INDIAN STATISTICAL INSTITUTE, KOLKATA

Statistical Structures in Data
Prof. Subhajit Dutta

Numerical Assignment Report:
Exploration of Univariate and Multivariate Data Using R

Datasets used:

- *Cars93*
- *Boston*
- *Air quality*
- *Diamonds*

- Gaurav Malani
24BM6JP19

DATASET 1. CARS93

(*R Library – 'Mass'*)

Univariate Analysis

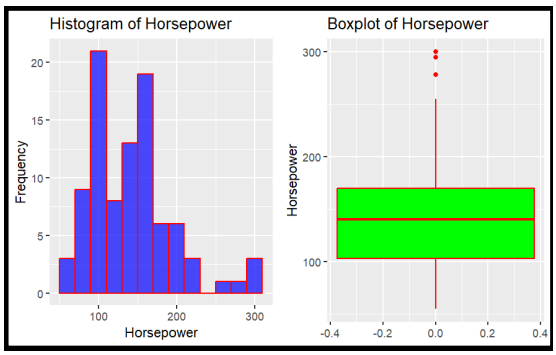
1. **About the Dataset:** The Cars93 dataset contains 93 observations and 27 variables related to car specifications, including price, performance, and fuel efficiency. It provides a comprehensive overview of cars available in 1993

2. **Summary statistics for 'Horsepower':**

Statistic <chr>	Value <dbl>
Mean	143.82796
Median	140.00000
Standard Deviation	52.37441
Minimum	55.00000
Maximum	300.00000

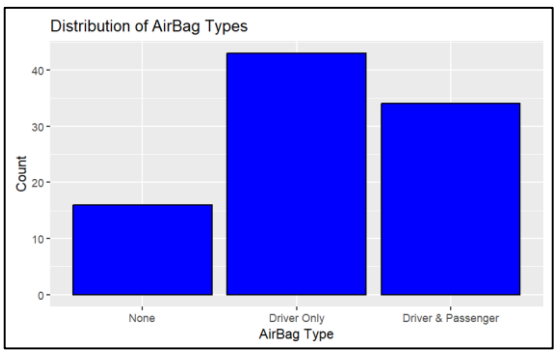
The summary statistics for the Horsepower variable in the Cars93 dataset indicate that the **average** horsepower is approximately **143.83**, with a median of **140**, suggesting a **fairly symmetric distribution**. The standard deviation of 52.37 shows **considerable variability**, with values ranging from a minimum of 55 to a maximum of 300, highlighting a wide range of engine power among the cars.

3. **Distribution Visualization:**



- **Distribution Shape:** The histogram shows a **slightly right-skewed distribution**, with most cars having horsepower in the range of 100 to 200.
- **Central Tendency**: The mean and median horsepower are close (143.83 and 140, respectively), indicating a relatively symmetric distribution with minimal skewness.
- Outliers: The boxplot **highlights several outliers above 250 horsepower**, representing high-performance cars that deviate significantly from the majority.
- Variability: The interquartile range (IQR) in the boxplot suggests that 50% of the cars have horsepower values between approximately 100 and 180, with a standard deviation of 52.37 **reflecting moderate variability.**

4. **Categorical Variable Analysis – 'Airbag Type':**



- The bar chart for the **AirBags** variable shows that **most cars** in the dataset are equipped with **"Driver Only" airbags**, followed by "Driver & Passenger" airbags. A smaller proportion of cars have no airbags, indicating a trend toward enhanced safety features in vehicles from the dataset.

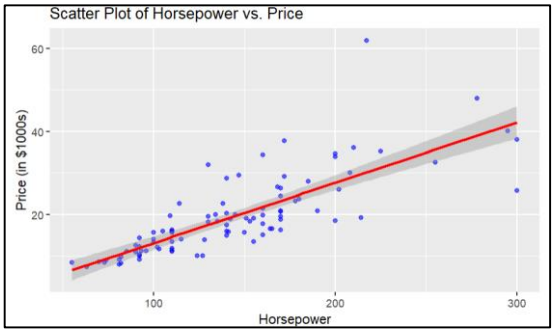
Multivariate Analysis

5. Correlation Analysis:

Variable1 <chr>	Variable2 <chr>	Correlation <dbl>
Horsepower	Price	0.7882176

- The correlation between Horsepower and Price is **0.788**, **indicating a strong positive relationship.** This suggests that as the horsepower of a car increases, its price tends to rise significantly, reflecting the impact of engine performance on car pricing.

6. Scatter Plot Visualization:



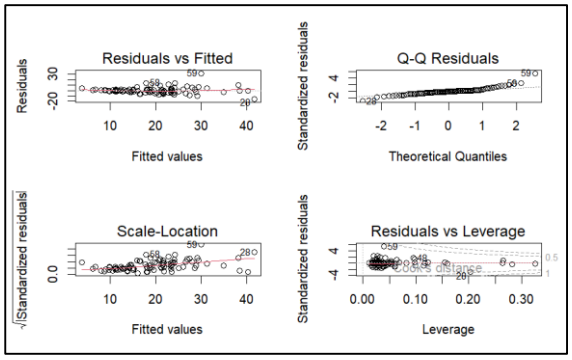
- **Positive Relationship**: The scatter plot shows a strong positive linear relationship between Horsepower and Price, indicating that cars with **higher horsepower tend to have higher prices.**
- **Trend Line**: The red regression line **fits the data well,** further confirming the strong correlation (r=0.788) between these variables.
- **Variability at Higher Horsepower**: While the relationship is consistent, there is slightly more variability in car prices as horsepower increases, particularly for cars with horsepower above 200.

7. Multiple Regression on 'Car Price':

Term <chr>	Estimate <dbl>	Std_Error <dbl>	t_Value <dbl>	P_Value <dbl>
(Intercept)	-3.23	10.55	-0.31	0.76
Horsepower	0.13	0.02	6.88	0.00
MPG.city	-0.08	0.21	-0.40	0.69
EngineSize	-0.80	1.18	-0.68	0.50
Weight	0.00	0.00	1.02	0.31

- **Significant Predictor: Horsepower has a highly significant positive effect on Price** (Estimate = 0.13, p-value < 0.01), indicating that an increase in horsepower is strongly associated with higher car prices.
- **Non-Significant Predictors**: Variables such as **MPG.city, EngineSize, and Weight have high p-values (> 0.05)**, suggesting they do not significantly contribute to predicting car prices in this model.
- Intercept: **The intercept (-3.23) is not statistically significant** (p-value = 0.76), meaning it does not provide meaningful information when all predictors are at their baseline values.

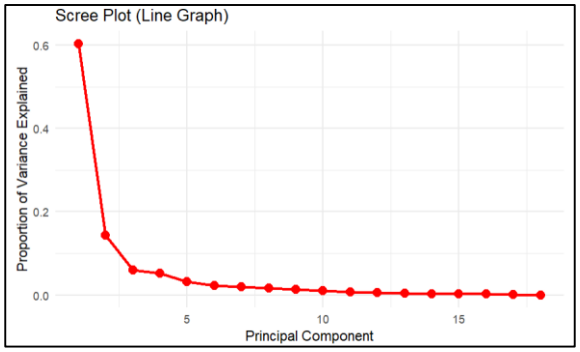
8. Model Diagnostics:



- **Residuals vs. Fitted Plot:** The residuals show some random scatter around zero, indicating that the **linearity assumption is mostly satisfied**. However, slight deviations suggest potential non-linearity or heteroscedasticity.
- Q-Q Plot: The residuals mostly **follow the theoretical quantile line**, suggesting that the normality assumption is reasonably met. However, there are minor deviations at the tails, indicating slight non-normality.
- Residuals vs. Leverage Plot: A few points with high leverage (e.g., observation 59) are visible, but they **do not exert excessive influence** on the model, as Cook's distance remains below 0.5 for most observations.

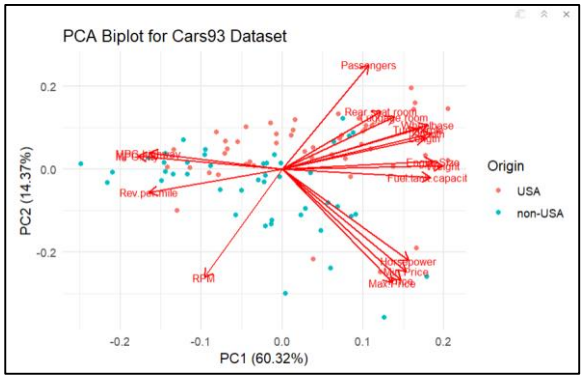
9. PCA

```
Importance of components:
                          PC1    PC2    PC3     PC4    PC5     PC6
Standard deviation     3.2951 1.6085 1.0367 0.97046 0.7506 0.63231
Proportion of Variance 0.6032 0.1437 0.0597 0.05232 0.0313 0.02221
Cumulative Proportion  0.6032 0.7469 0.8066 0.85896 0.8903 0.91248
                          PC7     PC8     PC9    PC10    PC11
Standard deviation     0.60061 0.53418 0.49519 0.44716 0.36079
Proportion of Variance 0.02004 0.01585 0.01362 0.01111 0.00723
Cumulative Proportion  0.93252 0.94837 0.96199 0.97310 0.98033
                          PC12    PC13    PC14    PC15    PC16
Standard deviation     0.31596 0.2876 0.24512 0.23013 0.18585
Proportion of Variance 0.00555 0.0046 0.00334 0.00294 0.00192
Cumulative Proportion  0.98588 0.9905 0.99381 0.99676 0.99867
                          PC17     PC18
Standard deviation     0.15447 0.002233
Proportion of Variance 0.00133 0.000000
Cumulative Proportion  1.00000 1.000000
```



- Variance Explained: The first principal component **(PC1) explains the majority of the variance (~60%),** indicating that most of the variability in the dataset is captured by this single component. The second component (**PC2) adds ~14%,** and subsequent components contribute progressively less.
- Scree Plot Analysis: The scree plot shows an **"elbow" after PC2**, suggesting that the first two components capture most of the meaningful variance, and additional components contribute diminishing returns.
- Dimensionality Reduction: Based on the cumulative proportion of variance, retaining the first **4-5 components would explain over 85%** of the total variance, making it a suitable choice for dimensionality reduction while preserving most of the information.

10. PCA Interpretation:



- **PC1 Dominance:** PC1 explains 60.32% of the variance, with **strong positive loadings** for variables like **Horsepower, Weight, and EngineSize**, indicating their significant contribution to car price and performance.
- **PC2 Differentiation**: PC2 (14.37% variance) highlights variables like **MPG.city and MPG.highway with high positive loadings**, capturing fuel efficiency, which contrasts with performance-related variables in PC1.
- Groupings: Cars from the **USA cluster more toward higher values of PC1** (higher horsepower and weight), while **non-USA cars align with PC2**, emphasizing fuel efficiency.

Conclusion:

Univariate Analysis: The histogram and boxplot for Horsepower revealed a slightly right-skewed distribution, with most cars having horsepower between 100 and 200. Outliers at higher horsepower values represent high-performance vehicles, indicating variability in engine power.

Multivariate Analysis: The scatter plot with regression demonstrated a strong positive relationship between Horsepower and Price, confirming that higher engine power is associated with higher car prices. The PCA biplot showed that the first two principal components explained 74.69% of the variance, with variables like Horsepower, Weight, and EngineSize contributing significantly to PC1, while fuel efficiency (MPG.city) dominated PC2.

Final Insight from Dataset: The Cars93 dataset analysis highlighted key relationships between car features and price. Univariate analysis showed variability in engine power, while multivariate regression confirmed Horsepower as a strong predictor of price. PCA effectively reduced dimensionality, emphasizing performance-related variables as primary drivers of variability, with clear groupings based on car origin (USA vs. non-USA)

DATASET 2. BOSTON HOUSING

(*R Library – 'Mass'*)

Univariate Analysis

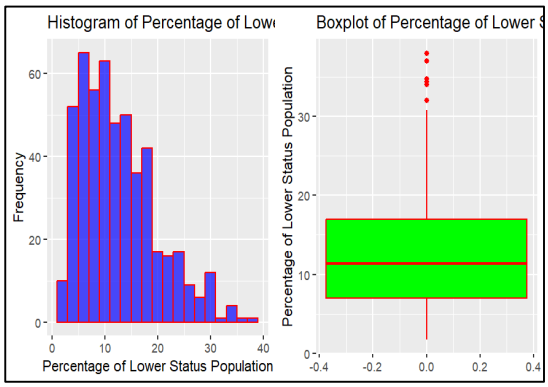
1. **About the Dataset:** The `Boston` dataset contains 506 observations and 14 variables related to housing prices and socioeconomic factors in Boston. It is widely used for regression and statistical modeling task.

2. **Summary statistics for 'lstat':**

Statistic <chr>	Value <dbl>
Mean	12.653063
Median	11.360000
Standard Deviation	7.141062
Minimum	1.730000
Maximum	37.970000

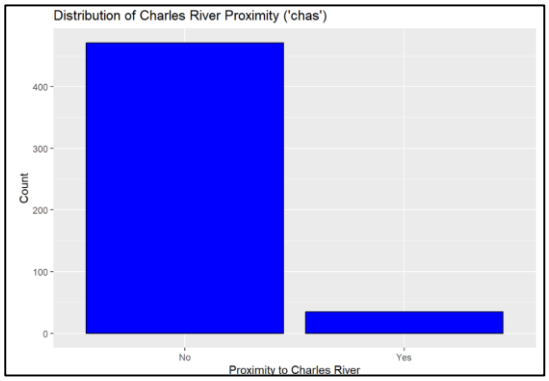
The summary statistics for the lstat variable (percentage of lower status population) indicate that the average percentage across neighborhoods is approximately **12.65%,** with a median of **11.36%,** suggesting a slightly right-skewed distribution. The standard deviation of **7.14%** shows moderate variability in the data. The minimum value of **1.73%** and maximum value of **37.97%** highlight a wide range, indicating significant differences in socioeconomic status across neighborhoods in the dataset.

3. **Distribution Visualization:**



- **Distribution Shape:** The histogram shows a **right-skewed distribution**, with most neighborhoods having a lower percentage of the lower status population (around 5-15%).
- **Outliers:** The boxplot highlights the presence of several **outliers on the higher end**, indicating some neighborhoods have significantly higher percentages of lower status populations.
- **Spread and Central Tendency**: The interquartile range (IQR) in the boxplot suggests that **50% of the data** lies between approximately **6% and 17%, with a median around 11%,** consistent with the summary statistics.
- Range: The histogram and boxplot confirm a wide range in lstat values, from as low as 1.73% to as high as 37.97%.

4. **Categorical Variable Analysis – 'Chas':**



- Proximity to Charles River (Bar Chart): The **majority of properties are not located near the Charles River**, as indicated by the overwhelming count in the "No" category compared to the much smaller "Yes" category.

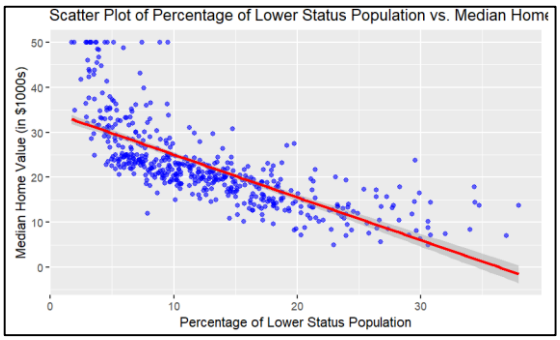
Multivariate Analysis

5. Correlation Analysis:

Variable1 <chr>	Variable2 <chr>	Correlation <dbl>
lstat	medv	-0.7376627

- The **strong negative correlation (−0.74)** between **lstat** (percentage of lower status population) and **medv** (median home value) indicates that as the percentage of **lower status population increases**, the **median home value tends to decrease** significantly.

6. Scatter Plot Visualization:



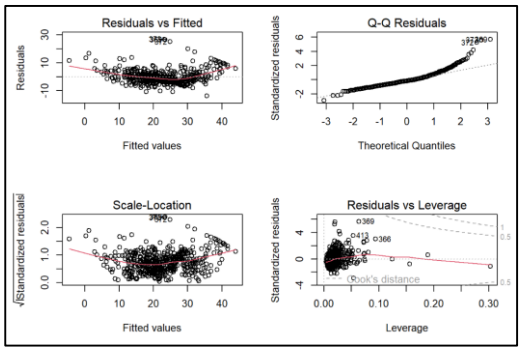
- **Negative Linear Relationship**: The scatter plot shows a clear negative linear trend, indicating that as the percentage of lower status population **(lstat) increases, the median home value (medv) decreases**.
- **Strength of Correlation:** The points are closely aligned with the regression line, reflecting a **strong negative correlation** (−0.7377), as confirmed by the correlation table.
- **Non-Uniform Spread**: The spread of points widens as lstat increases, suggesting **greater variability** in home values in neighborhoods with a higher percentage of lower status population.

7. Multiple Regression on 'medv':

Term <chr>	Estimate <dbl>	Std_Error <dbl>	t_Value <dbl>	P_Value <dbl>
(Intercept)	17.47	4.12	4.24	0.0000
rm	4.63	0.43	10.81	0.0000
lstat	-0.52	0.05	-10.18	0.0000
ptratio	-0.89	0.12	-7.50	0.0000
crim	-0.06	0.03	-1.96	0.0500
nox	-1.32	2.55	-0.52	0.6056

- **Significant Predictors**: Variables **rm (average number of rooms per dwelling), lstat (percentage of lower-status population), and ptratio (pupil-teacher ratio)** are highly significant predictors of medv (p-value < 0.05), indicating their strong influence on housing prices.
- **Positive Impact of rm**: The coefficient for rm is 4.63, i.e. an additional room in a dwelling increases the median home value by ~$4,630, holding other variables constant.
- **Negative Impact of lstat and ptratio**: Both lstat (-0.52) and ptratio (-0.89) have -ve coefficients, suggesting that higher percentages of lower-status populations and higher pupil-teacher ratios are associated with lower housing prices.

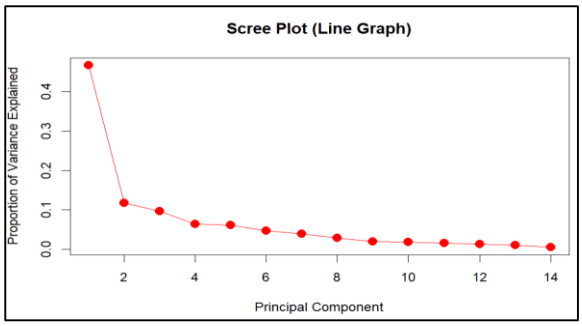
8. Model Diagnostics:



- **Residuals vs. Fitted:** The residuals show a slight pattern, suggesting **minor non-linearity** or heteroscedasticity in the model.
- **Q-Q Plot:** Residuals mostly **follow the theoretical quantile line**, but deviations at the tails indicate slight non-normality.
- **Residuals vs. Leverage**: A few high-leverage points are visible, but Cook's distance suggests they have limited influence on the model.

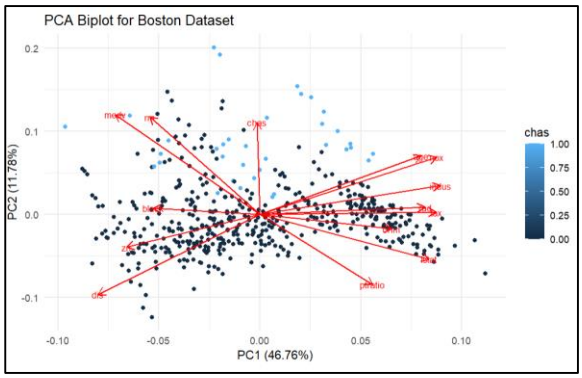
9. PCA

```
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7
Standard deviation     2.5585 1.2843 1.16142 0.94156 0.92244 0.81241 0.73172
Proportion of Variance 0.4676 0.1178 0.09635 0.06332 0.06078 0.04714 0.03824
Cumulative Proportion  0.4676 0.5854 0.68174 0.74507 0.80585 0.85299 0.89123
                          PC8     PC9    PC10    PC11    PC12    PC13    PC14
Standard deviation     0.63488 0.5266 0.50225 0.4613 0.42777 0.36607 0.24561
Proportion of Variance 0.02879 0.0198 0.01802 0.0152 0.01307 0.00957 0.00431
Cumulative Proportion  0.92003 0.9398 0.95785 0.9730 0.98612 0.99569 1.00000
```



- **PC1 explains the largest variance**: It accounts for **46.76%** of the total variance, making it the most significant component.
- **Five components explain over 80% of the variance:** The cumulative proportion reaches **80.59% by including PC1 to PC5,** sufficient for dimensionality reduction.
- **Elbow point at PC5 or PC6:** The scree plot shows **diminishing returns** in variance explained beyond **PC5,** marking it as the "elbow point."

10. PCA Interpretation:



- **Principal Components:** PC1 (46.76%) and PC2 (11.76%) explain **58.52%** of variance.
- **Red Arrows**: Represent variable contributions; direction shows correlation, length indicates strength.
- **Data Points:** Observations, colored by chas (Charles River proximity).
- **Patterns**: Similar observations cluster; chas = 1 concentrated in specific regions.

Conclusion

Univariate Analysis: The histogram and boxplot for lstat revealed a right-skewed distribution, with most neighborhoods having a lower percentage of the lower-status population (5-15%), and outliers at higher percentages. This indicates socioeconomic disparities across neighborhoods.

Multivariate Analysis: The scatter plot with regression analysis demonstrated a strong negative linear relationship between lstat and medv, confirming that higher percentages of lower-status populations are associated with lower home values. The PCA biplot showed that the first two principal components explained 58% of the variance, with lstat, rm, and nox being significant contributors, highlighting their importance in explaining variability in housing data.

Final insight from dataset: The analysis of the Boston dataset revealed key relationships and patterns. Univariate analysis highlighted the skewness and variability in variables like lstat and rm, with lstat showing a strong negative correlation with medv. Multivariate regression confirmed that rm, lstat, and ptratio are significant predictors of housing prices, with lstat having the strongest negative impact. Diagnostic plots suggested some non-linearity and heteroscedasticity in residuals. PCA reduced dimensionality effectively, with the first two components explaining 58% of the variance, emphasizing the importance of socioeconomic and environmental factors. Together, these analyses underscore the multifactorial nature of housing values in Boston.

DATASET 3. AIR QUALITY

(*R Library – 'Datasets'*)

Univariate Analysis:

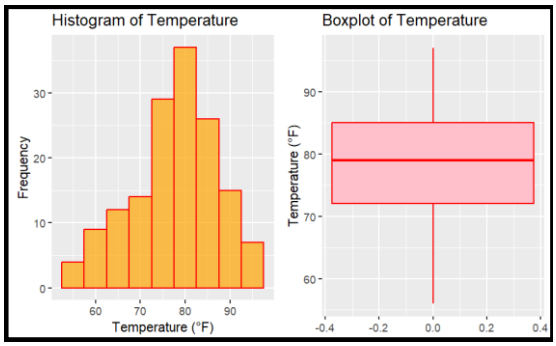
1. **About the Dataset**: The air quality dataset contains 153 observations and 6 variables related to air quality measurements in New York during May to September 1973. It includes variables such as ozone levels, solar radiation, wind speed, and temperature.

2. **Summary statistics for 'Temperature':**

Statistic <chr>	Value <dbl>
Mean	77.88235
Median	79.00000
Standard Deviation	9.46527
Minimum	56.00000
Maximum	97.00000

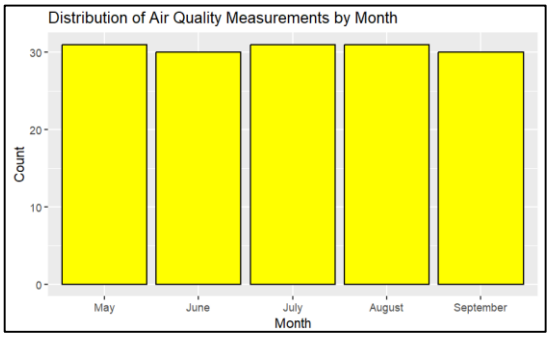
The summary statistics for the Temp variable in the air quality dataset show a **mean temperature of 77.88°F** and **a median of 79°F**, indicating a slightly **left-skewed distribution**. The standard deviation of 9.47°F reflects moderate variability, with temperatures ranging from a minimum of 56°F to a maximum of 97°F.

3. **Distribution Visualization:**



- **Distribution Shape:** The histogram shows a **roughly symmetric distribution**, with most temperatures clustered between 70°F and 85°F, indicating that these values are common during the recorded period.
- **Central Tendency:** The boxplot confirms a **median temperature of 79°F**, which aligns with the histogram's peak frequency.
- Variability: The interquartile range (IQR) shows that 50% of the temperatures fall between approximately **72°F and 85°F**, reflecting **moderate variability.**
- **Outliers:** There are **no visible outliers** in the temperature data, as all values fall within the whiskers of the boxplot.

4. **Categorical Variable Analysis – 'Month':**



- The bar chart for the Month variable in the air quality dataset shows an **equal distribution** of air quality measurements across the months of May, June, July, August, and September, with approximately 30 observations per month. This indicates **consistent data collection** throughout the recorded period.

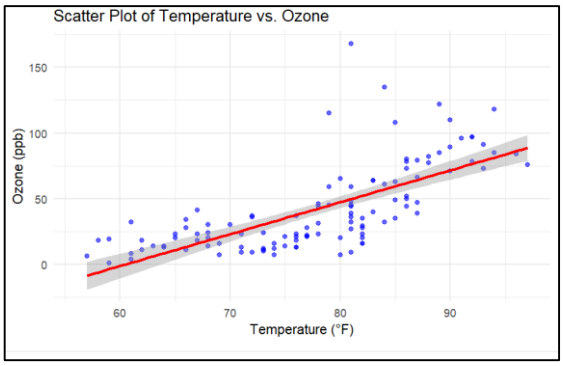
Multivariate Analysis

5. Correlation Analysis:

Variable1 <chr>	Variable2 <chr>	Correlation <dbl>
Temp	Ozone	0.6983603

- The correlation between Temp and Ozone is **0.698,** indicating a **strong positive relationship**. This suggests that higher temperatures are generally associated with increased ozone levels in the air quality dataset.

6. Scatter Plot Visualization:



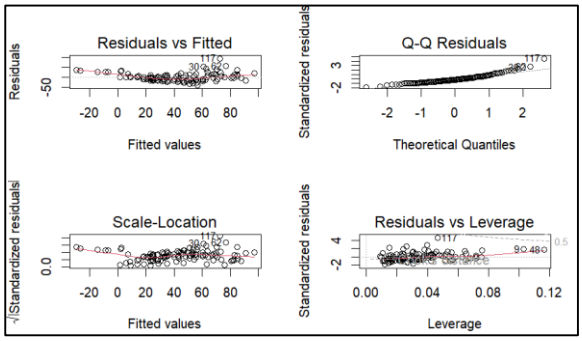
- **Positive Relationship:** The scatter plot shows a clear **positive trend**, indicating that higher temperatures are generally associated with higher ozone levels.
- **Variability at Higher Temperatures**: At temperatures above 85°F, there is **greater variability in ozone levels**, with some points showing very high concentrations.
- **Model Fit:** The red linear regression line fits the data well, **suggesting a strong linear relationship between temperature and ozone**, supported by the correlation value of 0.698.

7. Multiple Regression on 'Ozone level':

Term <chr>	Estimate <dbl>	Std_Error <dbl>	t_Value <dbl>	P_Value <dbl>
(Intercept)	-64.34	23.05	-2.79	0.0062
Temp	1.65	0.25	6.52	0.0000
Wind	-3.33	0.65	-5.09	0.0000
Solar.R	0.06	0.02	2.58	0.0112

- **Significant Predictors**: **All predictors (Temp, Wind, Solar.R) are statistically significant**, with p-values < 0.05, indicating they have a meaningful impact on ozone levels.
- **Positive Impact of Temperature**: The coefficient for Temp (1.65) suggests that for every 1°F increase in temperature, ozone levels increase by approximately 1.65 ppb, holding other variables constant.
- **Negative Impact of Wind**: The coefficient for Wind (-3.33) indicates that higher wind speeds are associated with a decrease in ozone levels, likely due to dispersion effects.

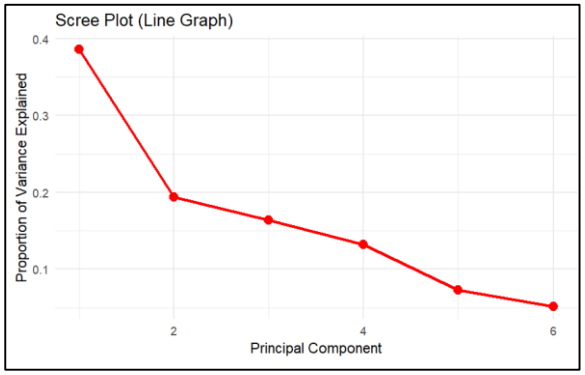
8. Model Diagnostics:



- **Residuals vs. Fitted Plot:** The residuals show no clear pattern, indicating that the **linearity assumption is reasonably satisfied**. However, slight deviations suggest some heteroscedasticity.
- **Q-Q Plot:** The residuals mostly align with the theoretical quantile line, confirming that the **normality assumption is largely met**, though there are minor deviations at the tails.
- **Residuals vs. Leverage Plot**: A few high-leverage points are visible (e.g., observation 117), but Cook's distance indicates that none of these points have an excessive influence on the model.

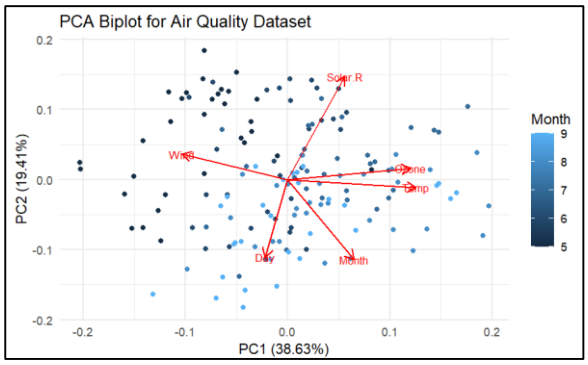
9. PCA:

```
Importance of components:
                          PC1    PC2    PC3    PC4     PC5     PC6
Standard deviation     1.5223 1.0792 0.9915 0.8891 0.65935 0.55634
Proportion of Variance 0.3862 0.1941 0.1638 0.1318 0.07246 0.05158
Cumulative Proportion  0.3862 0.5804 0.7442 0.8760 0.94842 1.00000
```



- **Variance Explained by PC1 and PC2:** The first principal component **(PC1) explains 38.62%** of the variance, and the second **(PC2) adds 21.42%,** meaning the first two components together capture **60.04% of the total variance.**
- **Diminishing Returns:** The scree plot shows a clear **"elbow" after PC2**, indicating that additional components contribute less variance, with PC3 explaining only 16.18%.
- **Dimensionality Reduction:** Retaining the **first three** components captures approximately **74.42% of the variance**, making them sufficient for most analyses while significantly reducing dimensionality.

10. PCA Interpretation:



- **Dominance of PC1:** PC1 explains 38.63% of the variance and is **strongly influenced by Solar.R, Ozone, and Temp,** indicating that these variables contribute significantly to overall variability.
- **Variable Relationships:** Solar.R, Ozone, and Temp have positive loadings on PC1, suggesting a strong positive correlation among them. Conversely, Wind has a negative loading, indicating an inverse relationship with these variables.
- **Month Grouping**: Observations are colored by Month, with some clustering visible, suggesting **seasonal patterns** in air quality measurements (e.g., higher ozone levels during warmer months).

Conclusion:

Univariate Analysis: The histogram and boxplot for Temp revealed a **roughly symmetric distribution**, with most temperatures ranging between 70°F and 85°F. **No outliers were observed**, and the variability in temperature was moderate, as shown by the interquartile range of approximately 13°F.

Multivariate Analysis: The scatter plot with regression demonstrated a **strong positive relationship between Temp and Ozone**, indicating that higher temperatures are associated with increased ozone levels. The PCA biplot showed that the **first two principal components explained 58.04%** of the variance, with Solar.R, Ozone, and Temp contributing significantly to PC1, while Wind had a strong negative loading.

Final Insight from Dataset: The air quality dataset analysis highlighted key relationships between weather variables and ozone levels. Univariate analysis showed **moderate variability in temperature**, while multivariate regression confirmed **Temp as a strong predictor of ozone concentration**. PCA effectively reduced dimensionality, emphasizing the importance of temperature, solar radiation, and wind in explaining air quality variability across months.

DATASET 4. <u>DIAMONDS</u>

(*R Library – 'ggplot2'*)

Univariate Analysis:

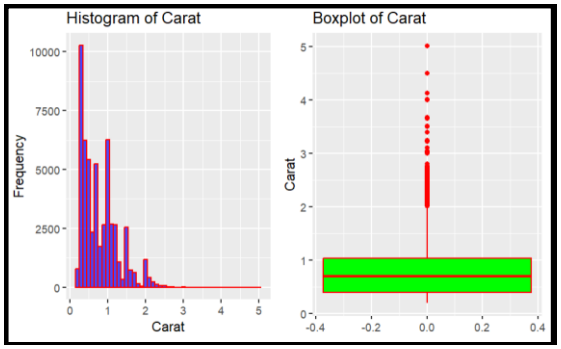
1. **About the Dataset:** The diamonds dataset contains 53,940 observations and 10 variables related to diamond characteristics, including carat, cut, color, clarity, and price. It provides a comprehensive overview of diamond quality and pricing

2. **Summary statistics for variable 'Carat':**

Statistic <chr>	Value <dbl>
Mean	0.7979397
Median	0.7000000
Standard Deviation	0.4740112
Minimum	0.2000000
Maximum	5.0100000

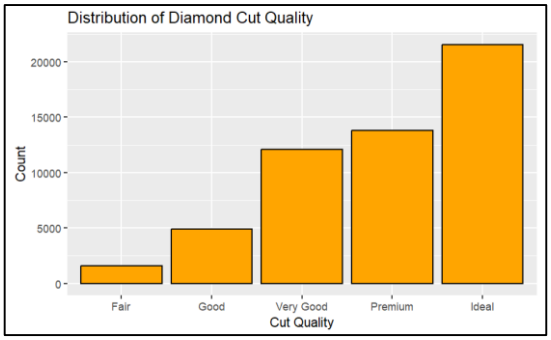
The summary statistics for the carat variable in the diamonds dataset show that the average diamond weight is approximately **0.80 carats**, with a **median of 0.70 carats**. The standard deviation of 0.47 indicates moderate variability, while the range spans from 0.20 to 5.01 carats, highlighting the **diversity in diamond sizes** within the dataset.

3. **Distribution Visualization:**



- **Distribution Shape:** The histogram shows **a right-skewed distribution**, with most diamonds having a carat weight between 0.2 and 1.0, indicating smaller diamonds dominate the dataset.
- **Outliers:** The boxplot highlights several **outliers above 3 carats**, representing rare and larger diamonds that deviate significantly from the majority.
- **Central Tendency:** The median carat weight is 0.7, as shown in the boxplot, aligning with the histogram's peak frequency around this range.
- **Variability:** The interquartile range (IQR) suggests that **50% of diamonds** have carat weights between approximately **0.4 and 1.0,** reflecting moderate variability within the dataset.

4. **Categorical Variable Analysis – 'Cut quality':**



- The bar chart for the cut variable in the diamonds dataset shows that **the majority of diamonds are of "Ideal" cut quality**, followed by "Premium" and "Very Good." **Fewer diamonds** fall into the "Good" and "Fair" categories, indicating a preference for higher-quality cuts in the dataset.

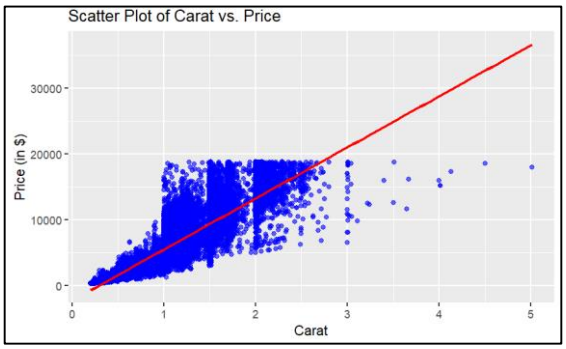
Multivariate Analysis

5. Correlation Analysis:

Variable1 <chr>	Variable2 <chr>	Correlation <dbl>
carat	price	0.9215913

- The correlation between carat and price **is 0.92,** indicating a **very strong positive relationship**. This suggests that as the carat weight of a diamond increases, its price tends to rise significantly.

6. Scatter Plot Visualization:



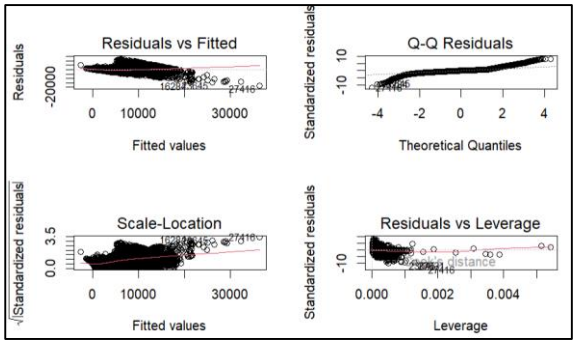
- **Strong Positive Relationship:** The scatter plot shows a **strong positive linear trend,** indicating that as the carat weight of a diamond increases, its price rises significantly.
- **Wide Price Range for Larger Diamonds**: Diamonds with higher carat weights (above 2 carats) exhibit **greater variability in price**, suggesting other factors like cut and clarity may influence pricing.
- Dense Cluster at Lower Carats: Most diamonds are clustered at lower carat weights (below 1.5 carats) with relatively lower prices, reflecting the **dominance of smaller diamonds** in the dataset.

7. Multiple Regression on 'Price' of diamond:

Term <chr>	Estimate <dbl>	Std_Error <dbl>	t_Value <dbl>	P_Value <dbl>
(Intercept)	13003.44	390.92	33.26	0
carat	7858.77	14.15	555.36	0
depth	-151.24	4.82	-31.38	0
table	-104.47	3.14	-33.26	0

- **Significant Predictor:** The **variable carat has a very strong positive impact** on price (Estimate = 7858.77, p-value < 0.001), meaning that an increase in carat weight significantly increases the diamond's price.
- Negative Contributions: **Both depth and table have negative coefficients** (-151.24 and -104.47, respectively), indicating that higher values of these variables are associated with slightly lower prices.
- Model Strength: The extremely small p-values (< 0.001) for all predictors confirm their statistical significance, **showing they strongly influence** the price of diamonds in this dataset.

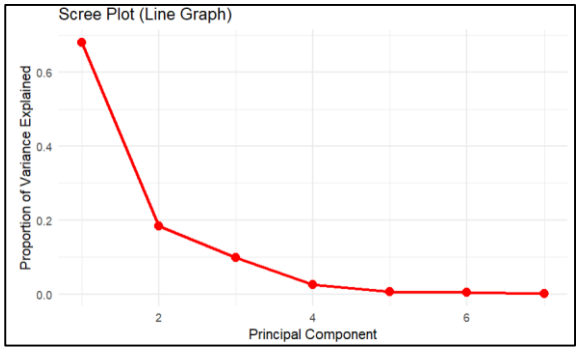
8. Model Diagnostics:



- **Residuals vs. Fitted Plot:** The residuals show a slight pattern, suggesting **some non-linearity** in the relationship between predictors and the response variable. This indicates the potential for model improvement.
- **Q-Q Plot**: The residuals mostly follow the theoretical quantile line, but deviations at the tails suggest slight **non-normality**, particularly for extreme values.
- **Residuals vs. Leverage Plot:** A few high-leverage points are visible, but Cook's distance indicates that none of these points have an excessive influence on the model.

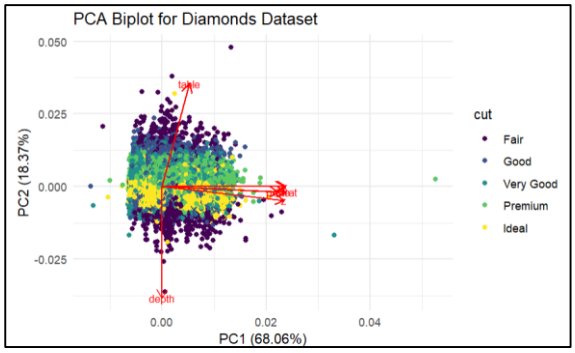
9. PCA:

```
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6
Standard deviation     2.1826 1.1340 0.83115 0.41684 0.20077 0.18151
Proportion of Variance 0.6806 0.1837 0.09869 0.02482 0.00576 0.00471
Cumulative Proportion  0.6806 0.8642 0.96294 0.98776 0.99352 0.99823
                           PC7
Standard deviation     0.11135
Proportion of Variance 0.00177
Cumulative Proportion  1.00000
```



- **Variance Explained by PC1:** The first principal component **(PC1) explains 60.6%** of the total variance, indicating that it captures the majority of variability in the dataset.
- **Cumulative Variance**: **The first two components together explain 84.2%** of the variance, suggesting that these two components are sufficient to represent most of the dataset's information.
- **Elbow Point:** The scree plot shows a clear **"elbow" after PC2**, with subsequent components contributing minimal additional variance, making dimensionality reduction to two components effective.

10. PCA Interpretation:



- **Dominance of PC1:** PC1 explains 68.06% of the variance, with **strong contributions from carat and price**, indicating that these variables primarily drive the variability in the dataset.
- **Variable Relationships**: The loadings show that **carat and price are positively correlated**, while depth has a negative correlation with PC1, suggesting an inverse relationship between depth and diamond size/price.
- **Grouping by Cut: Diamonds with "Ideal" and "Premium" cuts cluster closely,** reflecting similar characteristics, while "Fair" cut diamonds are more dispersed, indicating greater variability in their properties.

Conclusion:

Univariate Analysis: The histogram and boxplot for carat revealed a right-skewed distribution, with **most diamonds having weights between 0.2 and 1.0 carats**. Outliers at higher carat values represent larger, rarer diamonds, indicating variability in diamond sizes.

Multivariate Analysis: The scatter plot with regression demonstrated a **strong positive relationship between carat and price**, confirming that heavier diamonds are significantly more expensive. The PCA biplot showed that the **first two principal components explained 86.42%** of the variance, with carat and price contributing heavily to PC1, while depth and table influenced PC2.

Final Insight from Dataset: The diamonds dataset analysis highlighted key drivers of diamond pricing. **Univariate analysis showed variability in diamond sizes**, while **multivariate regression confirmed carat as the strongest predictor of price**. PCA effectively reduced dimensionality, emphasizing size (carat) as the primary driver of variability, with clear groupings by cut quality reflecting differences in diamond characteristics.