INDIAN STATISTICAL INSTITUTE, KOLKATA

POST-GRADUATE DIPLOMA IN BUSINESS ANALYTICS (Batch 10)

Numerical Assignment

Name: Yoogesh Kumar M
Roll No: 24BM6JP61

Table of Contents

Dataset I – Employee Data.

Univariate Analysis

Question 1.

The **employee_data** dataset contains **4,653 observations** and **9 variables**, offering anonymized insights into employee demographics, work history, and employment factors. Key highlights include:

- **JoiningYear**: Employees joined primarily between **2013 and 2017**, showcasing varied service lengths.
- **PaymentTier**: Most employees fall into **Tier 3** (lowest salary tier).
- **Age**: Workforce aged between **22 and 41 years**, reflecting a younger demographic.
- **ExperienceInCurrentDomain**: Employees typically have **2–4 years of experience**, indicating moderate expertise.
- **LeaveOrNot**: Around **34%** left the company, making this a critical **target variable** for retention studies.

This dataset aids HR professionals in analyzing **retention**, **salary structure**, and **workforce diversity**. **Dataset Link**: Employee Dataset

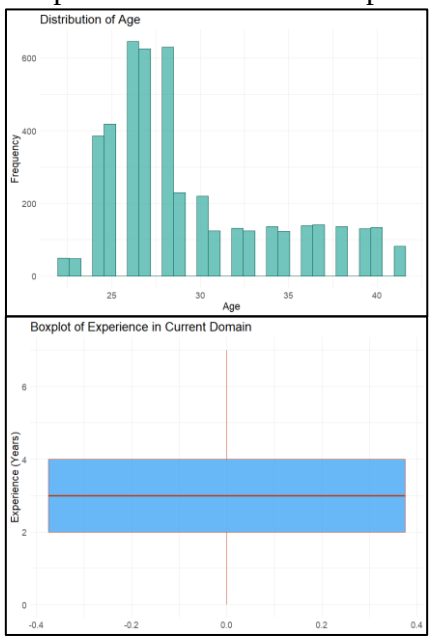
Question 2.

The summary statistics of the **numerical variables** in the dataset provide insights into employee characteristics, including their joining year, payment tier, age, experience, and employment status. Below are the key findings:

- **JoiningYear**: Employees joined primarily between **2013** and **2017**, with a mean and median of **2015**.
- **PaymentTier**: Most employees belong to **Tier 3** (lowest tier), with a mean of **2.70**.
- **Age**: Employee ages range from **22** to **41**, with a mean of **29.39** and a median of **28**.
- **ExperienceInCurrentDomain**: Employees typically have **2 to 4 years of experience**, with a mean of **2.91** and a maximum of **7 years**.
- **LeaveOrNot**: About **34% of employees** have left the company, as indicated by the mean value of **0.34**.

Question 3.

The **age distribution** and **experience in the current domain** provide key insights into the workforce composition and levels of experience.



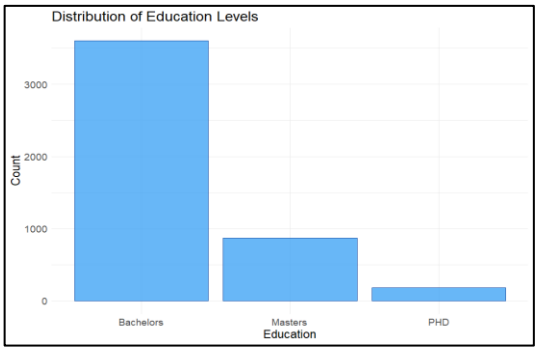
- **Histogram for Age**:
  - The distribution is **right-skewed**, with most employees aged between **25 and 35 years**.
  - The most frequent age group falls within **27–29 years**, indicating a **relatively young workforce**.
  - The distribution is smooth, with no apparent extreme values or outliers, showing a consistent spread of ages.
- **Boxplot for Experience in Current Domain**:
  - The majority of employees have **2 to 4 years** of experience, with a symmetric distribution and a **median of 3 years**.
  - The boxplot confirms the absence of outliers, suggesting **consistent experience levels** across the workforce.

Key Insights:

The analysis highlights a **young workforce** with moderate experience, providing opportunities for growth and development. The absence of outliers in both variables suggests a stable and balanced workforce composition.

Question 4.

The **bar plot for Education** reveals the distribution of employee qualifications, highlighting that the majority hold a **Bachelor's degree**, followed by a smaller proportion with **Master's degrees**, and an even smaller group with **PhDs**. Below are the key insights:



- **Bachelor's Degree**: Over **3,000 employees** hold a Bachelor's degree, making it the most common qualification in the workforce.
- **Master's Degree**: A moderate number of employees have a Master's degree, representing the second-largest group.
- **PhD**: Very few employees have a PhD, indicating limited demand for highly specialized or research-intensive roles.

Multivariate Analysis

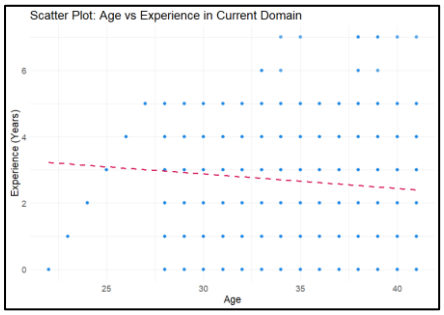
Question 5.

The **correlation analysis** between **ExperienceInCurrentDomain** and **Age** reveals a **weak negative relationship** with a Pearson coefficient of **-0.1346**, suggesting a slight inverse association between these variables.

- **Weak Negative Correlation**: As **Age** increases, **ExperienceInCurrentDomain** tends to decrease slightly.
- **Magnitude of Relationship**: The correlation is minimal, indicating a weak association.
- **Career Variability**: Highlights diverse career paths where older employees may not necessarily have extensive domain-specific experience.

Question 6.

The scatter plot visualizing the relationship between **Age** and **Experience in Current Domain** reveals a **weak negative correlation**, with older employees having slightly less experience in their current domain. Below are the key insights:



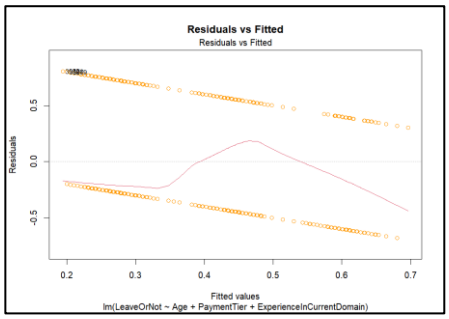
- The trend line indicates a **marginal downward slope**, reflecting a weak negative relationship.
- The variability in experience is consistent across all age groups, suggesting diverse career paths and transitions.
- Age does not strongly predict an employee's experience in their current domain.

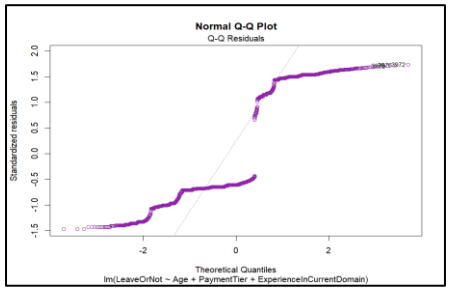
Question 7.

The regression model predicts **LeaveOrNot** based on **Age**, **PaymentTier**, and **ExperienceInCurrentDomain**, identifying key factors influencing employee retention. Below are the insights:

- **Age**: Older employees are slightly less likely to leave, with a small but significant negative impact.
- **PaymentTier**: Employees in higher salary tiers are significantly less likely to leave, making it the most influential predictor.
- **ExperienceInCurrentDomain**: More experience in the current domain slightly reduces the likelihood of leaving, though its impact is weaker.
- **Model Fit**: The model explains **4.27%** of the variability in attrition, highlighting the need for additional predictors to improve accuracy.

Question 8.





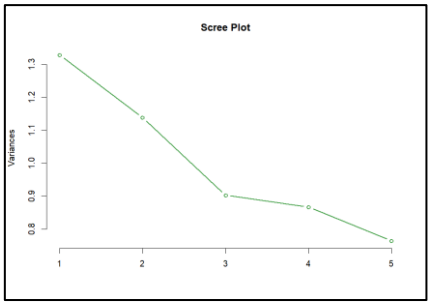
The diagnostic plots for the regression model highlight potential violations of key assumptions, including issues with **homoscedasticity** and **normality of residuals**, indicating areas where the model can be improved. Below are the key insights:

- **Homoscedasticity Violation**: The residuals vs. fitted plot shows non-random patterns, suggesting that residual variance is not constant across fitted values.
- **Non-Normal Residuals**: The Q-Q plot reveals deviations from normality, particularly in the tails, which may impact statistical inference.
- **Model Limitations**: The results suggest the need for variable transformations, additional predictors, or alternative modeling approaches to improve the fit and address these violations.

Created a random Forest with "Accuracy: 86.13 %"

Advanced Analysis

Question 9.

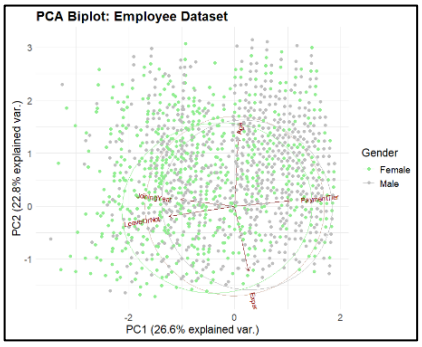


The PCA analysis reveals that the first two principal components capture a significant portion of the variance, making them optimal for dimensionality reduction. Below are the key findings:

- **Explained Variance**: The first two components together explain **49.33%** of the total variance.
- **Scree Plot Observation**: A sharp decline in variance after the second component indicates the "elbow point."
- **Optimal Components**: Selecting **two components** balances dimensionality reduction with the retention of meaningful information.

Question 10.

The PCA biplot highlights the relationships between variables and observations, with **PC1** and **PC2** explaining a combined **49.4%** of the variance. Below are the key insights:



- **Variable Loadings**: **PaymentTier** and **ExperienceInCurrentDomain** strongly influence **PC1**, while **Age** and **JoiningYear** contribute more to **PC2**.
- **LeaveOrNot**: Shows moderate association, aligning closely with **JoiningYear**.
- **Group Patterns**: Gender-based clusters (**Male** and **Female**) show significant overlap, indicating no strong separation by gender.
- **Key Influences**: Payment tier and experience are critical in explaining variability, while age and joining year capture secondary dimensions.

Conclusion

The analysis reveals that most employees are aged 25–35, hold Bachelor's degrees, and have 2–4 years of experience. PaymentTier is the strongest predictor of retention, followed by age and experience. PCA captured 49.33% variance, highlighting PaymentTier and experience. Insights guide HR on retention, salary structure, and planning.

Dataset II – Adult Census Income.

Univariate Analysis

Question 1.

The Adult Census Income Dataset contains **32,561 observations** and **15 variables**, offering insights from the **1994 Census Bureau** to predict if an individual earns **over $50K annually**.

- **Numerical Variables**:
  - Age: 17–90 years.
  - fnlwgt: 12,285–1,484,705 (population weight).
  - Education.num: 1–16 (Preschool to Doctorate).
  - Capital Gain/Loss: Mostly zeros; outliers (e.g., 99,999 gain, 4,356 loss).
  - Hours per Week: 1–99 (median: 40).
- **Categorical Variables**:
  Workclass, marital status, occupation, race, sex, native country, and income (<=50K or >50K).

Dataset Link: Adult Census Income

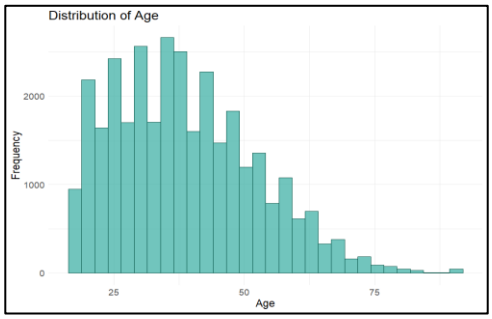
Question 2.

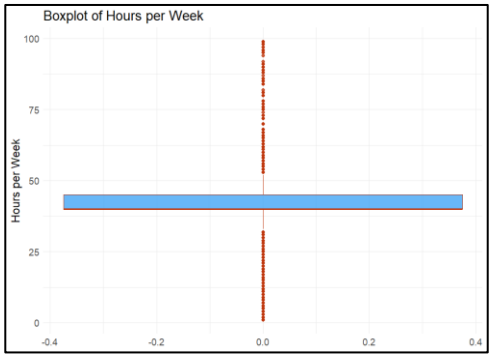
The summary statistics of the **Adult Census Income Dataset** reveal key demographic and work-related characteristics of the population. Below are the key insights:

- **Age**: Ranges from **17 to 90 years**, with a mean of **38.58** and a median of **37**, indicating a primarily adult population.
- **fnlwgt**: Spans **12,285 to 1,484,705**, with a mean of **189,778**, representing population weighting based on demographic controls.
- **Education.num**: Values range from **1 to 16**, with a mean of **10.08**, reflecting that most individuals have completed at least high school.
- **Capital Gain/Loss**: Most values are **0**, with outliers like a gain of **99,999** and a loss of **4,356**.
- **Hours per Week**: Ranges from **1 to 99 hours**, with a mean of **40.44** and a median of **40**, indicating most individuals work full-time.

Question 3.

The **Adult Census Income dataset** provides insights into key numerical variables, **Age** and **Hours per Week**, using visualizations such as histograms and boxplots to uncover population and working trends.



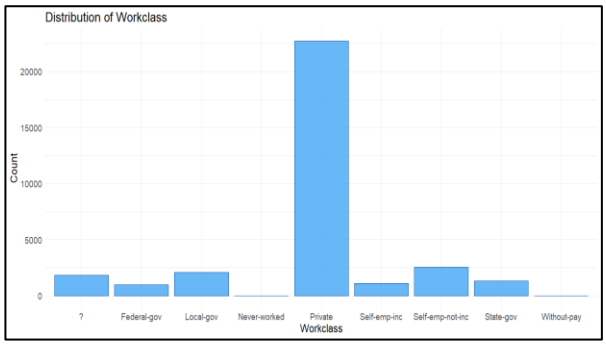


- **Histogram for Age**: The age distribution is **right-skewed**, with most individuals aged **25 to 50 years**, representing the working-age population. Frequencies decline significantly for both younger (<25) and older (>60) age groups, indicating lower representation in these age brackets.
- **Boxplot for Hours per Week**: The boxplot highlights a concentrated working range around **40–45 hours per week**, consistent with standard full-time employment. However, there are numerous **outliers** below 30 hours and above 50 hours, suggesting part-time workers or individuals working overtime. The extreme values above **80 hours per week** may reflect anomalies or individuals with unusually high workloads.

Insights: These visualizations effectively illustrate age distribution and work-hour trends, offering a better understanding of population demographics and employment patterns within the dataset.

Question 4.

The **'Workclass' variable** in the **Adult Census Income dataset** provides insights into employment distribution across various sectors. The bar plot reveals key patterns in the workforce:



- **Private Sector**: Dominates employment, with over 20,000 individuals working in this sector.
- **Government Jobs**: Fewer individuals are employed in **Federal-gov**, **Local-gov**, and **State-gov** roles.
- **Self-Employment**: Moderate representation in **Self-employed categories** (incorporated and not incorporated).
- **Missing Data**: The presence of **"?" values** indicates missing or unclassified data that requires cleaning.
- **Rare Categories**: Minimal entries for **Without-pay** and **Never-worked**, showing their limited prevalence.

Multivariate Analysis

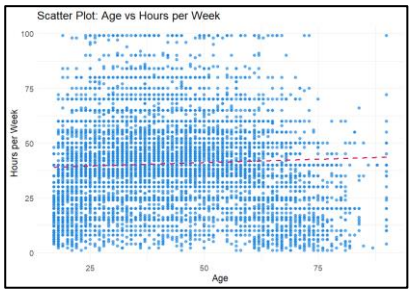
Question 5.

The Pearson correlation analysis between **Age** and **Hours per Week** reveals a very weak positive relationship, with a coefficient of **0.0688**, suggesting minimal influence of age on the number of hours worked weekly. Below are the key insights:

- **Weak Positive Correlation**: The correlation coefficient of 0.0688 indicates a slight direct relationship between the two variables.
- **Minimal Influence**: Age has little to no impact on weekly work hours, suggesting other factors like occupation or job type may play a larger role.
- **Non-Significant Relationship**: The weak correlation highlights that work hours are not strongly dependent on age.

Question 6.

The scatter plot between **Age** and **Hours per Week** in the **Adult Census Income dataset** reveals no strong relationship between the variables, with a consistent trend across age groups. Below are the key insights:



- **Flat Trend**: The trend line is nearly horizontal, indicating no significant correlation between age and hours worked.
- **Standard Work Hours**: Most individuals work around 40 hours per week, regardless of age, reflecting typical full-time schedules.
- **Outliers**: Some extreme values show individuals working unusually high or low hours, though these are rare exceptions.

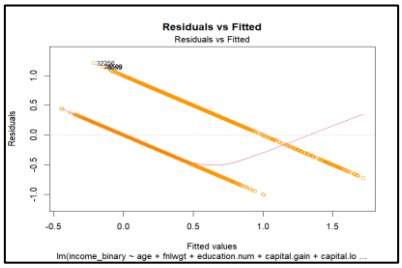
Question 7.

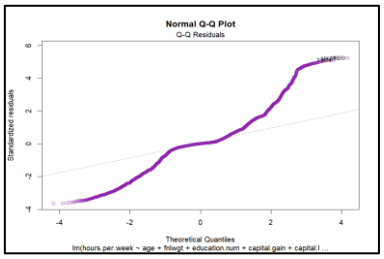
The **multiple regression model** explores how **age**, **fnlwgt**, **education.num**, **capital.gain**, and **capital.loss** influence **hours worked per week** in the Adult Census Income dataset. While the model identifies significant predictors, it explains only a small portion of the variation in working hours.

- **Age**: Older individuals tend to work slightly more hours per week (**positive significant impact**).
- **Education.num**: Higher education levels are strongly associated with longer working hours.
- **Capital.gain**: Individuals with capital gains tend to work marginally longer hours.
- **Capital.loss**: Reporting losses is linked to a slight increase in work hours.
- **Fnlwgt**: No significant relationship with hours worked, suggesting limited predictive power.
- **Model Performance**: Explains 3.1% of the variation in working hours (**$R^2 = 0.03093$**) with significant predictors, except for **fnlwgt**.

Question 8.

The **diagnostic plots** for the regression model reveal violations of key assumptions, indicating model fit issues.





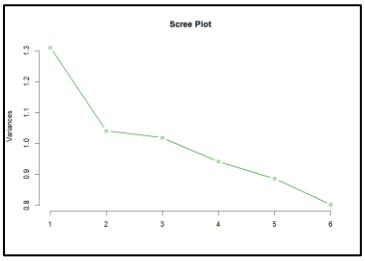
- **Residuals vs Fitted Plot (Homoscedasticity):**
  - Residuals split into two distinct lines instead of scattering randomly around zero.
  - This violates the **constant variance assumption**, suggesting poor model fit.
- **Normal Q-Q Plot (Normality of Residuals):**
  - Residuals deviate significantly from the diagonal line, especially at the tails.
  - This violates the **normality assumption**, indicating heavy-tailed residuals.
- **Model Limitations:**

Variable transformations, additional predictors, or alternative models are needed to address these issues. Used Random Forest and got Accuracy: 86.13 %

Advanced Analysis

Question 9.

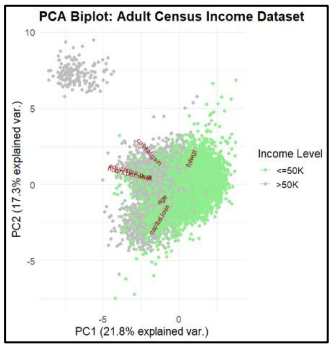
Principal Component Analysis (PCA) was conducted on the numerical variables to analyze dimensionality reduction and the variance explained by each component. The scree plot and explained variance provide guidance on selecting optimal components.



- **Explained Variance**: The first principal component (PC1) explains **21.84%**, and the first two components combined explain **39.19%** of the variance.
- **Cumulative Variance**: The first four components together capture **71.87%** of the total variance.
- **Scree Plot Insight**: The "elbow point" after the second component indicates diminishing returns in variance explained by additional components.
- **Recommendation**: Choosing **2 to 4 components** strikes a balance between dimensionality reduction and retaining meaningful information.

Question 10.

The **PCA biplot** of the **Adult Census Income dataset** visualizes relationships between variables and observations, with the first two principal components explaining **39.1% of the variance**.



- **Variable Loadings**:
  - Capital Gain and Capital Loss strongly contribute to **PC1**.
  - Education Num and Hours per Week heavily influence **PC2**.
- **Group Patterns**:
  - Income levels (<=50K and >50K) form distinct clusters, with <=50K being denser.
  - Clustering highlights the role of financial and work-related factors.
- **Key Insights:**

Financial and work-related variables are critical for variance explanation, useful for income segmentation and prediction.

Conclusion

The **Adult Census Income dataset** reveals key socio-economic and occupational insights, highlighting a predominantly working-age population with typical full-time hours. Private sector employment dominates, while financial and educational factors significantly influence income levels. Weak correlations between variables like age and hours worked suggest other predictors may better explain income classification. PCA confirms financial attributes and education as major contributors to variability, offering opportunities for segmentation and predictive analysis.

Dataset III – Car Price Prediction.

Univariate Analysis

Question 1.

The **Car Price Prediction Dataset** contains **205 observations** and **26 variables**, providing insights into **car attributes** and their impact on pricing.

- **Numerical Variables**: Key metrics include **wheelbase** (86.6–120.9), **carlength** (141.1–208.1), **curbweight** (1,488–4,066 lbs), **enginesize** (61–326), **horsepower** (48–288), **citympg** (13–49), and **price** ($5,118–$45,400).
- **Categorical Variables**: Includes **CarName**, **fueltype** ("gas"/"diesel"), **aspiration** ("std"/"turbo"), **carbody**, **drivewheel**, **enginetype**, and **fuelsystem**.

This dataset supports **exploratory data analysis (EDA)** and **multiple linear regression modeling** to identify **significant predictors of car prices** and understand **pricing dynamics** in the American automobile market.

Dataset Link: Car Price Prediction

Question 2.

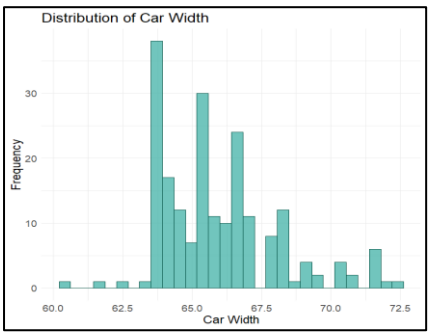
The **Car Price Prediction Dataset** provides detailed insights into key **numerical variables** like car dimensions, weight, engine performance, and pricing. The statistics reveal significant variability across these attributes, essential for analyzing car valuation.

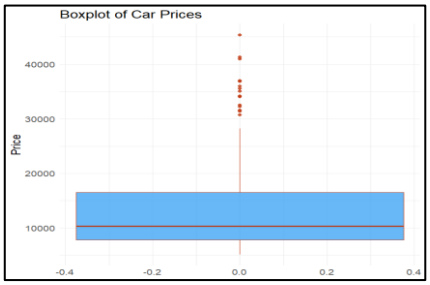
- **Wheelbase**: Ranges from **86.6 to 120.9 inches**, with a mean of **98.76** and a median of **97.0 inches**.
- **Carlength**: Varies between **141.1 and 208.1 inches**, with a mean of **174.0 inches**.
- **Carwidth**: Spans from **60.3 to 72.3 inches**, with a mean of **65.91 inches**.
- **Curbweight**: Ranges from **1,488 to 4,066 lbs**, with a mean of **2,556 lbs**.
- **Enginesize**: Shows a wide range between **61 and 326**, with a mean of **126.9**.
- **Horsepower**: Ranges from **48 to 288 hp**, with a mean of **104.1 hp**.
- **Price**: Varies significantly from **$5,118 to $45,400**, with a mean of **$13,277**.

The **wide variability** in attributes like **enginesize**, **horsepower**, and **price** highlights critical factors influencing car performance and market valuation

Question 3.

The **car width** and **car price** distributions reveal important patterns, variability, and outliers in the dataset, offering insights into car design and pricing trends.



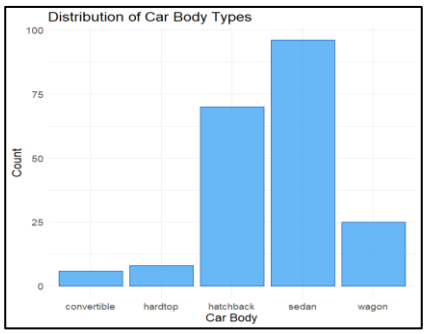


- **Car Width (Histogram)**:
  - The distribution is **bimodal**, with peaks around **64–65 inches** and **66–67 inches**, showing two distinct groupings.
  - Most car widths fall between **64 and 67 inches**, representing typical car designs.
  - A few widths exceeding **70 inches** could be considered outliers, indicating exceptionally wide vehicles.
- **Car Price (Boxplot)**:
  - The price distribution is **right-skewed**, with a median of approximately **$10,295**.
  - Most car prices are concentrated between **$5,118 and $16,503** (IQR).
  - High-end outliers above **$30,000** reflect luxury or premium cars, with some exceeding **$40,000**.

Insights: The bimodal car width distribution suggests variability in car design, while the skewed car price distribution highlights affordability for most cars and a small segment of luxury vehicles.

Question 4.

The analysis of the **car body types** highlights the dominance of certain designs in the market, reflecting consumer preferences and industry trends.



- **Sedan:** Most popular body type, with nearly **100 cars**.
- **Hatchback:** Second most common, with around **70 cars**, showing strong appeal.
- **Wagon:** Moderately preferred, with about **25 cars**.
- **Hardtop and Convertible:** Least frequent, indicating **low demand** for these styles.

Overall, the dominance of **sedans** and **hatchbacks** suggests a market preference for **functional and compact car designs**.

Multivariate Analysis

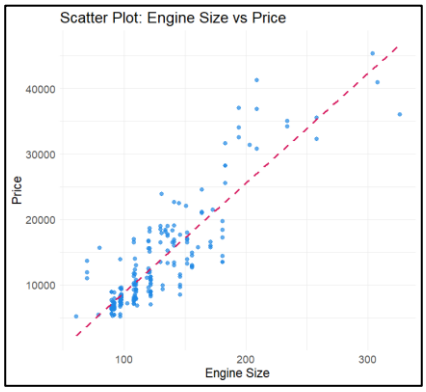
Question 5.

The **correlation analysis** between **Engine Size** and **Price** reveals a **strong positive relationship**, indicating that as engine size increases, car prices tend to rise significantly. Below are the key points:

- **Pearson Correlation Coefficient:** The value is **0.8741**, showing a strong positive correlation.
- **Positive Trend:** Larger engine sizes are directly associated with higher car prices.
- **Key Insight:** Engine size is a significant predictor of car price, aligning with the general perception that larger engines contribute to higher performance and cost.

Question 6.

The **scatter plot** between **Engine Size** and **Price** reveals a strong positive relationship, indicating that cars with larger engines generally have higher prices. The linear trend line further validates this observation.



- **Positive Correlation:** As **Engine Size** increases, **Price** tends to rise proportionally.
- **Trend Line:** The upward-sloping dashed line confirms the strong linear relationship.
- **Alignment:** Data points are closely clustered along the trend line, showing a strong association.
- **Insight:** Engine size is a significant factor in determining car prices, making it a crucial predictor for pricing models.

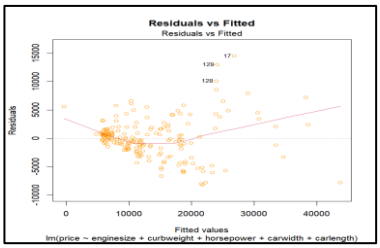
Question 7.

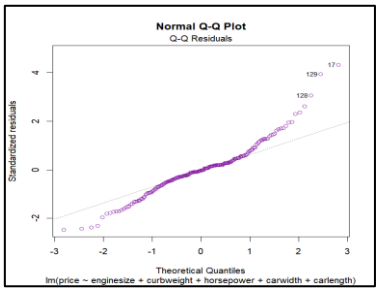
The **multiple regression model** predicts car price using **Engine Size**, **Curb Weight**, **Horsepower**, **Car Width**, and **Car Length** as predictors, explaining **82.02% (R^2 = 0.8202)** of price variability.

- **Strong Predictors**:
  - Engine Size: Positively significant (83.61), larger engines increase prices.
  - Horsepower: Significant (47.66), higher horsepower raises prices.
  - Car Width: Significant (634.82), wider cars cost more.
- **Moderate Impact**:
  - Curb Weight: Marginal significance (p = 0.067).
- **Insignificant**:
  - Car Length (p = 0.493).
- **Model Fit**: Adjusted R^2 = 0.8157. Residuals suggest scope for improvements or additional predictors.

Question 8.

The diagnostic plots assess the assumptions of **homoscedasticity** and **normality** in the regression model predicting car prices. The results indicate deviations in residual behavior, suggesting areas for model improvement.





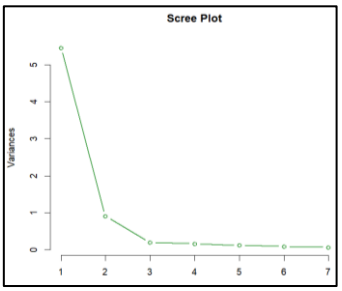
- **Residuals vs Fitted Plot**:
  - Shows **heteroscedasticity** with a curved pattern, indicating non-constant variance of residuals.
  - Scatter at higher fitted values suggests potential **non-linearity** or missing predictors.
- **Normal Q-Q Plot**:
  - Residuals deviate from normality, particularly in the **tails**, indicating the presence of outliers.
  - Observations like **17, 128, and 129** contribute significantly to deviations.

Insights: The model partially violates assumptions of homoscedasticity and normality, suggesting the need for variable transformations, inclusion of additional predictors, or addressing outliers for improved model fit.

Advanced Analysis

Question 9.

PCA was performed on the numerical variables to reduce dimensionality while retaining maximum variance.

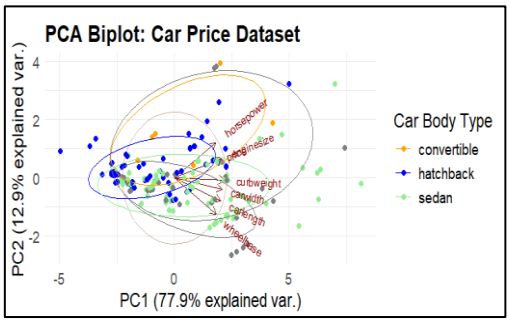


Based on the **scree plot** and explained variance, two principal components are sufficient as they capture the majority of the variance.

- **PC1** explains **77.94%** of the variance.
- **PC2** adds **12.9%**, making the cumulative variance **90.84%**.
- The **scree plot** shows a clear **elbow point** at PC2, beyond which the explained variance diminishes significantly.
- **Decision**: Selecting **two components** effectively balances dimensionality reduction and variance retention.

Question 10.

The **PCA biplot** highlights relationships between numerical variables and car body types, with **PC1 (77.9%)** and **PC2 (12.9%)** explaining most variability.



- **Variable Loadings**:
  - Horsepower and Engine Size significantly contribute to **PC1**.
  - Wheelbase, Curb Weight, and Car Width influence **PC1** and **PC2**.
- **Group Patterns**:
  - Sedans (green): Greater dispersion, indicating high variability.
  - Hatchbacks (blue): Tightly clustered, showing similar characteristics.
  - Convertibles (orange): Distinct small cluster, reflecting unique features.
- **Insights**:
  High horsepower and engine size align with higher prices, especially for convertibles, with clear differences among body types.

Conclusion

The **Car Price Prediction Dataset** analysis identifies **Engine Size**, **Horsepower**, and **Car Width** as key price drivers, while **Car Length** has minimal impact. Prices are right-skewed with high-end outliers, and sedans dominate the market. A strong correlation (0.8741) exists between **Engine Size** and **Price**. PCA explains **90.84% variance**, with horsepower and engine size as major contributors, showing clear groupings by car body types and market trends.

Dataset IV – Retail Sales and Customer Demographics.

Univariate Analysis

Question 1.

The **Retail Sales and Customer Demographics Dataset** contains **1,000 observations** and **9 variables**, focusing on retail operations and customer behaviors.

- **Numerical Variables**:
  - Age: 18–64 (median: 42, mean: 41.39).
  - Quantity: 1–4 items (mean: 2.51).
  - Price per Unit: $25–$500.
  - Total Amount: $25–$2,000 (median: $135).
- **Categorical Variables**:
  Product Category, Gender, Customer ID, Transaction ID, and Date.
- **Insights**:
  The dataset enables analysis of sales trends, product preferences, and demographic influences, offering opportunities to identify purchasing patterns and generate actionable business insights.

Dataset Link: Retail Sales and Customer Demographics

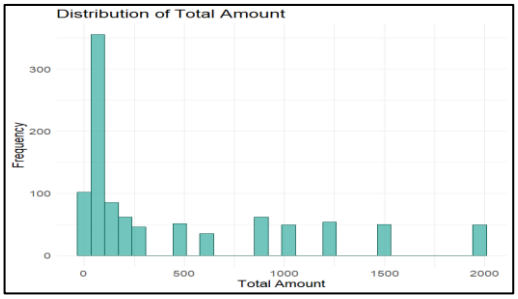
Question 2.

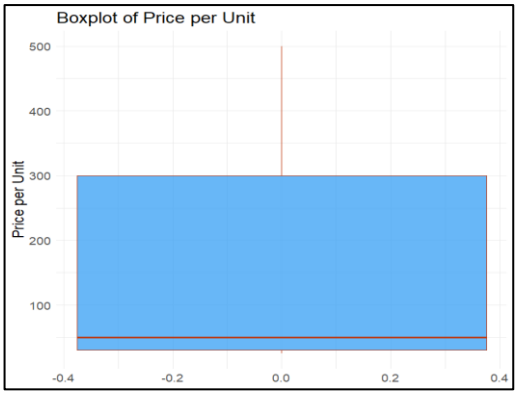
The summary statistics for the **Retail Sales Dataset** provide insights into key numerical variables, including **Quantity**, **Price per Unit**, and **Total Amount**, highlighting their central tendencies and variability.

1. **Quantity**: Ranges from **1 to 4** with a **mean of 2.514** and **median of 3**, indicating most transactions involve 2–3 items.
2. **Price per Unit**: Spans **$25 to $500**, with a **median of $50** and a **mean of $179.9**, showing a right-skewed distribution driven by higher-priced items.
3. **Total Amount**: Varies between **$25 and $2,000**, with a **median of $135** and a **mean of $456**, reflecting significant variability influenced by larger purchases or high unit prices.

Question 3.

The analysis of **Total Amount** and **Price per Unit** reveals notable patterns in transaction values and pricing behaviors, highlighting skewness, variability, and significant outliers.





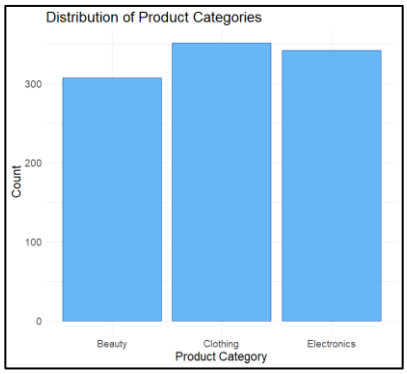
- **Total Amount (Histogram)**:
  - The distribution is **highly right-skewed**, with the majority of transactions below **500**.
  - A smaller number of transactions exceed **1,000**, indicating higher-value purchases, with outliers reaching up to **2,000**. This suggests occasional large transactions within the dataset.
- **Price per Unit (Boxplot)**:
  - Prices range between **25 and 500**, with a **median below 100**, indicating that most items are lower priced.
  - Higher-priced items appear as outliers above **300**, reflecting premium products or exceptions.

Key Insights:

The data highlights **diverse transaction sizes** and **pricing behaviors** with a significant concentration of small to mid-range purchases. The presence of high-value outliers in both variables suggests the importance of analyzing premium pricing strategies and large transactions separately for better business insights and sales optimization.

Question 4.

The **distribution analysis** of the *Product Category* variable reveals consumer preferences across key categories:



- **Clothing** is the most popular category, with the highest count of transactions.
- **Electronics** closely follows Clothing, indicating a strong market demand.
- **Beauty** products have slightly fewer transactions but still exhibit significant consumer interest.

Key Insight: Clothing, Electronics, and Beauty represent three major product categories, with Clothing leading slightly in popularity. This highlights balanced demand across diverse product segments.

Multivariate Analysis

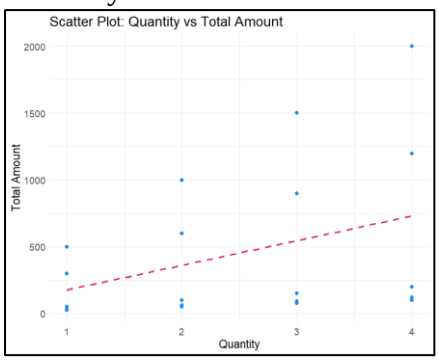
Question 5.

The **correlation analysis** between *Quantity* and *Total Amount* reveals a moderate positive relationship, as indicated by the **Pearson correlation coefficient** of **0.374**.

- **Positive Correlation**: An increase in the quantity purchased tends to increase the total amount spent.
- **Moderate Strength**: The correlation is not very strong, suggesting that while Quantity influences Total Amount, other factors like *Price per Unit* also play a significant role.
- **Key Insight**: Higher quantities contribute to increased spending, but the relationship is influenced by variations in product pricing.

Question 6.

The **scatter plot** visualizes the relationship between *Quantity* and *Total Amount*, with a trend line included for clarity.



• **Positive Relationship**: The upward-sloping trend line indicates a positive correlation between Quantity and Total Amount—higher quantities generally lead to higher spending.
• **Scattered Data Points**: While the trend is visible, there is considerable spread in the data, suggesting that factors like *Price per Unit* influence the Total Amount.
• **Moderate Trend**: The linear relationship is moderate, as seen from the less tightly clustered points around the trend line.
The relationship between Quantity and Total Amount is positive but not perfectly linear, indicating that the price per unit varies across transactions.

Question 7.

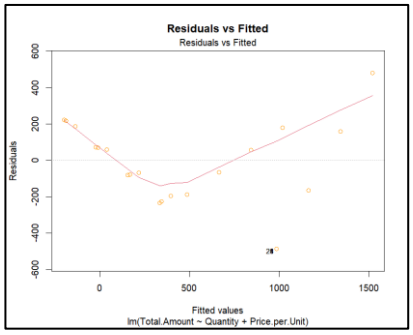
The **multiple regression model** predicts **Total Amount** using **Quantity** and **Price per Unit** as predictors, explaining **85.43%** of the variability (Adjusted R^2).

- **Model Performance**:
  - **F-statistic**: 2929 (p-value < 2.2e-16), confirming model significance.
- **Predictor Coefficients**:
  - **Quantity**: Each additional unit increases Total Amount by **177.44**.
  - **Price per Unit**: A 1-unit price increase raises Total Amount by **2.50**.
- **Residual Analysis**:
  Residuals range between **-486.6** and **481.1** (standard error: 213.8), showing minor variability.

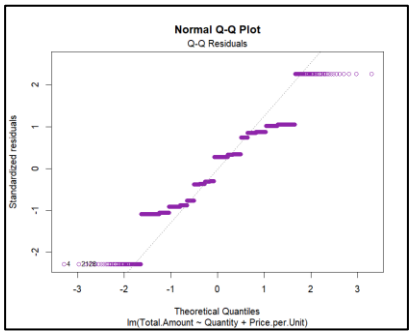
Insights: Quantity and Price per Unit are significant drivers of sales value.

Question 8.

The model diagnostics and PCA analysis evaluate the regression fit and dimensionality reduction effectiveness. The diagnostic plots highlight issues with residual patterns, while PCA efficiently reduces dimensionality with minimal information loss.



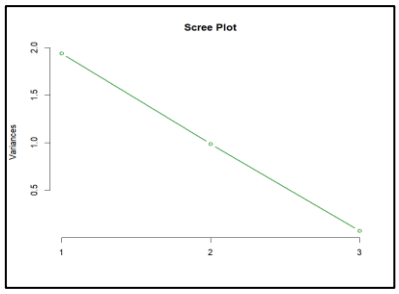
1. **Residuals vs Fitted Plot**:
   - Residuals show a curved pattern, violating homoscedasticity.
   - Increasing variance at higher fitted values suggests potential non-linearity or missing predictors.
2. **Normal Q-Q Plot**:
   - Residuals deviate from normality, especially in the tails.
   - Outliers such as 4, 218, and 232 contribute to these deviations.



Advanced Analysis

Question 9.

The **PCA** on the numerical variables (**Quantity**, **Price per Unit**, and **Total Amount**) shows that the first two principal components explain nearly all the variance.

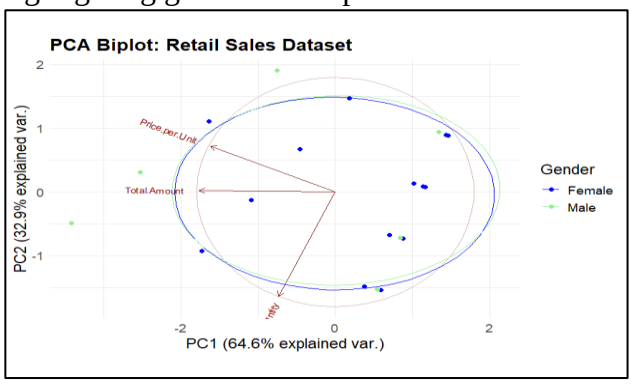


- **Explained Variance**:
  - **PC1**: 64.56% of total variance.
  - **PC2**: Adds 32.90%, totaling **97.47%**.
- **Scree Plot**:
  - The elbow point occurs at **PC2**, where explained variance significantly drops.
  - **PC3** contributes only **2.53%**, indicating minimal impact.
- **Component Selection**:

Using two components optimally captures **97.47%** of the variance, balancing **dimensionality reduction** and information retention.

Question 10.

The **PCA biplot** visualizes relationships between **Quantity**, **Price per Unit**, and **Total Amount**, highlighting gender-based patterns.



- **Variable Loadings**:
  - **Total Amount** and **Price per Unit** align strongly with **PC1** (64.6% variance).
  - **Quantity** aligns more with **PC2** (32.9% variance).
- **Gender-Based Patterns**:
  - **Females (blue)**: Tightly clustered, indicating consistent purchasing behavior.
  - **Males (light green)**: Greater dispersion, showing transaction variability.
- **Insights**:

Total Amount and Price per Unit are positively correlated. Gender separation reveals subtle spending differences, useful for **targeted marketing strategies**.

Conclusion

The analysis of the **Retail Sales and Customer Demographics Dataset** highlights key insights into purchasing behaviors and sales trends. **Univariate analysis** shows **Total Amount** and **Price per Unit** are right-skewed with outliers, while **Clothing** leads product categories. **Multivariate analysis** reveals a moderate correlation (0.374) between Quantity and Total Amount, with regression explaining **85.43%** of variability. **PCA** captures **97.47%** of variance, showing gender-based differences: females display consistent spending, while males show greater variability, aiding **sales optimization** and **targeted marketing**.