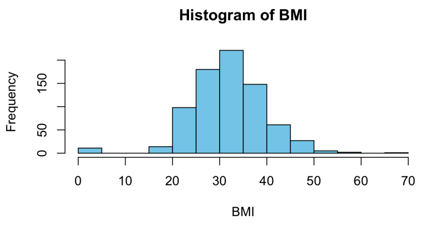
Indian Statistical Institute, Kolkata
PGDBA 2024-2026

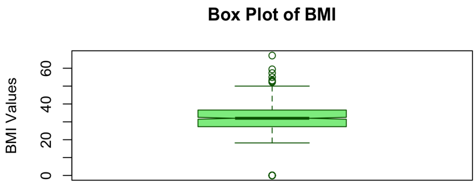
Statistical Structures in Data
Numerical Assignment

Submitted by
Dabbikar Likhith
24BM6JP16
PGDBA 2024-26

Dataset 1 (Diabetes Dataset)

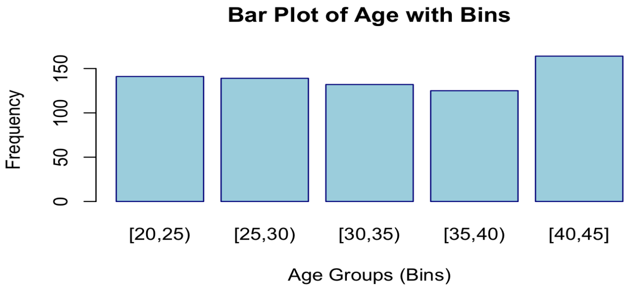
1. **Overview of the data**
    - The Data that is taken from Diabetes data. The structure of the data is 768 x 9. There are 768 observations and 9 variables.
2. **Summary Statistics**
    - I have chosen a numerical variable 'BMI' which is Body mass index. Mean of the variable BMI is 31.99258. Median is 32 , standard deviation is 7.88416 , minimum value of the variable is 0, maximum value of the variable is 67.1.
3. **Distribution Visualisation**





- **Histogram Observations**: The histogram shows a unimodal distribution with a slight positive skew, with most BMI values concentrated in the 30–40 range.
- **Box Plot Observations**: The box plot highlights outliers above 50 BMI and below 10 BMI, indicating extreme values.
- **Data Distribution**: The data is approximately normal but has thin tails and some noticeable outliers.
- **General Insight**: Most BMI values are within a healthy range, with the outliers warranting further investigation for validity.

4. **Categorical variable analysis:**
    - The variable chosen for the categorical analysis is Age of the persons



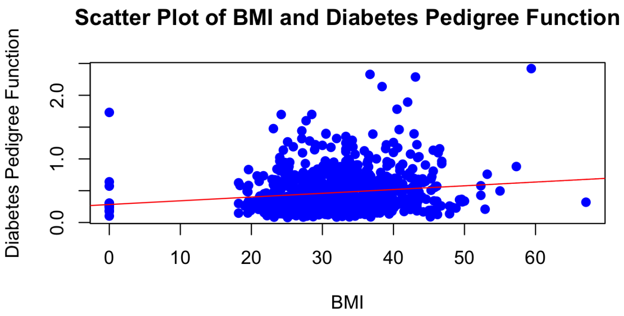
-
- **Uniform Distribution**: The age distribution is relatively uniform across bins, indicating balanced representation among age groups.
- **Age Range**: The dataset spans ages 20 to 45, divided into equal-width bins of 5 years.
- **Balanced Groups**: No specific age group dominates in frequency, supporting balanced analysis.
- **Higher Age Bin Frequency**: A slight increase in the 40–45 age group suggests more individuals in the older demographic.

Multivariate Analysis

5. Correlation Analysis:

- **Selected Variables**: BMI measures weight relative to height, while Diabetes Pedigree Function indicates hereditary diabetes risk.
- **Correlation Coefficient**: The Pearson correlation coefficient between BMI and Diabetes Pedigree Function is 0.1406.
- **Low Correlation**: The low coefficient suggests a weak or negligible relationship between the two variables.
- **Independent Changes**: Variations in one variable are not strongly associated with changes in the other.

6. Scatter Plot Visualization:

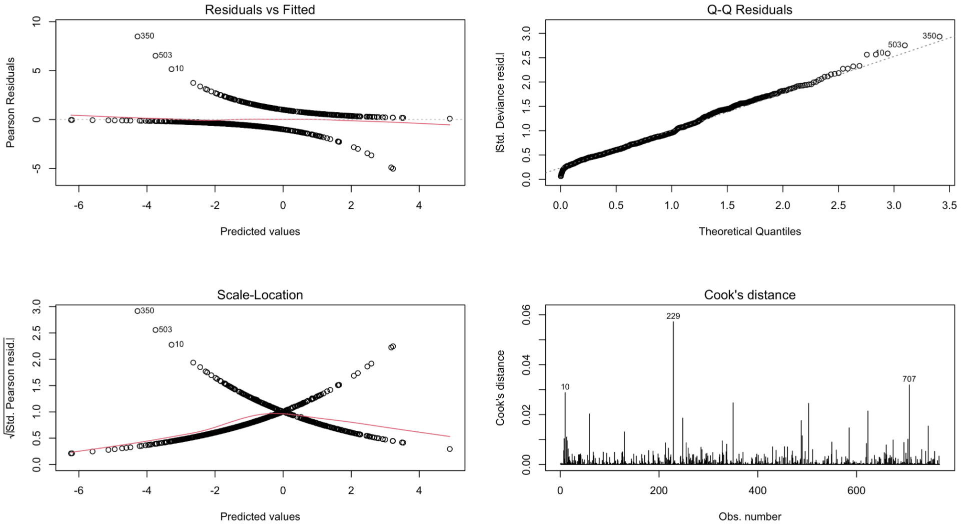


- The scatter plot indicates a weak positive relationship between BMI and Diabetes Pedigree Function, showing minimal direct correlation between body weight and genetic diabetes risk.
- Significant variability in Diabetes Pedigree Function values for a given BMI suggests that other factors besides BMI influence genetic predisposition to diabetes.

7. **Summary of the Logistic Regression Model**

- **Model Overview**: A multiple logistic regression model was used to predict diabetes (binary outcome: 1 = Diabetes, 0 = No Diabetes) with key predictors including Glucose, BMI, Pregnancies, and DiabetesPedigreeFunction.
- **Significant Predictors**: Glucose and BMI are the strongest predictors, with Pregnancies and DiabetesPedigreeFunction also showing significant positive effects. BloodPressure has a small but significant negative effect.
- **Non-Significant Predictors**: SkinThickness, Insulin, and Age were found to have no significant contribution to predicting diabetes in this model.
- **Model Fit and Performance**: The reduction in deviance (993.48 to 723.45) indicates substantial improvement with predictors, and the AIC of 741.45 suggests reasonable model performance. The model achieved an accuracy of 78.26%.
- **Coefficient Interpretation**: Each unit increase in Glucose and BMI significantly increases the likelihood of diabetes, while other predictors like SkinThickness and Insulin contribute minimally.
- **Practical Implications**: Prioritize Glucose and BMI in diabetes prevention and intervention strategies, while less impactful variables may require further study or exclusion in future models.
-

8. Model Diagnostics:

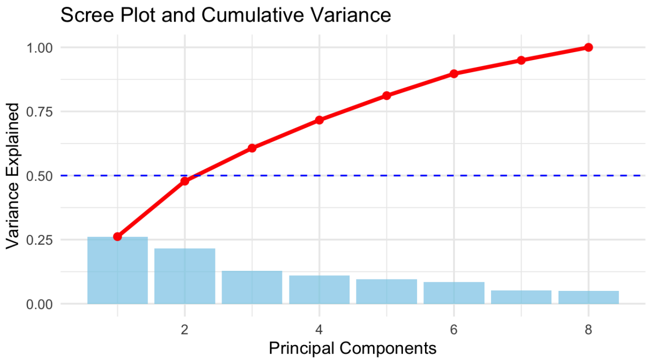


1. **Residuals vs Fitted Plot**: Residuals show a curved pattern, indicating heteroscedasticity and potential non-linear effects not captured by the model.
2. **Normal Q-Q Plot**: Residuals largely follow the diagonal line but deviate at the tails, suggesting slight non-normality due to outliers.

3. **Scale-Location Plot**: A curved trend highlights non-constant variance, violating homoscedasticity assumptions and suggesting the need for variable transformations.
4. **Cook's Distance Plot**: Observations 10, 229, and 707 have high Cook's distances, identifying them as influential points requiring review.
5. **Homoscedasticity**: Residual plots indicate heteroscedasticity, which could be addressed through transformations or weighted regression.
6. **Model Improvements**: Normality deviations and influential points suggest that handling outliers and re-evaluating model assumptions could enhance robustness.

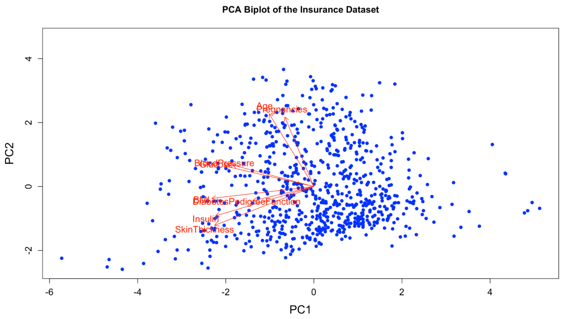
Advanced Analysis

9. Principal Component Analysis:



- **Observation**: The scree plot shows an elbow point at 2 components, where the explained variance begins to level off, indicating that the first 2 components capture most of the meaningful variance.
- **Elbow Method**: The elbow point reflects the optimal balance between variance explained and model simplicity. Beyond 2 components, additional components contribute minimal variance.
- **Variance Explained**: The first 2 components explain a substantial portion of the data's variance, making them sufficient for dimensionality reduction.
- **Conclusion**: Based on the scree plot, 2 components should be chosen to retain most of the data's structure while minimizing dimensionality.

10. Plotting the biplot:

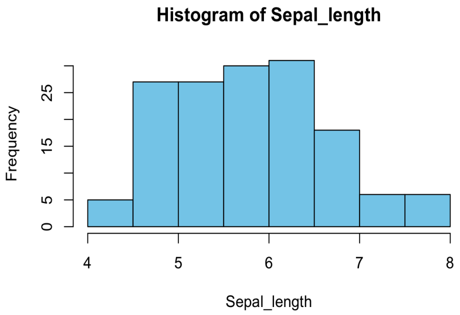


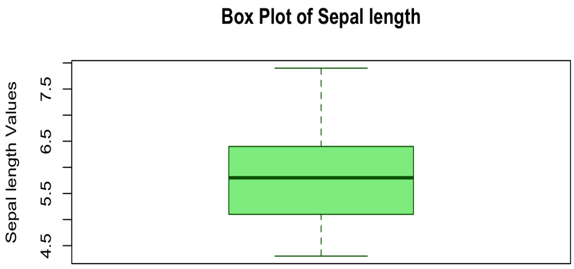
1. **Key Contributors to PC1**: Variables like BMI, BloodPressure, and DiabetesPedigreeFunction significantly influence the first principal component.
2. **Influence of Age and Pregnancies**: Age and Pregnancies are aligned and contribute to both PC1 and PC2, showing their joint importance in variability.
3. **Correlation Between Variables**: Insulin and SkinThickness are closely grouped, indicating a strong correlation and similar behavior.
4. **Data Distribution**: Data points are mainly spread along PC1, showing it captures most variability, with PC2 adding some differentiation.

Dataset 2 (IRIS Dataset)

1. **Overview of the data**
    - ○ The Data that is taken from iris data. The structure of the data is 150 x 5 There are 150 observations and 5 variables.

2. **Summary Statistics**
    - ○ I have chosen a numerical variable ‘sepal_length’ which represents the length of a flower's sepal, measured in centimeters. Mean of the variable ‘sepal_length is 5.8433. Median is 5.8, standard deviation is 0.8281 , minimum value of the variable is 4.3, maximum value of the variable is 7.9.
3. **Distribution Visualisation:**



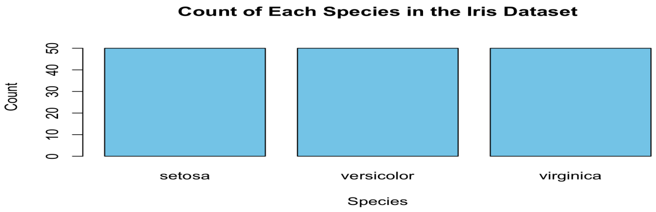


- ○ **Histogram Analysis**: The distribution of Sepal Length is nearly symmetric with a slight right skew. Most values fall between 5 cm and 6.5 cm, with fewer values at the extremes.
- ○ **Box Plot Analysis**: The box plot shows no visible outliers, as all points are within the whiskers. The median lies near the center of the IQR, confirming a balanced and symmetric dataset for Sepal Length.

Summary:

- The **distribution of Sepal Length** is roughly symmetric and does not show significant outliers.
- The data is evenly spread, with no extreme deviations from the central tendency.

4. **Categorical variable Analysis:**
    - ○ The variable chosen for the categorical analysis is Species



○

Insights from the Bar Plot of Species in the Iris Dataset:

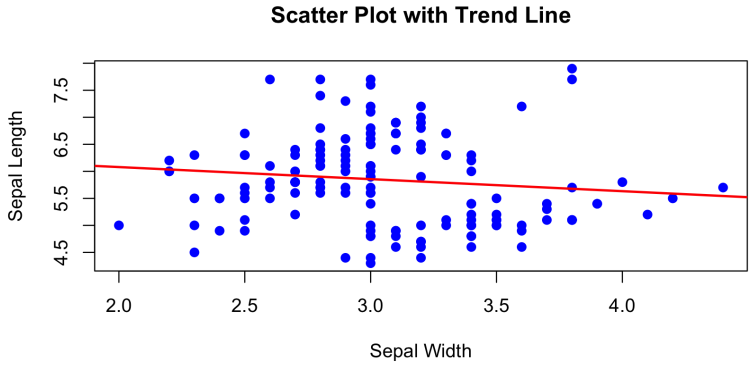
1. **Balanced Dataset**:
    - ○ The bar plot indicates that each species (*setosa*, *versicolor*, and *virginica*) has an equal count of **50 observations**.
    - ○ This ensures the dataset is balanced and suitable for classification tasks without concerns of class imbalance.
2. **Uniform Distribution**:
    - ○ All species are equally represented, making the dataset ideal for unbiased analysis and modeling.
3. **Applicability for Machine Learning**:
    - ○ The balanced distribution across species ensures that classification models trained on this dataset will not be skewed or biased towards a particular class.

Multivariate Analysis

5. Correlation Analysis:

- The Numerical variables that I have selected are Sepal length which represents the length of a flower's sepal, measured in centimeters and **Sepal width** which represents the width of a flower's sepal in centimeters.
- The Pearson correlation coefficient between the variables **Sepal Length** and **Sepal Width** is **-0.1175698**, which signifies a weak negative correlation. This indicates that changes in one variable are not strongly associated with consistent changes in the other, and there is a slight inverse relationship between them.

6. Scatter Plot visualisation:

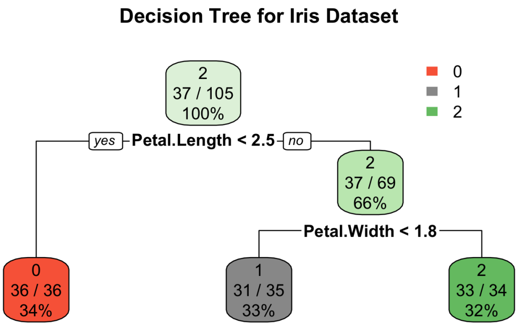


Inference from the Scatter Plot

- The scatter plot indicates a **weak negative relationship** between Sepal Width and Sepal Length, as shown by the downward slope of the trend line. This is consistent with the Pearson correlation coefficient of **-0.1175698**, suggesting that an increase in Sepal Width is associated with a slight decrease in Sepal Length. However, the relationship is not strong, as the points are widely scattered around the trend line.
- The distribution of points appears evenly spread across the range of Sepal Width, with no significant clusters or outliers.

7. **Implementing a Decision tree algorithm**

- Implementing a Decision tree algorithm because the dataset chosen consists of classification variable

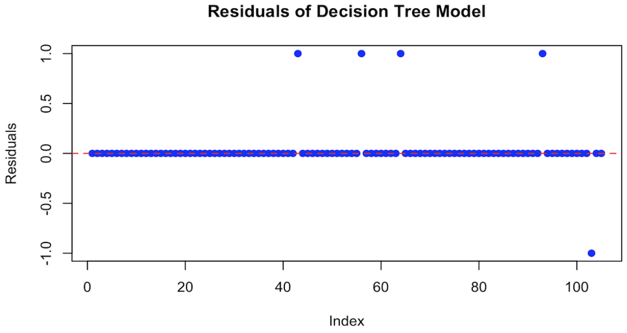


-
  -

Accuracy of the model = 97.28

- **Key Variables**: Petal.Width (34%) and Petal.Length (32%) are the most important features for classification.
- **Splits**: The root node splits on Petal.Length < 2.45 (classifying Setosa), and the right node splits on Petal.Width < 1.75 (distinguishing Versicolor and Virginica).
- **Accuracy**: Terminal nodes predict species with high probabilities, such as 97.1% for Virginica.
- **Model Efficiency**: Two splits reduce the error significantly (from 1.0 to 0.0735), with petal-related features driving the classification.

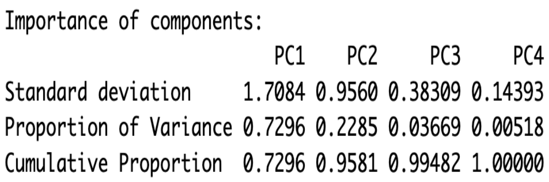
8. **Model Diagnostics:**



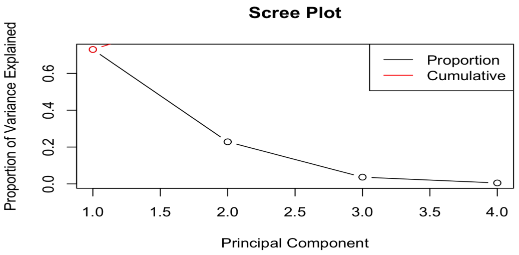
- Residuals are mostly concentrated around zero, showing the decision tree captures most patterns in the data.
- A few higher residuals suggest misclassified instances where the model struggles to separate classes.
- The lack of a clear pattern indicates no significant overfitting or underfitting, but further refinement may help address misclassifications.

Advanced Analysis

9. PCA plot

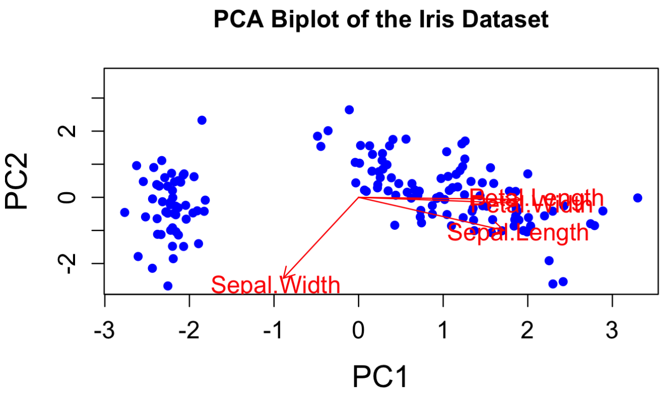


```
Importance of components:
                          PC1    PC2     PC3     PC4
Standard deviation     1.7084 0.9560 0.38309 0.14393
Proportion of Variance 0.7296 0.2285 0.03669 0.00518
Cumulative Proportion  0.7296 0.9581 0.99482 1.00000
```



- Based on the scree plot, the elbow method suggests selecting 3 principal components.
- The first three components explain a significant portion of the variance in the dataset.
- Including more components adds little additional explanatory power while increasing complexity.
- The "elbow" is clearly visible at the 3rd component, where the slope levels off.
- Choosing 3 components ensures a balance between simplicity and retaining variance.

10. **PCA Interpretation:**



Loadings of PC1 and PC2:

- **Petal.Length** and **Petal.Width** have strong positive loadings on PC1, indicating they contribute significantly to the variance along this component.
- **Sepal.Width** has a negative loading on PC1, showing an inverse relationship with Petal variables.
- **Sepal.Length** contributes moderately to both PC1 and PC2.

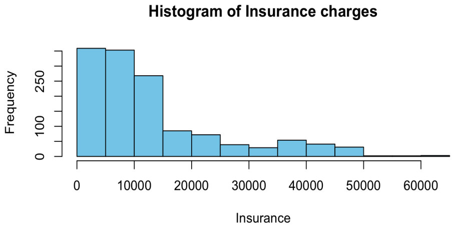
Patterns and Groupings:

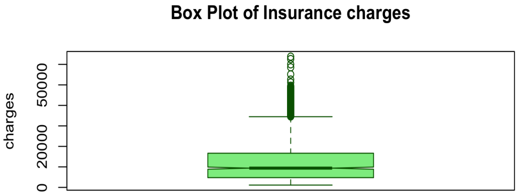
- Data points are well-separated along PC1, primarily distinguishing species based on Petal features.
- Groupings can be observed, where one cluster represents **Setosa** (on the left) and the others likely correspond to **versicolor** and **virginica.**
- Petal measurements play a dominant role in separating the species, while Sepal measurements provide secondary variation.

Dataset 3 (Insurance Dataset)

1. **Overview of the data**
   - ○ The Data that is taken from the insurance dataset. The structure of the data is 1338 x 7 There are 1338 observations and 7 variables.

2. **Summary Statistics**
   - ○ I have chosen a numerical variable '**charges**' which signifies the insurance charges. Mean of the variable charges is 13270.42. Median is 9382.033, standard deviation is 12110.01 , minimum value of the variable is 1121.874, maximum value of the variable is 63770.43.
3. **Distribution Visualisation**:

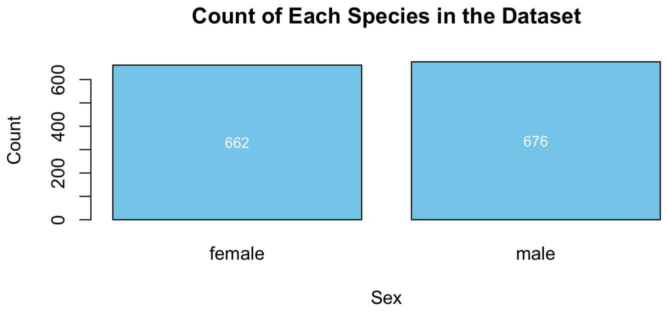




Description of the Distribution and Outliers:

1. The histogram shows a right-skewed distribution, with most insurance charges below 20,000 and fewer individuals with significantly higher charges.
2. The box plot reveals several outliers in the upper range, likely linked to factors like high BMI, older age, or smoking.
3. These outliers warrant further investigation to determine their causes.

5. **Categorical variable Analysis:**
   - ○ The variable chosen for the categorical analysis is Sex of the Individuals



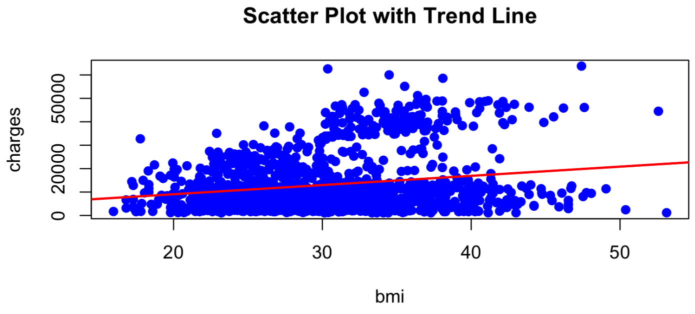
- ○
   - ○ The dataset has 662 females and 676 males.
   - ○ The gender distribution is nearly equal, making the data balanced for analysis.

Multivariate Analysis

5. Correlation Analysis:

- The Numerical variables that I have selected are charges (insurance charges) relative to **bmi** to indicate if an individual has a healthy weight relative to height
- The Pearson correlation coefficient between the variables that is observed is 0.198341 which signifies that the variables are highly uncorrelated.
- Changes in one variable are not strongly associated with consistent changes in the other.
- Changes in one variable are not strongly associated with consistent changes in the other.

6. **Scatter Plot visualisation:**



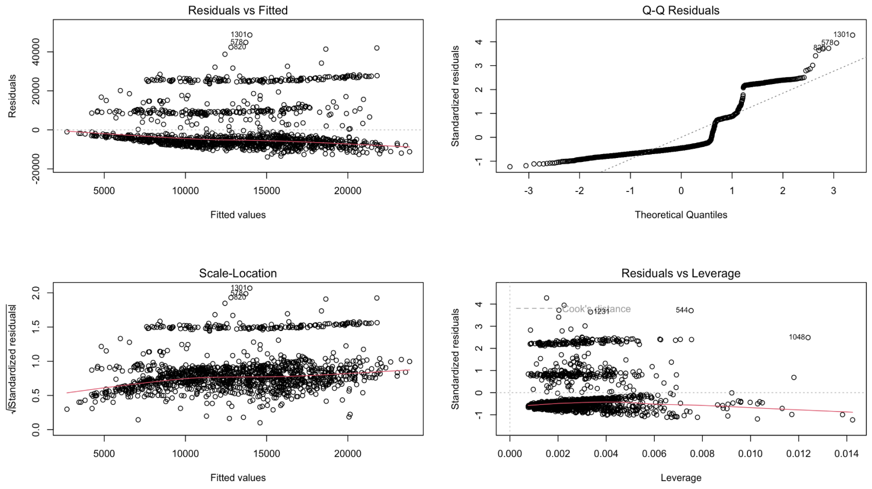
Inferences from the Scatter Plot:

- There is a **positive relationship** between BMI and insurance charges, as indicated by the trend line. This suggests that individuals with higher BMI tend to have higher insurance charges.
- The spread of points increases with BMI, showing **greater variability in charges** for individuals with higher BMI values. This may indicate that other factors, like age or smoking status, also contribute significantly to charges.

7. **Implementing a Multiple linear regression model:**

- A multiple linear regression predicts insurance charges using age, BMI, and number of children.
- **Age**: Each additional year increases charges by **239.99 units**.
  BMI: Each unit increase adds **332.08 units**.
  Children: Each additional child adds **542.86 units**, with a smaller effect compared to age and BMI.
- The model is statistically significant but explains only **12% of the variation** (R-squared = 0.12).
- Other factors like smoking or region may improve the model's predictive power.

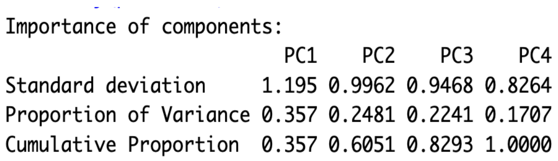
8. Model Diagnostics:



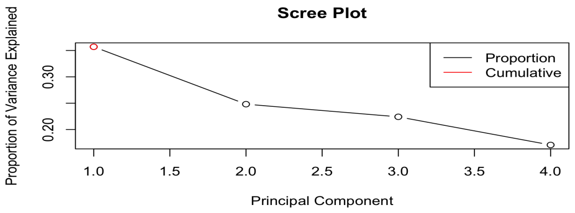
1. **Residuals vs Fitted Plot**: Residuals show a pattern, indicating non-linearity or missing variables, with increasing variance suggesting heteroscedasticity.
2. **Q-Q Plot**: Residuals deviate from normality at the tails, with outliers like point 1301.
3. **Scale-Location Plot**: A clear pattern in standardized residuals confirms heteroscedasticity.
4. **Residuals vs Leverage Plot**: High-leverage points (e.g., 1301, 1048, 544) are influential and require investigation.

Advanced Analysis

9. PCA plot



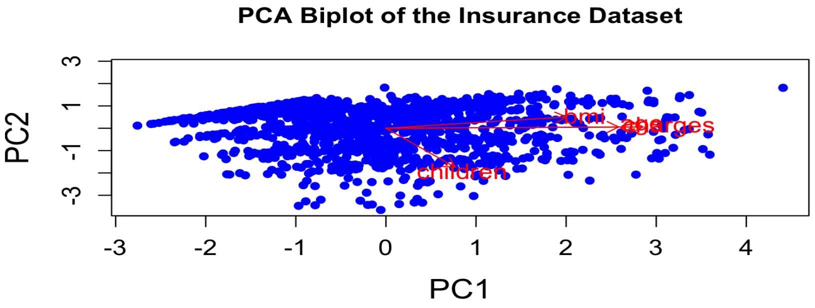
```
Importance of components:
                         PC1    PC2    PC3    PC4
Standard deviation     1.195 0.9962 0.9468 0.8264
Proportion of Variance 0.357 0.2481 0.2241 0.1707
Cumulative Proportion  0.357 0.6051 0.8293 1.0000
```



Based on the scree plot, the elbow method suggests selecting **2 principal components**.

- The first two components explain a significant portion of the variance in the dataset.
- Including more components beyond the second adds little additional explanatory power while increasing complexity.
- The "elbow" is clearly visible at the 2nd component, where the slope begins to level off.
- Choosing **2 components** ensures a balance between simplicity and retaining the maximum variance.

10. PCA Interpretation:

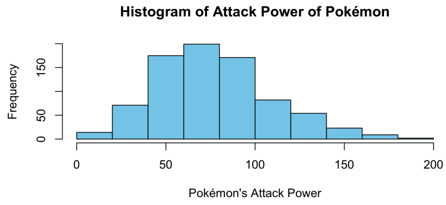


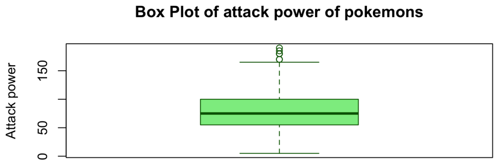
- **Main Variables**: **`charges`** and `bmi` contribute most to PC1, explaining the majority of the variance.
- **Smaller Contribution**: **`children`** contribute less and have a weaker relationship with other variables.
- **Positive Relationships**: **`charges, bmi, and age`** show positive correlations.
- **Data Spread**: Most points cluster near the origin, with some variability across the dataset.
- **Key Insight**: **`charges and bmi`** are the most significant variables in the dataset..

Dataset 4 (Pokemon Dataset)

1. Overview of the data
   - ○ The Pokémon dataset contains information about various Pokémon, including attributes such as name, type, stats (e.g., attack, defense, speed), and generation. It is often used for analysis of strengths, weaknesses, and distributions across different types and categories. The dataset provides valuable insights into patterns in Pokémon characteristics and game design.

2. Summary Statistics
   - ○ I have chosen a numerical variable '**Attack**' which signifies the attack power of the pokemons. Mean of the attack variable is 79.00125. Median is 75, standard deviation is 32.45737 , minimum value of the variable is 5, maximum value of the variable is 190.
3. Distribution Visualisation:





Histogram :

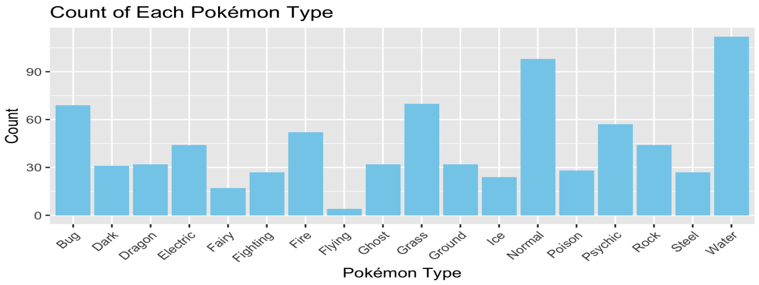
1. The attack power of most Pokémon is concentrated between 50 and 100, forming a peak in the histogram.
2. The distribution of attack power appears roughly symmetric, with fewer Pokémon at the lower and higher ends of the range.
3. The range of attack power spans from 0 to 200, with a gradual decrease in frequency for higher values.

Boxplot:

1. The boxplot shows the median attack power is around 75, with a relatively balanced interquartile range.
2. Outliers are visible above 150, indicating some Pokémon with significantly higher attack power compared to the majority.
3. The whiskers of the boxplot extend from near 0 to just below 150, capturing the majority of the dataset without the outliers.

6. Categorical variable Analysis:

- The variable chosen for the categorical analysis is type of the Pokemon



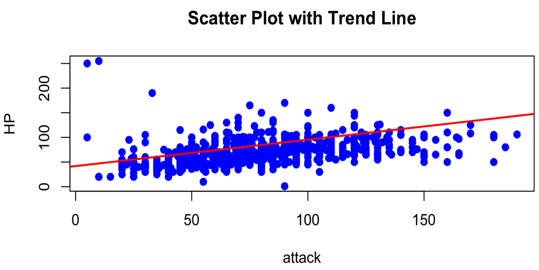
-
    - **Water-type Pokémon** are the most abundant in the dataset, followed closely by **Normal-type Pokémon**, indicating their popularity or prevalence in the Pokémon universe.
    - **Bug-type Pokémon** also show a relatively high count, whereas types like **Ghost** and **Fairy** are significantly less common, reflecting their rarity.
    - There is a diverse distribution of Pokémon types, showcasing variety, but with noticeable imbalances in their representation.

Multivariate Analysis

5. Correlation Analysis:

- The Numerical variables that I have selected are attack (attacking power) relative to **health power of the pokemon**
- The Pearson correlation coefficient between the variables that is observed is 0.42236 which signifies that the variables are weakly correlated.
- Changes in one variable are not strongly associated with consistent changes in the other but as they are positively correlated with each other as one increases the other increases slightly (directly proportional)

6. **Scatter Plot visualisation:**



- The scatter plot shows a positive relationship between attack power and HP, indicating higher attack power is generally associated with higher HP.
- There is a weak to moderate correlation, with a significant spread and many points scattered widely around the trend line.
- A few Pokémon with extremely high HP deviate as outliers from the general trend.

7. Implementation of Multiple linear regression model

Intercept: The baseline Speed is **34.81** when all predictors are zero.

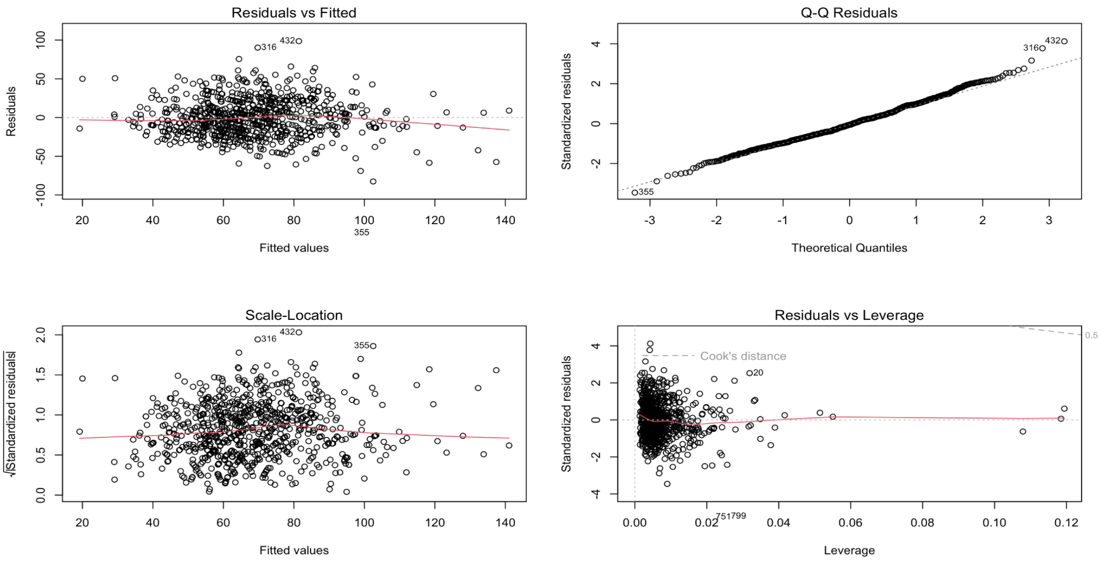
Key Predictors:

- **Attack**: Each unit increase raises Speed by **0.33** (p < 2e-16).
- **HP**: Each unit increase lowers Speed by **0.11** (p = 0.0034).
- **Defense**: Each unit increase lowers Speed by **0.27** (p < 2.87e-14).
- **Special Attack**: Each unit increase raises Speed by **0.30** (p < 2e-16).
- **Special Defense**: Each unit increase raises Speed by **0.18** (p = 1.21e-05).

Model Performance: Explains **32.44%** of Speed variance (R-squared = 0.3244) with a residual error of **23.96**.

Overall Significance: The model is statistically significant (F-statistic = **76.25**, p < 2.2e-16), showing predictors significantly impact Speed.

8. Model Diagnostics:

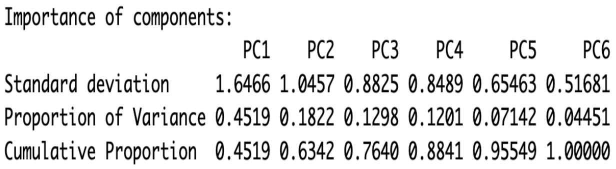


- Residuals vs. Fitted Plot: Residuals are scattered around zero, suggesting no major linearity issues, but slight heteroscedasticity is visible as variance increases with fitted values.
- Q-Q Plot: Residuals mostly follow the 45-degree line, indicating reasonable normality, with a few outliers at the tails.
- Scale-Location Plot: Points are evenly spread, but minor curvature suggests slight heteroscedasticity and non-constant residual variance.

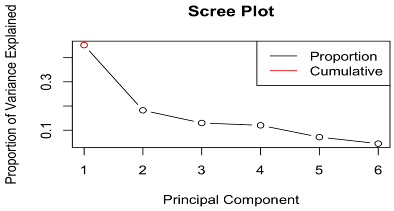
- Residuals vs. Leverage Plot: Most points have low leverage, but a few high-leverage points may influence the model and need further investigation.

Advanced Analysis

9. PCA plot



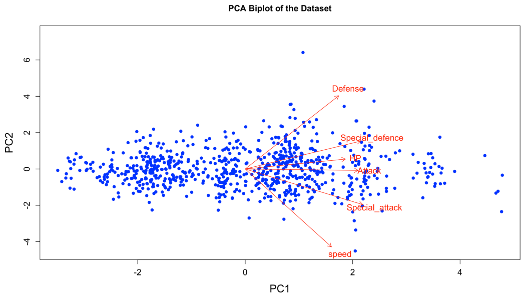
```
Importance of components:
                          PC1    PC2    PC3    PC4     PC5     PC6
Standard deviation     1.6466 1.0457 0.8825 0.8489 0.65463 0.51681
Proportion of Variance 0.4519 0.1822 0.1298 0.1201 0.07142 0.04451
Cumulative Proportion  0.4519 0.6342 0.7640 0.8841 0.95549 1.00000
```



Based on the scree plot, the elbow method suggests selecting **2 principal components**.

- The first two components explain a significant portion of the variance in the dataset.
- Including more components beyond the second adds little additional explanatory power while increasing complexity.
- The "elbow" is clearly visible at the 2nd component, where the slope begins to level off.
- Choosing **2 components** ensures a balance between simplicity and retaining the maximum variance.

10. PCA Interpretation:



- **Principal Component Contributions**: The first principal component (PC1) captures significant variance in variables like **`Defense`, `Special_attack`, and `Special_defence`,** as evident from their strong projections along PC1.
- **Variable Relationships**: `Attack` and HP appear to contribute moderately to both PC1 and PC2, indicating they have balanced influence in explaining variance across these two components.
- **Direction and Length of Arrows**: The length of the arrows indicates the magnitude of variance each variable contributes, with `Defense` showing the longest arrow, suggesting it explains more variance compared to the other variables.
- **Cluster Distributions**: The data points are widely spread along PC1 and PC2, with no distinct clustering patterns visible, indicating a diverse distribution of Pokémon characteristics across the dataset.