Statistical Structures in Data Project

Under the guidance of Prof. Subhajit Dutta

Name: Urja Kumari
Roll No: 24BM6JP58

Project Overview:

This project involves the analysis of four datasets, including identifying correlations between variables and visualizing distributions to better understand the data. It also requires the application of advanced techniques such as regression and PCA for dimensionality reduction.

Dataset 1: Obesity Estimation dataset: This dataset includes data for the estimation of obesity levels in individuals from the countries of Mexico, Peru, and Colombia, based on their eating habits and physical condition. It has been taken from the UCI repository.

Dataset 2: The data set refers to clients of a wholesale distributor. It includes the annual spending in monetary units (m.u.) on diverse product categories. It has been taken from the UCI repository.

Dataset 3: The `air quality` dataset in R is a built-in dataset that contains daily air quality measurements in New York City from May to September 1973.

Dataset 4: The 'Tips' dataset contains information about the tips given at a restaurant and includes various features related to the transaction. This dataset is available under the seaborn library in Python.

DATASET-1

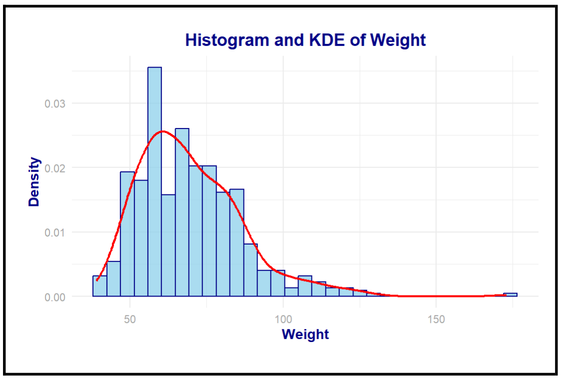
- The dataset-1 consists of 2111 observations and 17 features. Around 500 observations are considered for further analysis due to some discrepancies present afterward.
- There are no missing values in the dataset along any feature.
- There are a total of 9 categorical variables and 8 numerical variables.
- The two numerical variables considered for further analysis like doing distribution analysis, visualization purposes, etc are '**Height**' and '**Weight**'.
- The categorical Variable being considered for analysis like for understanding the distribution, presence of outliers, etc, is '**NObeyesdad**' which is representing the weight level of a person.

Summary Statistics of a few Numerical Variables ("Weight", "Height", "Age")

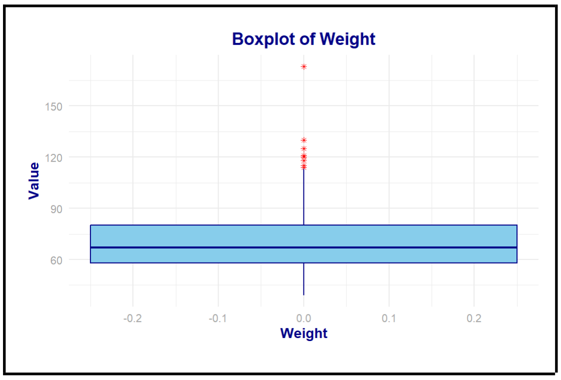
Statistics	Weight	Height	Age
Mean	69.57	1.69	23.15
Median	67	1.69	21
Standard Deviation	17.01	0.10	6.72
Minimum	39	1.45	14
Maximum	173	1.98	61

DISTRIBUTION VISUALISATIONS

Histogram of Numerical Variable ("Weight"):

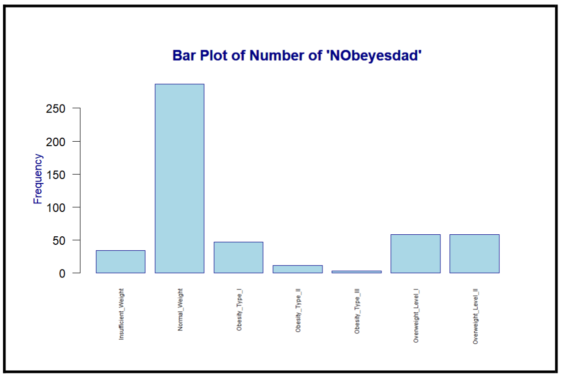


Boxplot of Numerical Variable (“Weight”):



- The distribution of ‘Weight’ is positively skewed or Right Skewed (visible from the Kernel Density Estimation Plot) with a skewness value of 1.19.
- There are a few outliers present along the Weight Variable having values of more than 110 kgs as evident from the boxplot of the Variable. One of the values seems very far away from the other values

Barplot of Categorical Variable (“NObeyesdad”):

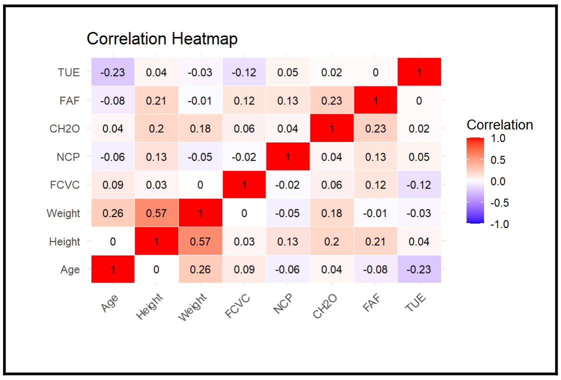


The barplot shows that most people are of Normal weight, followed by people who are moderately overweight and severely overweight.

MULTIVARIATE ANALYSIS

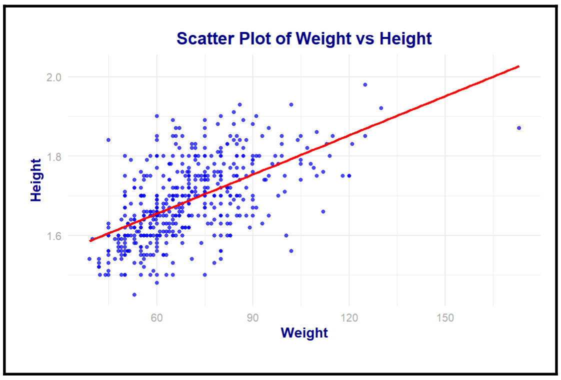
- The Pearson Correlation Coefficient between two numerical variables (“Weight” and “Height”): **0.57**

Correlation Matrix Heatmap:



- The heatmap reveals the pairwise correlations between variables.
- Strong positive correlations are observed between Height and Weight (correlation = 0.57) and Age and Weight (correlation = 0.26).
- Variables like TUE and Age show a negative correlation (-0.23), indicating an inverse relationship.
- Most variables exhibit weak or negligible correlations, suggesting limited linear relationships among them.

Scatterplot For Visualising Relationship between ‘Height’ and ‘Weight’ :



The scatter plot shows a positive linear relationship between Weight and Height, as indicated by the red regression line. This suggests that as weight increases, height tends to increase as well, although there is some variability around the trend line.

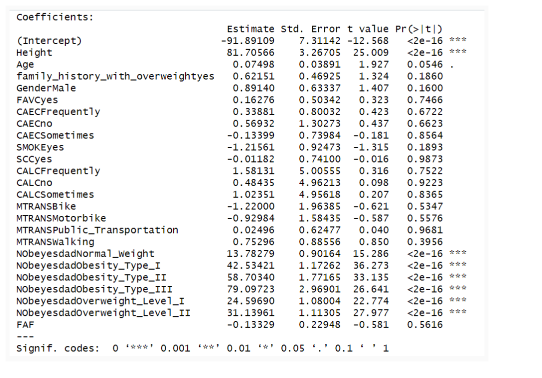
Fitting Multiple Linear Regression:

A multiple linear regression analysis is performed on the continuous numerical variable "Weight" using continuous predictors such as "Age" and "Height," categorical predictors like "FAVC," "CAEC," "SMOKE," "SCC," "CALC," "MTRANS," and "NObeyesdad, ’family_history_with_overweight, as well as the discrete numerical variable "FAF." The categorical variables are converted into factors, as R treats them distinctly during analysis, similar to how one-hot encoding would be applied.

Note:

- **FAVC :** Do you eat high caloric food frequently? (Categories: Yes/No)
- **CAEC :**Do you eat any food between meals? (Categories: Sometimes/ Frequently/ Always /No)
- **SMOKE:** Do you smoke? (Categories: Yes/No)
- **SCC:** Do you monitor the calories you eat daily? (Categories: Yes/No)
- **CALC:** How often do you drink alcohol? (Categories: Sometimes/ Frequently/ Always /No)
- **MTRANS:** Which transportation do you usually use? (Categories: Automobile/ Bike/ Motorbike/ Public_Transportation/ Walking)
- **FAF:** How often do you have physical activity? (Continuous)

Model Coefficients Estimates



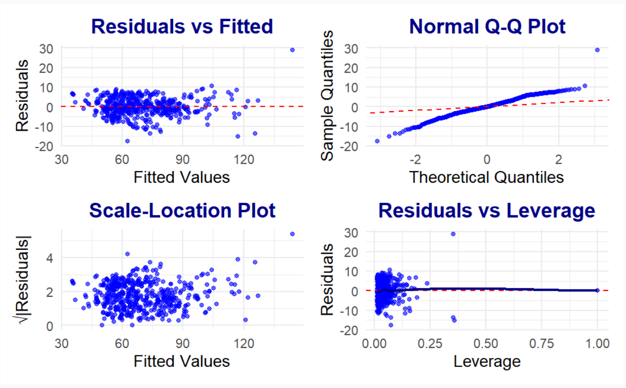
```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)    
(Intercept)                  -91.89109    7.31142 -12.568   <2e-16 ***
Height                        81.70566    3.26705  25.009   <2e-16 ***
Age                            0.07498    0.03891   1.927   0.0546 .  
family_history_with_overweightyes   0.62151    0.46925   1.324   0.1860    
GenderMale                     0.89140    0.63337   1.407   0.1600    
FAVCyes                        0.16276    0.50342   0.323   0.7466    
CAECFrequently                 0.33881    0.80032   0.423   0.6722    
CAECno                         0.56932    1.30273   0.437   0.6623    
CAECSometimes                 -0.13399    0.73984  -0.181   0.8564    
SMOKEyes                      -1.21561    0.92473  -1.315   0.1893    
SCCyes                        -0.01182    0.74100  -0.016   0.9873    
CALCFrequently                 1.58131    5.00555   0.316   0.7522    
CALCno                         0.48435    4.96213   0.098   0.9223    
CALCSometimes                  1.02351    4.95618   0.207   0.8365    
MTRANSBike                    -1.22000    1.96385  -0.621   0.5347    
MTRANSMotorbike               -0.92984    1.58435  -0.587   0.5576    
MTRANSPublic_Transportation    0.02496    0.62477   0.040   0.9681    
MTRANSWalking                  0.75296    0.88556   0.850   0.3956    
NObeyesdadNormal_Weight       13.78279    0.90164  15.286   <2e-16 ***
NObeyesdadObesity_Type_I      42.53421    1.17262  36.273   <2e-16 ***
NObeyesdadObesity_Type_II     58.70340    1.77165  33.135   <2e-16 ***
NObeyesdadObesity_Type_III    79.09723    2.96901  26.641   <2e-16 ***
NObeyesdadOverweight_Level_I  24.59690    1.08004  22.774   <2e-16 ***
NObeyesdadOverweight_Level_II 31.13961    1.11305  27.977   <2e-16 ***
FAF                           -0.13329    0.22948  -0.581   0.5616    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Model Summary	Values
Residual Standard Error	4.87
Multiple R-squared	0.92
Adjusted R-squared	0.91
F-Statistic	233.2 on 24 and 473 DF
p-value	< 2.2e-16

Interpretation:

1. **Model Strength:** The model explains 92.21% of the variance in weight, demonstrating excellent fit.
2. **Statistical Significance**: The overall model is highly significant with p<2.2e−16.
3. **Key Predictors: Height** and **NObeyesdad** categories (obesity levels) are the strongest predictors of weight.
4. **Non-Significant Predictors**: Variables like Age, FAVC, CAEC, and MTRANS are not significant and may be optimized or excluded.
5. **Prediction Accuracy:** Residual standard error (RSE) of 4.87 units suggests good prediction accuracy.
6. **Future Improvements:** Dropping non-significant variables and exploring interactions or non-linear relationships could enhance the model.
7. **Practical Use:** The model is valuable for health studies, especially in weight prediction and obesity analysis.

DIAGNOSTIC PLOTS:



Key Observations:

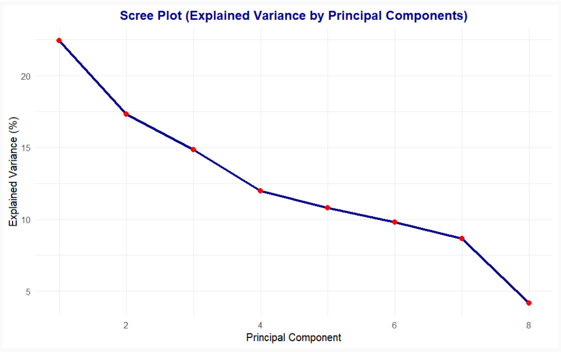
- Residuals scatter randomly around the red line (zero). Suggests linearity and constant variance are mostly met. Few points deviate, indicating potential outliers/influential observations.
- From the **Q-Q plot**, we see that most points align with the red dashed line and residuals are approximately normal. Tail deviations hint at possible outliers' presence.
- **Scale Location Plot** checks homoscedasticity. Residuals are spread fairly across fitted values which suggests constant variance.Mild variability at higher fitted values indicates possible heteroscedasticity.
- **Residuals vs Leverage plot** detect influential observations. Most residuals cluster near zero and have low leverage.

Performing PCA:

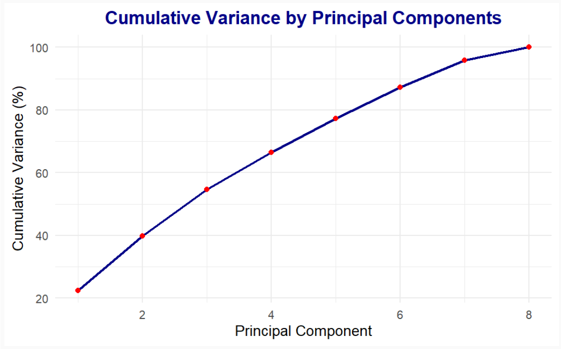
Principal Component Analysis (PCA) is a dimensionality reduction technique that transforms high-dimensional data into a smaller set of uncorrelated components while retaining most of the data's variance**.**

PCA is applied to 8 numerical features after scaling them, resulting in the generation of 8 principal components.

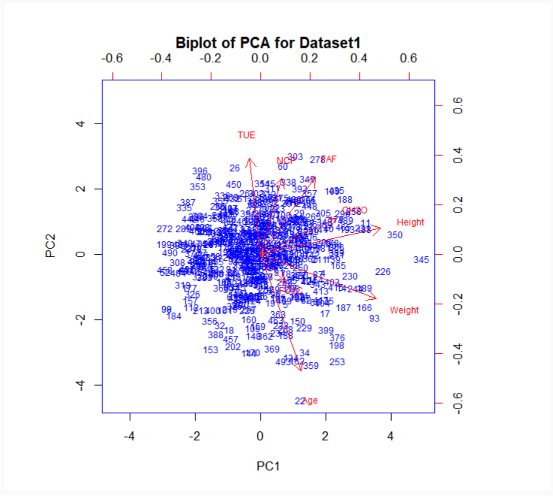
The **scree plot** shows the percentage of explained variance for each principal component (PC). The first few components explain a higher variance, with the contribution decreasing as we move to higher PCs.

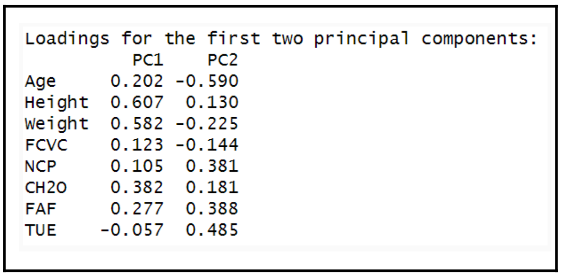


Based on the plot, we notice a clear elbow around **PC 3 or PC 4.** After this point, the variance explained by additional components becomes negligible.



From the cumulative variance plot it is visible that if we choose first 5 PCs then 80% of the variance present in the data will be explained by those eigenvalues.





```
Loadings for the first two principal components:
          PC1    PC2
Age      0.202 -0.590
Height   0.607  0.130
Weight   0.582 -0.225
FCVC     0.123 -0.144
NCP      0.105  0.381
CH2O     0.382  0.181
FAF      0.277  0.388
TUE     -0.057  0.485
```

Interpretation from the Loadings Matrix

* Height (0.607) and Weight (0.582) show strong positive loadings. CH2O (0.382) and FAF (0.277) contribute positively but to a lesser extent. PC1 reflects variations related to physical traits (Height and Weight) and water consumption (CH2O).

* Age (-0.590) has a strong negative loading, indicating an inverse relationship. TUE (0.485), FAF (0.388), and NCP (0.381) contribute positively. PC2 represents behavioral and lifestyle factors, suggesting older individuals may display distinct behaviors compared to younger ones.

DATASET-2

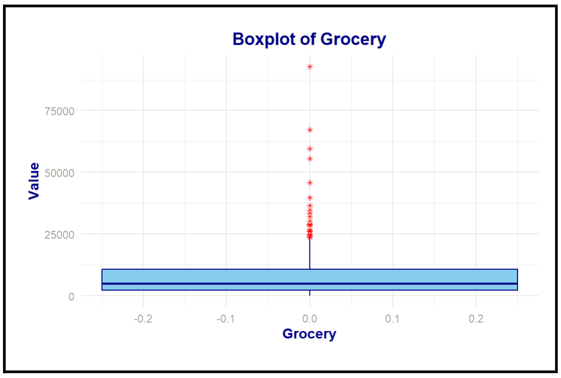
- The dataset consists of 440 observations and 8 features.
- There are no missing values in the dataset along any feature.
- There are 2 numerical categorical variables and 6 continuous variables.
- The two numerical variables considered for further distribution analysis, presence of outliers, visualization purposes, etc are ‘**Grocery’** and **‘Milk**’.
- The categorical Variable being considered for analysis like for understanding the distribution and relationship with other features is ‘**Region’**.

Summary Statistics of Numerical Variables (“Milk”, “Grocery”, “Fresh”,” Frozen”)

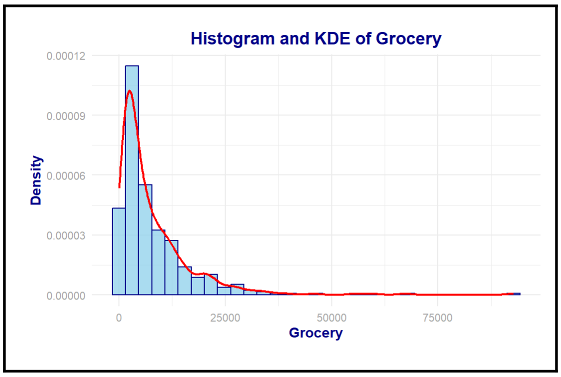
Statistics	Milk	Grocery	Fresh	Frozen
Mean	5796.27	7951.28	12000.3	3071.93
Median	3627	4755.5	8504	1526
Standard Deviation	7380.38	9503.16	12647.3	4854.67
Minimum	55	3	3	25
Maximum	73498	92780	112151	60869

DISTRIBUTION VISUALISATIONS

Boxplot of Numerical Variable (“Grocery”):

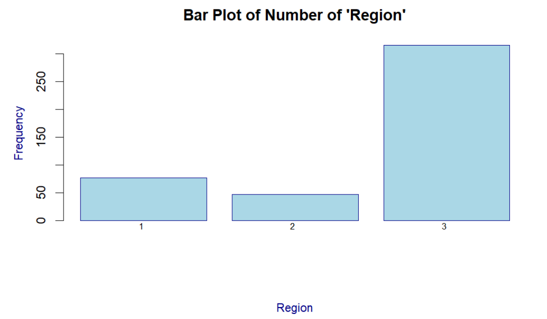


Histogram of Numerical Variable (“Grocery”):



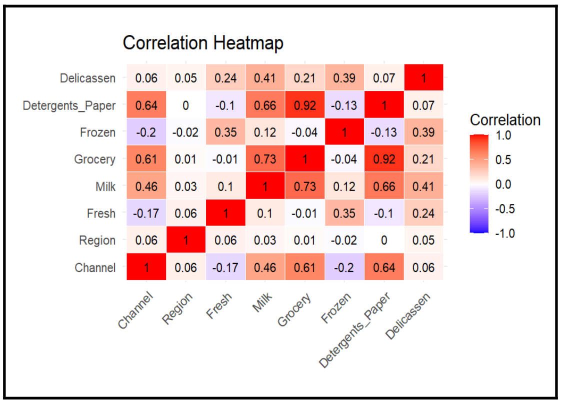
- The distribution of the ‘Grocery’ feature is Positively Skewed (RightSkewed) with a skewness value of 3.56.
- **Outliers** are present, with some values significantly higher than the mean, suggesting the need for further investigation or potential transformation.
- The **'Milk'** feature is also **positively skewed**, with a skewness value of **4.03**, indicating an even greater skew compared to 'Grocery'.

Barplot of Categorical Variable (“Region”)



The barplot indicates that the majority of clients are from Region 3, suggesting that enhanced facilities and services could be prioritized for this region.

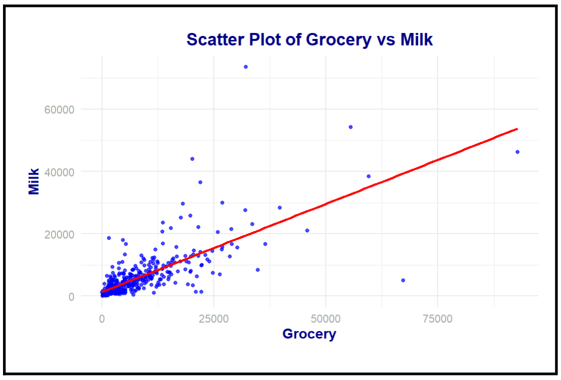
Correlation Matrix:



	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
Delicassen	0.06	0.05	0.24	0.41	0.21	0.39	0.07	1
Detergents_Paper	0.64	0	-0.1	0.66	0.92	-0.13	1	0.07
Frozen	-0.2	-0.02	0.35	0.12	-0.04	1	-0.13	0.39
Grocery	0.61	0.01	-0.01	0.73	1	-0.04	0.92	0.21
Milk	0.46	0.03	0.1	1	0.73	0.12	0.66	0.41
Fresh	-0.17	0.06	1	0.1	-0.01	0.35	-0.1	0.24
Region	0.06	1	0.06	0.03	0.01	-0.02	0	0.05
Channel	1	0.06	-0.17	0.46	0.61	-0.2	0.64	0.06

* The correlation coefficient between Milk and Grocery: **0.73**

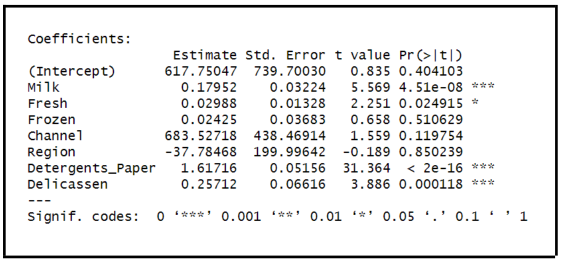
Scatterplot between continuous variables (‘Milk’ and ‘Grocery’):



Multiple Linear Regression

A multiple linear regression analysis was conducted to investigate the relationship between the continuous numerical variable "Grocery" and several predictors, including continuous variables ('Milk', 'Fresh', 'Frozen', 'Detergents_Paper', 'Delicassen') and categorical variables ('Channel', 'Region').

Model Coefficient Estimates:



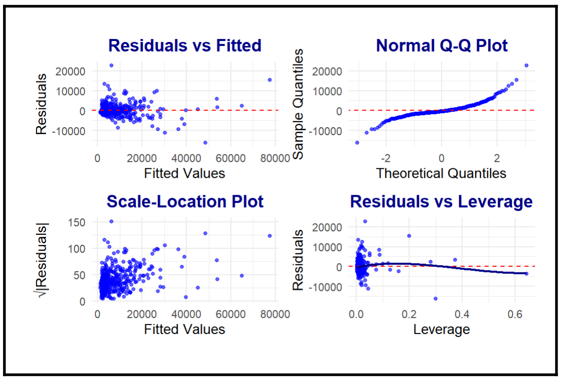
```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      617.75047  739.70030   0.835 0.404103    
Milk               0.17952    0.03224   5.569 4.51e-08 ***
Fresh              0.02988    0.01328   2.251 0.024915 *  
Frozen             0.02425    0.03683   0.658 0.510629    
Channel          683.52718  438.46914   1.559 0.119754    
Region           -37.78468  199.99642  -0.189 0.850239    
Detergents_Paper   1.61716    0.05156  31.364  < 2e-16 ***
Delicassen         0.25712    0.06616   3.886 0.000118 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Model Summary	Values
Residual Standard Error	3221
Multiple R-squared	0.887
Adjusted R-squared	0.885
F-Statistic	484.1 on 7 and 432 DF
p-value	< 2.2e-16

Key Observations:

1. **Significant Predictors**: Milk, Fresh, Detergents_Paper, and Delicassen have significant positive effects on grocery sales, with Milk showing the highest significance (p < 0.001).
2. **Insignificant Predictors**: The Channel and Region variables do not significantly influence grocery sales, as indicated by their high p-values.
3. **Model Fit**: The model explains approximately 88.69% of the variability in grocery sales (Multiple R-squared), indicating a strong fit.
4. **Residual Standard Error**: The residual standard error of 3221 suggests that the average prediction error is relatively high, but the model still performs well based on the R-squared value.
5. **Overall Model Significance**: The F-statistic of 484.1 with a p-value < 2.2e-16 confirms that the model is statistically significant, indicating that at least one predictor variable has a non-zero effect on grocery sales.

Diagnostic Plot



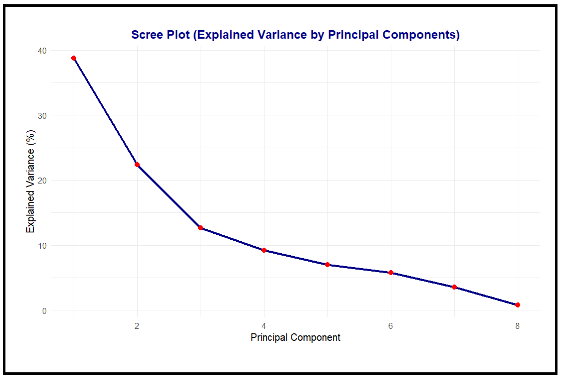
Interpretation:

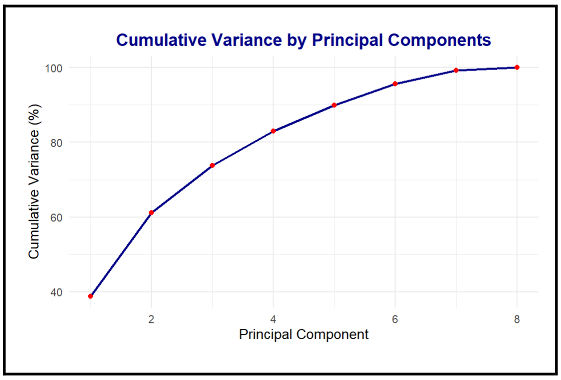
- The residuals show clustering and spread, indicating possible non-linearity or heteroscedasticity. The linear model may not fully capture the relationship between predictors and 'Grocery.
- Deviations from the diagonal line, especially at the tails, suggest residuals deviate from normality, potentially affecting statistical inference reliability.
- Residual variance increases with fitted values, indicating heteroscedasticity, which may impact the model's accuracy.
- Some points have high leverage or large residuals, suggesting influential observations that require further investigation.

PCA Implementation:

PCA is applied to 8 numerical features after scaling them, resulting in the generation of 8 principal components.

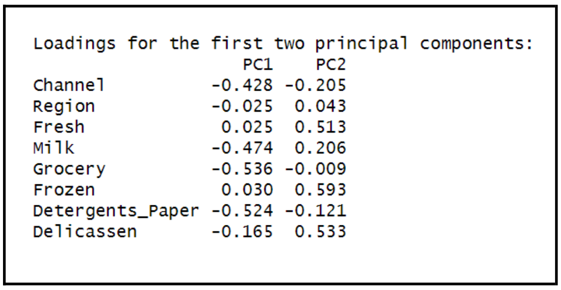
Scree Plot





Interpretation:

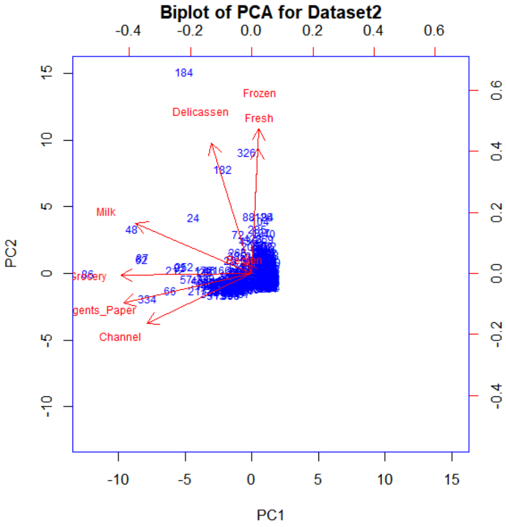
- The scree plot suggests an elbow point around the second or third principal component.
- The cumulative variance plot shows that the first five or six principal components capture a significant portion of the total variance.
- Based on both plots, reducing the dimensionality to the first 5 or 6 principal components is a reasonable approach. However we need to look at the trade off between accuracy and model simplicity.



```
Loadings for the first two principal components:
                   PC1    PC2
Channel         -0.428 -0.205
Region          -0.025  0.043
Fresh            0.025  0.513
Milk            -0.474  0.206
Grocery         -0.536 -0.009
Frozen           0.030  0.593
Detergents_Paper -0.524 -0.121
Delicassen      -0.165  0.533
```

Interpretations:

- **PC1:** Strong negative loadings for Grocery (-0.536), Detergents_Paper (-0.524), and Milk (-0.474) indicate that this component captures a trend of lower consumption in these categories, suggesting a "low consumption" or "cost-sensitive" consumer behavior.
- **PC2:** Strong positive loadings for Fresh (0.513), Frozen (0.593), and Delicassen (0.533) highlight a preference for fresh and frozen products, indicating a "health-conscious" or "fresh food preference" among consumers**.**



DATASET -3

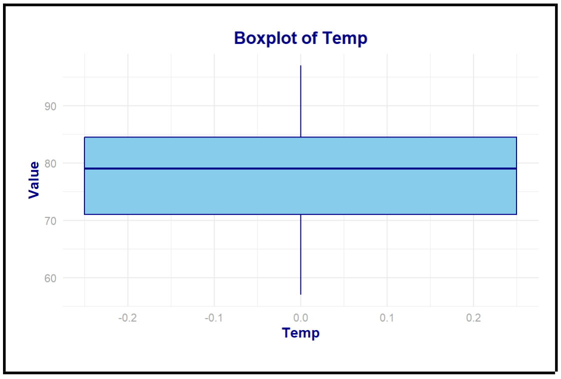
- The dataset consists of 153 observations and 6 features..
- There are a few missing values present in the dataset along the features Ozone and Solar.R. The observations containing the missing values are dropped from the data frame as a preprocessing step.
- There are 2 categorical variables, namely ‘Month’ and ‘Day’, and 4 continuous variables.
- The two numerical variables considered for further analysis like for doing distribution analysis, visualization purposes, etc are ‘Temp” and ‘Ozone”.
- The categorical Variable being considered for analysis like for understanding the distribution, presence of outliers, etc, is ‘Month’ and it only consists of 5 values, i.e., months 5,6, 7, 8, and 9.

Summary Statistics of Numerical Variables (“Temp”, “Wind”, “Ozone”,” Solar.R”)

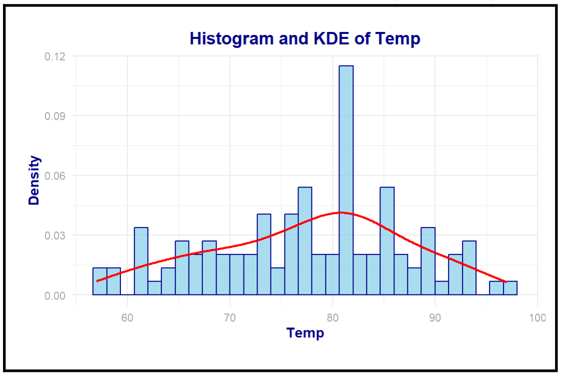
Statistics	Temp	Wind	Solar.R	Ozone
Mean	77.79	9.94	184.80	42.10
Median	79	9.7	207	31
Standard Deviation	9.53	3.56	91.15	33.28
Minimum	57	2.3	7	1
Maximum	97	20.7	334	168

DISTRIBUTION VISUALISATIONS

Boxplot of Numerical Variable (“Temp”):



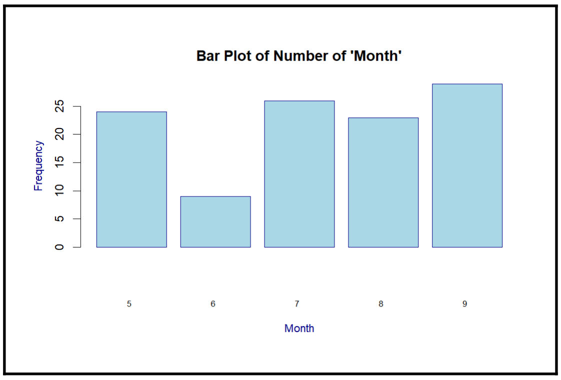
Boxplot of Numerical Variable (“Wind”):



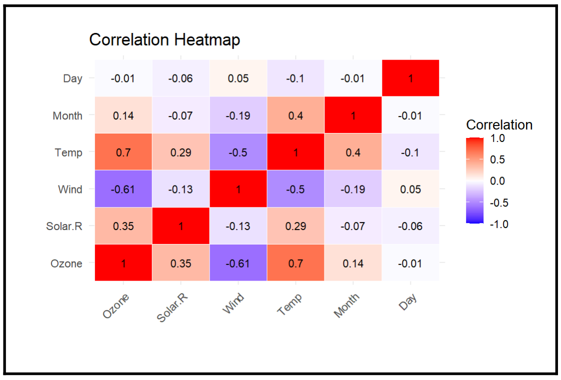
Interpretation:

- The distribution of Temp is Approximately Symmetric with a skewness value of -0.22 .
- The distribution of Ozone is Positively Skewed (Right-Skewed) with a skewness value of 1.23
- No outliers present along the temperature variable. There are outliers present along the ‘Wind’ feature as can be seen from the boxplots of the variables.

Barplot of the Categorical Variable ‘Month’

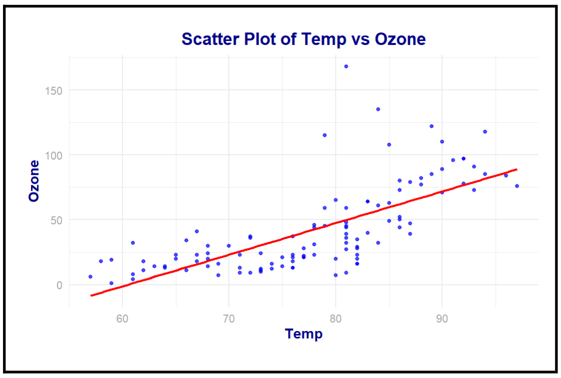


The barplot indicates a relatively even distribution of data across the months of May (5th), July (7th), August (8th), and September (9th). This suggests that the dataset provides balanced observations for these months, which can help ensure that any analysis or conclusions drawn are not biased toward a specific month.



- The Pearson Correlation Coefficient between Temp and Ozone: **0.70**
- There is a strong positive correlation between ozone levels and solar radiation, indicating that higher solar radiation contributes to increased ozone levels.
- A strong negative correlation exists between wind speed and temperature, suggesting that as wind speed increases, temperature tends to decrease.
- Moderate positive correlations are observed between ozone and temperature, and between solar radiation and temperature, implying that warmer temperatures might be associated with higher ozone levels and increased solar radiation.
- Most other variables exhibit weak or no correlation with each other, suggesting

limited linear relationships between them.

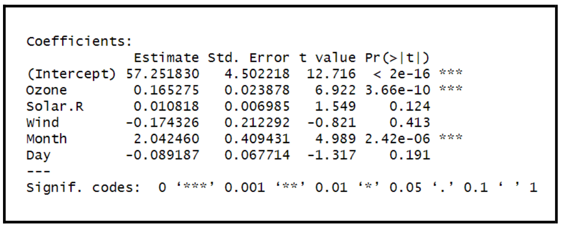


The scatter plot shows a positive linear relationship between temperature and ozone levels with some variability around the trend line. This suggests that temperature might be a contributing factor to ozone levels.

Multiple Linear Regression:

A multiple linear regression analysis was conducted to investigate the relationship between the continuous numerical variable "temp" and several predictors, like Ozone, Solar.R, Wind, and categorical variables (Month, Day).

Model Coefficients Estimates:



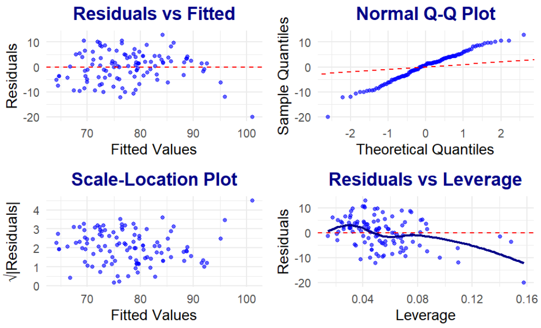
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 57.251830   4.502218  12.716  < 2e-16 ***
Ozone        0.165275   0.023878   6.922 3.66e-10 ***
Solar.R      0.010818   0.006985   1.549    0.124
Wind        -0.174326   0.212292  -0.821    0.413
Month        2.042460   0.409431   4.989 2.42e-06 ***
Day         -0.089187   0.067714  -1.317    0.191
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Model Summary	**Values**
Residual Standard Error	3221
Multiple R-squared	0.887
Adjusted R-squared	0.885
F-Statistic	484.1 on 7 and 432 DF
p-value	< 2.2e-16

Interpretations about the model:

- Ozone and Month significantly predict Temperature, with p-values of 3.66e-10 and 2.42e-06, respectively, showing positive relationships. Therefore we can focus on Ozone levels and seasonal patterns (Month) when building more complex models or creating actionable insights.
- Solar.R, Wind, and Day are not statistically significant, indicating weak linear relationships with Temperature.
- The model explains about 60% of the variability in Temperature (R-squared = 0.6013), and the overall model is statistically significant (*p-value < 2.2e-16*).
- A Residual Standard Error of 6.159 suggests moderate prediction accuracy, but there is room for improvement to capture unexplained variability.

Diagnostic Plots



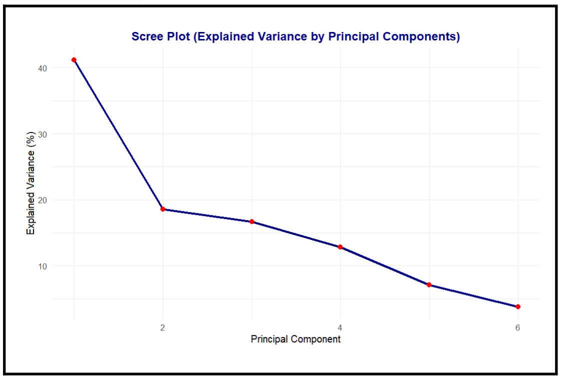
Interpretation from Diagnostic Plots:

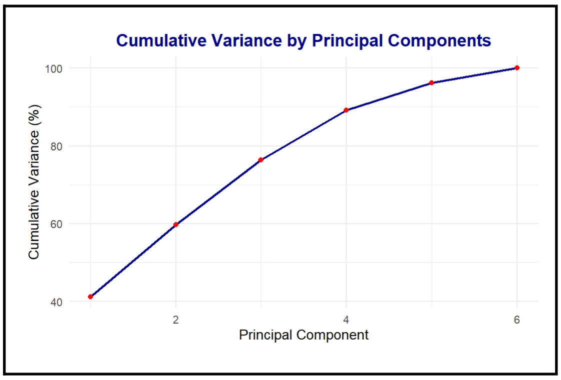
- Model Fit: The model appears to be a good fit for the data, as the residuals are randomly scattered and follow a normal distribution.
- Constant Variance: The model's assumption of constant variance is met, as indicated by the lack of patterns in the residuals vs fitted and scale-location plots.
- No Outliers: There are no influential outliers or high-leverage points that could significantly impact the model's predictions.
- Normality of Residuals: The normal Q-Q plot suggests that the residuals are normally distributed, which is a crucial assumption for hypothesis testing and confidence interval estimation**.**

Implementing PCA

PCA is applied to all 6 numerical features after scaling them, resulting in the generation of 6 principal components.

Scree Plot:

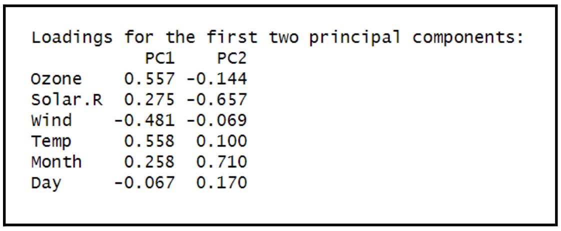




Interpretation for PCA:

- The scree plot indicates an elbow point at the second principal component, suggesting that the first two components capture a significant portion of the data's variance.
- The first two principal components together explain approximately 60% of the total variance.
- Reducing the dimensionality to the first two principal components might be reasonable to capture most of the original data's information.

Loadings of first 2 PCs:

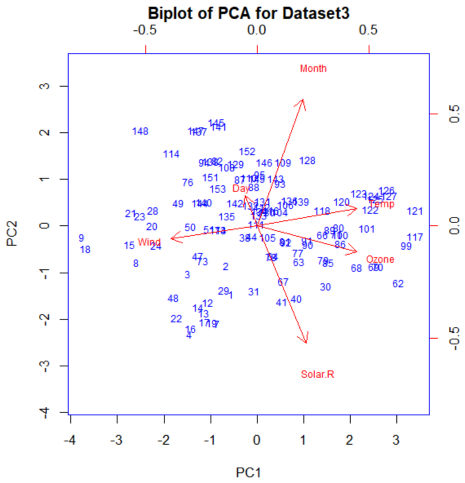


```
Loadings for the first two principal components:
           PC1     PC2
Ozone     0.557  -0.144
Solar.R   0.275  -0.657
Wind     -0.481  -0.069
Temp      0.558   0.100
Month     0.258   0.710
Day      -0.067   0.170
```

Interpretation:

- **PC1:** This component primarily represents a general pattern of atmospheric conditions, with ozone, temperature, and solar radiation positively correlated and wind negatively correlated.
- **PC2**: PC2 seems to capture seasonal variations, with the month of the year being the dominant factor influencing this component.

Biplot for the Dataset-3



Interpretation:

- Variable Relationships: Ozone, Solar.R, and Temp are positively correlated, while Wind is negatively correlated. Month is the primary driver of variation along PC2.
- Sample Clusters: Samples with similar values for Ozone, Solar.R, and Temp tend to cluster together.
- Principal Component Interpretation: PC1 represents general atmospheric conditions, while PC2 captures seasonal variations.

DATASET-4

- The dataset consists of 244 observations and 6 features.
- There are no missing values present in the dataset along any feature. Therefore preprocessing step is required for this dataset.
- There are 5 categorical variables and 2 continuous variables.
- The two numerical variables considered for further analysis for doing distribution analysis, visualization purposes, etc are 'total_bill'' and 'tip''.
- The categorical Variable being considered for analysis like for

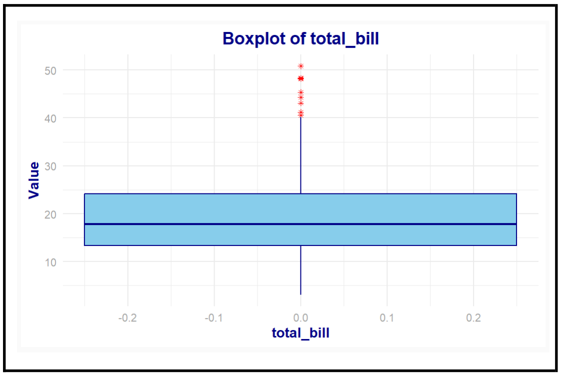
understanding the distribution, presence of outliers, etc, is ‘day’ and it only consists of 4 values, i.e., ‘sun’, ‘sat’, thur’, and ‘fri’.

.

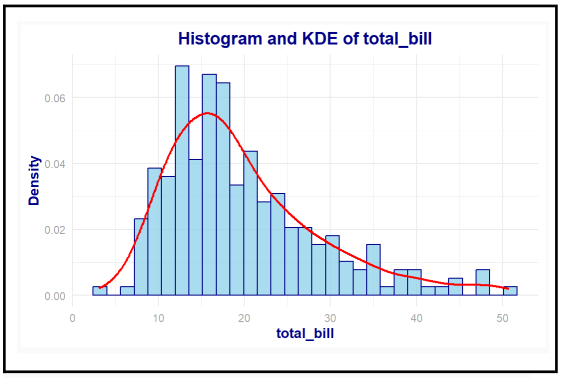
Summary Statistics of Numerical Variables:

Statistics	Total_bill	Tip	Size
Mean	19.79	3.00	2.57
Median	17.80	2.9	2
Standard Deviation	8.90	1.38	0.95
Minimum	3.07	1	1
Maximum	50.81	10	6

Boxplot of a Continuous Variable (‘total_bill’)



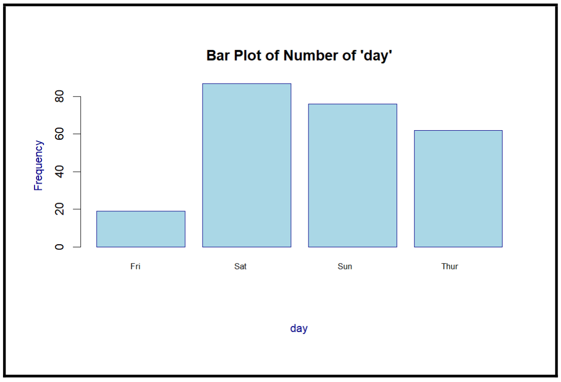
Histogram and Shape of a numerical variable (‘total_bill’)



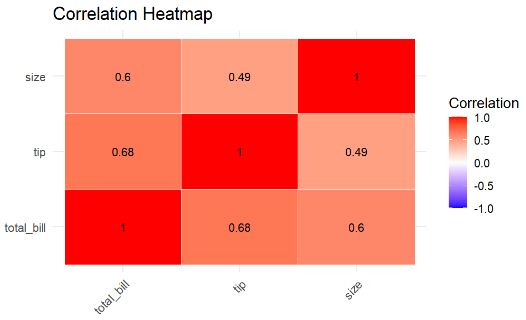
Interpretation for the variable ‘total_bill’

- The distribution of `total_bill` is right-skewed with a skewness of 1.12, which suggests that most customers tend to spend less, with a smaller proportion of customers spending significantly more.
- The peak of the distribution appears to be around the 15-20 range, indicating that this is the most common range for total bill amounts.
- The distribution is relatively spread out, suggesting that there is considerable variation in the total bill amounts.
- While there aren't any extreme outliers visible in the plot, there are some data points towards the higher end of the distribution that might be considered potential outliers. These could be due to factors like larger groups dining together or special occasions.

Barplot of a Categorical Variable:

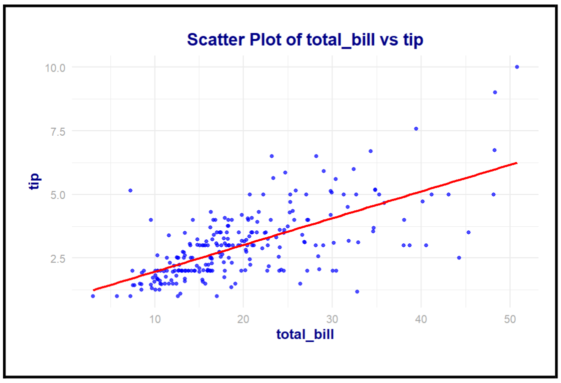


- The bar plot clearly shows a higher frequency of observations on Saturdays compared to both weekdays (Thursday and Friday). This reinforces the notion of a strong "weekend effect" in customer visits.



- The Pearson Correlation Coefficient between tip and total_bill : **0.68**
- Total bill and tip, and total bill and size are strongly positively correlated. This means larger bills tend to have larger tips and larger parties tend to spend more.

Scatterplot between Total_bill and Tip

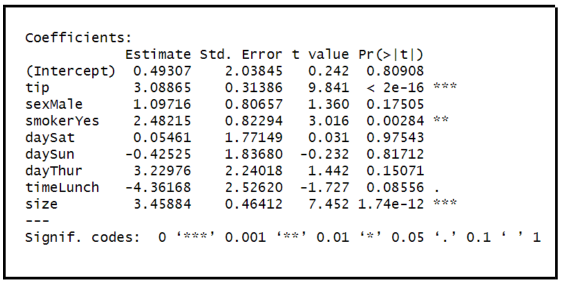


- The plot clearly shows a positive linear relationship between tip and total bill. This means that as the total bill amount increases, the tip amount also tends to increase.
- While there's a clear positive trend, the points are not perfectly aligned on the line. This indicates some variability in the relationship, with some bills leading to higher tips than others, even for similar bill amounts.

Multiple Linear Regression:

A multiple linear regression analysis was conducted to investigate the relationship between the variable ‘total_bill’ and several predictors, like tip, sex, smoker, day, time, size.

Model Coefficient Estimates:



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.49307    2.03845   0.242  0.80908    
tip          3.08865    0.31386   9.841  < 2e-16 ***
sexMale      1.09716    0.80657   1.360  0.17505    
smokerYes    2.48215    0.82294   3.016  0.00284 ** 
daySat       0.05461    1.77149   0.031  0.97543    
daySun      -0.42525    1.83680  -0.232  0.81712    
dayThur      3.22976    2.24018   1.442  0.15071    
timeLunch   -4.36168    2.52620  -1.727  0.08556 .  
size         3.45884    0.46412   7.452 1.74e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

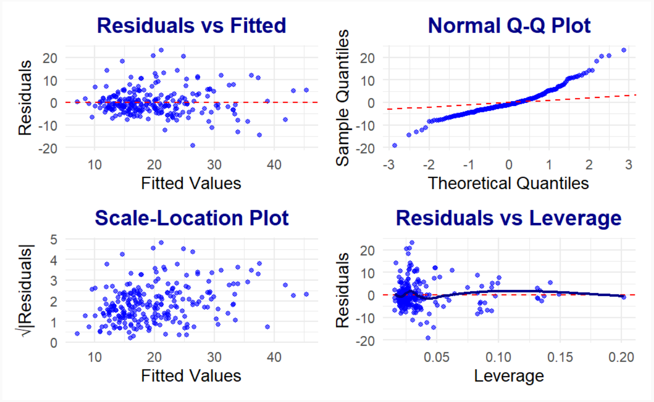
Model Summary:

Model Summary	Values
Residual Standard Error	5.856
Multiple R-squared	0.5816
Adjusted R-squared	0.5673
F-Statistic	40.83 on 8 and 235 DF
p-value	< 2.2e-16

Interpretation:

- **Significant Predictors:** The analysis reveals that tip, smokerYes , and size are significant predictors of total_bill, indicating that higher tips, smoking status, and larger party sizes lead to increased total bills.
- **Non-Significant Predictors**: Predictors such as sex, day, and time show no strong evidence of affecting total bills, with p-values indicating insignificance.
- **Model Fit and Significance:** The model explains approximately 58.16% of the variability in total bills (Multiple R-squared = 0.5816), with an overall significant regression model (F-statistic = 40.83, p < 2.2e-16), suggesting that at least one predictor significantly contributes to the outcome.
- **Residual Analysis:** The Residual Standard Error of 5.856 indicates variability in total bills not captured by the model, highlighting areas for potential model improvement or the inclusion of additional predictors.

Diagnostic Plot:



Interpretation:

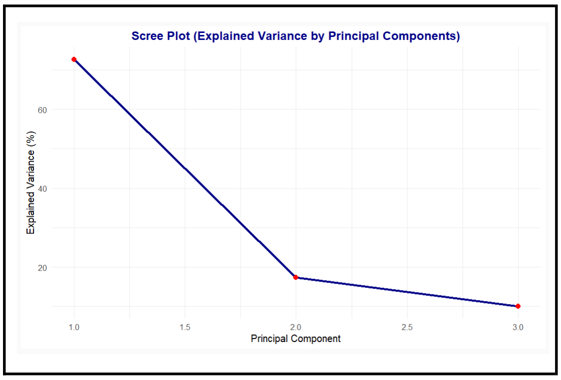
- The model appears to be a good fit for the data, as the residuals are randomly scattered and approximately normally distributed.
- The model's assumption of constant variance is met, as indicated by the lack of patterns in the residuals vs fitted and scale-location plots.
- There are no influential outliers or high-leverage points that could

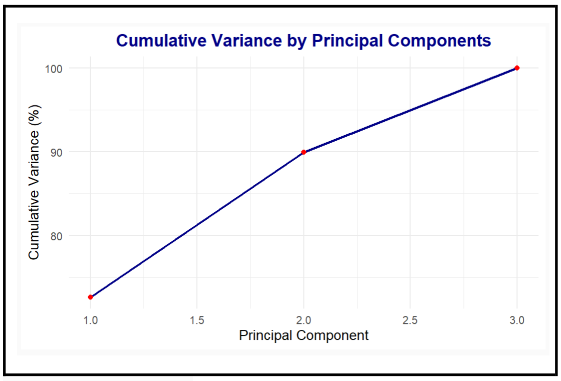
significantly impact the model's predictions.

- The normal Q-Q plot suggests that the residuals are approximately normally distributed, which is a crucial assumption for hypothesis testing and confidence interval estimation.

PCA Implementation:

PCA is applied to 3 numerical features (‘Total_bill’, ‘tip’, and ‘size’) after scaling them, resulting in the generation of 3 principal components.

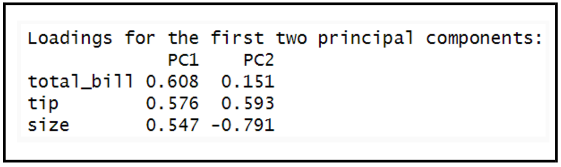




Interpretation:

- The scree plot exhibits an elbow point at the second principal component, suggesting that the first two components capture a significant portion of the total variance, approximately 80% of the total variance
- Reducing the dimensionality to the first two principal components might be reasonable to capture most of the original data's information.

Loadings for the first two PCs

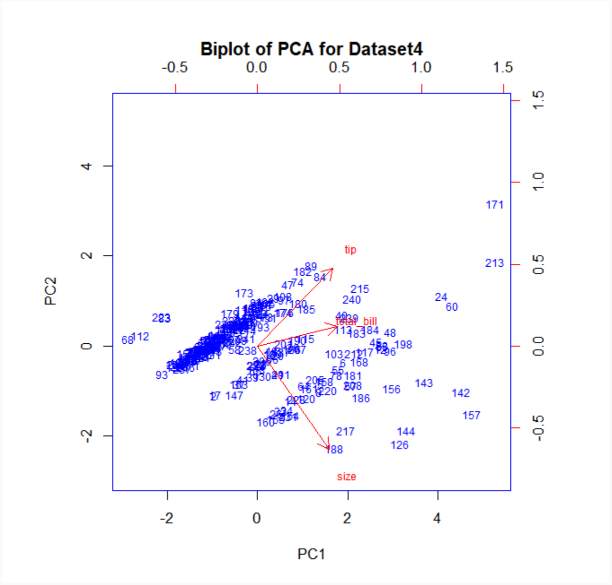


```
Loadings for the first two principal components:
            PC1    PC2
total_bill 0.608  0.151
tip        0.576  0.593
size       0.547 -0.791
```

Interpretation from the Loadings of first 2PCs:

PC1: All three variables (total_bill, tip, and size) have positive loadings on PC1. This suggests that PC1 represents a general pattern of spending, where higher values of these variables tend to occur together.

PC2: Tip has a high positive loading on PC2, indicating that it is strongly correlated with this component. Size has a high negative loading on PC2, suggesting that it is negatively correlated with tip. This might indicate that larger parties tend to tip less per person.



Interpretation:

- PC1: This component seems to represent a general pattern of spending, with variables like total bill and size being positively correlated.
- PC2: PC2 appears to capture additional information related to tipping behavior, with Tip being the dominant factor influencing this component.
- The samples appear to form clusters based on their total bill and size. Samples with similar total bill and size values tend to cluster together.
- Some samples, like those with high Tip values, might be considered outliers as they are located far from the main cluster.
- Total_bill and Size are positively correlated, as they are located in the same direction on the PC1 axis. Tip is also positively correlated with these two variables.