

Indian Statistical Institute, Kolkata

PGDBA 2024-2026



SSD ASSIGNMENT

Prof Subhajith Dutta

Submitted by

Adwaith Aravind A

24BM6JP02

PGDBA 2024-26

Univariate Analysis

1.Data Overview

Dataset	Short Description	No. of Observation	No. of Variable	Some features used for analysis	Analysis Objective
mtcars	Mileage of cars	32	11	*mpg,*cyl,*hp	Dependency of fuel efficiency on other features
Iris	Plant charecteristics	150	5	sepal.length, petal.length	Relationship among plant features
airquality	Quality of air	153	6	ozone,solar. R,wind	Relationship between meteorological factors and air pollution
LifeSaving Cycle	Economic Savings	50	6	*sr,*dpi, pop15, pop75	Savings based on demographic/economic factor of 50 countries

2. Summary Statistics

Numerical Variable	Dataset	Mean	Median	Standard Deviation	Minimum	Maximum
mpg	mtcar	20.09	19.2	6.027	10.4	33.9
Sepal length	Iris	5.84	5.8	.828	4.3	7.9
Ozone	airquality	42.12	42.12	28.69	1	168
sr	LifeSaving Cycle	9.671	10.51	4.48	0.6	21.1

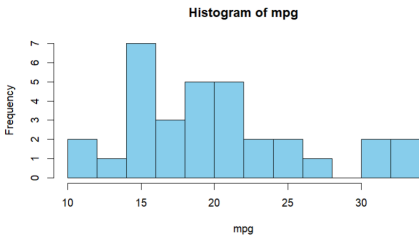
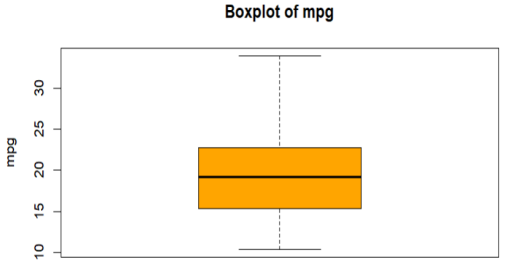
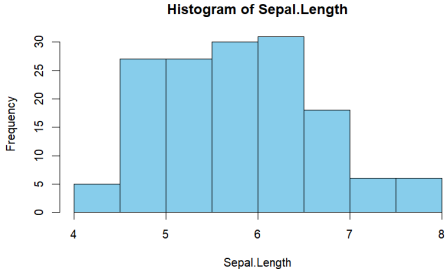
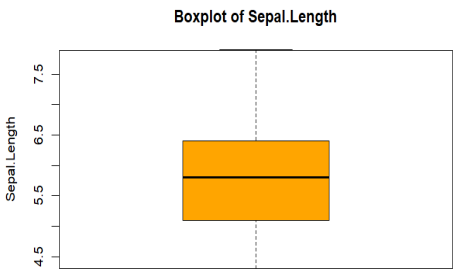
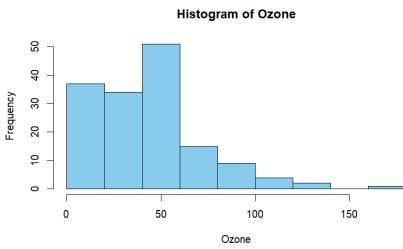
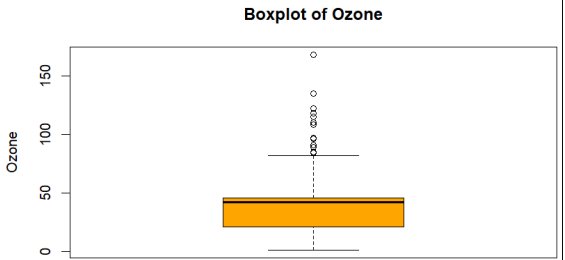
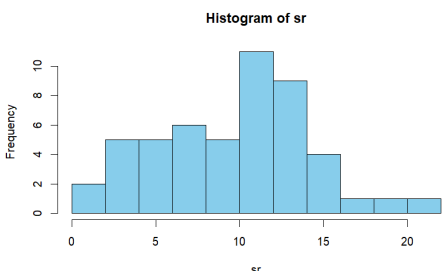
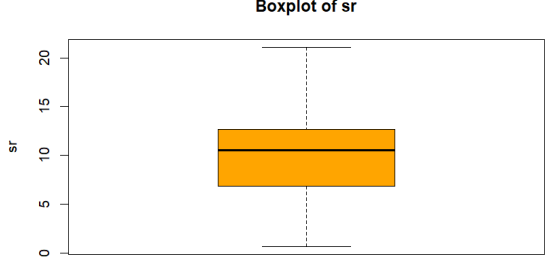
Interpretation:

- mtcars (mpg): The mean (20.09) and median (19.2) indicate a roughly symmetric distribution, though the standard deviation (6.03) suggests significant variation in fuel efficiency across vehicles.
- Iris (Sepal.Length): Mean (5.84) and median (5.8) are close, indicating symmetry. A smaller standard deviation (0.828) shows less variability in sepal lengths across samples.
- airquality (Ozone): A higher standard deviation (28.69) relative to the mean (42.12) suggests wide variation in ozone levels, indicative of outliers or skewed data.
- LifeCycleSavings (sr): A mean (9.671) lower than the median (10.51) suggests a left-skewed distribution, with some countries having extremely low savings ratios.

*mpg = miles per gallon , *cyl = no. of cylinder ,*hp=horse power ,

*sr = saving ratio per capital , *dpi= per capita disposable income,*ddpi= growth rate of dpi

3.Distribution Visualisation

Variable	Histogram	Boxplot
mtcars (mpg): The histogram and boxplot shows a <u>positive skew</u> , with most cars clustering in the 15-25 mpg range. The boxplot reveals no outliers.	 <p>Histogram of mpg</p>	 <p>Boxplot of mpg</p>
Iris (Sepal.Length): Displays a <u>near-normal distribution</u> with <u>no significant outliers</u> , consistent with natural biological data	 <p>Histogram of Sepal.Length</p>	 <p>Boxplot of Sepal.Length</p>
airquality (Ozone): The histogram shows a <u>right-skewed distribution</u> with several <u>high ozone levels as outliers</u> , as confirmed by the boxplot.	 <p>Histogram of Ozone</p>	 <p>Boxplot of Ozone</p>
LifeCycleSavings (sr): The histogram highlights the concentration of most savings ratios below 15%, with a few countries showing higher values. <u>slightly right-skewed</u> , as the tail of the histogram extends further on the right side. No outliers.	 <p>Histogram of sr</p>	 <p>Boxplot of sr</p>

*mpg = miles per gallon , *cyl = no. of cylinder , *hp=horse power ,

*sr = saving ratio per capital , *dpi= per capita disposable income, *ddpi= growth rate of dpi

4. Categorical Variable Analysis

	Barplot												
cars in mtcars have 4, 6, or 8 cylinders, with 8-cylinder cars dominating , 6-cylinder cars are less	<p>Distribution of cyl</p> <table border="1"> <thead> <tr> <th>cyl</th> <th>Frequency</th> </tr> </thead> <tbody> <tr> <td>4</td> <td>11</td> </tr> <tr> <td>6</td> <td>7</td> </tr> <tr> <td>8</td> <td>14</td> </tr> </tbody> </table>	cyl	Frequency	4	11	6	7	8	14				
cyl	Frequency												
4	11												
6	7												
8	14												
Each species of Iris flower in Iris dataset has exactly 50 observations, making it evenly distributed for analysis	<p>Distribution of Species</p> <table border="1"> <thead> <tr> <th>Species</th> <th>Frequency</th> </tr> </thead> <tbody> <tr> <td>setosa</td> <td>50</td> </tr> <tr> <td>versicolor</td> <td>50</td> </tr> <tr> <td>virginica</td> <td>50</td> </tr> </tbody> </table>	Species	Frequency	setosa	50	versicolor	50	virginica	50				
Species	Frequency												
setosa	50												
versicolor	50												
virginica	50												
The airquality dataset in R provides measurements of air quality parameters in New York during May to September 1973. Observations span five months (May to September), allowing for seasonal analysis of air quality. The Ozone and Solar.R variables had missing values found during initial analysis but after imputation it is evenly distributed	<p>Distribution of Month</p> <table border="1"> <thead> <tr> <th>Month</th> <th>Frequency</th> </tr> </thead> <tbody> <tr> <td>5</td> <td>30</td> </tr> <tr> <td>6</td> <td>30</td> </tr> <tr> <td>7</td> <td>30</td> </tr> <tr> <td>8</td> <td>30</td> </tr> <tr> <td>9</td> <td>30</td> </tr> </tbody> </table>	Month	Frequency	5	30	6	30	7	30	8	30	9	30
Month	Frequency												
5	30												
6	30												
7	30												
8	30												
9	30												
The dataset explores relationships between economic and demographic variables and the savings behavior of nations. It didn't had categorical datasets, but to check whether the dataset captured data evenly from all range of savings ratio ,data made into five categories based on their savings ratio. Medium saving ratio(10-15) is dominating in the dataset,then low(5-10),very low(<5),high(15-20) and very high(>20) in the order. *(a-b) shows saving ratio in the range a to b.	<p>Distribution of SavingCategory</p> <table border="1"> <thead> <tr> <th>SavingCategory</th> <th>Frequency</th> </tr> </thead> <tbody> <tr> <td>Very Low</td> <td>10</td> </tr> <tr> <td>Low</td> <td>13</td> </tr> <tr> <td>Medium</td> <td>22</td> </tr> <tr> <td>High</td> <td>3</td> </tr> <tr> <td>Very High</td> <td>1</td> </tr> </tbody> </table>	SavingCategory	Frequency	Very Low	10	Low	13	Medium	22	High	3	Very High	1
SavingCategory	Frequency												
Very Low	10												
Low	13												
Medium	22												
High	3												
Very High	1												

*mpg = miles per gallon , *cyl = no. of cylinder , *hp=horse power ,

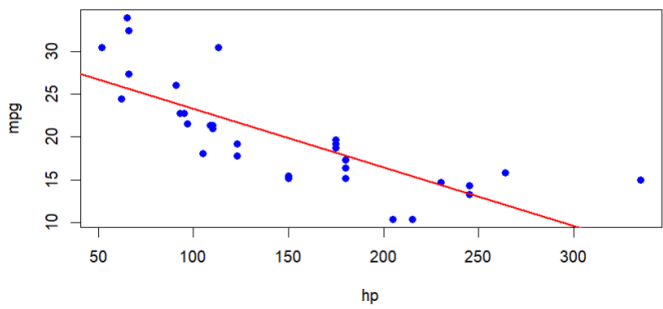
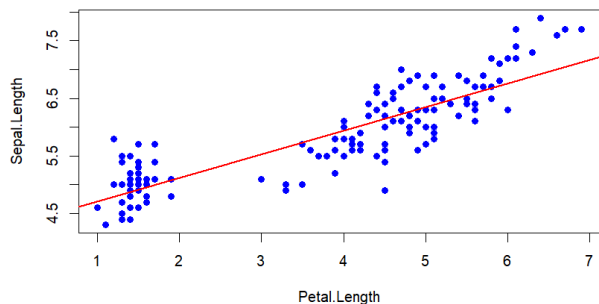
*sr = saving ratio per capital , *dpi= per capita disposable income,*ddpi= growth rate of dpi

Multivariate Analysis

5. Correlation Analysis

Variable	Dataset	Pearson correlation	Relationship between these two variables
Mpg vs hp	mtcar	-0.7761684	A strong negative Pearson correlation (-0.776) indicates that cars with higher horsepower generally have lower fuel efficiency.
Sepal.length vs Petal.Length	Iris	0.87175	A high positive correlation (0.871) confirms a linear relationship between these variables, reflecting proportional growth in plant features.
Ozone vs Temp	airquality	0.6087	A moderate positive correlation (0.609) suggests warmer days tend to have higher ozone levels.
Sr vs dpi	LifeCycleSavings	0.22035	A weak positive correlation (0.22) between savings ratio and disposable income, possibly affected by other economic factors.

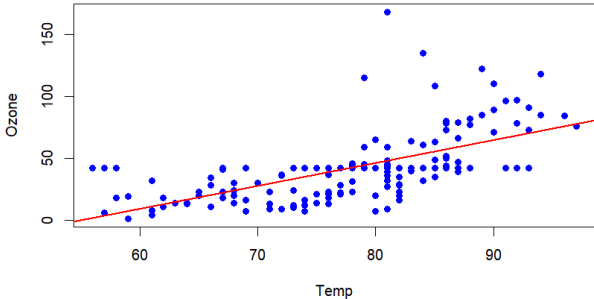
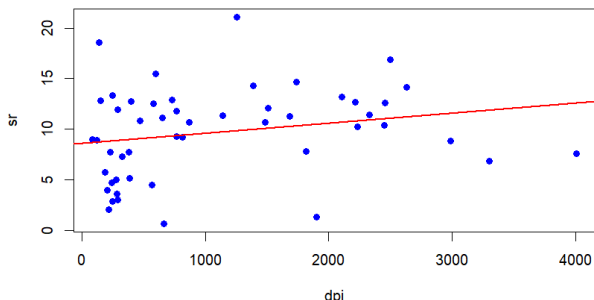
6.Scatter Plot Visualisation

<p style="text-align: center;">Scatter Plot of mpg vs hp</p> 	<p>mtcars</p> <p>the negative trend between mpg and hp is visually apparent even with a low observation dataset mtcars. As the observations increases, I expect a more fit line .A strong negative correlation statistically support my expectation.</p> <p>Points are not tightly clustered .It is explanotory since manufacturing technology and quality varies with different manufacturers.</p>
<p style="text-align: center;">Scatter Plot of Sepal.Length vs Petal.Length</p> 	<p>Iris</p> <p>the positive trend between sepal and petal lengths is clear, with points clustering tightly around the trend line showing less variability</p>

*mpg = miles per gallon , *cyl = no. of cylinder ,*hp=horse power ,

*sr = saving ratio per capital , *dpi= per capita disposable income,*ddpi= growth rate of dpi

Multivariate Analysis

<p style="text-align: center;">Scatter Plot of Ozone vs Temp</p>  <p>A scatter plot showing the relationship between Temperature (Temp) on the x-axis and Ozone levels on the y-axis. The x-axis ranges from approximately 55 to 95, and the y-axis ranges from 0 to 150. The data points are blue dots, and a red regression line shows a positive correlation. There are several outliers, particularly at higher temperature values.</p>	<p>airquality shows a less tightly clustered but still noticeable positive trend between ozone levels and temperature. Some points are in significant distance from fit line but most of the points tend to move in the same direction. These outliers were also visible from barplot and histogram. Need more investigation to know more about these outliers.</p>
<p style="text-align: center;">Scatter Plot of sr vs dpi</p>  <p>A scatter plot showing the relationship between per capita disposable income (dpi) on the x-axis and the savings ratio (sr) on the y-axis. The x-axis ranges from 0 to 4000, and the y-axis ranges from 0 to 20. The data points are blue dots, and a red regression line shows a weak positive correlation. The data points are widely scattered, indicating high variability.</p>	<p>Lifecycle Saving A weak linear trend supported by weak positive pearson correlation. The data points are widely scattered, indicating high variability in the savings ratio even at similar income levels. Also outliers are significant. This suggests that other factors besides dpi might influence the savings ratio (e.g., population demographics, cultural factors, or economic policies).</p>

7. Multiple Regression - Summary in table

Metric	mtcars	iris	airquality	LifeSaving Cycle	Interpretation of coefficient
Dependent Variable	mpg	sepal.length	Ozone	sr	The variable being predicted
Independent Variables. ($\beta_1, \beta_2, \beta_3$)	hp, wt	sepal.width, petal.length, petal.width,	Temp, Wind, Solar.R	Dpi, Pop15, pop75	The predictors. ($\beta_1, \beta_2, \beta_3$) are given in the same order in respective column data
R ² value	0.8268	0.8586	0.48	0.2744	Proportion of variance in dependent variable explained by independent variable.
Adjusted R ² value	0.8148	0.8557	0.4696	0.227	R-squared adjusted for the number of predictors.
F-Statistic	69.21 (df:2 & 29)	295.5 (df:3 & 146)	45.85 (df:3 & 149)	5.797 (df:3 & 46)	Test statistic for overall significance of the regression model.
p-value(F-Statistic)	9×10^{-12}	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	0.001898	Indicates that the overall model is statistically significant as the p-value is much smaller than 0.05.
Coefficients					Regression coefficients for each

*mpg = miles per gallon , *cyl = no. of cylinder , *hp=horse power ,

*sr = saving ratio per capital , *dpi= per capita disposable income, *ddpi= growth rate of dpi

Multivariate Analysis

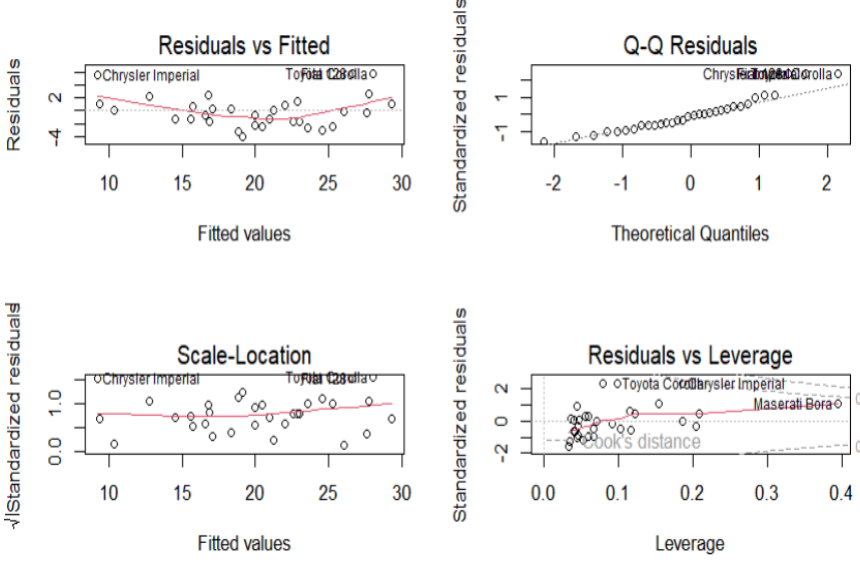
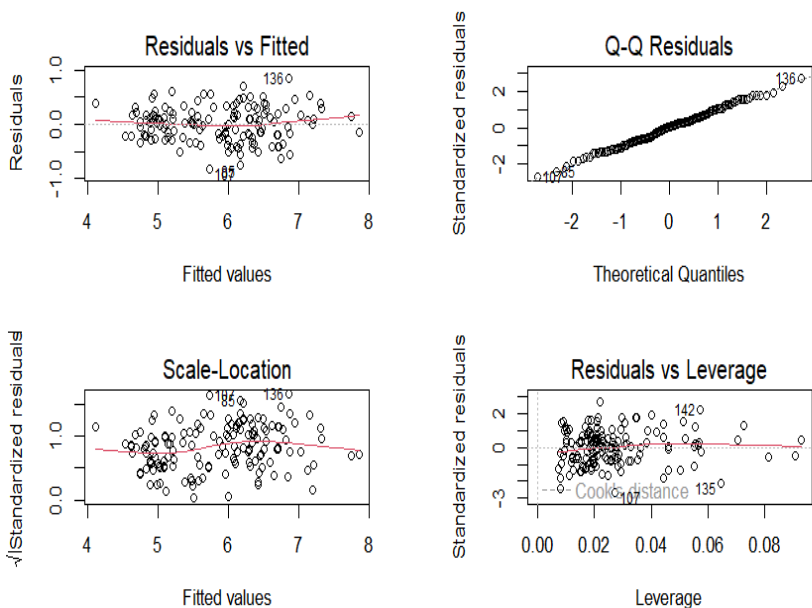
					predictor.
β_1	-0.03	0.6508	1.24	-0.000834	Effect of a 1-unit increase in β_1 on dependent variable holding other constant.
β_2	-3.88	0.709	-2.72	-0.49214	Effect of a 1-unit increase in β_2 on dependent variable holding other constant.
β_3	0	-0.556	0.058	-1.5677	Effect of a 1-unit increase in β_3 on dependent variable holding other constant.
Intercept	37.23	1.856	-38.22	31.45738	Predicted value when predictors are zero.
Residual standard Error	2.593 (df=29)	0.3145(df=29)	20.9 (df=149)	3.93 (df=46)	Measure of the variation of predicted value that remains unexplained by the model
p-value for β_1	0.00145	2×10^{-16}	1.9×10^{-8}	0.3759	Indicates has a statistically β_1 significant relationship with predicted variable.
p-value for β_2	1.12×10^{-6}	2×10^{-16}	1.5×10^{-6}	0.00186	Indicates has a statistically β_2 significant relationship with predicted variable.
p-value for β_3		2.4×10^{-5}	1.5×10^{-6}	0.16861	Indicates has a statistically β_3 significant relationship with predicted variable.

Statistic	Dataset	Min	Q1	Median	Q3	Max
Residuals	mtcars	-3.941	-1.6	-0.182	1.05	5.854
Residuals	iris	-0.82816	-0.21989	-0.01875	0.19709	0.8457
Residuals	airquality	-38.618	-14.491	-5.054	12.27	101.176
Residuals	LifeSaving Cycle	-8.6464	-2.567	-0.1188	2.28	10.3653

*mpg = miles per gallon , *cyl = no. of cylinder ,*hp=horse power ,

*sr = saving ratio per capital , *dpi= per capita disposable income,*ddpi= growth rate of dpi

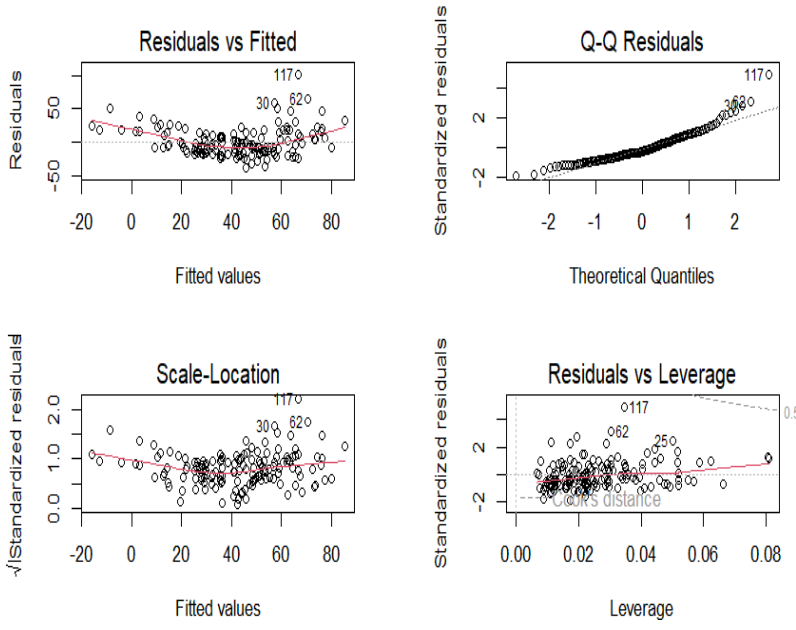
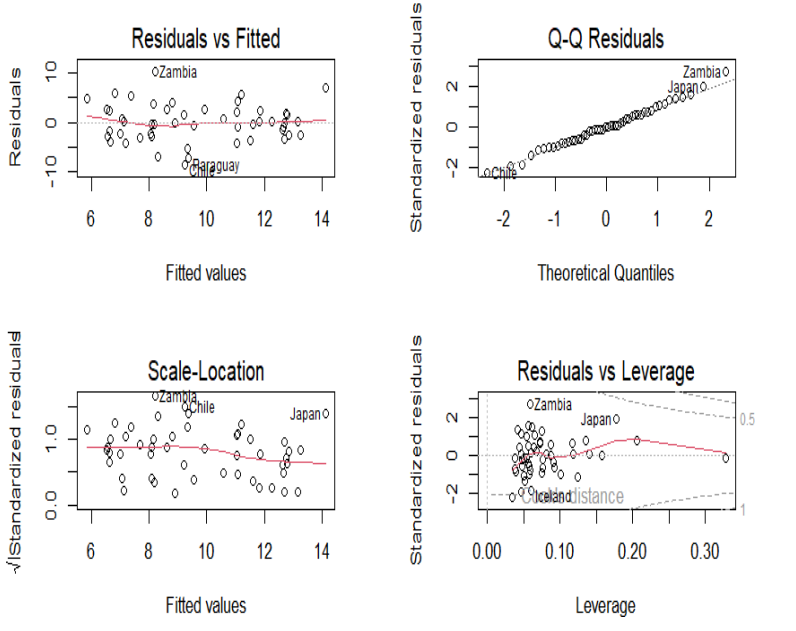
8. Model Diagnostics

	<p>Mtcars</p> <ul style="list-style-type: none"> The model mostly satisfies the assumption of homoscedasticity, but non-linearity or slight heteroscedasticity at the extremes. Most residuals follow the 45-degree line, but there are deviations at the tails The residuals are approximately normal, but non-normality at the tails. Maserati Bora have high leverage and Chrysler Imperial & Toyota Corolla are outliers. They are influential points. Model fits reasonably well.
	<p>Iris</p> <ul style="list-style-type: none"> Homoscedacity is met. Two points show a larger deviation but do not indicate major issues. The residuals are approximately normal. Minor deviations in the tails indicate slight non-normality but are not severe enough to invalidate the model. Points like 136, 135, and 142 have relatively high leverage, but none exceed Cook's distance threshold, suggesting they are not overly influential. Model fit the data well

*mpg = miles per gallon , *cyl = no. of cylinder , *hp=horse power ,

*sr = saving ratio per capital , *dpi= per capita disposable income, *ddpi= growth rate of dpi

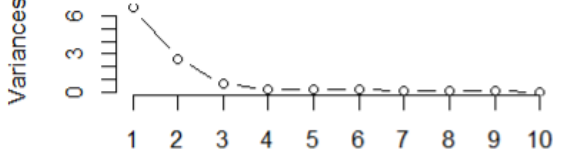
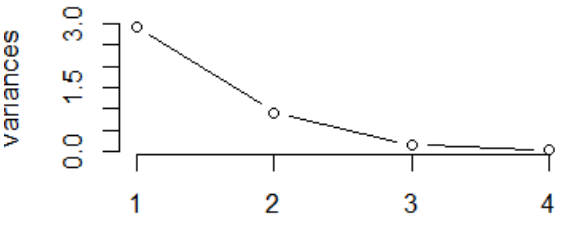
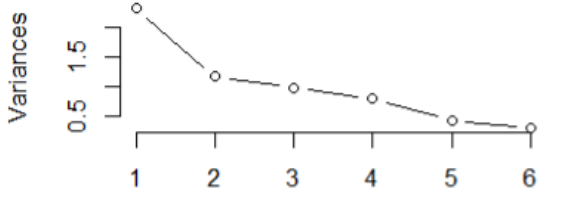
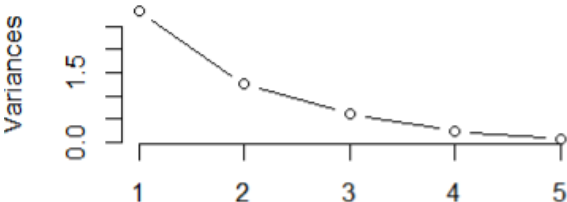
Multivariate Analysis

 <p>The diagnostic plots for the Air Quality model are as follows:</p> <ul style="list-style-type: none"> Residuals vs Fitted: Shows residuals on the y-axis (ranging from -50 to 50) against fitted values on the x-axis (ranging from -20 to 80). A red smoothing line is shown. Points 117, 30, and 62 are labeled. Q-Q Residuals: Shows standardized residuals on the y-axis (ranging from -2 to 2) against theoretical quantiles on the x-axis (ranging from -2 to 2). A red line indicates the expected normal distribution. Point 117 is labeled. Scale-Location: Shows the square root of standardized residuals on the y-axis (ranging from 0.0 to 2.0) against fitted values on the x-axis (ranging from -20 to 80). A red smoothing line is shown. Points 117, 30, and 62 are labeled. Residuals vs Leverage: Shows standardized residuals on the y-axis (ranging from -2 to 2) against leverage on the x-axis (ranging from 0.00 to 0.08). A red smoothing line and Cook's distance contours are shown. Points 117, 62, and 25 are labeled. 	<p>Air Quality</p> <ul style="list-style-type: none"> The assumption of homoscedasticity is largely satisfied. There is no significant curvature, suggesting the linearity assumption is also reasonable. The slight curvature in the red line suggests potential non-linearity at certain fitted value ranges The residuals are approximately normal. Minor deviations in the tails indicate slight non-normality but are not severe enough to invalidate the model. The model fits reasonably well
 <p>The diagnostic plots for the LifeSavingCycle model are as follows:</p> <ul style="list-style-type: none"> Residuals vs Fitted: Shows residuals on the y-axis (ranging from -10 to 10) against fitted values on the x-axis (ranging from 6 to 14). A red smoothing line is shown. Points Zambia, Chile, and Paraguay are labeled. Q-Q Residuals: Shows standardized residuals on the y-axis (ranging from -2 to 2) against theoretical quantiles on the x-axis (ranging from -2 to 2). A red line indicates the expected normal distribution. Points Zambia and Japan are labeled. Scale-Location: Shows the square root of standardized residuals on the y-axis (ranging from 0.0 to 1.0) against fitted values on the x-axis (ranging from 6 to 14). A red smoothing line is shown. Points Zambia, Chile, and Japan are labeled. Residuals vs Leverage: Shows standardized residuals on the y-axis (ranging from -2 to 2) against leverage on the x-axis (ranging from 0.00 to 0.30). A red smoothing line and Cook's distance contours are shown. Points Zambia, Japan, and Iceland are labeled. 	<p>LifeSavingCycle</p> <ul style="list-style-type: none"> There might be some non-linearity as the residuals exhibit slight patterns for certain fitted values From Q-Q plot, deviations (e.g., "Zambia") at the tails suggest the residuals might not perfectly follow a normal distribution. the variance seems to increase slightly with the fitted values (non-constant variance) points like "Zambia" and "Japan" have higher leverage, suggesting they might be influential. the linear model may not fully capture the relationship.

*mpg = miles per gallon , *cyl = no. of cylinder , *hp=horse power ,

*sr = saving ratio per capital , *dpi= per capita disposable income,*ddpi= growth rate of dpi

9. Principal Component Analysis (PCA)

<p>Mtcars</p> <p>The plot shows a sharp decline in variance after the second component, forming an "elbow" at PC2. The "elbow" in the plot indicates the optimal number of components to retain. In this case, it suggests retaining 2 components for the analysis.</p>	<p style="text-align: center;">Scree Plot</p> 
<p>Iris</p> <p>The variance explained decreases sharply after the second component, forming a clear "elbow" at Component 2. Components 3 and 4 contribute very little additional variance, suggesting they may not be significant.</p>	<p style="text-align: center;">Scree Plot</p> 
<p>Air Quality</p> <p>The variance drops steeply after the first component and then flattens out gradually. The "elbow" in the plot seems to occur at Component 2. Beyond the second component, the explained variance reduces significantly and stabilizes, contributing marginally to the total variance.</p>	<p style="text-align: center;">Scree Plot</p> 
<p>LifeSavingCycle</p> <p>There is a steep drop in explained variance after the first component. The "elbow" appears at Component 2, where the curve begins to flatten. Beyond the second component, the additional variance explained by each component is marginal and stabilizes.</p>	<p style="text-align: center;">Scree Plot</p> 

10. PCA Interpretation

Components of the Biplot:

1. Axes:
 - The x-axis (PC1) and y-axis (PC2) represent the first two principal components, which capture the most variance in the data.
 - Each axis is scaled based on the contribution of the principal components.

Advanced Analysis

2. Points:

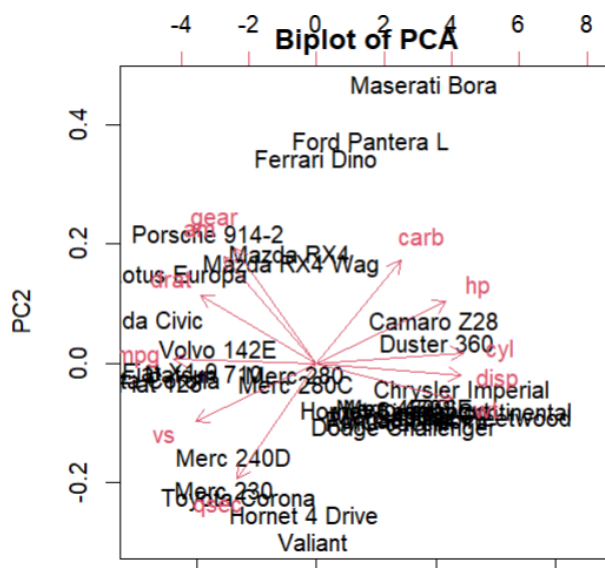
- The points represent the observations
- Observations close together are similar in terms of the original variables.

3. Vectors (Red Arrows):

- These represent the original variables and their loadings in the principal component space.
- The length and direction of the arrows indicate the contribution of each variable to the principal components.
- Variables pointing in the same direction are positively correlated, while those pointing in opposite directions are negatively correlated.

4. Projections:

- You can project the observations onto the variable vectors to interpret their relationship. For example, a car closer to the "hp" (horsepower) vector likely has higher horsepower.

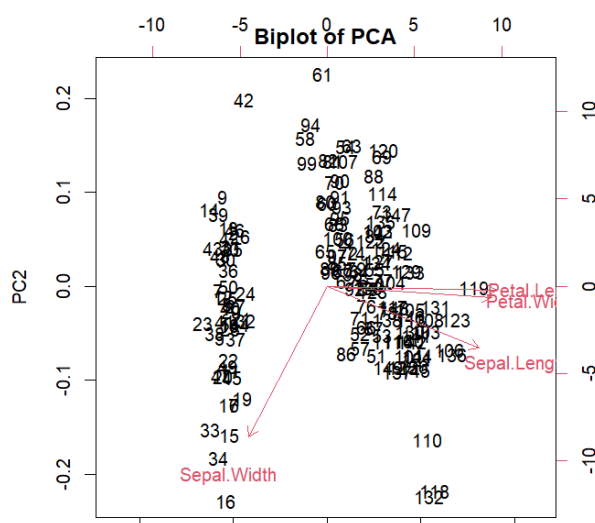


mtcars

The "hp" (horsepower) and "carb" (carburetors) variables strongly influence PC1, as their vectors are long and aligned with the PC1 axis. Variables like "vs" and "gear" have a stronger influence on PC2.

Positive correlation: Variables like "hp" and "carb" point in similar directions, indicating they are positively correlated.

Negative correlation: Variables like "mpg" (miles per gallon) and "hp" point in opposite directions, indicating they are negatively correlated.



Iris

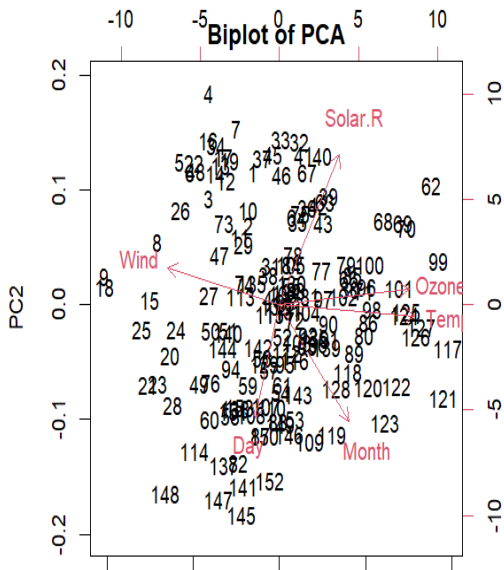
Points are scattered, with some clustering near the origin and others spread along PC1 and PC2.

Petal.Length and Petal.Width: Have strong contributions to PC1 (longer vectors along PC1) and are positively correlated as their vectors point in the same direction.

Observations like 110 and 119 are strongly influenced by Petal.Length and Petal.Width (near these vectors).

Observations like 16 and 33 are more influenced by Sepal.Width.

Advanced Analysis



AirQuality

Principal Component 1 (PC1): Dominated by Ozone and Temp, it likely represents a factor related to air quality or temperature effects.

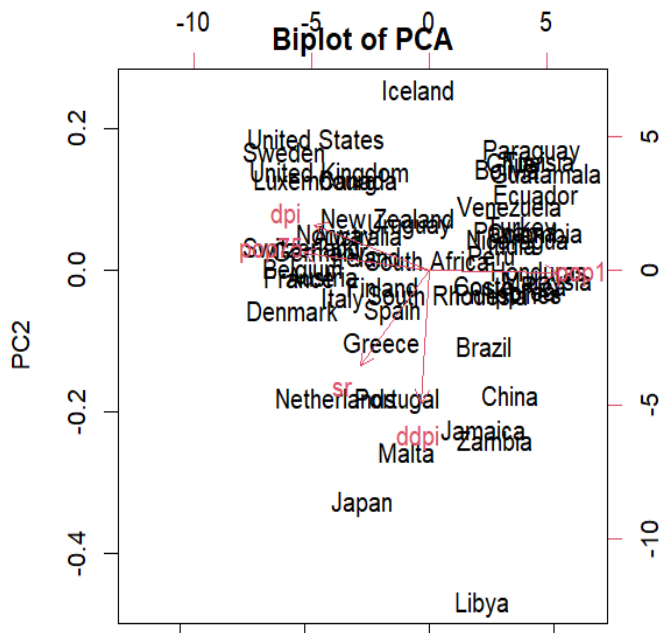
Principal Component 2 (PC2): Dominated by Wind, it likely represents wind-related effects on the observations.

Ozone and Temp are positively correlated, as their arrows point in the same direction.

Wind is negatively correlated with Ozone and Temp, as its arrow points in the opposite direction. Observations in the top-right quadrant (e.g., 62 and 99) likely have high values for Ozone and Temp.

Observations like 15 and 18 are associated with high values of Wind.

Points like 4, 145, and 148 might be outliers due to their distance from the center.



LifeSavingCycle

dpi strongly contributes to PC1, as indicated by the long vector along the x-axis.

ddpi also contributes but at a slightly different angle, influencing both PC1 and PC2.

Iceland, United States, and other developed countries cluster near the origin, suggesting average or balanced contributions from the variables. Japan is positioned negatively along PC1, likely indicating low values for "dpi." Libya is an outlier along PC2..

Conclusion

- **Mtcars:**
 - Fuel efficiency (mpg) is symmetric but varies significantly across vehicles.
 - No outliers; distribution slightly skewed.
 - Strong negative correlation between mpg and horsepower.
 - The regression model performs well ($R^2 = 0.83$), with horsepower and weight as significant predictors. Outliers like "Maserati Bora" influence results.
- **Iris:**
 - Sepal and petal dimensions show proportional growth with minimal variability.
 - Near-normal distribution with no significant outliers.
 - Strong positive correlation between sepal and petal lengths.
 - Excellent fit ($R^2 = 0.86$), with minimal deviations in residuals, making it ideal for predictive modeling.
- **Air Quality:**
 - Ozone levels are right-skewed with high variability and several outliers.
 - High ozone levels act as outliers.
 - Moderate positive correlation between ozone and temperature.
 - Reasonable fit ($R^2 = 0.48$), but variability and outliers like observation 145 suggest room for improvement.
- **LifeCycleSavings:**
 - Savings ratios are slightly skewed, with most values below 15%.
 - No significant outliers; right-skewed distribution.
 - Weak positive correlation between savings and disposable income.
 - Weak fit ($R^2 = 0.27$) due to high variability and influential points like "Zambia" and "Japan." Additional factors should be explored.