SSD Assignment Report

Student Name: Sayali Malshikare
Registration No. : 24BM6JP50
Date: 8th December 2024

Dataset 1: diamonds
R – Package: ggplot2

The diamonds dataset from the ggplot2 package contains data about various diamond characteristics, including carat, cut, colour, clarity, weight and price, which are essential for understanding their relationship with pricing of diamonds.

Univariate Analysis

1. **Data Overview:**
   - **Structure:** The dataset consists of 53,940 observations (rows) and 10 variables (columns).
   - **Variables**:
     - Numerical variables: carat, depth, table, price, x (length), y (width), z (depth)
     - Categorical variables: cut, colour, clarity (all ordered factors)

2. **Summary Statistics:**
   For the numerical variable 'price', the following summary statistics were computed:
   - **Mean**: **3932.8 , Median**: **2401** : The median is considerably lower than the mean, which suggests that the distribution is skewed to the right (i.e., there are a number of expensive diamonds that are pulling the mean upwards).
   - **Standard Deviation**: **3989.44** : A large standard deviation indicates considerable variability in the prices of diamonds, likely due to the wide range of carat sizes, cuts, and other factors.
   - **Minimum**: **326, Maximum: 18823** : The highest price in the dataset, indicating the presence of very expensive diamonds.

3. **Distribution Visualization:**

3.1 Histogram

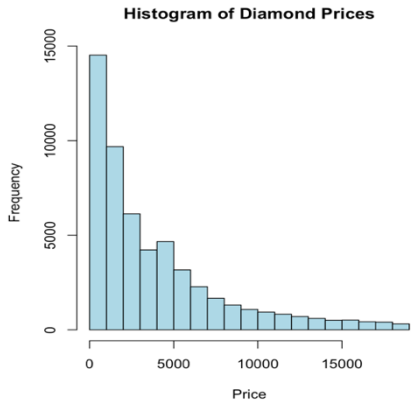


Figure 1

The R output of the histogram of variable 'Price' is as seen in Figure 1.
The distribution appears to be **right-skewed** (positively skewed), with the majority of the diamond prices concentrated on the lower end.
This is consistent with the observation that the median price is significantly lower than the mean price. A few high-price diamonds are causing the right tail to extend, indicating that while most diamonds are priced within a moderate range, a small number of extremely expensive diamonds are pulling the average higher.

3.2 Boxplot

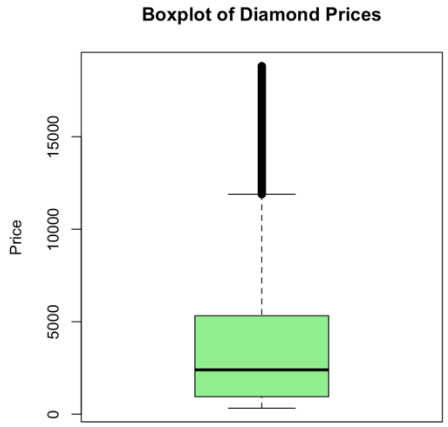


Figure 2

- **Shape**: The distribution is **right-skewed**. This suggests that while most diamonds are priced in the lower to mid-range, there is a group of higher-priced diamonds that are pulling the distribution to the right.
- **Outliers**: There are several **outliers** marked as points outside the whiskers of the boxplot. Given the large number of observations (53,940), these outliers are likely to be very expensive diamonds, further supporting the idea that the dataset contains a mix of more affordable and high-end diamonds.
- **Median and IQR**: The median price is closer to the lower quartile, indicating that a significant portion of the diamonds have prices below the median.

4. **Categorical Variable Analysis**:
   - The categorical variable 'Clarity' , suggesting how clear the diamond was in terms of 8 categories from (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best)), was selected for analysis through bar plot as shown in Figure 3.
   - The most common clarity grades are **"SI1"** and **"VS2"**, indicating a balance between quality and affordability in the market. Diamonds with **"I1"** clarity are the least frequent, as they are lower quality with visible inclusions. The **"FL"** (Flawless) clarity category is rare, reflecting its higher price point and rarity in the dataset.

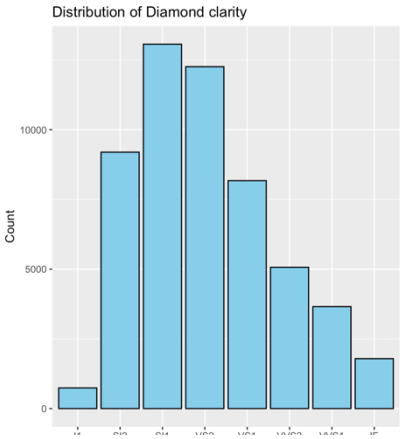


Figure 3

Multivariate Analysis

5. **Correlation Analysis:**
   The Pearson correlation coefficient between the two selected variables **'carat'** and **'price'** is **0.92**, indicating a strong positive linear relationship between the two variables.

6. **Scatter Plot Visualization**

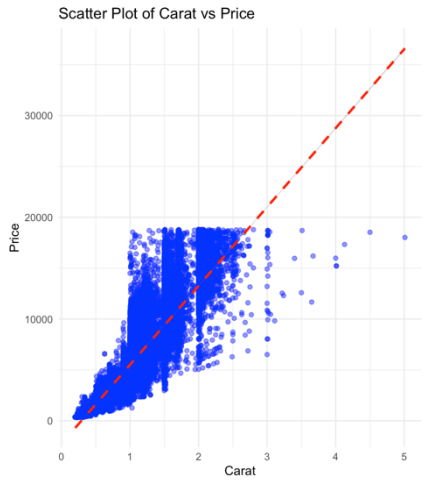


Figure 4

The scatter plot as shown in /figure 4, visualizes the relationship between carat and price. Here's the interpretation:

- **Relationship**: There is a strong positive linear relationship between carat and price, which aligns with the Pearson correlation result. As the carat size increases, the price of the diamond also increases.
- **Trend Line**: The trend line further emphasizes this positive relationship, showing that larger diamonds tend to be significantly more expensive than smaller ones.
- **Spread**: The data points are more spread out for higher carat values, indicating greater price variability for larger diamonds, whereas smaller diamonds show a more concentrated range of prices.

7. **Multiple Regression:** The multiple linear regression model was used to predict 'price; based on 'carat' and 'depth'. Here's the interpretation of the results:

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4045.3	286.2	14.13	<2e-16
carat	7765.1	14.1	554.3	<2e-16
depth	-102.1	4.6	-22.1	<2e-16

Residuals:

Min	1Q	Median	3Q	Max
-18238.9	-801.6	-19.6	546.3	12683.7

The **Multiple R-squared** value of **0.8507** means that 85.07% of the variability in diamond prices is explained by the carat and depth variables, indicating a good model fit.
Significance: All predictors (carat and depth) have a **p-value** less than 0.001, which means they are statistically significant at the 0.05 level.

8. Model Diagnostics

Residual vs Fitted Plot (Homoscedasticity Check):

- The residuals in this plot are not randomly scattered which shows that there still exists some heteroscedasticity in the residuals.

Normal Q-Q Plot (Normality of Residuals):

- The residuals mostly follow a straight line, implying that they are normally distributed but the Minor deviations suggest that there could be slight skewness or outliers, but overall, the assumption of normality holds.

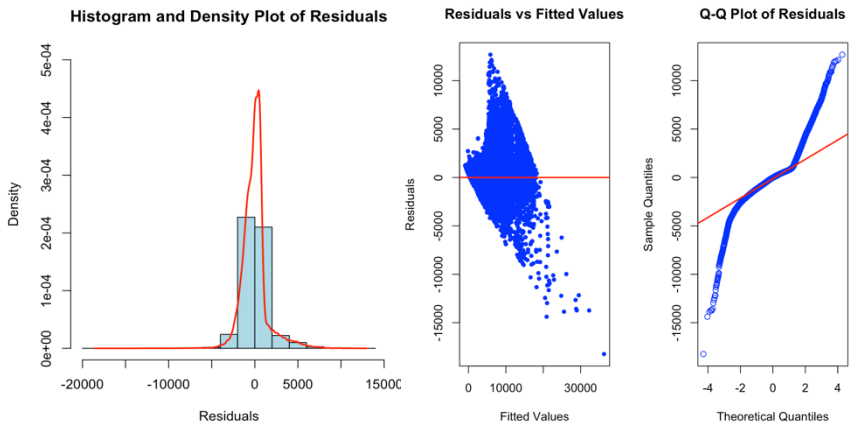


Figure 5

Advance Analysis

9. Principle Component Analysis:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Standard deviation	2.18	1.13	0.83	0.42	0.20	0.18	0.11
Proportion of Variance	0.68	0.18	0.10	0.02	0.01	0.00	0.00
Cumulative Proportion	0.68	0.86	0.96	0.99	0.99	1	1

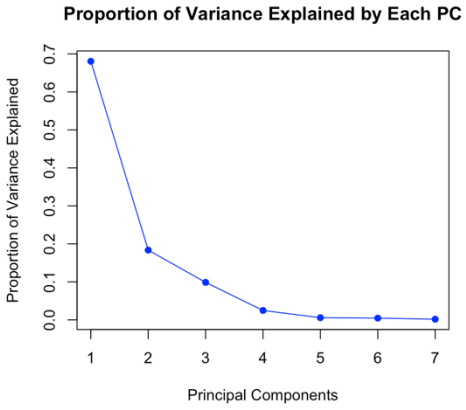


Figure 6

PCA is a technique used to reduce the dimensionality of data while preserving most of the variance in the data. The PCA analysis reveals the proportion of variance explained by each (PC) in the dataset as shown above in the summary:

- PC1 explains **68.06%** of the variance, indicating it is the most significant component in capturing the variability in the dataset.
- PC2 explains **18.37%,** and together PC1 and PC2 account for **86.42%** of the total variance, suggesting that these two components capture most of the relevant information in the dataset.
- Subsequent PCs (PC3 to PC7) explain much smaller proportions of the variance, with PC3 accounting for only 9.87%.

Scree Plot: The Scree plot visually confirms that most of the variance **(96%)** is captured by the first 3 components, (elbow at 3).Hence first 3 components can be chosen.

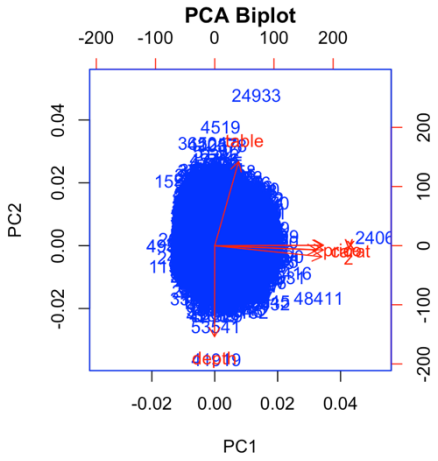
10. PCA Interpretation:

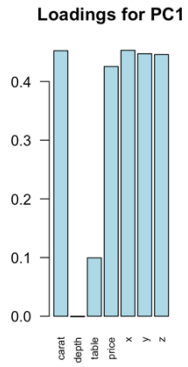
The biplot helps visualize how the different variables are related to the overall variance captured by the first two principal components.

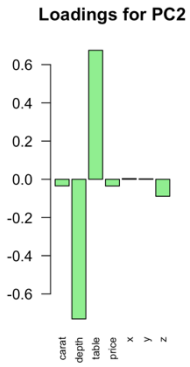
- Variables like carat, price, x, y and z have strong positive contributions to PC1, indicating their high influence on the first component. These variables have the highest loadings, suggesting they contribute significantly to the first principal component
- Variables like depth and table contribute significantly to PC2

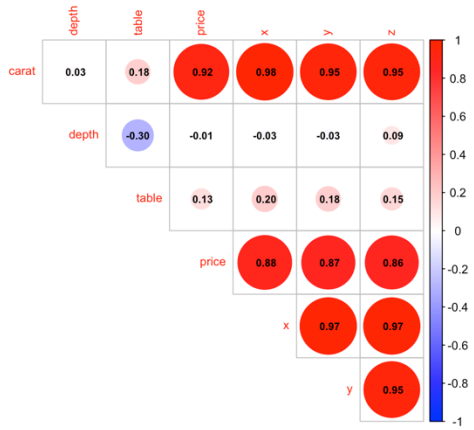
The same can be viewed from the PC1 and PC2 Loadings table and graph as shown below. The heat map of correlation matrix gives the correlation between each variable.

	carat	depth	table	price	x	y	z
PC1	0.452	-0.001	0.100	0.426	0.453	0.447	0.446
PC2	-0.035	-0.731	0.675	-0.035	0.004	0.002	-0.089









Conclusion: In the Diamond dataset, univariate analysis revealed that price is right-skewed, with a high mean compared to the median, indicating a few extremely expensive diamonds. Multivariate analysis showed a strong positive correlation between carat and price, suggesting that larger diamonds are more expensive. The regression model indicated that carat and depth significantly predict price. PCA identified PC1 as capturing size and value-related variance, while PC2captured diamond proportions like table and depth. These insights emphasize the key factors driving diamond prices and the role of size and proportions in their characteristics.

Dataset 2: kanga
R – Package: faraway
The kanga dataset from the faraway library contains information about the physical measurements of kangaroos, including variables such as species, sex, and various body measurements like basilar. length, palate.width, mandible.depth, and others, with some missing values in certain fields.

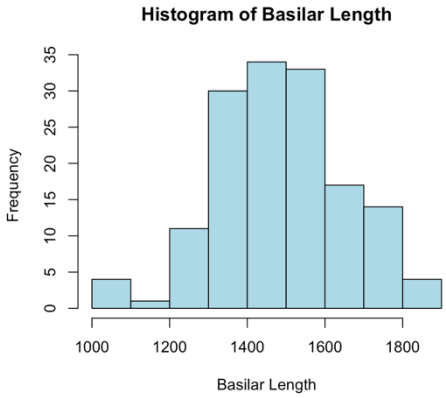
Univariate Analysis

1. **Data Overview:**
   - **Structure:** The dataset consists of 148 observations (rows) and 20 variables (columns).
   - As the dataset had few missing values in numerical variables, the missing values were imputed using mean imputation.
   - **Variables**:
     - Numerical variables: basilar.length, palate.width, palate.length, mandible.length , mandible.width ,mandible.depth, occipitonasal.length, nasal.length, nasal.width, squamosal.depth, lacrymal.width, zygomatic.width, orbital.width rostral.width, occipital.depth, crest.width, foramina.length, ramus.height
     - Categorical variables: species, sex

2. **Summary Statistics:** For the numerical variable ‘basilar’, the following summary statistics were computed:
   - **Mean**: **1490.05**
   - **Median**: The median value of **1486.5** is slightly lower than the mean, indicating that the distribution is slightly right-skewed (with a few large values pulling the mean higher).

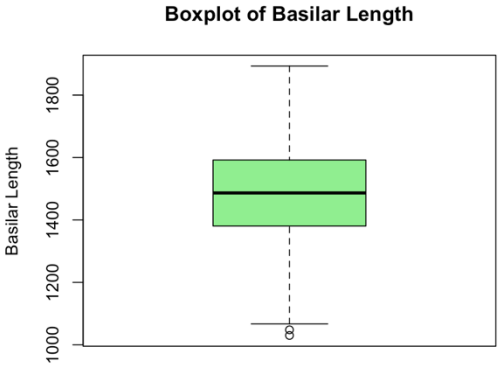
- **Standard Deviation**: The standard deviation of **164.47** indicates a relatively wide spread around the mean, suggesting considerable variation in basilar length across individuals.
- **Minimum**: **1030**, **Maximum**: **1893** indicating some individuals have smaller body sizes while some are much larger than the average.

3. Distribution Visualization



3.1 Histogram:

- **Shape of the distribution**: **Left-skewed**, indicating that most of the data is clustered around the higher basilar lengths, while smaller values are less frequent.
- **Potential outliers**: The left tail suggests the presence of some small values that could be considered **outliers**.



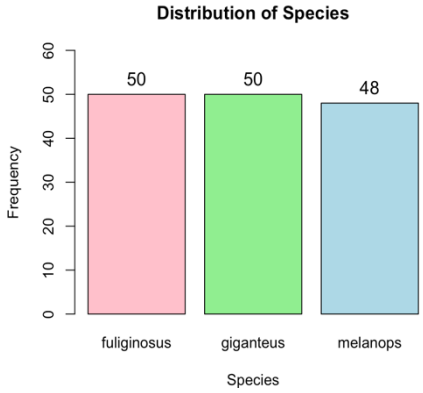
3.2 Box Plot

The **Box Plot** for **Basilar Length** further confirms the **left-skewed** nature of the distribution:

- Median: The median is closer to the upper quartile, indicating that most of the data lies in the higher range of basilar lengths.
- Interquartile Range (IQR): The spread of the box shows that the middle 50% of the data is clustered towards the higher values.
- Whiskers: The whiskers extend towards the lower values, with the left whisker being longer, indicating that the lower part of the distribution has more spread.
- Outliers: Outliers on the lower end of the distribution suggest that there are a few observations with much smaller basilar lengths than the majority.

4. Categorical Variable Analysis

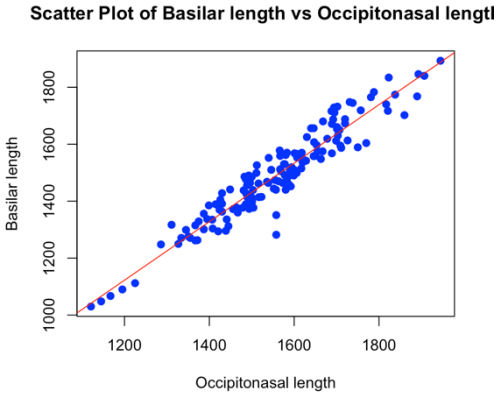
The Bar Plot for the species variable visualizes the distribution of the different species,(fuliginous, giganteus and melanops) in the dataset, fuliginous is the most frequent species with a higher concentration of observations, followed by gigenteus and melanops



Multivariate Analysis

5. **Correlation Analysis:** The Pearson correlation coefficient between the two selected variables **'basilar length'** and **'Occipitonasal Length'** is **0.95**, indicating a strong positive linear relationship between the two variables.

6. **Scatter Plot Visualization:** The Scatter Plot between Basilar Length and Occipitonasal Length illustrates the relationship between these two variables:
   - **Relationship:** The scatter plot suggests a strong positive correlation between Basilar Length and Occipitonasal Length, as the points tend to increase in value together.
   - **Trend Line:** The linear trend suggests that these two body dimensions of the kangaroos are closely related and likely grow proportionally.

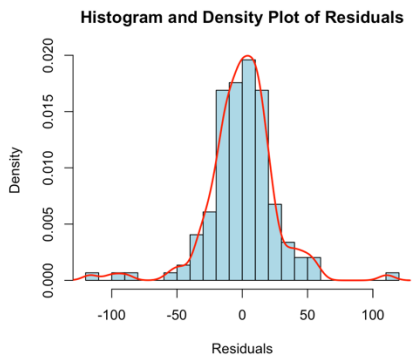


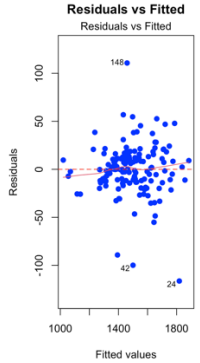
7. **Multiple Regression :** The multiple linear regression model was used to predict 'Basilar Length' using 'Occipitonasal Length' and 'Palate Length' as predictors. Here's the interpretation of the results:

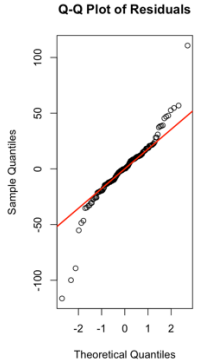
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	79.42117	24.609	3.227	0.00155
occipitonasal_length	0.20679	0.04508	4.588	9.62E-06
palate_length	1.06639	0.05551	19.21	2.00E-16

The Multiple R-squared value of 0.9741 indicates that approximately 97.41% of the variance in Basilar Length is explained by the model.
Both Occipitonasal Length and Palate Length are significant predictors of Basilar Length.

8. **Model Diagnostics :**







Residual vs Fitted Plot (Homoscedasticity Check):

- Interpretation: The residuals in this plot are spread fairly randomly across the fitted values, with no apparent patterns, and the assumption of homoscedasticity is likely met.

Normal Q-Q Plot (Normality of Residuals):

- Interpretation: The Q-Q plot shows the residuals plotted against a normal distribution. Most of the points lie close to the diagonal line, suggesting that the residuals are approximately normally distributed with a heavy tail

Advance Analysis

9. **Principle Component Analysis**

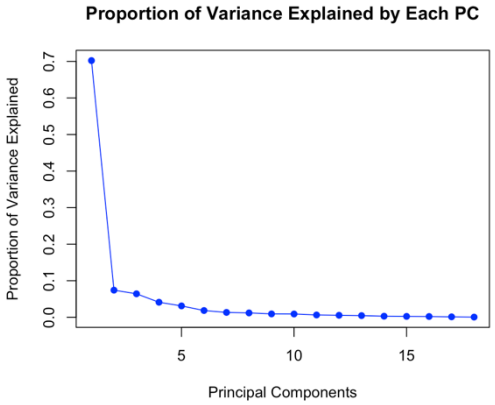
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16	PC17	PC18
Standard deviation	3.555	1.155	1.075	0.860	0.748	0.576	0.489	0.463	0.408	0.402	0.337	0.313	0.286	0.229	0.208	0.194	0.145	0.090
Proportion of Variance	0.702	0.074	0.064	0.041	0.031	0.018	0.013	0.012	0.009	0.009	0.006	0.005	0.005	0.003	0.002	0.002	0.001	0.000
Cumulative Proportion	0.702	0.776	0.841	0.882	0.913	0.931	0.945	0.956	0.966	0.975	0.981	0.986	0.991	0.994	0.996	0.998	1	1

PCA is a technique used to reduce the dimensionality of data while preserving most of the variance in the data. The PCA analysis reveals the proportion of variance explained by each (PC) in the dataset as shown above in the summary:

- PC1 explains the most variance (70.2%) (stands out as the most important component), followed by PC2 (7.4%) and PC3 (6.4%). The first two principal components (PC1 and PC2) together explain 77.6% of the total variance in the dataset.
- The cumulative proportion of variance increases as we include more components, and by PC4, 88.2% of the variance is explained. After the fourth component, the increase in explained variance becomes smaller, with the remaining components explaining less than 1% each.

Scree Plot

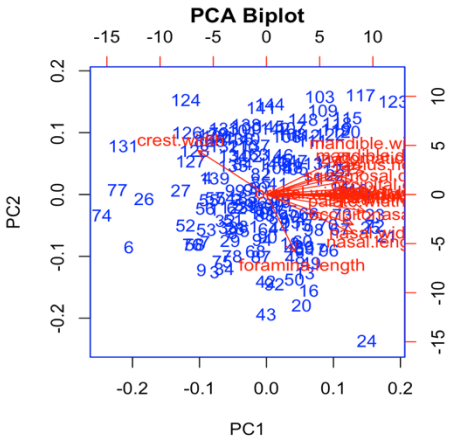
- PC1 would be the dominant component with a steep drop-off, followed by smaller contributions from subsequent components.
- The elbow method could suggest that PC1, PC2, and PC3 are the most significant components, which explains the bulk of the variation in the dataset.

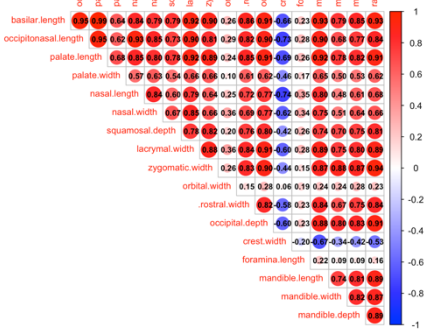


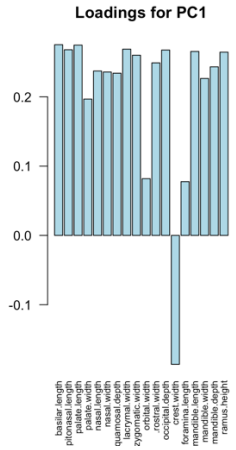
10. PCA Interpretation:

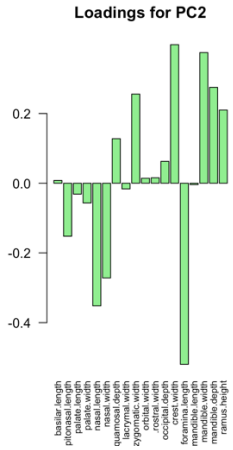
The biplot helps visualize how the different variables are related to the overall variance captured by the first two principal components.

- From the covariance matrix and the PC loadings graph and the table it is evident that all the Variables except orbital width and foramina largely contribute to PC1, indicating their high influence on the first component. These variables have the highest loadings, suggesting they contribute significantly to the first principal component









Conclusion : In the Kanga dataset, univariate analysis revealed that variables like basilar length show a nearly normal distribution with a slight right skew, and species distribution is balanced across groups. Multivariate analysis highlighted strong correlations between variables like occipitonasal length and basilar length, suggesting they are related in determining the size of the kangaroo. The regression model showed significant predictive power for basilar length using occipitonasal length and palate length. PCA identified PC1,PC2 and PC3 as capturing the majority of the variance, highlighting size-related characteristics and proportional measurements in kangaroo anatomy.

Dataset 3: decathalon2
R – Package: factoextra

The decathlon2 dataset contains performance metrics for athletes competing in decathlon events. The variables represent various track and field disciplines, as well as scores and rankings for each athlete. This dataset is useful for analysing the performance of decathletes across different disciplines and understanding correlations between events.

Univariate Analysis

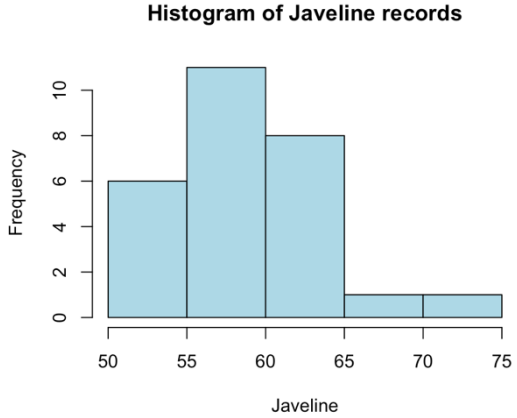
1. **Data Overview:**
   - **Structure:** The dataset consists of 27 observations (rows) and 13 variables (columns).
   - **Variables**:
     - Numerical variables: X100m, Long.jump, Shot.put, High.jump, X400m, X110m.hurdle, Discus, Pole.vault, Javeline, X1500m, Rank, Points

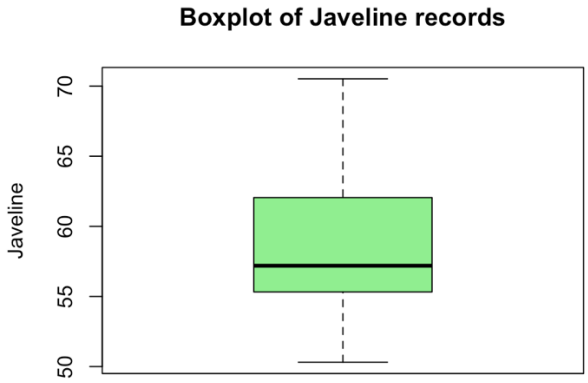
- Categorical variables: Competition

2. **Summary Statistics:** For the numerical variable 'javeline', the following summary statistics were computed:
   - **Mean**: **58.32 meters**.
   - **Median**: The median javelin throw distance is **57.19 meters**, which is slightly lower than the mean, suggesting a possible right-skewed distribution.
   - **Standard Deviation**: The standard deviation is **5.23 meters**, indicating moderate variability in the javelin throw distances among the athletes.
   - **Minimum**: **50.31 meters**. **Maximum**: **70.52 meters**.

   The javelin throw distances show a moderate spread, with most throws concentrated around 58 meters. The difference between the minimum and maximum values indicates a notable range in performance levels. The mean being slightly higher than the median hints at a potential skewness toward higher distances.

3. **Distribution Visualizations:**

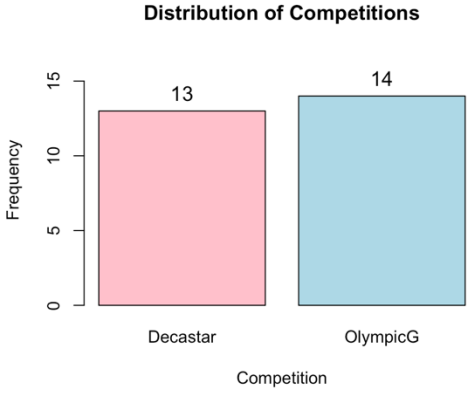




Histogram Visualization:
- The histogram shows that the distribution of javelin throws is slightly **right-skewed**.
- Most of the values are clustered between 55 and 65 meters.
- There are fewer observations above 65 meters, with a small tail extending towards 70+ meters.

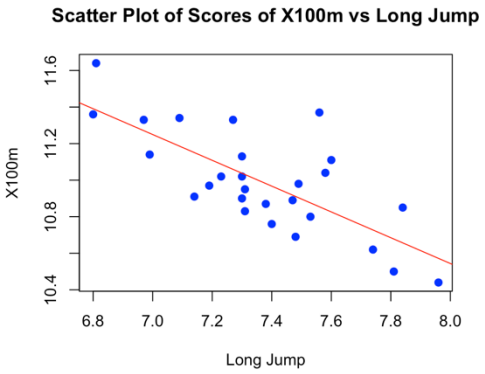
Box Plot Visualization:
- The interquartile range (IQR) is between approximately 55 meters and 65 meters.
- There are no clear outliers beyond the whiskers, which suggests the data is relatively consistent.
- The median is close to the lower quartile, reinforcing that the distribution is right skewed.

4. **Categorical Variable Analysis:**
   From the bar plot, we can see that there are 13 athletes in the Decastar competition and 14 athletes in the OlympicG competition.
   This distribution indicates a fairly balanced representation between the two competition types, suggesting that the dataset contains similar numbers of athletes from both events.



<u>Multivariate Analysis</u>

5. **Correlation Analysis :** The Pearson correlation coefficient between the variables **'X100m'** and **'Long Jump'** is **-0.75**, indicating a strong negative linear relationship between the two variables.



6. **Scatter Plot Visualization :** The Scatter Plot between X100m and Long Jump illustrates the relationship between these two variables:
   - **Relationship:** The scatter plot suggests a strong negative correlation between X100m and Long Jump
   - **Trend Line:** The linear trend suggests that athletes who perform better (faster) in the 100-meter sprint tend to achieve longer distances in the long jump event

7. **Multiple Regression:** The multiple linear regression model was used to predict scores of X100m using scores of long jump and shot put as predictors. Here's the interpretation of the results:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	16.44028	1.01651	16.173	2.08E-14
long_jump	-0.66666	0.14052	-4.744	7.96E-05
Shot.put	-0.03701	0.04946	-0.748	0.462

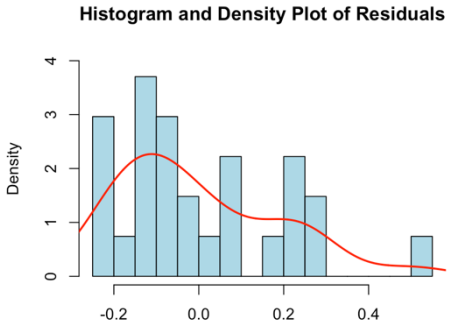
Interpretation of Coefficients:

1. Long.jump Coefficient: For every 1-meter increase in Long.jump, the 100-meter sprint time decreases by 0.67 seconds on average ($p < 0.001$). This predictor is highly significant, indicating that better long jump performance is associated with faster sprint times.
2. Shot.put Coefficient: The coefficient for Shot.put is -0.037 with a p-value of 0.462, meaning it is not statistically significant. This suggests that shot put performance does not meaningfully impact the 100-meter sprint time.

   The regression model shows that **long jump performance** is a significant predictor of 100-meter sprint times, while shot put performance does not significantly influence sprint times. The model explains a moderate portion of the variability in sprint times.

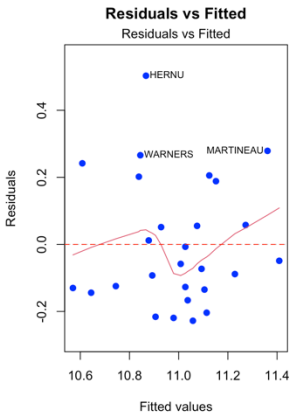
8. **Model Diagnostics:**

1. Histogram and Density Plot of Residuals



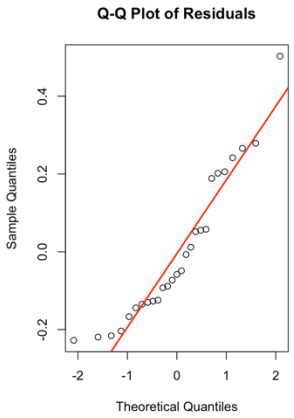
- The **red density curve** overlaid on the histogram helps visualize the shape of the residual distribution.
- The residuals appear to be **approximately normal**, though there may be slight deviations.

2. Residuals vs Fitted Plot



- The **Residuals vs Fitted** plot helps check for **homoscedasticity** (constant variance of residuals).
- The points are spread randomly around the horizontal line (red dashed line at 0), indicating no clear pattern or funnel shape.
- This suggests that the assumption of **homoscedasticity** holds reasonably well.

3. Q-Q Plot of Residuals



- The **Q-Q Plot** (Quantile-Quantile plot) checks the normality of residuals.
- The points generally fall along the red reference line, suggesting the residuals are **approximately normally distributed**.
- There are no extreme deviations, confirming that the normality assumption is met.

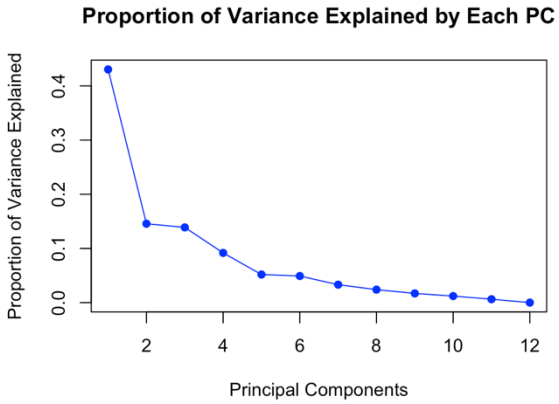
Advance Analysis

9. Principle Component Analysis:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12
Standard deviation	2.27	1.32	1.29	1.05	0.79	0.77	0.63	0.54	0.45	0.38	0.27	0.01
Proportion of Variance	0.43	0.15	0.14	0.09	0.05	0.05	0.03	0.02	0.02	0.01	0.01	0.00
Cumulative Proportion	0.43	0.58	0.71	0.81	0.86	0.91	0.94	0.96	0.98	0.99	1	1

Variance Explained:

- PC1 explains 43% of the variance, indicating it captures the most information., PC2 explains 15%, and PC3 explains 14% of the variance.
- The first three principal components (PC1, PC2, and PC3) together explain 71% of the total variance.
- The cumulative proportion of explained variance reaches 81% with PC4, and 91% with PC6.

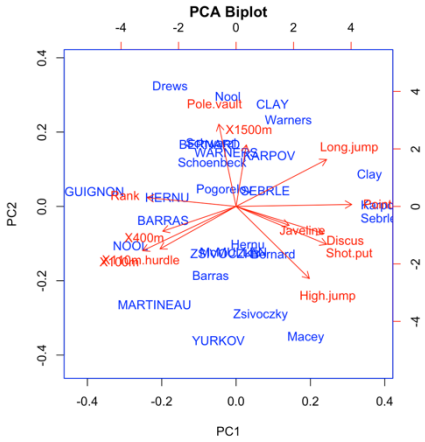


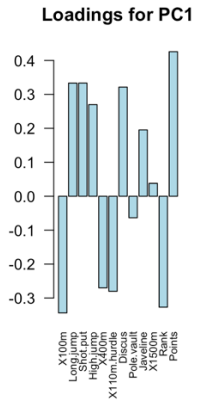
Scree Plot Interpretation:

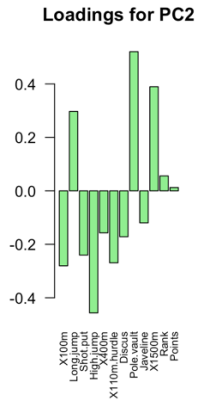
- Beyond **PC4**, the explained variance decreases gradually, indicating diminishing returns.
- Therefore, selecting the first 4 to 5 components is reasonable, as they capture a significant portion of the variance (around 81-91%).

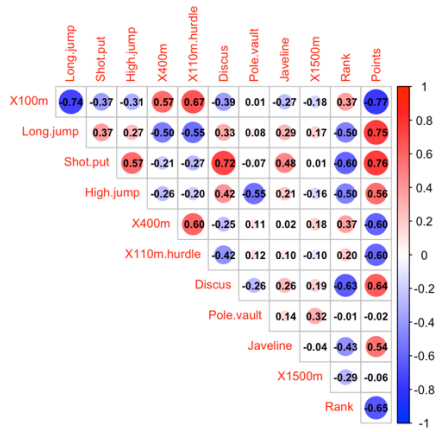
10. PCA Interpretation: The biplot helps visualize how the different variables are related to the overall variance captured by the first two principal components. The loadings of the first 2 principle components are shown below:

	X100m	Long jump	Shot put	High jump	X400m	X110m hurdle	Discus	Pole vault	Javeline	X1500m	Rank	Points
PC1	-0.34	0.33	0.33	0.27	-0.27	-0.28	0.32	-0.06	0.20	0.04	-0.33	0.43
PC2	-0.28	0.30	-0.24	-0.46	-0.16	-0.27	-0.17	0.52	-0.12	0.39	0.06	0.01









Conclusion: The univariate analysis revealed key statistics of individual athletic performances, such as the mean and standard deviation of events like the Javelin throw. Multivariate analysis using PCA showed that the first two components explained a substantial proportion of variance, with the variables "X100m" and "Long.jump" contributing most to PC1. Correlation analysis identified significant relationships between performance metrics. Overall, the analysis highlighted important factors influencing athletic performance, revealing correlations and key components that drive variation across the dataset.

Dataset 4: mtcars

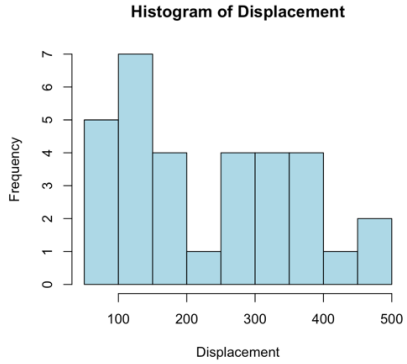
The mtcars dataset comprises data extracted from the 1974 Motor Trend US magazine, providing specifications of 32 car models across 11 attributes. It has a mix of continuous and discrete / categorical variables.

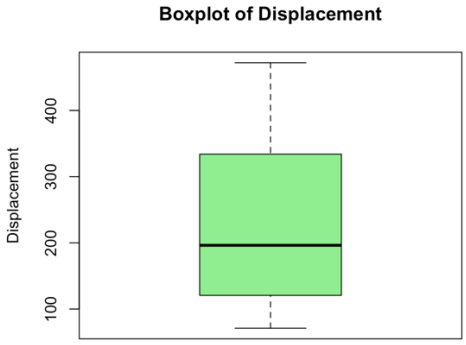
Univariate Analysis

1. **Data Overview:**
   - **Structure:** The dataset consists of 32 observations (rows) and 11 variables (columns).
   - **Variables**:
     - Numerical variables: mpg (mile/gallon), disp, hp(gross horse power), drat, weight, qsec
     - Categorical variables: Cylinders, Engine type, Transmission, No. of Gears, Carburators

2. **Summary Statistics:** For the numerical variable 'disp', the following summary statistics were computed:
   - **Mean Displacement:** 230.72 cubic inches
   - **Median Displacement:** 196.30 cubic inches, Suggests that half the cars have engine sizes smaller than 196.30 cubic inches, highlighting a slight skew towards higher displacement.
   - **Standard Deviation:** 123.94 cubic inches, Reflects substantial variability in engine sizes across the dataset.
   - **Minimum Displacement**: 71.10 cubic inches, **Maximum Displacement**: 472.00 cubic inches
     The dataset features a wide range of engine displacements, from compact to high-powered engines, with considerable diversity around the mean.

3. **Distribution Visualization:**





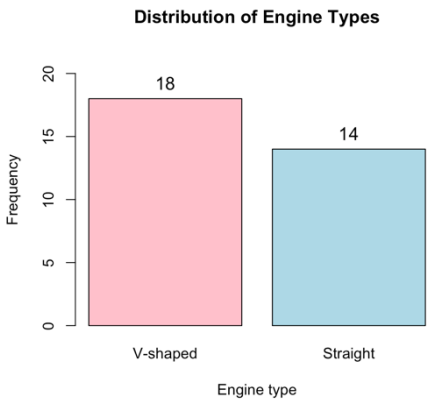
Histogram

- Skewness: The histogram appears to be slightly right-skewed, as there are more values concentrated in the lower ranges (e.g., around 100), with fewer observations as the displacement increases. However, the skewness is not very pronounced.
- Spread: The 'disp' values range from 71.1 to 472, and the histogram shows a relatively even distribution across these values. There are several peaks, but none of them are overly sharp or extreme, suggesting a moderately spread-out data range.

Box Plot

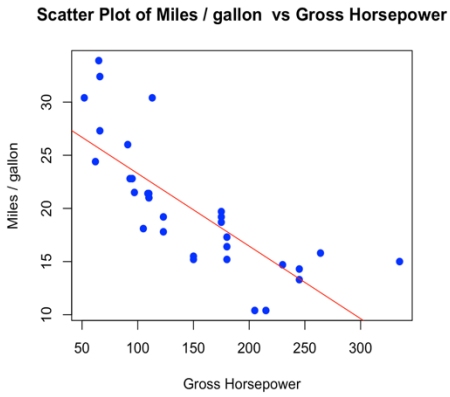
The distribution appears slightly right skewed with no extreme outliers, and most of the data is clustered within the IQR.

4. **Categorical Variable Analysis:**
   - From the bar plot, we can see that there are 32 car models with 2 types of engines : V- Shaped and Straight type Engine.
   - This distribution indicates a dominance of V- Shaped engines vs the straight types engines in different car models studied.



Multivariate Analysis

5. **Correlation Analysis:** The Pearson correlation coefficient between the two selected variables **'Miles per Gallon'** and **'Gross HP'** is **-0.78**, indicating a strong negative linear relationship between the two variables.



6. **Scatter Plot Visualization:**
   - **Downward Trend**: In the scatter plot, we observe a general downward slope, meaning as Gross Horsepower (hp) increases, Miles per Gallon (mpg) decreases. This visual trend aligns with the negative correlation.
   - **Strength of the Relationship**: Since the Pearson correlation is strong (**-0.776**), the scatter plot shows a relatively tight, linear clustering of points, though with some dispersion. This means that while the relationship is strong, there is still some variation due to factors not captured by horsepower.

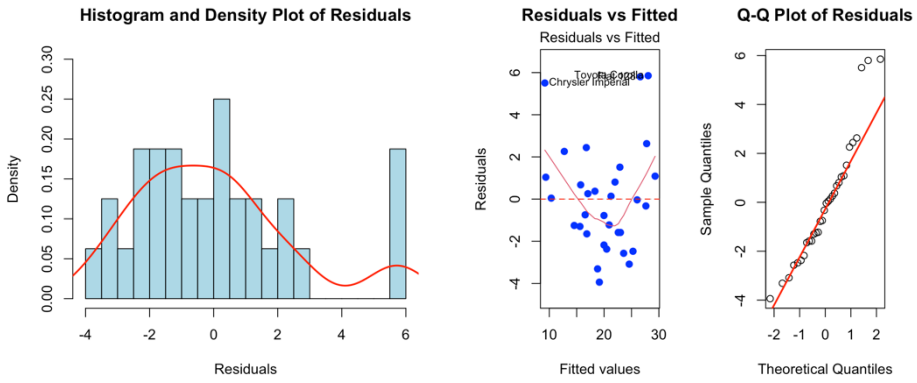
7. **Multiple Regression:** The multiple linear regression model was used to predict mileage in mpg using gross hp and weight as predictors. Here's the interpretation of the results:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	37.23	1.60	23.29	2.00E-16
hp	-0.03	0.01	-3.52	0.00145
wt	-3.88	0.63	-6.13	1.12E-06

Significance of Coefficients:
- hp: The p-value for hp is 0.00145, which is less than 0.05, indicating that horsepower is a statistically significant predictor of mpg.
- wt: The p-value for wt is 1.12e-06, which is also less than 0.05, indicating that weight is a statistically significant predictor of mpg.

Multiple R-squared: **0.8268**: This means that approximately 82.68% of the variance in mpg is explained by the model, which is a good fit. Also, **Adjusted R-squared**: **0.8148**: Adjusted R-squared accounts for the number of predictors, so this value confirms that the model is robust and does not overfit the data.
Both horsepower and weight are statistically significant predictors of mpg. The model explains 82.68% of the variability in mpg, which indicates a strong relationship between these predictors and fuel efficiency. The negative coefficients suggest that higher horsepower and weight both reduce mpg.

8. **Model Diagnostics:**



Residuals vs Fitted Plot
- The **Residuals vs Fitted** plot helps check for homoscedasticity (constant variance of residuals).
- The points are spread randomly around the horizontal line (red dashed line at 0), indicating no clear pattern or funnel shape.
- This suggests that the assumption of homoscedasticity holds reasonably well.

Q-Q Plot of Residuals
- The points generally fall along the red reference line, suggesting the residuals are approximately normally distributed.
- There are no extreme deviations, confirming that the normality assumption is met.

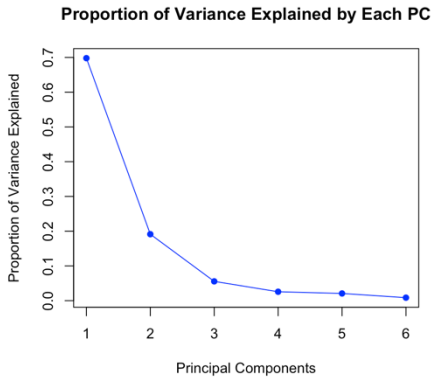
Advance Analysis

9. Principle Component Analysis:

	PC1	PC2	PC3	PC4	PC5	PC6
Standard deviation	2.05	1.07	0.58	0.39	0.35	0.23
Proportion of Variance	0.70	0.19	0.06	0.03	0.02	0.01
Cumulative Proportion	0.70	0.89	0.94	0.97	0.99	1.00

The output shows the importance of each principal component. The first principal component (PC1) has a standard deviation of 2.0463 and explains 69.79% of the total variance. The second principal component (PC2) has a standard deviation of 1.0715 and explains an additional 19.13% of the variance. Together, PC1 and PC2 account for 88.92% of the variance. Including the third component (PC3) brings the cumulative proportion to 94.48%, which covers most of the dataset's variability.

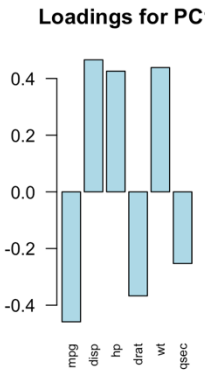
Scree Plot : The scree plot shows a sharp drop after PC1 and PC2, suggesting that these two components capture most of the information. After PC3, the variance explained by subsequent components is minimal, indicating diminishing returns. Hence first 3 components can be selected.

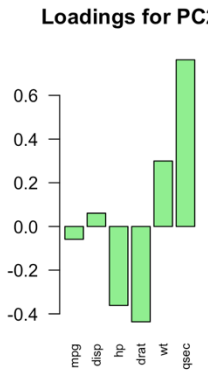


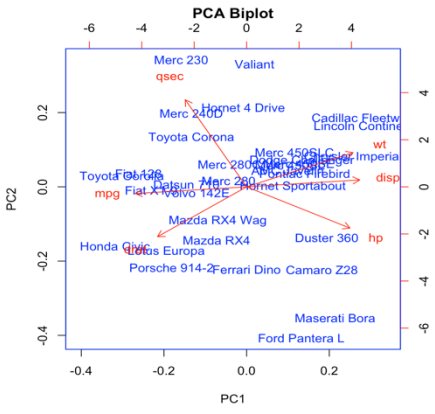
10. PCA Interpretation :

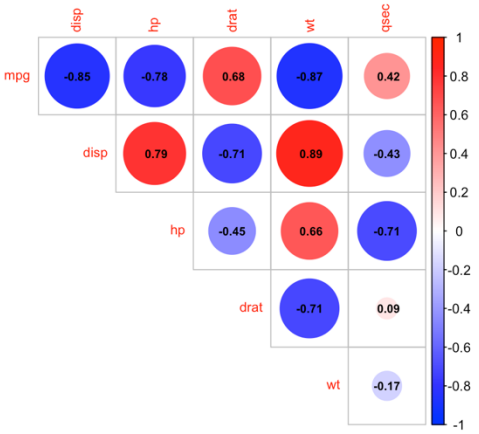
- Variables like mpg, disp, hp, wt have strong contributions to PC1, indicating their high influence on the first component. These variables have the highest loadings, suggesting they contribute significantly to the first principal component. This suggests PC1 captures the trade-off between engine size/power and fuel efficiency.
- Variables like drat and qsec contribute significantly to PC2. PC2 differentiates vehicles based on speed performance and rear axle ratio.
- The same can be viewed from the PC1 and PC2 Loadings table and graph as shown below.
- The heat map of correlation matrix shows how strongly each pair of variables is correlated. These patterns reinforce the insights gained from the PCA loadings.

	mpg	disp	hp	drat	wt	qsec
PC1	-0.46	0.47	0.43	-0.37	0.44	-0.25
PC2	-0.06	0.06	-0.36	-0.44	0.30	0.76









Conclusion: The univariate analysis highlighted key statistics of individual variables revealing variability in fuel efficiency and vehicle weight. Multivariate analysis identified significant correlations, such as mpg being negatively correlated with disp and hp. The regression model indicated disp and wt as strong predictors of mpg. PCA reduced dimensionality effectively, with the first 3 components capturing 95% of the variance. The analysis showed clear groupings: heavier, powerful cars have lower fuel efficiency, while lighter cars are more efficient. These insights provide a comprehensive understanding of vehicle performance and characteristics in the mtcars dataset.