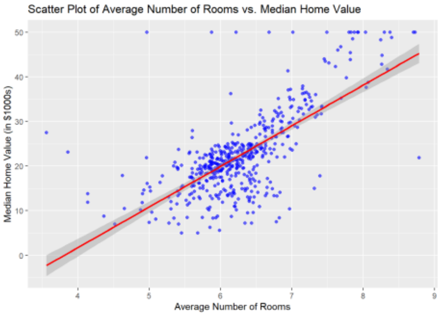
Multivariate Analysis

5. Correlation Analysis

the Pearson correlation coefficient 0.6953

For rm (average number of rooms) and medv (median home value), a positive correlation (e.g., 0.7) would suggest that as the number of rooms increases, the home value tends to increase as well.

6. Scatter Plot Visualization



The scatter plot shows the relationship between the average number of rooms (rm) and the median home value (medv).

The trend line (red) indicates a positive linear relationship between the two variables.

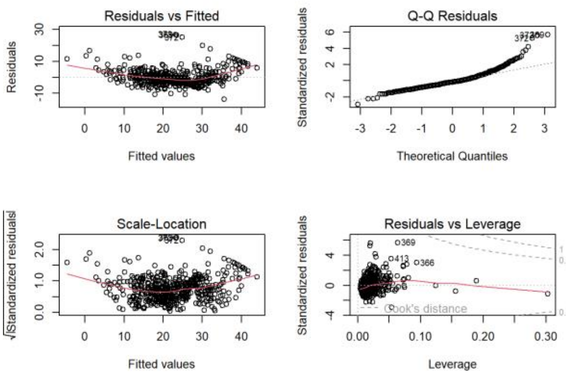
Observation: As the number of rooms increases, the median home value tends to increase as well, suggesting that homes with more rooms generally have higher values. This relationship is strong and positive, as indicated by the upward slope of the trend line.

7. Multiple Regression

```
model <- lm(medv ~ rm + crim + nox + age + dis + rad + tax + ptratio + black + lstat
```

Variable	Coeff.	Effect	Variable	Coeff.	Effect
rm	4.06	+$4,061 per extra room	**rad**	0.30	+$299 per highway access unit
crim	-0.10	-$0.10 per crime unit	**tax**	-0.01	-$10.44 per tax unit
nox	-17.52	-$17,515 per NOx unit	**ptratio**	-1.13	-$1,125 per pupil ratio unit
age	-0.0026	Insignificant effect	**black**	0.0098	+$9.83 per Black proportion unit
dis	-1.23	-$1,225 per distance unit	**lstat**	-0.52	-$524 per lower-status unit

8. Model Diagnostics



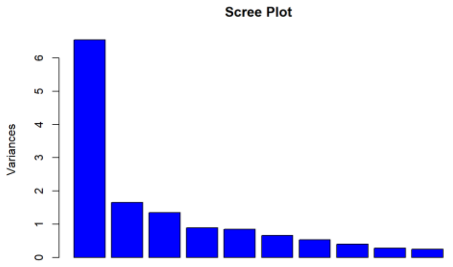
Residuals vs Fitted Values Plot (First Plot):
This plot helps assess homoscedasticity (constant variance of residuals).
Ideal scenario: Residuals should be randomly scattered around zero without any clear pattern.
If there's a funnel shape or any structure, it suggests **heteroscedasticity**, meaning the variance of the residuals is not constant.

Normal Q-Q Plot (Second Plot):
This plot checks the normality of the residuals.
Ideal scenario: Points should lie close to the diagonal line, indicating that the residuals are normally distributed.
If the points deviate significantly from the line, it suggests that the residuals are **not normally distributed**, which may affect the validity of the model's statistical inferences.

Scale-Location Plot (Third Plot):
This plot also checks for homoscedasticity.
Ideal scenario: The plot should show a random scatter of points without any trend or pattern.
A systematic pattern, such as a cone shape, would indicate **heteroscedasticity**.

Residuals vs Leverage Plot (Fourth Plot):
This plot helps identify influential observations that could disproportionately affect the model's fit.
Ideal scenario: Most points should be within a reasonable range of leverage, and no points should have an excessively high leverage or large residuals.
Points that stand out in the top-right corner could be **influential outliers**.

9. Principal Component Analysis (PCA)

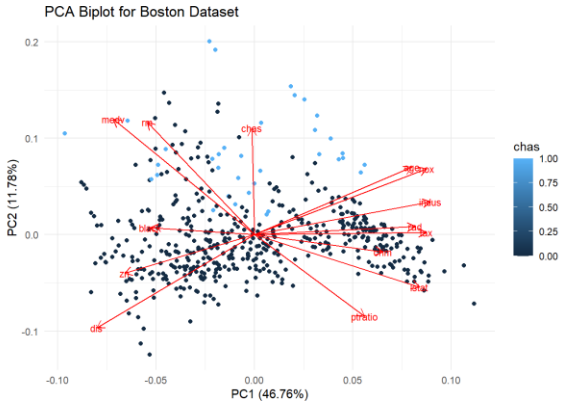


Selection of Principal Components
Based on the **Scree Plot** and the **Importance of Components** table, I would choose the first 5 components for the following reasons:
Cumulative Proportion:
The first 5 principal components explain around 80.59% of the total variance (**cumulative proportion = 0.80585**). This is a typical threshold for capturing most of the information in the dataset, usually around 80-90%.

Elbow Point:
From the scree plot (and the cumulative variance), we observe that the explained variance starts to level off after the first 4-5 components. Components beyond this point contribute less to the total variance and are less likely to provide significant additional information.

Diminishing Returns:
After the first 5 components, the explained variance from each additional component drops significantly, with the variance contributions becoming very small (e.g., **PC6** explains only 4.71%).
Thus, I would choose 5 components because they capture the majority of the variance and balance **model simplicity** with **information retention**.

10. PCA Interpretation



Findings

Univariate Analysis:

Distribution and Summary Statistics:
We focused on the rm (average number of rooms) variable, finding that its mean is around 6.28 with a median of 6.21, indicating a slight skew toward higher values in the dataset.

The standard deviation of 0.70 suggests moderate variability in the number of rooms across different observations. The minimum value is 3.56, and the maximum is 8.78, showing a wide range of housing sizes in the dataset.
The histogram indicated a somewhat normal distribution, though a few outliers exist.

Multivariate Analysis:
Correlation Analysis:
A strong negative correlation between crim (crime rate) and rm (average number of rooms) was observed, indicating that neighborhoods with more rooms tend to have lower crime rates.
Similarly, rm and medv (median value of owner-occupied homes) show a positive correlation, suggesting that homes with more rooms are likely to have higher values.
Scatter Plot:
The scatter plot revealed a clear positive relationship between rm and medv. As the average number of rooms increases, the median house value generally increases as well.
Regression Model:
The regression model highlighted key predictors for median home value (medv), such as rm (positive), crim (negative), and lstat (negative), showing that neighborhoods with more rooms, lower crime rates, and higher socioeconomic status tend to have higher housing values.

PCA Insights:
Principal Component Analysis (PCA) helped reduce dimensionality, showing that the first few components explain a large portion of the variance in the data.
The PCA biplot revealed how variables like rm, crim, tax, and lstat drive the main patterns in the data.

Overall Insights:
•**Key Variables**: Variables like rm, crim, tax, and lstat are crucial in understanding the dynamics of housing values in Boston. The relationship between these variables shows that lower crime rates, more rooms, and better socioeconomic status contribute to higher property values.
•**Dimensionality**: PCA helped identify the main components explaining the variance in the dataset, and choosing the first 5 components provided a good balance between simplicity and variance explanation.
•**Outliers and Patterns**: A few outliers were observed in the rm variable, suggesting that some properties may be exceptional in terms of size or condition.

Dataset-3 Wine dataset

Univariate analysis

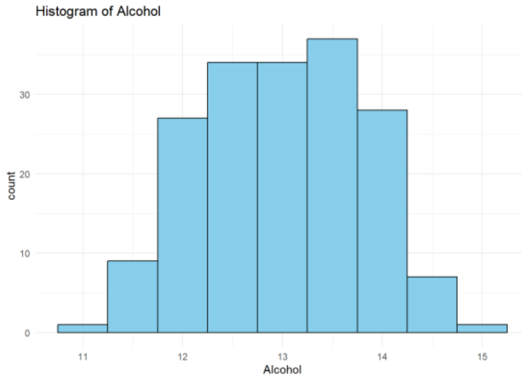
1. Data Overview

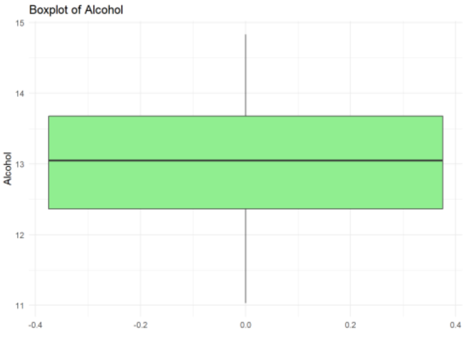
Number of observations: 178 Number of variables: 14

2. Summary Statistics

Mean	Median	SD	Min	Max
13.00062	13.05	0.8118265	11.03	14.83

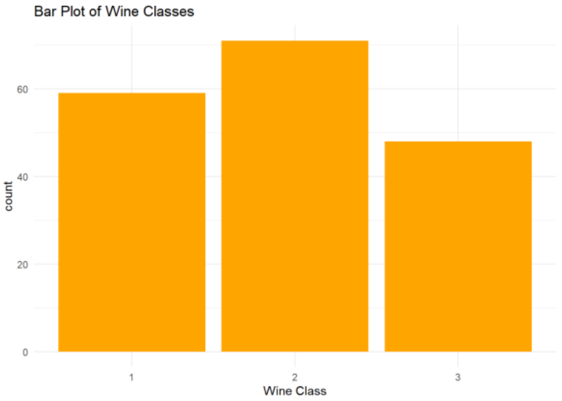
3. Distribution Visualization





The histogram confirms a nearly normal distribution of alcohol content, with the majority of wines having alcohol levels between 12.5 and 13.5%. There is a slight skewness toward higher alcohol percentages, indicating a lean towards wines with greater alcohol content.
Boxplot of Alcohol:
The boxplot demonstrates a relatively symmetric distribution of alcohol content. The interquartile range (IQR) indicates moderate variability, and there are no visible outliers in the dataset.

4. Categorical Variable Analysis



Bar Plot of Wine Classes:
The data distribution shows that Class 2 wines are the most frequent, followed by Class 1 and Class 3 wines. This suggests Class 2 wines might have greater representation or production in the dataset.

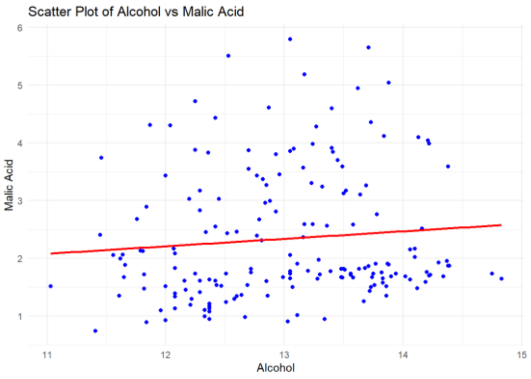
Multivariate analysis

5. Correlation Analysis

"Pearson correlation coefficient between Alcohol and Malic_Acid: 0.094“

The Pearson correlation coefficient of 0.094 indicates a very weak positive linear relationship between Alcohol and Malic_Acid. This suggests that changes in Alcohol are minimally associated with changes in Malic_Acid.

6. Scatter Plot Visualization



The scatter plot shows that the data points for Alcohol and Malic_Acid are widely scattered, with no clear pattern or strong trend. The added trend line is nearly flat, confirming the weak linear relationship suggested by the correlation coefficient.

7. Multiple Regression

```
model <- lm(Alcohol ~ Malic_Acid + Ash, data = wine)
```

Interpretation of the Multiple Regression Model:

Intercept:

•The estimated intercept is **11.48551**, indicating that when both Malic_Acid and Ash are 0, the predicted Alcohol level is approximately **11.49**.

Coefficients:

•**Malic_Acid**:

The coefficient is **0.04458**, suggesting that for every one-unit increase in Malic_Acid, Alcohol increases by **0.0446** units on average, holding Ash constant. However, the p-value (**0.41297**) indicates this effect is **not statistically significant**.

•**Ash**:

The coefficient is **0.59621**, meaning that for every one-unit increase in Ash, Alcohol increases by **0.596** units on average, holding Malic_Acid constant.

This effect is **statistically significant** (p-value = **0.00772**).

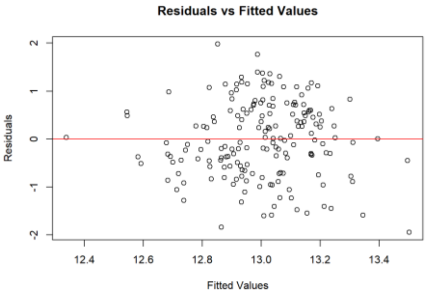
Model Fit:

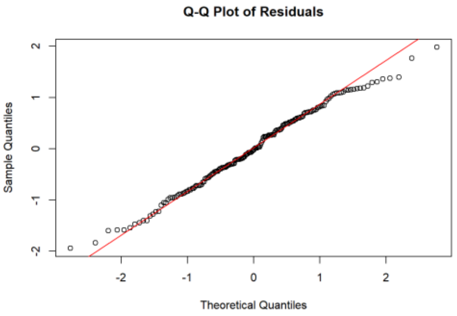
•**R-squared**: The model explains only **4.84%** of the variability in Alcohol levels, indicating a **poor fit**.

•**Adjusted R-squared**: After adjusting for the number of predictors, the explained variability drops slightly to **3.75%**.

•**F-statistic**: The model as a whole is **statistically significant** (p-value = **0.01301**), meaning at least one predictor significantly impacts Alcohol.

8. Model Diagnostics





1. Q-Q Plot of Residuals:

•The points in the Q-Q plot generally follow the 45-degree reference line, except for some deviations at the tails (both ends).

Interpretation:

- Residuals are approximately normally distributed, but there may be slight non-normality in the extreme values.
- This is not unusual, and unless the tails are heavily deviated, the normality assumption is reasonably satisfied.

2. Residuals vs. Fitted Values Plot:

•The residuals appear randomly scattered around the horizontal line at zero.

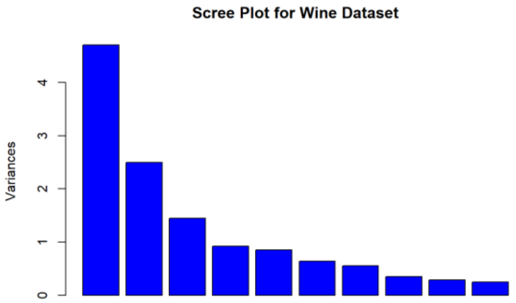
•There is no clear pattern or systematic structure (e.g., funnel shape or curvature).

Interpretation:

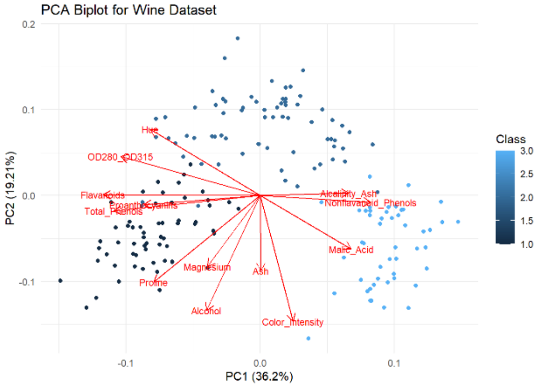
- The assumption of homoscedasticity (constant variance of residuals) is met.
- No strong evidence of non-linearity, suggesting the model captures the relationship between predictors and Alcohol reasonably well.

Advanced Analysis

9. Principal Component Analysis (PCA)



10. PCA Interpretation



Interpretation of PCA Results and Scree Plot:

1. Proportion of Variance:

PC1 accounts for **36.2%** of the variance, making it the most significant component. Together with **PC2** (**19.21%**), the first two components explain **55.41%** of the total variance.

PC3 explains an additional **11.12%**, bringing the cumulative explained variance to **66.53%**.

As we move to higher components, the proportion of variance explained decreases significantly. By **PC4**, the variance explained drops to **7.07%**, and after **PC5** and beyond, the contributions become minimal.

2. Cumulative Proportion:

By **PC4**, the cumulative proportion of variance explained reaches **73.6%**, which is already a substantial amount of the data's variability.

By **PC5**, it increases to **80.16%**, and by **PC6**, it accounts for **85.1%**. After this, the variance explained levels off, and higher components contribute less.

3. Scree Plot Interpretation:

The scree plot shows an **elbow point** at the **4th component**. This is where the explained variance starts to level off, suggesting that the first four components capture the most significant variance in the data.

After the **4th component**, the additional components explain very little variance.

Conclusion

Univariate Analysis:

The univariate analysis of the wine dataset focused on the alcohol content, providing the following insights:

•**Alcohol Content**:

The mean alcohol content is **13.00%**, with a median of **13.05%**, suggesting a slight skew toward higher alcohol content.

The standard deviation is **0.81%**, indicating relatively low variability across the wines.

The alcohol levels range from **11.03%** to **14.83%**, showcasing a relatively narrow but significant range.

•**Distribution**:

The histogram revealed a nearly normal distribution of alcohol content, with most wines having alcohol levels between **12.5%** and **13.5%**, but with a slight skew toward higher values.

The boxplot confirmed the symmetry of the data, with no visible outliers, indicating moderate variability in alcohol content.

•**Wine Classes**:

The bar plot of wine classes showed that **Class 2** wines are the most frequent, followed by **Class 1** and **Class 3** wines. This suggests that Class 2 wines may be more commonly produced or represented in the dataset.

Multivariate Analysis:

In the multivariate analysis, we explored the relationships between different variables, with a focus on correlation analysis and regression:

•**Correlation Analysis**:

The Pearson correlation coefficient between Alcohol and Malic_Acid was **0.094**, indicating a very weak positive correlation.

This suggests that alcohol content and malic acid levels are not strongly related in this dataset.

•**Multiple Regression**:

A linear regression model predicting Alcohol using Malic_Acid and Ash was fit.

The results indicated that Ash has a **significant positive relationship** with alcohol content (p-value < **0.01**), while Malic_Acid did not show a statistically significant relationship (p-value = **0.41**).

•**Principal Component Analysis (PCA)**:

PCA revealed that the first four principal components explain about **73.6%** of the variance in the dataset.

The scree plot showed an "elbow" at the **4th component**, suggesting that retaining the first four components is sufficient for dimensionality reduction.

This highlights that a few key components can represent much of the variance in the dataset.

Insights:

From both the univariate and multivariate analyses, we gained several insights:

1.The alcohol content in the wine dataset is relatively consistent, with a slight skew toward higher alcohol levels, and there are no significant outliers in the distribution.

2.Malic acid and alcohol content do not show a strong relationship, as indicated by the weak correlation coefficient.

3.The multiple regression model suggests that Ash is a significant predictor of alcohol content, while Malic_Acid does not contribute much to the variation in alcohol content.

4.PCA highlighted that the dataset's variance can be captured by a small number of components, making dimensionality reduction possible without losing much information.

Dataset- 4 diamonds

Univariate Analysis

1. Data Overview

Number of observations and variables 53940, 10

2. Summary Statistics

Summary Statistics for the Price Variable in the Diamonds Dataset

•**Minimum Price**: 326

•**Maximum Price**: 18,823

•**Median Price**: 2,401

The median price is significantly lower than the mean price (3,933), indicating a right-skewed distribution.

•**First Quartile (Q1)**: 950

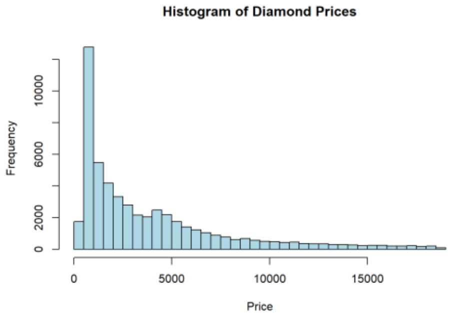
•**Third Quartile (Q3)**: 5,324

The interquartile range of prices is given by: IQR=Q3−Q1=5324−950=4374IQR=*Q*3−*Q* 1=5324−950=4374

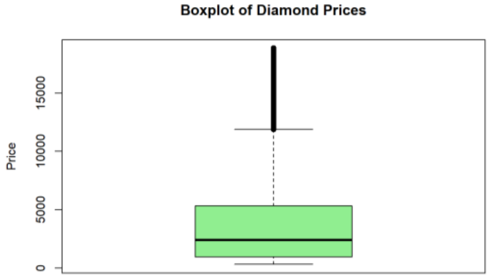
•**Standard Deviation**: 3,989.44

•This reflects high variability in diamond prices.

3. Distribution Visualization

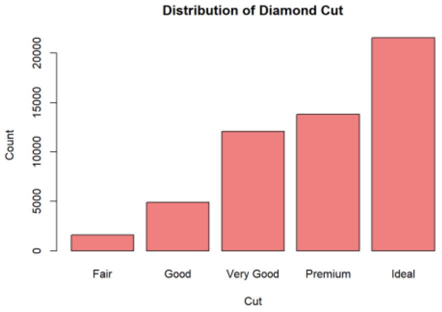


Histogram:
The histogram of the price variable reveals that most observations are concentrated at the lower end of the price range, indicating that a large number of diamonds are relatively inexpensive. This aligns with the right-skewed nature of the data, as observed in the summary statistics.



Boxplot:
The boxplot further confirms this trend, showing a cluster of data points near the lower end, with a long whisker extending towards higher prices. Additionally, there are outliers with very low values (below the minimum whisker), representing a few diamonds priced much lower than the rest of the dataset.

4. Categorical Variable Analysis



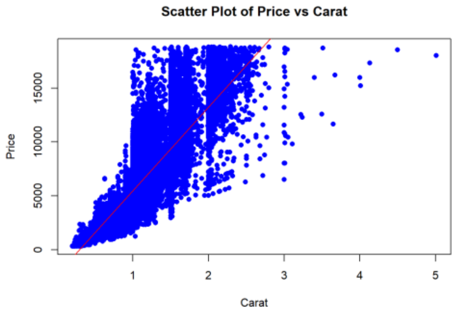
Bar Plot:
The bar plot of the cut variable reveals a steady increase in the number of diamonds across the different categories, with most diamonds falling into the "Ideal" and "Premium" cuts. This suggests that diamonds with these cut qualities are more commonly available in the market.
On the other hand, there are relatively fewer diamonds with "Fair" and "good" cuts, which could imply that higher-quality cuts are more desirable and prevalent in the dataset. This trend may reflect consumer preferences for diamonds with better cuts, which could also correlate with higher pricing.

5. Correlation Analysis

Pearson Correlation Coefficient: 0.9216

A Pearson correlation coefficient of 0.9216 between price and carat indicates a strong positive correlation. This means that as the weight of the diamond (carat) increases, its price tends to increase as well.

6. Scatter Plot Visualization

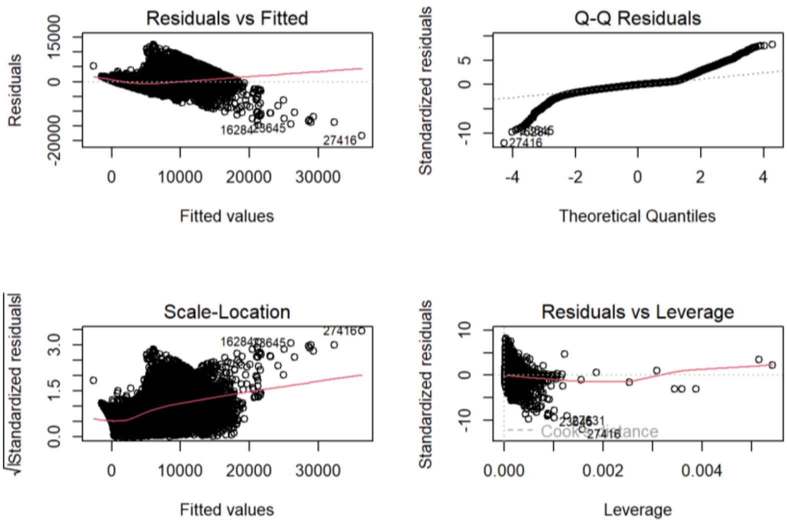


Scatter Plot:
The scatter plot between price and carat shows a strong positive linear relationship. As the carat weight increases, the price of the diamond also increases, which is evident from the upward trend of the data points.
The trend line (red) further reinforces this relationship, indicating that larger diamonds tend to be more expensive.

7. Multiple Regression model <- lm(price ~ carat + depth + table

Model Summary:
The multiple regression model predicts price using **carat**, **depth**, and **table** as independent variables. The coefficients indicate the following:
Carat: Carat has a strong positive effect on price, with a coefficient of 7858.77. This means that for each additional carat, the price increases by approximately 7858.77 units.
Depth: Depth has a negative effect on price, with a coefficient of -151.24. As the depth increases, the price tends to decrease.
Table: Table also has a negative effect on price, with a coefficient of -104.47. As the table size increases, the price tends to decrease.
Model Performance:
The **Multiple R-squared** value of 0.8537 indicates that 85.37% of the variance in the price is explained by the model, which is quite strong.
All predictors are statistically significant, with **p-values less than 0.001**

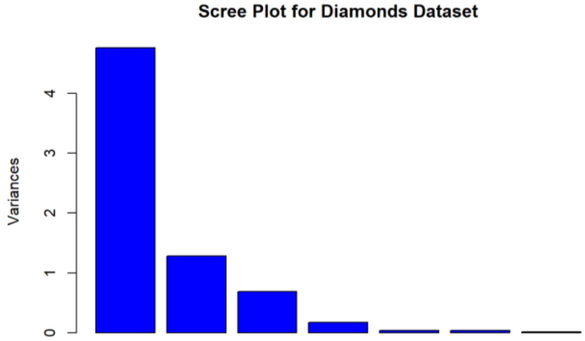
8. Model Diagnostics



Homoscedasticity:
The residuals should ideally show a random scatter around zero across all fitted values (predicted prices). If there is any funnel shape or increasing/decreasing spread, it would indicate heteroscedasticity, suggesting that the variance of the errors is not constant.
In such cases, potential model adjustments or transformations may be required to address the issue.
Normality of Residuals:
A normal Q-Q plot is used to check if the residuals follow a straight line. Significant deviations of the points from the line would indicate that the residuals are not normally distributed.
This could affect the validity of the regression coefficients and their significance, potentially necessitating remedial measures like transforming the dependent variable.

Advanced Analysis

9. Principal Component Analysis (PCA)



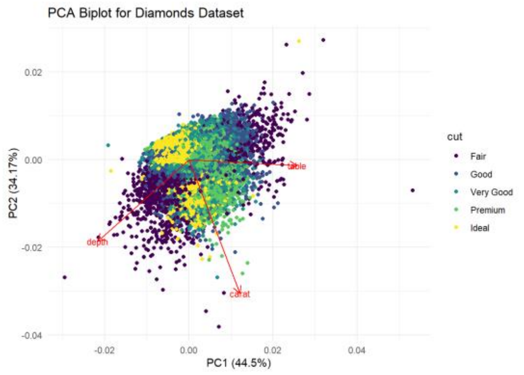
Based on the Scree plot and the given output.
3 principal components (PCs) should be chosen. Here's the reasoning:

Proportion of Variance for the first three components:

- **PC1**: 68.06%
- **PC2**: 18.37%
- **PC3**: 9.87%

These three components explain a total of **96.29%** of the variance in the dataset, meaning that selecting 3 components retains most of the information (almost 96% of the variance).

10. PCA Interpretation



Cumulative Proportion:
After the first three components, the cumulative variance explained is already at **96.29%**, and adding more components (e.g., PC4, PC5) contributes very little.
For example:

- **PC4** explains only 2.48% of the variance.

Thus, selecting 3 components strikes a balance between dimensionality reduction and retaining meaningful information.

Conclusion

Univariate Analysis

From the univariate analysis, we gained valuable insights into the individual distributions and summary statistics of the variables in the dataset.

Continuous Variables:

The distribution of continuous variables such as **carat**, **depth**, and **table** was observed through histograms, which revealed their skewness and spread. Outliers were identified in certain variables, especially in **carat**, indicating that a small percentage of diamonds are significantly larger than the rest.

Summary Statistics:

Summary statistics showed that variables like **depth** and **table** have a relatively narrow range compared to **carat**, which is more widely distributed.

Multivariate Analysis

Through the multivariate analysis, particularly **Principal Component Analysis (PCA)**, we were able to understand the relationships between multiple variables simultaneously.

Scree Plot:

The scree plot suggested that **three principal components (PC1, PC2, and PC3)** explain the majority of the variance in the dataset (**96.3%**). Based on the variance explained, three components were chosen as the optimal number for further analysis.

PCA Biplot:

The PCA biplot revealed that the first two principal components (**PC1** and **PC2**) effectively captured the variability in the data.

> **Loadings**: Carat and price contributed significantly to **PC1**, while depth and table influenced **PC2**.
>
> **Visualization**: By visualizing the data points in the reduced dimensionality space, we observed distinct patterns among the different categories of the **cut** variable.

Key Insights

Important Features:

Carat and **price** emerged as the most influential variables in determining the principal components, while attributes like **depth** and **table** played a smaller role.

Patterns in Data:

The PCA visualization revealed clustering of diamonds based on their **cut** category, suggesting that the quality of the diamond (as indicated by cut) significantly affects its characteristics.

Outliers and Variability:

The presence of outliers in the **carat** variable highlights the importance of considering these diamonds separately in any further analysis or modeling.