Statistical Structures in Data — Prashant Bisht
Prof. Subhajit Dutta — 24BM6JP42

Dataset 1: Diamonds Dataset

1. Overview of dataset:

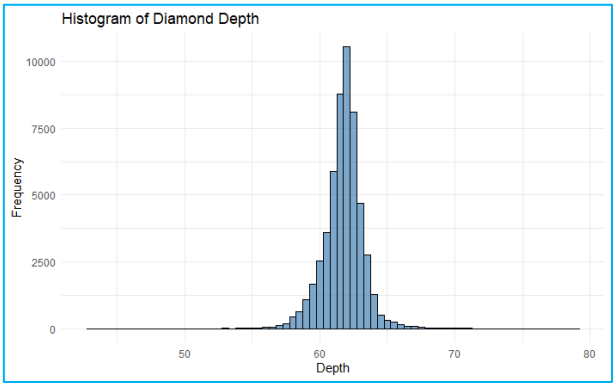
carat	cut	color	clarity	depth	table	price	x	y	z
0.23	Ideal	E	SI2	61.5	55	326	3.95	3.98	2.43
0.21	Premium	E	SI1	59.8	61	326	3.89	3.84	2.31
0.23	Good	E	VS1	56.9	65	327	4.05	4.07	2.31
0.29	Premium	I	VS2	62.4	58	334	4.2	4.23	2.63
0.31	Good	J	SI2	63.3	58	335	4.34	4.35	2.75
0.24	Very Good	J	VVS2	62.8	57	336	3.94	3.96	2.48

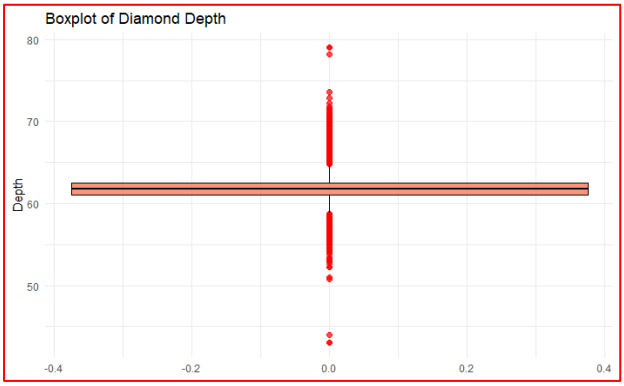
Number of observations: 53940 Number of Variables: 10

2. Summary Statistics for key variable: depth

Mean = 61.7494 **Median** = 61.8 **Standard depth** = 1.432621 **Minimum** = 43 **Maximum** = 79

3. Distribution Visualization: depth





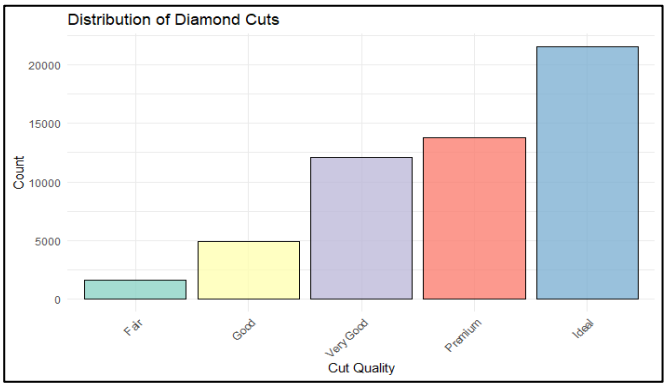
The histogram shows a bell-shaped distribution, indicating that the diamond depth follows a normal distribution centered around 60. There are potential outliers on both tails (depths below 55 and above 70), but they are relatively few compared to the central values.

The boxplot confirms a **symmetric distribution** of diamond depth centered around 60, with a narrow interquartile range. **Potential outliers** are present on both ends: depths below 55 and above 70.

4. Categorical Variable Analysis:

Distribution of diamond cuts

The **"Ideal"** cut is most frequent, followed by **"Premium"** and **"Very Good"**, indicating customer preference for high-quality cuts. **"Fair"** cuts are the least common, suggesting lower demand.

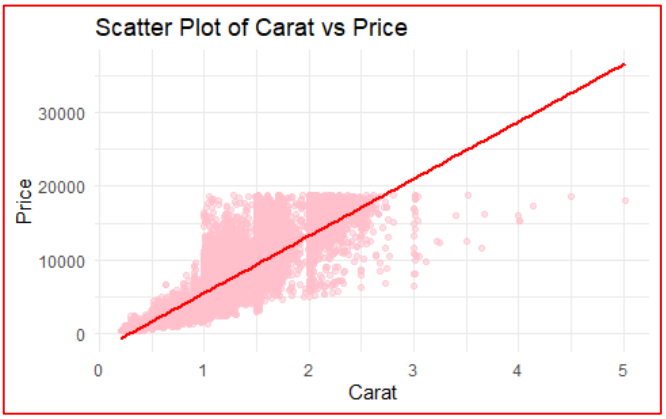


Multivariate Analysis

5. Correlation Analysis: Pearson Correlation Coefficient between carat and price: 0.9215913

Summary: The Pearson correlation suggests a **strong positive relationship** between carat and price.

6. Scatterplot Visualization:



The scatterplot shows a strong positive relationship between carat and total price, consistent with the Pearson correlation coefficient of 0.921. The trendline highlights this direct relationship.

7. Multiple Regression:

```
Call:
lm(formula = price ~ carat + cut + depth, data = diamonds)

Residuals:
     Min       1Q   Median       3Q      Max
-17475.7   -790.8    -38.8    522.6  12694.8

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  434.754    301.924   1.440     0.15
carat       7873.249     13.967 563.691  < 2e-16 ***
cut.L       1148.315     27.518  41.730  < 2e-16 ***
cut.Q       -471.615     23.750 -19.857  < 2e-16 ***
cut.C        366.133     20.195  18.130  < 2e-16 ***
cut^4         87.579     16.271   5.382 7.38e-08 ***
depth        -50.418      4.848 -10.401  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1510 on 53933 degrees of freedom
Multiple R-squared:  0.8568,    Adjusted R-squared:  0.8567
F-statistic: 5.377e+04 on 6 and 53933 DF,  p-value: < 2.2e-16
```

Key Insights:

Carat: Has the most substantial impact on price, making it the most influential predictor. Highly significant.

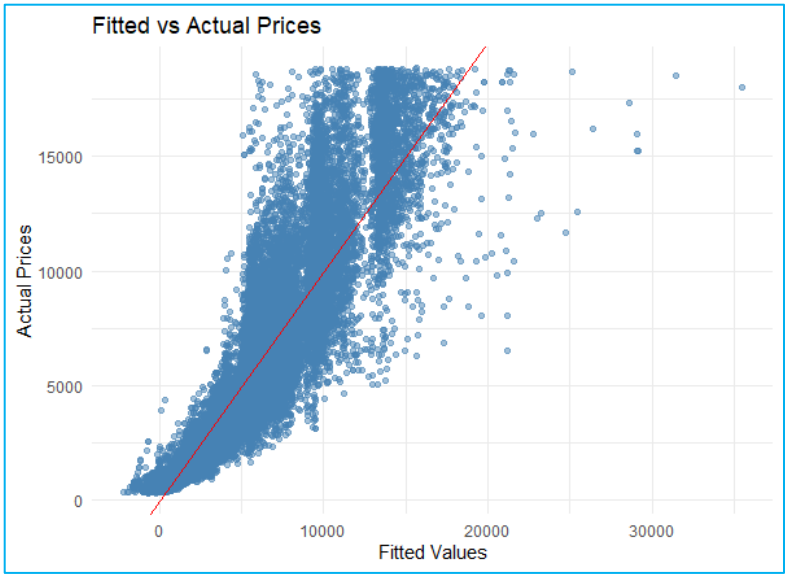
Cut: has a non-linear relationship with price, with varying levels influencing price in different ways.

Intercept (434.754): The baseline price when all predictors are zero. Not statistically significant (p = 0.15).

Depth: negatively affects price, though its impact is relatively smaller compared to carat and cut.

The model explains a significant amount of variance in price (R^2 = 0.857), and the predictors are highly statistically significant overall (F-statistic p-value is less).

Final Insights: (Fitted line)

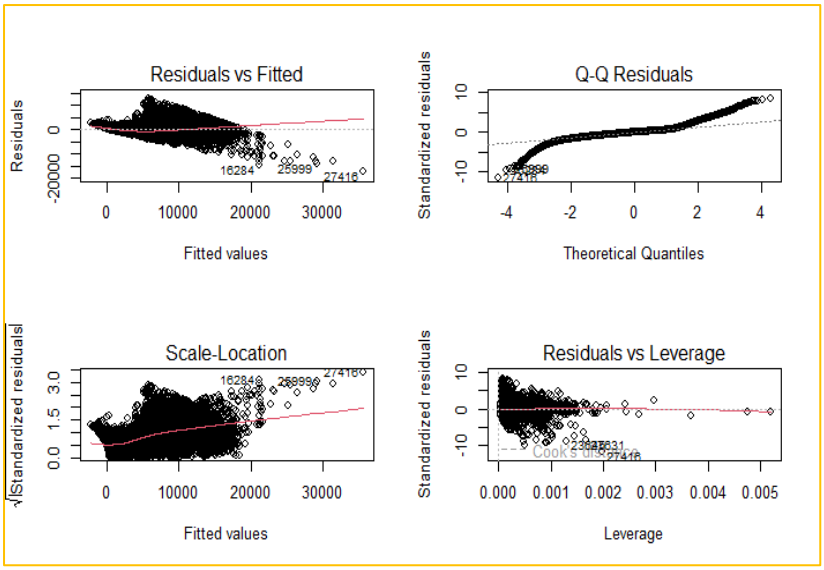


The fitted plot shows a generally strong alignment between fitted and actual prices, indicating a good fit of the model.

However, there is some dispersion and non-linearity in higher price ranges, suggesting the model may underpredict or overpredict for some expensive diamonds.

This model is a good fit for predicting diamond prices, with carat being the dominant driver.

8. Model Diagnostics:



Homoscedasticity Check:

Residuals vs. Fitted Plot: The residuals exhibit a funnel-shaped pattern, with increasing variance as fitted values increase. This suggests **heteroscedasticity** (non-constant variance).

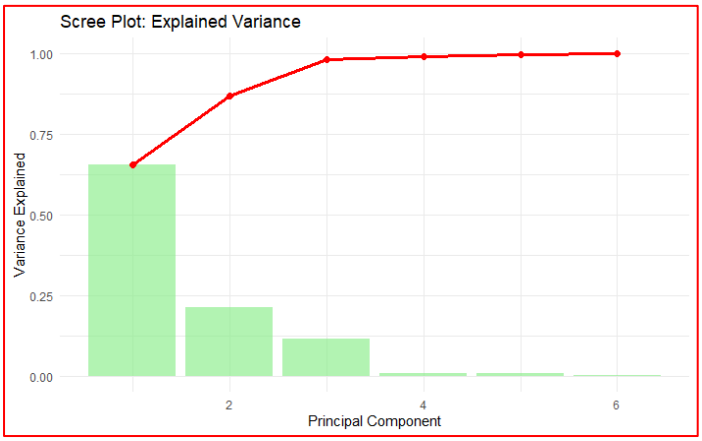
Scale-Location Plot: The red line curves upward, confirming the increasing spread of residuals, further supporting heteroscedasticity.

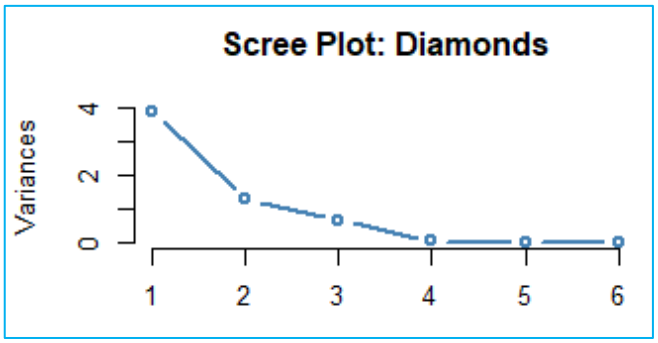
Normality Check

Normality of Residuals: [Q-Q Plot] The residuals deviate from the diagonal line at both ends, indicating that they are **not normally distributed** and that the model struggles with extreme values.

Residuals vs. Leverage Plot: A few points have high leverage, meaning these points have a strong influence on the model fit and could be outliers.

9. Principal Component Analysis:

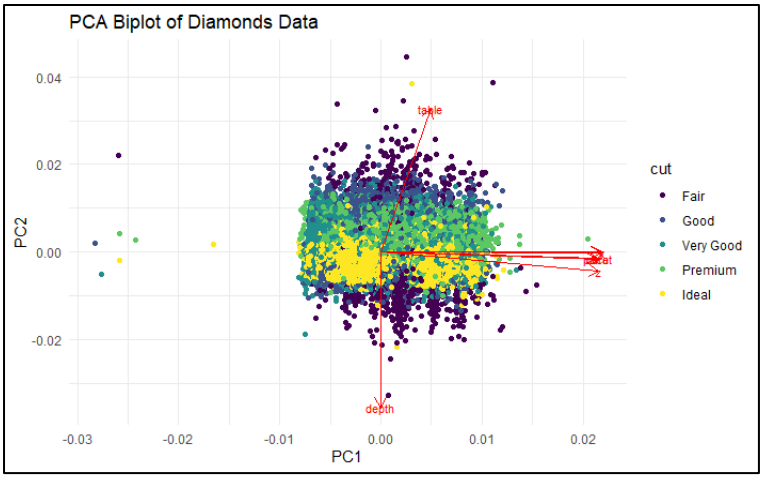




Number of components chosen = 3. Because the scree plot shows an elbow or a sharp decline transition to gentle slope. And 90%+ of the variability seems to be explained at around 3 PCs.

10. PCA Interpretation:

Observed Patterns and Groupings: Cut Quality:



Ideal cuts (yellow): Tightly clustered near the center, indicating uniformity in size and proportion attributes for diamonds with this cut.
Fair cuts (dark blue): Spread further from the center, showing greater variability in quality.
Other cuts (Good, Very Good, Premium): Form intermediate groupings, reflecting moderate consistency in their features.

Principal Component Influence:

- PC1 (Horizontal Axis): Dominated by carat, x, y, and z, it differentiates diamonds by size. Larger diamonds will have higher PC1 scores, while smaller ones have lower PC1 scores.
- PC2 (Vertical Axis): Influenced by depth and table. It likely reflects shape and proportion-related attributes that are not directly linked to size.

Outliers: Some points are farther from the center, representing diamonds with extreme size (carat, x, y, z) or unusual proportions (depth, table).

Statistical Structures in Data — Prashant Bisht
Prof. Subhajit Dutta — 24BM6JP42

Dataset 2: IRIS Dataset

1. Overview of dataset:

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa

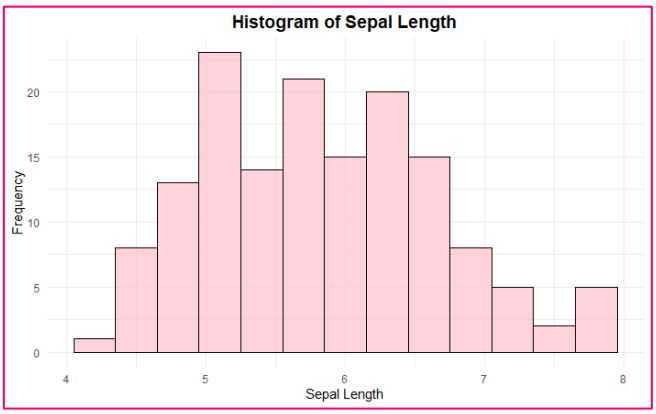
Number of observations: 150 Number of Variables: 5

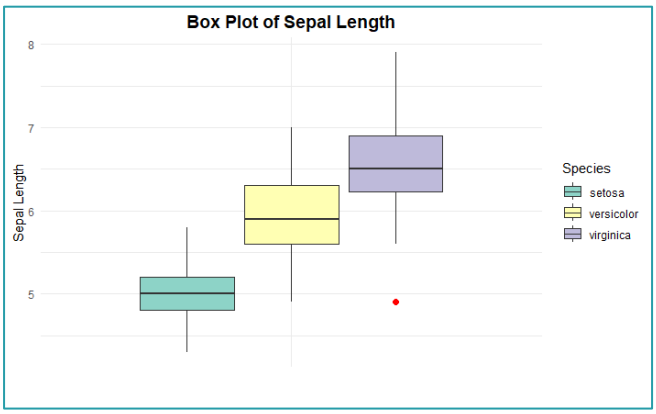
2. Summary Statistics for key variable: Sepal.Length

Mean = 61.7494 **Median** = 61.8 **Standard depth** = 1.432621 **Minimum** = 43 **Maximum** = 79

The values suggest that the sepals' lengths are moderately spread around the mean, with a range of 3.6 units (7.9 - 4.3). The distribution appears to be symmetric based on the closeness of the mean and median.

3. Distribution Visualization: Sepal.Length





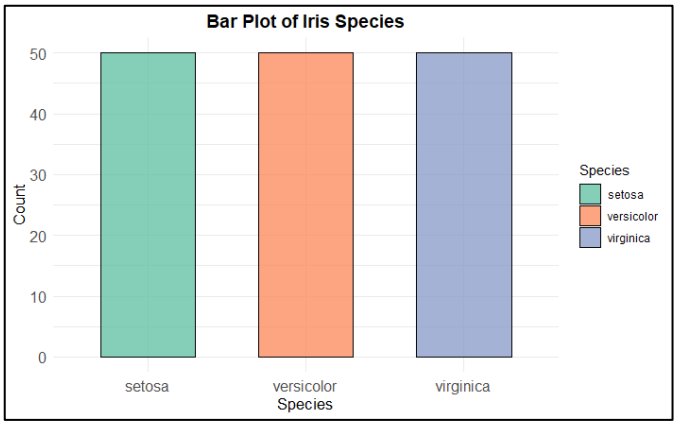
The histogram of Sepal Length is right skewed and almost normally distributed. The Box- plot for virginica and versicolor is right skewed and has an outlier as well.

4. Categorical Variable Analysis:

Distribution of IRIS species:

All the 3 categories have the same count of 50.

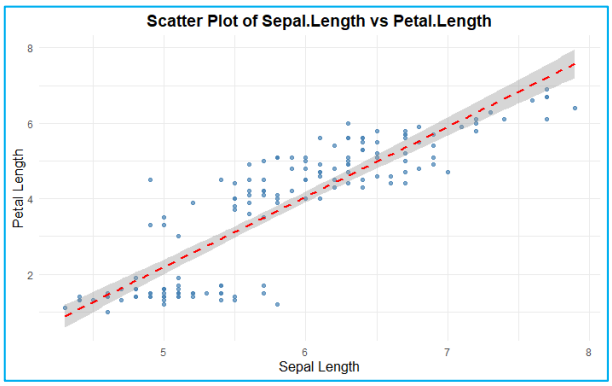
So, the distribution of categorical variable species is uniform.



Multivariate Analysis

5. Correlation Analysis: Pearson Correlation Coefficient between Sepal.Length and Petal.Length: 0.8717538

***Summary*:** The Pearson correlation suggests a **strong positive relationship** between Sepal.Length and Petal.Length.



6. Scatterplot Visualization:

The scatterplot shows a strong relationship between Sepal.Length and Petal.Length, consistent with the Pearson correlation coefficient of 0.871. The trendline highlights this direct relationship.

7. Multiple Regression:

```
Call:
lm(formula = Sepal.Length ~ Sepal.Width + Petal.Length + Petal.Width,
    data = iris)

Residuals:
     Min       1Q   Median       3Q      Max
-0.82816 -0.21989  0.01875  0.19709  0.84570

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.85600    0.25078   7.401 9.85e-12 ***
Sepal.Width   0.65084    0.06665   9.765  < 2e-16 ***
Petal.Length  0.70913    0.05672  12.502  < 2e-16 ***
Petal.Width  -0.55648    0.12755  -4.363 2.41e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3145 on 146 degrees of freedom
Multiple R-squared:  0.8586,    Adjusted R-squared:  0.8557
F-statistic: 295.5 on 3 and 146 DF,  p-value: < 2.2e-16
```

Coefficients Interpretation:

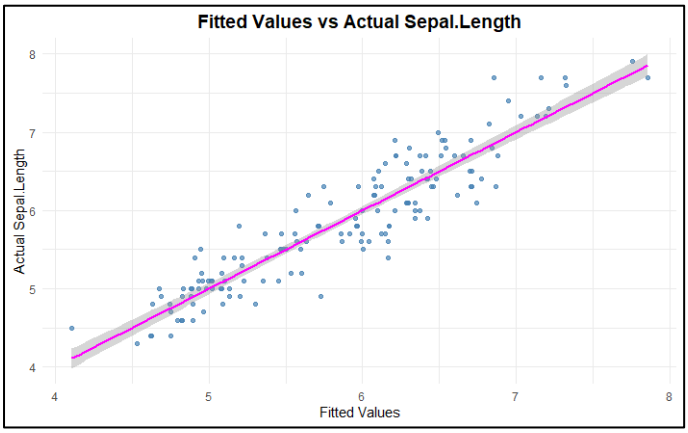
***Intercept (1.856)*:** When all predictors (Sepal.Width, Petal.Length, and Petal.Width) are zero, the predicted Sepal.Length is 1.856 units. This is the baseline value.

***Sepal.Width (0.65084)*:** A unit increase in Sepal.Width increases Sepal.Length by approximately 0.65 units, holding other predictors constant. Highly significant ($p < 2e\text{-}16$).

***Petal.Length (0.70913)*:** A unit increase in Petal.Length increases Sepal.Length by about 0.71 units, holding other predictors constant. Highly significant ($p < 2e\text{-}16$).

***Petal.Width (-0.55648)*:** A unit increase in Petal.Width decreases Sepal.Length by approximately 0.56 units, holding other predictors constant. Highly significant ($p < 0.001$).

Fitted line



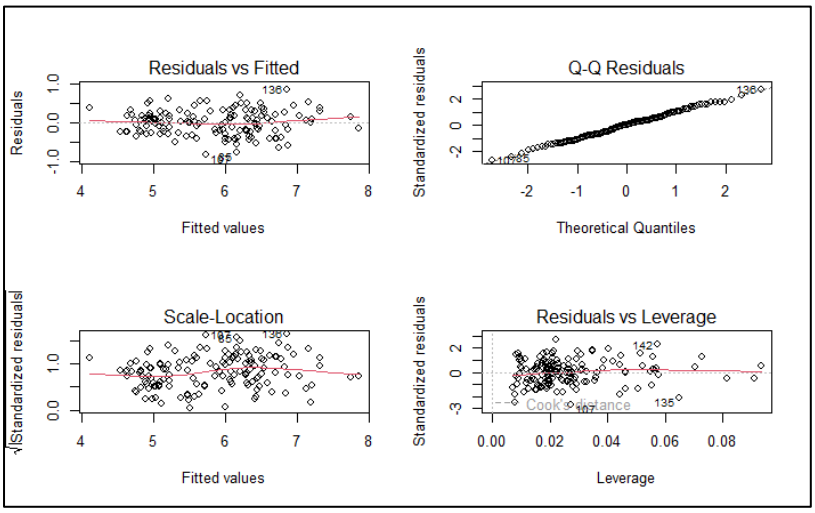
Key Insights:

1. The model explains **85.86%** of the variance in Sepal.Length ($R^2 = 0.8586$), indicating a strong fit.

2. **Petal.Length** has the strongest positive influence, followed by Sepal.Width. Interestingly, **Petal.Width** has a negative effect, suggesting an inverse relationship with Sepal.Length.

3. The small **Residual Standard Error (0.3145)** and highly significant F-statistic ($p < 2.2e\text{-}16$) confirm the model's reliability.

8. Model Diagnostics:



Homoscedasticity Check:

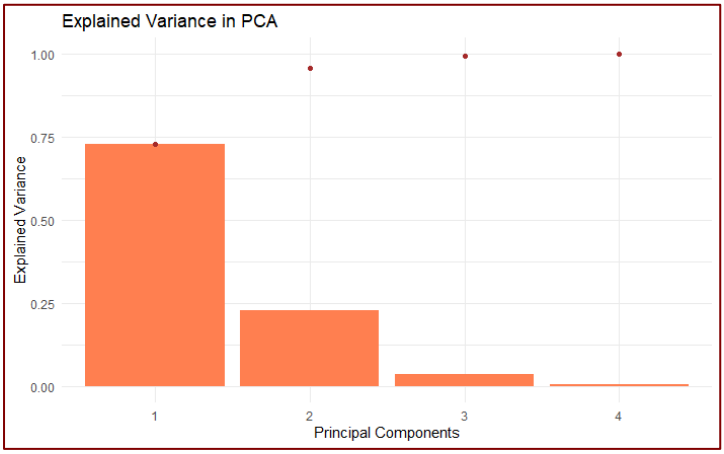
The residuals appear randomly scattered around 0 without any discernible pattern. This suggests that the assumption of **homoscedasticity** is reasonably satisfied.

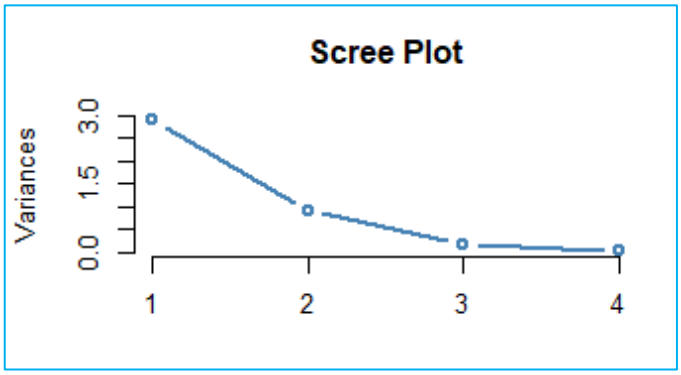
The standardized residuals appear fairly evenly spread across fitted values, and the red line is relatively flat.

Normality Check: The residuals mostly follow the diagonal line, indicating that the residuals are **approximately normally distributed**. Some minor deviations at the tails, but they are not severe.

Overall Model Assessment: Most points have low leverage, and there are no residuals with extremely high Cook's distance. A few observations may be influential (e.g., points near Cook's distance lines), but they don't dominate the model fit.

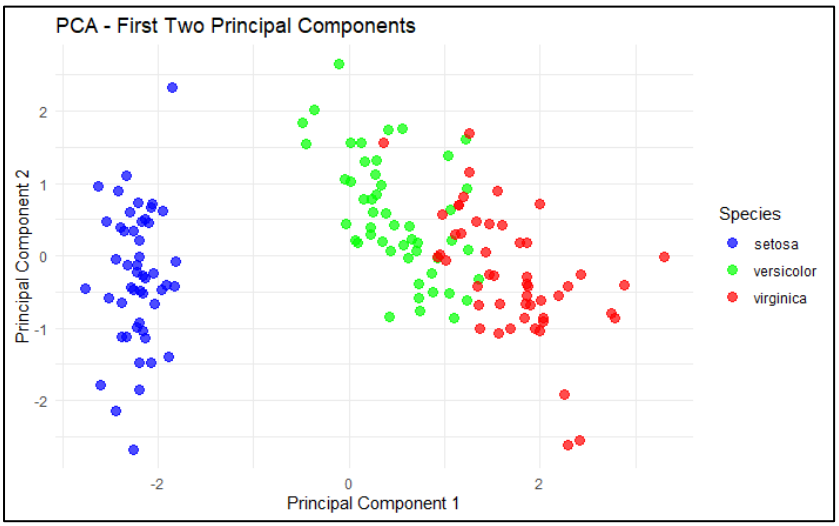
Principal Component Analysis:





Number of components chosen = 2. Because the scree plot shows an elbow or a sharp decline transition to gentle slope. And 95%+ of the variability seems to be explained at around 3 PCs.

9. PCA Interpretations:



Loadings of the First 2 Principal Components:

PC1: The first principal component (PC1) appears to separate the 'setosa' species from the other two species. This suggests that PC1 captures variation related to features that differentiate 'setosa' from 'versicolor' and 'virginica'.

PC2: The second principal component (PC2) seems to differentiate between 'versicolor' and 'virginica'. This indicates that PC2 captures variation related to features that distinguish these two species.

Clear Separation of 'setosa': The 'setosa' species is distinctly separated from the other two species along the PC1 axis. This suggests that 'setosa' has unique characteristics that set it apart from 'versicolor' and 'virginica'.

Overlap between 'versicolor' and 'virginica': There is some overlap between the 'versicolor' and 'virginica' species, particularly along the PC2 axis. This indicates that these two species share some similarities, making it more difficult to distinguish them based on the captured variation.

Statistical Structures in Data — Prashant Bisht
Prof. Subhajit Dutta — 24BM6JP42

Dataset 3: MPG Dataset

1. Overview of dataset:

manufacturer	model	displ	year	cyl	trans	drv	cty	hwy	fl	class
audi	a4	1.8	1999	4	auto(l5)	f	18	29	p	compact
audi	a4	1.8	1999	4	manual(m5)	f	21	29	p	compact
audi	a4	2	2008	4	manual(m6)	f	20	31	p	compact
audi	a4	2	2008	4	auto(av)	f	21	30	p	compact
audi	a4	2.8	1999	6	auto(l5)	f	16	26	p	compact

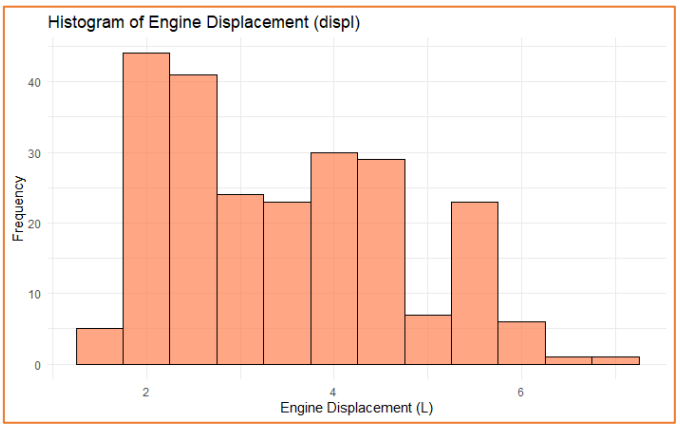
Number of observations: 234 Number of Variables: 11

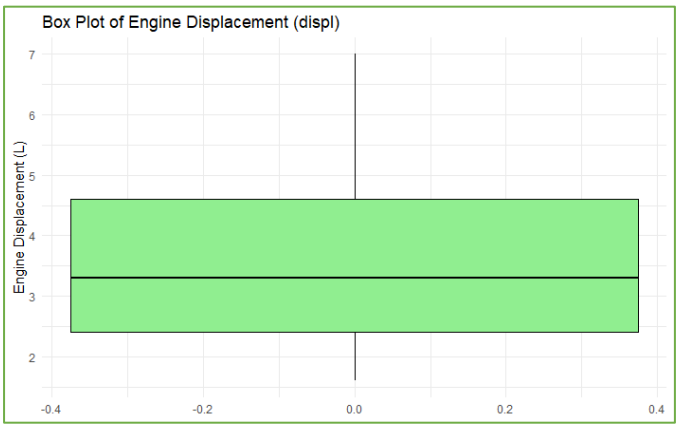
2. Summary Statistics for key variable: displ (Engine Displacement)

Mean = 3.4717 **Median** = 3.3 **Standard depth** = 1.291959 **Minimum** = 1.6 **Maximum** = 7

The dataset exhibits moderate variability in engine displacement, with values ranging from 1.6 to 7 Liters. The slight difference between the mean and median suggests the presence of larger engine sizes that slightly skew the distribution.

3. Distribution Visualization: displ (Engine Displacement)





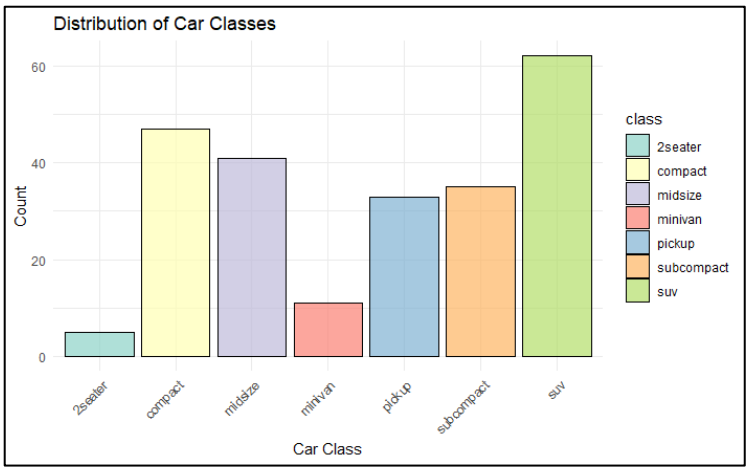
The distribution of engine displacement is **right-skewed** with a **peak** around 2.5 L. There is a **long tail** towards higher displacement values, suggesting the presence of some **outliers** with larger engines.

The distribution of engine displacement appears to be **roughly symmetric** with a **central tendency** around 3.5 L. There is a **single outlier** with a displacement value of approximately 7 L, indicated by the vertical line extending above the box plot.

4. Categorical Variable Analysis:

The distribution appears to be **multimodal**, with distinct peaks for different car classes. This indicates that there are several preferred car classes among the population represented in the dataset.

The plot shows the distribution of car classes. **SUVs** are the most common class, followed by **compact** and **2-seaters**. **Minivans** are the least common class. This visualization suggests that SUVs are the most popular choice among car buyers in this dataset.



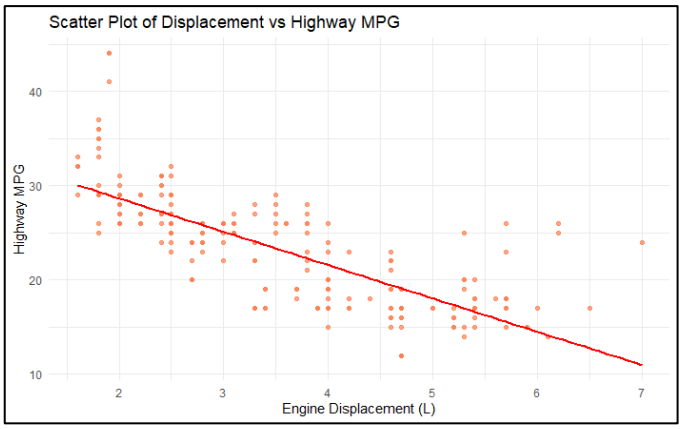
Multivariate Analysis

5. Correlation Analysis: Pearson Correlation Coefficient between displ and hwy: -0.766

***Summary*:** The Pearson correlation suggests a **negative relationship** between displ and hwy.

6. Scatterplot Visualization

The scatter plot reveals a **clear negative linear relationship** between engine displacement and highway MPG. This means that as engine displacement increases, highway MPG tends to decrease. This is likely due to the fact that larger engines consume more fuel, leading to lower fuel efficiency.



7. Multiple Regression

```
Call:
lm(formula = hwy ~ displ + cty + year, data = mpg)

Residuals:
    Min      1Q  Median      3Q     Max
-5.0465 -1.2159 -0.0645  1.1809  4.1827

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -104.86157   51.60443  -2.032   0.0433 *
displ         -0.09476    0.14980  -0.633   0.5276
cty            1.31658    0.04501  29.253   <2e-16 ***
year           0.05312    0.02585   2.055   0.0410 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.744 on 230 degrees of freedom
Multiple R-squared:  0.9153,    Adjusted R-squared:  0.9142
F-statistic:   829 on 3 and 230 DF,  p-value: < 2.2e-16
```

Coefficients Interpretations:

***Intercept*:** -104.862. This represents the predicted highway MPG when displacement, city MPG, and year are all 0, which is unlikely in a real-world scenario.

Displacement: -0.095. This indicates that for every 1-unit increase in displacement, highway MPG is predicted to decrease by 0.095 units, holding other variables constant.

***City MPG*:** 1.317. This suggests that for every 1-unit increase in city MPG, highway MPG is predicted to increase by 1.317 units, holding other variables constant.

***Year*:** 0.053. This implies that for each year, highway MPG is predicted to increase by 0.053 units, keeping other variables constant.

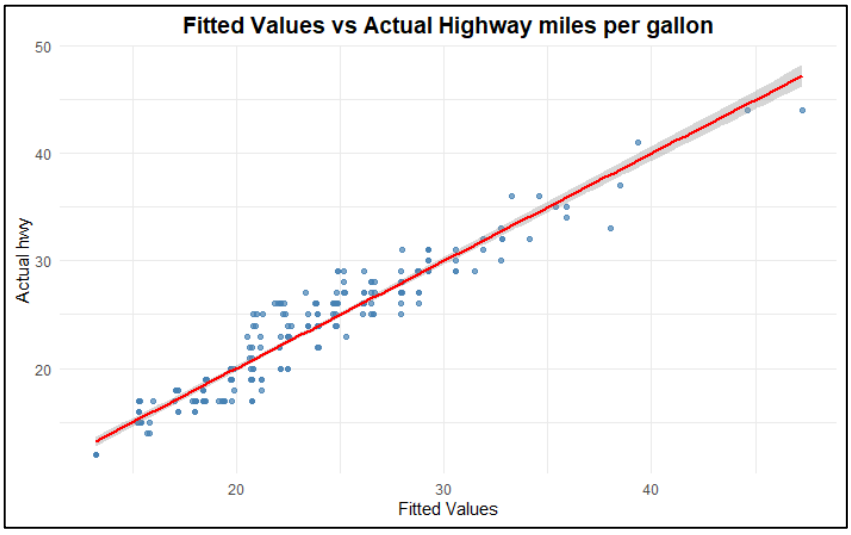
Key Insights:

Displacement: Larger engines generally lead to lower fuel efficiency.

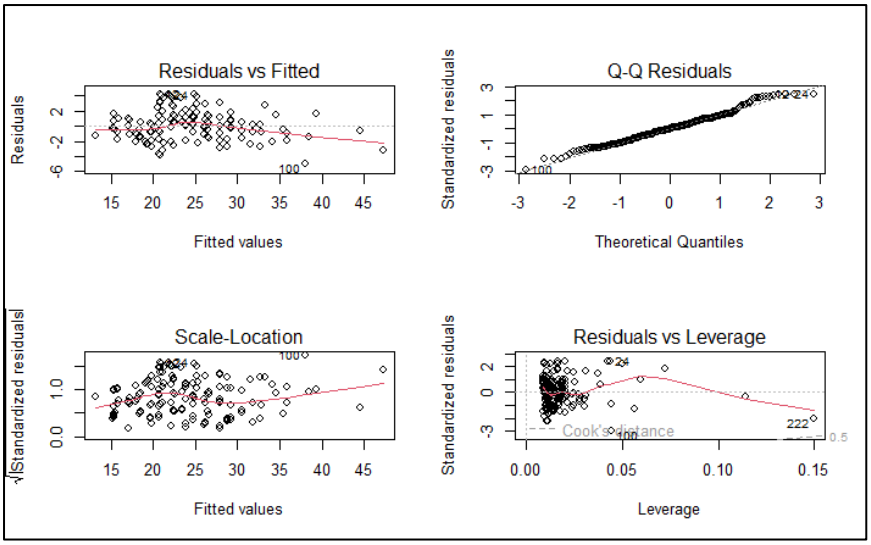
City MPG: Cars with better city fuel efficiency tend to have better highway fuel efficiency as well.

Year: Newer cars tend to have better fuel efficiency, likely due to technological advancements.

Overall, the model suggests that fuel efficiency is influenced by engine size, city fuel efficiency, and the age of the car.



8. Model Diagnostics:

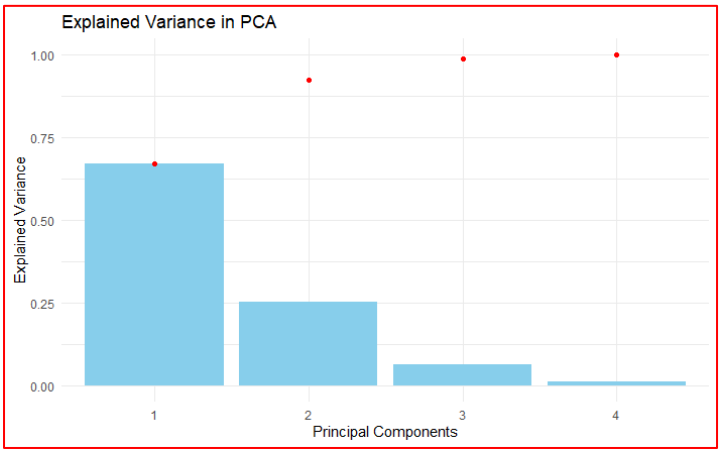


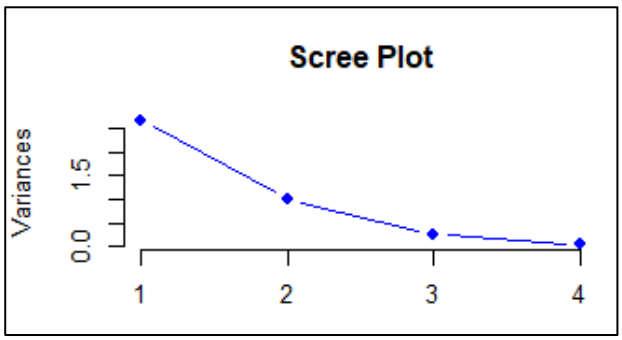
Homoscedasticity Check: Residuals vs Fitted: The plot shows a slight upward trend, suggesting that the variance of residuals might increase with fitted values. This indicates potential heteroscedasticity.

Normality of Residuals Check: Q-Q Plot: The plot shows that the points deviate from the straight line, especially in the tails. This suggests that the residuals might not be normally distributed.

There are a few points with high leverage, indicated by the points on the right side of the plot. These points might be influencing the model's fit. Overall, the diagnostic plots indicate that the model might not be a perfect fit for the data.

9. Principal Component Analysis:



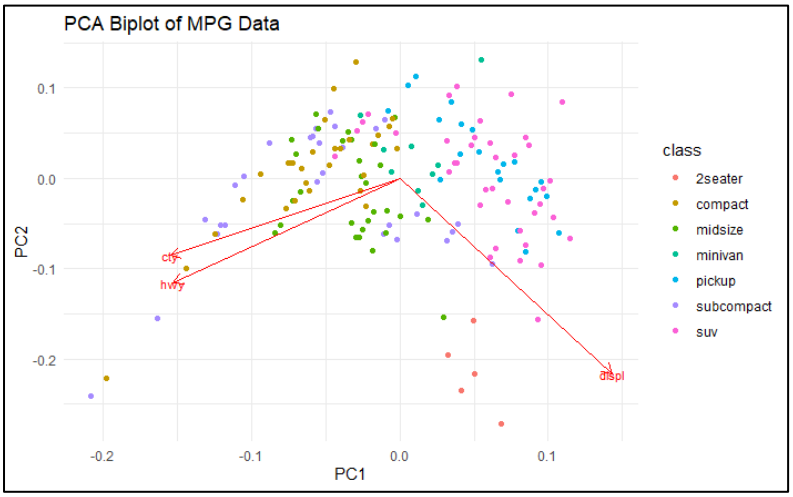


Number of components chosen = 3. Around 98%+ of the variability seems to be explained at around 3 PCs.

10. PCA Interpretations:

Variable Relationships:

- displ (engine displacement) is negatively correlated with cty (city MPG) and hwy (highway MPG).
- cty and hwy are strongly positively correlated (arrows point in the same direction).



Vehicle Class Groupings:

- Compact, subcompact, and midsize cars cluster near higher cty and hwy values, indicating better fuel efficiency.
- SUVs and pickups are closer to displ, reflecting larger engine sizes and lower fuel efficiency.

Principal Component Influence:
PC1 differentiates vehicles by engine size (displ) and efficiency (cty, hwy).
PC2 shows secondary variation, likely influenced by additional factors like vehicle weight or design.

Statistical Structures in Data — Prashant Bisht
Prof. Subhajit Dutta — 24BM6JP42

Dataset 4: Gapminder

1. Overview of dataset:

country	continent	year	lifeExp	pop	gdpPercap
Afghanistan	Asia	1952	28.8	8425333	779
Afghanistan	Asia	1957	30.3	9240934	821
Afghanistan	Asia	1962	32	10267083	853
Afghanistan	Asia	1967	34	11537966	836
Afghanistan	Asia	1972	36.1	13079460	740
Afghanistan	Asia	1977	38.4	14880372	786

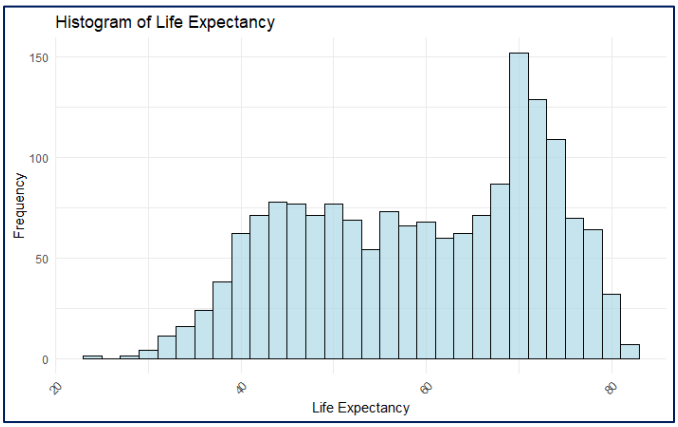
Number of observations: 1704 Number of Variables: 6

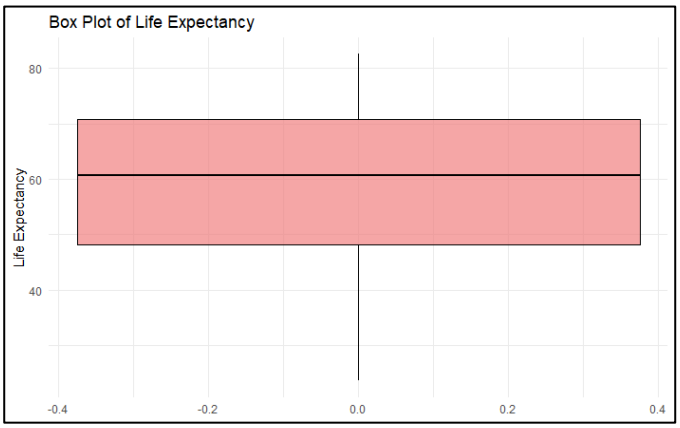
2. Summary Statistics for key variable: lifeExp (Life Expectancy)

Mean = 59.47444 **Median** = 60.712 **Standard depth** = 12.917 **Minimum** = 23.59 **Maximum** = 82.6

Overall, the results suggest that life expectancy varies significantly across different countries. There are countries with very low life expectancies and countries with very high life expectancies.

3. Distribution Visualization: lifeExp (Life Expectancy)

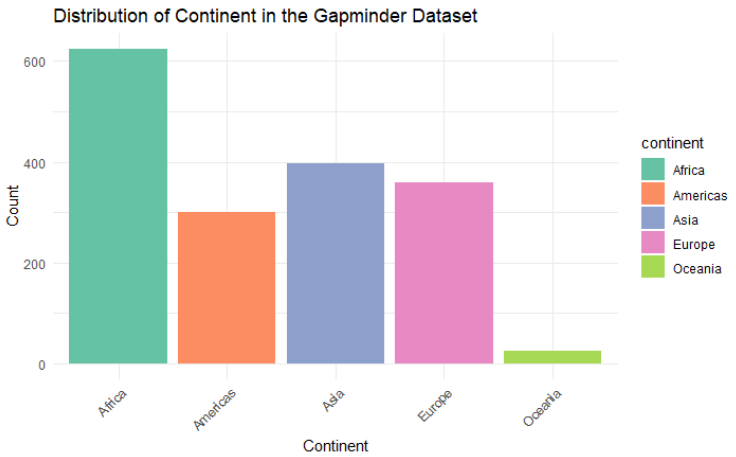




The distribution of life expectancy is **right-skewed** with a **peak** around 70 years. There is a **long tail** towards lower life expectancy values, suggesting the presence of some **outliers** with lower life expectancies.

The distribution of life expectancy appears to be **roughly symmetric** with a **central tendency** around 60 years. There is a **single outlier** with a life expectancy value of approximately 80 years, indicated by the vertical line extending above the box plot.

4. Categorical Variable Analysis:



Africa has the highest number of countries, followed by **Americas** and **Asia**. **Oceania** has the fewest number of countries. This visualization suggests that the dataset has a higher representation of countries from Africa and the Americas compared to other continents.

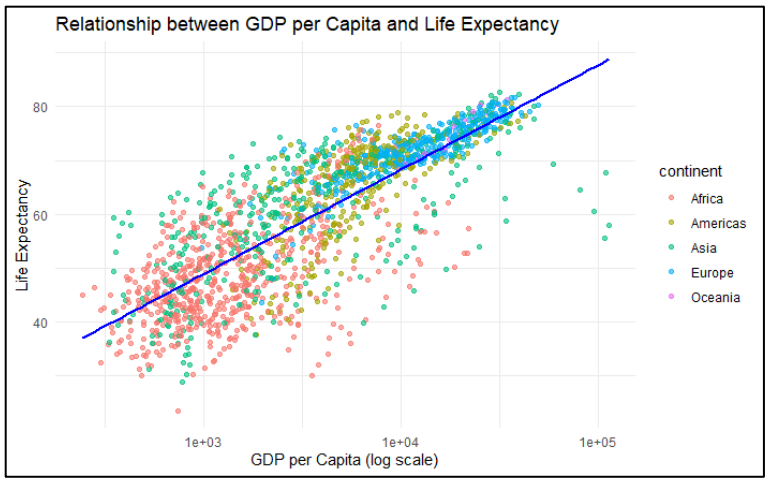
Multivariate Analysis

5. Correlation Analysis: Pearson Correlation Coefficient between gdpPercap, lifeExp: 0.584

***Summary*:** The Pearson correlation suggests a **positive relationship** between gdpPercap and lifeExp

6. Scatterplot Visualization:

The scatter plot reveals a **clear positive linear relationship** between GDP per capita and life expectancy. This means that as GDP per capita increases, life expectancy tends to increase as well. This is likely due to the fact that higher GDP per capita often leads to better healthcare, education, and overall living standards, which contribute to increased life expectancy.



7. Multiple Regression:

```
Call:
lm(formula = lifeExp ~ gdpPercap + pop, data = gapminder)

Residuals:
    Min      1Q  Median      3Q     Max
-82.754  -7.745   2.055   8.212  18.534

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.365e+01  3.225e-01  166.36  < 2e-16 ***
gdpPercap   7.676e-04  2.568e-05   29.89  < 2e-16 ***
pop         9.728e-09  2.385e-09    4.08 4.72e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.44 on 1701 degrees of freedom
Multiple R-squared:  0.3471,    Adjusted R-squared:  0.3463
F-statistic: 452.2 on 2 and 1701 DF,  p-value: < 2.2e-16
```

Coefficients Interpretations:

Intercept: 53.65. This represents the predicted life expectancy when GDP per capita and population are both 0, which is unlikely in a real-world scenario.

GDP per capita: 7.676e-04. This indicates that for every 1-unit increase in GDP per capita, life expectancy is predicted to increase by 0.0007676 units, holding population constant.

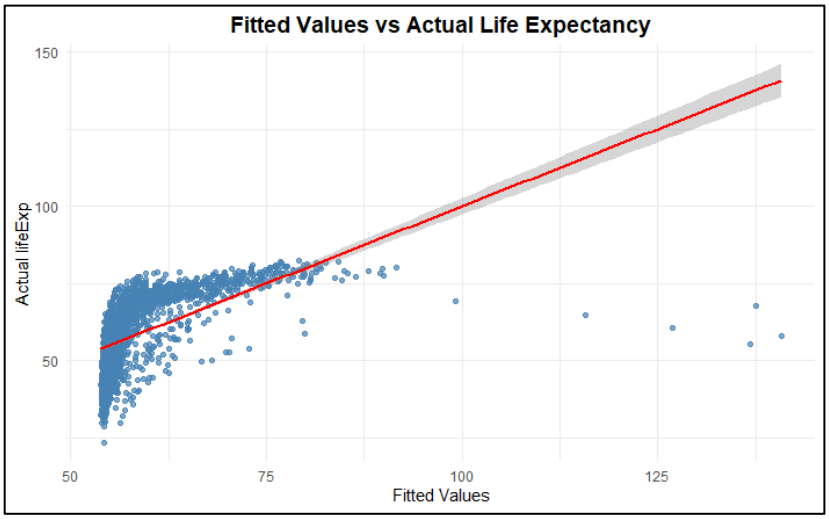
Population: 9.728e-09. This suggests that for every 1-unit increase in population, life expectancy is predicted to increase by 9.728e-09 units, holding GDP per capita constant.

Key Insights:

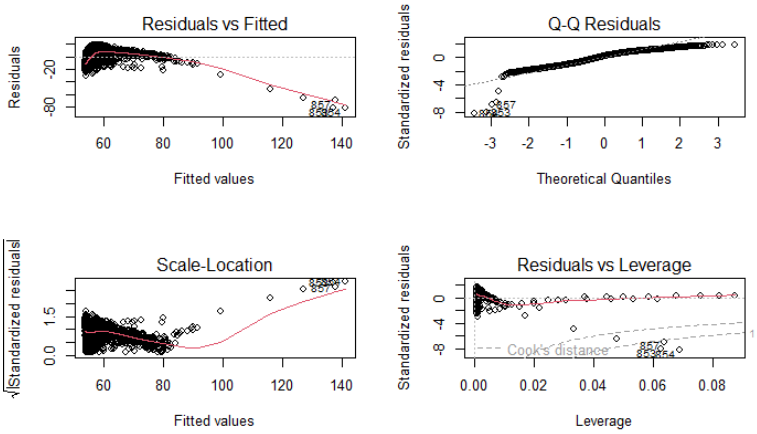
The **Adjusted R-squared** is 0.3463. This means that approximately 34.63% of the variation in life expectancy can be explained by the combined effects of GDP per capita and population.

While the model is statistically significant, as indicated by the p-value, it explains only a moderate portion of the variation in life expectancy.

Overall, the model suggests that GDP per capita is a strong predictor of life expectancy. Population has a negligible impact on life expectancy in this model.



8. Model Diagnostics:



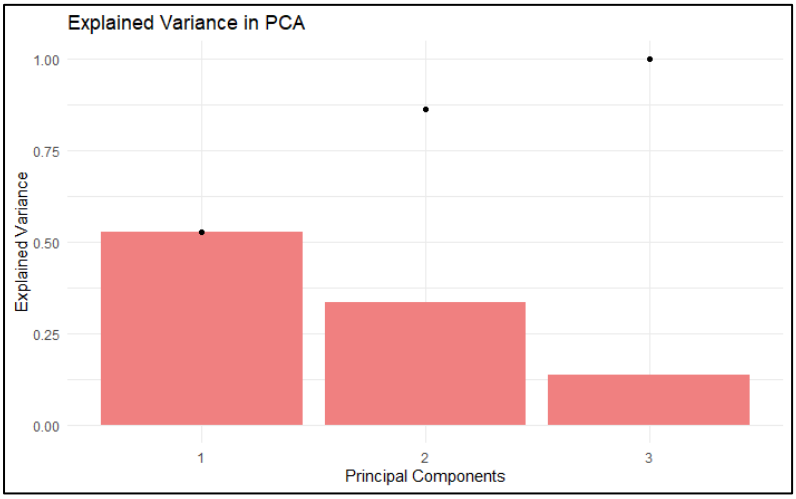
[Homoscedasticity Check:](#) **Residuals vs Fitted:** The residuals show a clear pattern and fan shape, suggesting **non-constant variance** (heteroscedasticity). This indicates that the assumption of homoscedasticity is violated.

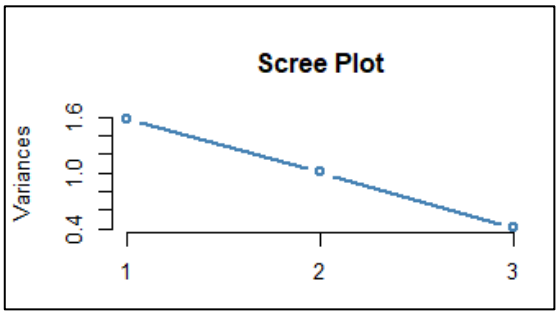
[Normality Check:](#) Q-Q Plot
The residuals deviate significantly from the straight line, especially at the tails, indicating that the residuals are not normally distributed.

[Scale-Location Plot](#)**:** The red line shows an upward trend, and the spread of the square root of standardized residuals is not uniform. This reinforces the finding of **heteroscedasticity**.

[Residuals vs Leverage](#)**:** Some points lie near or beyond the Cook's distance lines, suggesting the presence of **influential data points** that may have a strong impact on the model fit.

9. Principal Component Analysis:

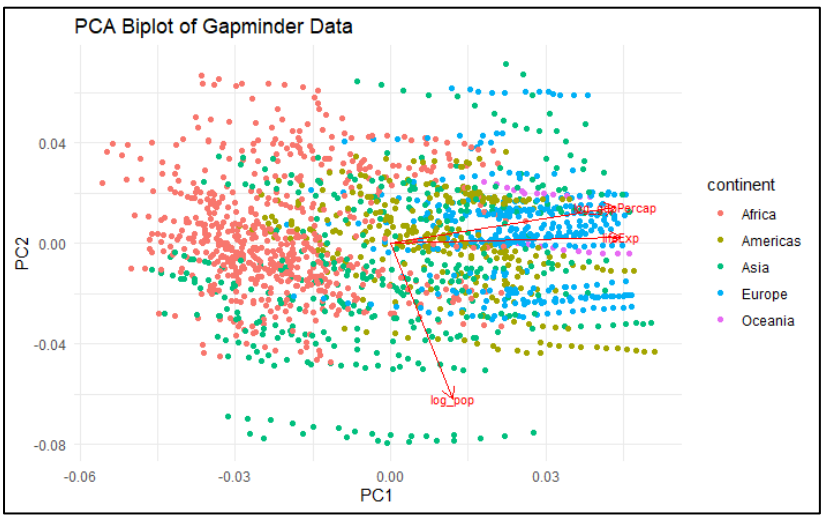




Number of components chosen = 3. Around 98%+ of the variability seems to be explained at around 3 PCs.

10. PCA Interpretations: Continental Clusters:

- Africa (red): Concentrated bottom-left, likely lower development.
- Europe (blue): Top-right, representing higher development.
- Other continents (Asia, Americas, Oceania) are spread, showing varying patterns.



[PC1 and PC2:](#)
PC1 (horizontal): Reflects development (e.g., GDP per capita, life expectancy).
PC2 (vertical): Likely influenced by population size.

[Variable Relationships:](#)
Life expectancy (LifeExp) and GDP per capita (GDPpcap) positively correlate (arrows align with PC1).
Log(population): Associated with PC2, weaker correlation with others.

Outliers: Distant points from the origin represent countries with extreme values.