INDIAN STATISTICAL INSTITUTE

POST GRADUATE DIPLOMA IN BUSINESS ANALYTICS

Statistical Structures in Data

Numerical Assignment

Name: Anirudh A

Roll No: 24BM6JP06

Contents

Dataset 1: Air Quality

Univariate Analysis

1. Data Overview

There are a total of 153 observations with 6 variables.

```
> str(airquality)
'data.frame':   153 obs. of  6 variables:
 $ Ozone  : int  41 36 12 18 NA 28 23 19 8 NA ...
 $ Solar.R: int  190 118 149 313 NA NA 299 99 19 194 ...
 $ Wind   : num  7.4 8 12.6 11.5 14.3 14.9 8.6 13.8 20.1 8.6 ...
 $ Temp   : int  67 72 74 62 56 66 65 59 61 69 ...
 $ Month  : int  5 5 5 5 5 5 5 5 5 5 ...
 $ Day    : int  1 2 3 4 5 6 7 8 9 10 ...
```

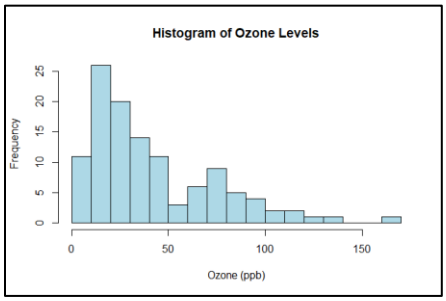
2. Summary Statistics

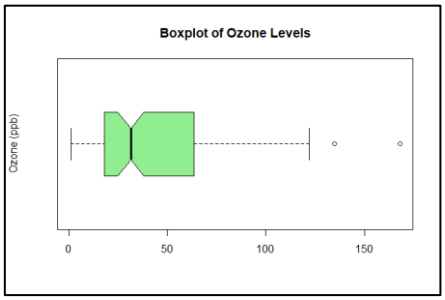
Numerical variable chosen: Ozone

Metric	Value
Mean	42.12931
Median	31.5
Standard Deviation	32.98788
Minimum	1
Maximum	168

- The mean is significantly higher than the median, which suggests that the dataset might be right skewed.
- The standard deviation of 32.99 is relatively high. Given the range of the dataset (1 - 168), it is expected. Most values would be clustered around the median, while a few extreme values create a large spread/dispersion in the data.

3. Distribution Visualization

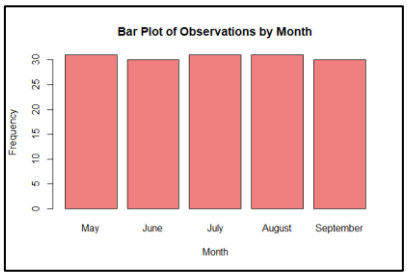




- The histogram shows a right-skewed distribution.
- There are 2 outliers in the chosen variable (Ozone), as shown by the boxplot. This has been verified with R code by manually calculating the number of outliers.

4. Categorical Variable Analysis

Categorical variable chosen: Month



- The bar plot shows the count of data points collected each month, from May to September.
- The bars are of roughly equal height (months with 31 days are slightly taller due to an additional observation), implying data was collected consistently across all months.

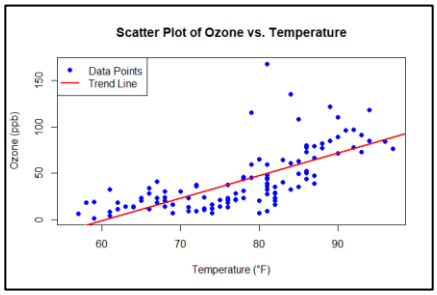
Multivariate Analysis

5. Correlation Analysis

Variables taken: Ozone and Temp

- **Correlation coefficient**: 0.6983603
- The correlation coefficient is **positive** and is **> 0.5**, implying a **moderately strong positive correlation**.
- Temperature can explain some of the variation in ozone levels, but not all of it.
- In practice, positive correlation between temperature and ozone is often expected because **higher temperatures can accelerate photochemical reactions** that lead to ozone formation, particularly in urban areas.

6. Scatter Plot Visualization



- The scatterplot displays a positive trend, implying that as temperature increases ozone levels also tend to increase, confirming the positive correlation observed previously.
- Even though there is a clear upward trend, the points are not forming a perfect straight line indicating that there is still variability not explained by temperature alone.

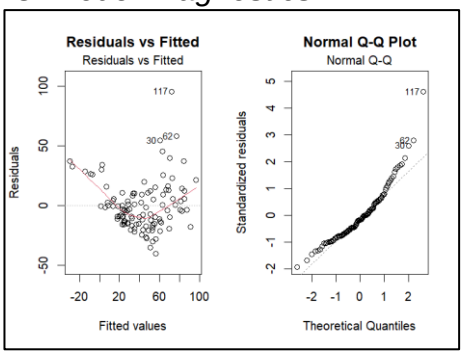
7. Multiple Regression

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -64.34208   23.05472  -2.791  0.00623 **
Temp          1.65209    0.25353   6.516 2.42e-09 ***
Wind         -3.33359    0.65441  -5.094 1.52e-06 ***
Solar.R       0.05982    0.02319   2.580  0.01124 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.18 on 107 degrees of freedom
  (42 observations deleted due to missingness)
Multiple R-squared:  0.6059,    Adjusted R-squared:  0.5948
F-statistic: 54.83 on 3 and 107 DF,  p-value: < 2.2e-16
```

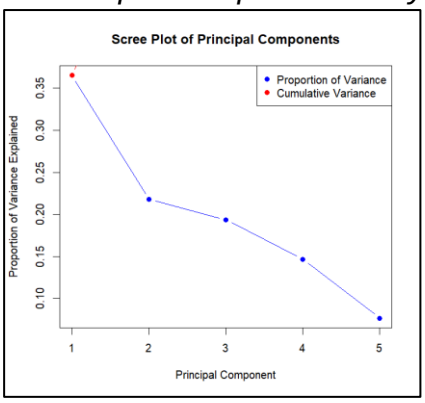
- Temperature: Statistically very significant (p<0.001) positive effect. For 1°F increase in temperature, ozone raises by 1.65 units.
- Wind: Statistically very significant (p<0.001) negative effect. For 1 unit increase in wind speed, ozone falls by 3.33 units.
- Solar.R: Statistically significant(p<0.05) positive effect.
- R squared of **0.6059** indicates that about **60.59%** of the variability in the dependent variable is explained by the model.

8. Model Diagnostics



- There is some non-linearity in the model, suggested by the slight curvature in the "Residuals vs Fitted" plot.
- The Q-Q plot shows deviations in the tails, indicating potential **non-normality**.
- The residuals generally appear to have constant variance **homoscedastic**), though there's a mild suggestion of increasing variance with higher fitted values.

9. Principal Component Analysis



- The first two principal components explain almost **60%** of the total variance.
- The curve appears to have an "elbow" around the 2nd/3rd component.
- The fourth and fifth components each explain a relatively small portion of the variance, as seen by the flatter slope. These components contribute less unique information to the dataset.
- Based on the elbow plot, **2/3 components** can be selected because they capture the largest changes in variance.

10. PCA Interpretation

- PC1 - **Temp** has a high negative loading. Wind has a high positive loading.
- PC2 - **Day** and **Solar.R** are the key contributors to this component. **Day's** vertical alignment indicates its independence from other variables.
- Patterns & Groupings: **Solar.R** and **Temp** are positively correlated while Wind is negatively correlated to both of them. Day is vertically aligned with respect to all 3 other variables, indicating it is independent of the other features and a temporal component.

Conclusion

- The AirQuality dataset reveals important trends in air pollution and weather. From univariate analysis, variables like solar and ozone exhibit variability with key outliers and a distribution that is right-skewed for Ozone. The month variable is used to capture seasonal patterns, with uneven distribution across the months to ensure seasonality.
- From the multivariate analysis, temperature and ozone have high positive associations, while wind and ozone have a significant negative link. Though the regression model accounts for 61% of the variation in ozone levels, it suggests that non-linear components would be needed to improve fit. By selecting 2 principal components (1 which reflects the correlations between Ozone, Temperature, and Wind, and 1 which reflects seasonal fluctuations via Day), PCA simplifies the dataset.

Dataset 2: Wine Quality

Univariate Analysis

1. Data Overview

There are a total of 1599 observations with 12 variables.

```
> str(winequality)
'data.frame':   1599 obs. of  12 variables:
 $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
 $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
 $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
 $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
 $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071 ...
 $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
 $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
 $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
 $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...
 $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...
 $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
 $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...
```

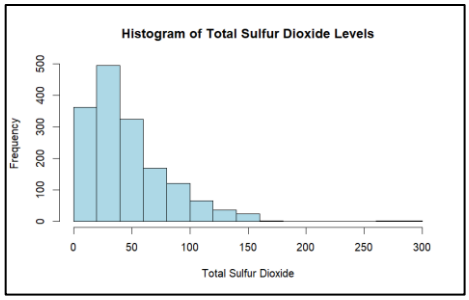
2. Summary Statistics

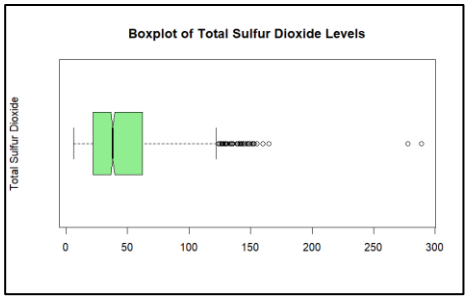
Numerical variable chosen: Total Sulfur Dioxide

Metric	Value
Mean	46.46779
Median	38
Standard Deviation	32.89532
Minimum	6
Maximum	289

- The mean being significantly higher than the median, suggests a **right-skewed distribution**, where few wines have much higher sulfur dioxide levels.
- The relatively large standard deviation compared to the mean indicates substantial variability in sulfur dioxide levels among the samples. Given the range of the dataset (6 - 289), it is expected. Most values would be clustered around the median, while a few extreme values create a large spread/dispersion in the data.

3. Distribution Visualization





- Total Sulfur Dioxide Levels exhibit a right-skewed distribution with several outliers at the higher end of the range.
- The data is predominantly concentrated between 0 and 100, while a small number of extreme values (outliers) extend the range up to approximately 300.

4. Categorical Variable Analysis

Categorical variable chosen: Quality

- The bar plot shows the count of data points by quality, from 3 to 8.
- The dataset primarily consists of wines with medium quality (5 and 6), while wines with low (3, 4) or high (7, 8) quality are less. This indicates that most wines in the dataset are average in quality.

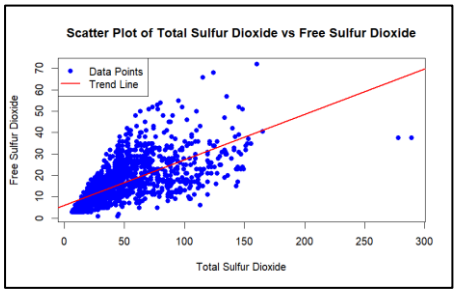
Multivariate Analysis

5. Correlation Analysis

Variables taken: Total Sulfur Dioxide and Free Sulfur Dioxide

- **Correlation coefficient**: 0.6676665
- The correlation coefficient is **positive** and is **> 0.5**, implying a **moderately strong positive correlation**.
- Free Sulfur Dioxide is likely a component of Total Sulfur Dioxide, so it makes sense that changes in one would reflect changes in the other.

6. Scatter Plot Visualization



- The scatterplot displays a positive trend, implying that as total sulfur dioxide increases free sulfur dioxide also tends to increase, confirming the positive correlation observed previously.
- Even though there is a clear upward trend, the points are not forming a perfect straight line indicating that there is still variability not explained by free sulfur dioxide alone.

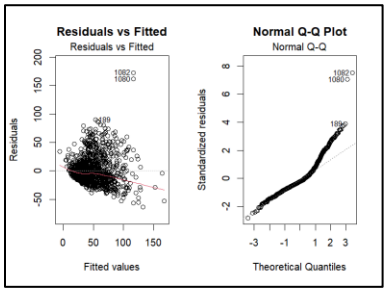
7. Multiple Regression

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         172.70843   20.00985   8.631   <2e-16 ***
fixed.acidity        -5.27611    0.53361  -9.888   <2e-16 ***
volatile.acidity     38.02875    3.93626   9.661   <2e-16 ***
citric.acid          44.64856    4.76119   9.378   <2e-16 ***
free.sulfur.dioxide   2.06840    0.05598  36.947   <2e-16 ***
pH                  -44.50223    5.18237  -8.587   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 23.07 on 1593 degrees of freedom
Multiple R-squared:  0.5099,    Adjusted R-squared:  0.5083
F-statistic: 331.4 on 5 and 1593 DF,  p-value: < 2.2e-16
```

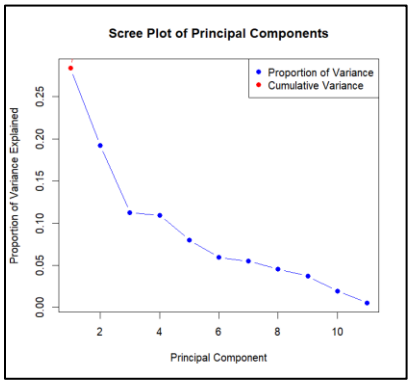
- All the variables available were used to fit the model, and the ones which had the least p-value were chosen as the predictors.
- R squared of **0.5099** indicates that about **50.99%** of the variability in the dependent variable is explained by the variables used.

8. Model Diagnostics



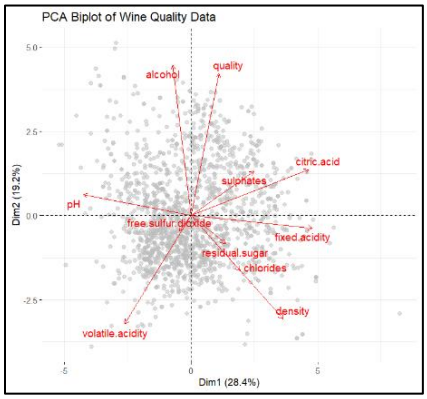
- Residuals show some patterns at lower and higher fitted values. This implies **heteroscedasticity** (variance not entirely constant).
- The Q-Q plot shows significant deviations at the tails. This indicates **deviation from normality**, especially due to potential outliers.

9. Principal Component Analysis



- The first two principal components explain almost **50%** of the total variance.
- The curve appears to have an "elbow" around the 3rd/4th component.
- The remaining components each explain a relatively small portion of the variance, as seen by the flatter slope. These components contribute less unique information to the dataset.
- Based on the elbow plot, 3/4 components can be selected because they capture the largest changes in variance.

10. PCA Interpretation



- PC1 - **Sulphates, citric acid** and **fixed acidity** have strong positive loadings, with pH having a strong negative loading indicating that these variables contribute significantly to the variation along this component.
- PC2 - Variables such as **alcohol** and **quality** dominate PC2 with positive loadings, with **density** and **volatile acidity** having considerable negative loadings.
- Patterns & Groupings: Wines with higher sulphates, fixed acidity, and citric acid tend to have lower volatile acidity and lower sulfur dioxide levels. The plot suggests distinct groupings of wines based on their chemical properties, which can be useful for classifying wines into quality categories.

Conclusion

- The univariate analysis of the UCL wine quality red dataset provides insights into the distribution and characteristics of individual variables. Visualization techniques like histograms and boxplots illustrated the distribution's positive skewness and highlighted extreme values, indicating the presence of outliers. Bar plots for wine quality revealed the overall spread of wines in the dataset, quality-wise.
- The multivariate analysis highlights a strong positive correlation between free sulfur dioxide and total sulfur dioxide, but the regression model was able to explain only 51% of the variability. Diagnostic checks confirm the presence of heteroscedasticity and non-linearity, highlighting the need to incorporate non-linearity into the model and that linear models cannot sufficiently explain the variance.

Dataset 3: Seoul Bike Data

Univariate Analysis

1. Data Overview

There is a total of 8760 observations with 14 variables.

```
> str(bike)
'data.frame':   8760 obs. of  14 variables:
 $ Date               : chr  "01-12-17" "01-12-17" "01-12-17" "01-12-17" ...
 $ Rented.Bike.Count  : int  254 204 173 107 78 100 181 460 930 490 ...
 $ Hour               : int  0 1 2 3 4 5 6 7 8 9 ...
 $ Temperature        : num  -5.2 -5.5 -6 -6.2 -6 -6.4 -6.6 -7.4 -7.6 -6.5 ...
 $ Humidity           : int  37 38 39 40 36 37 35 38 37 27 ...
 $ WindSpeed          : num  2.2 0.8 1 0.9 2.3 1.5 1.3 0.9 1.1 0.5 ...
 $ Visibility         : int  2000 2000 2000 2000 2000 2000 2000 2000 2000 1928 ...
 $ DewPointTemperature: num  -17.6 -17.6 -17.7 -17.6 -18.6 -18.7 -19.5 -19.3 -19.8 -22.4 ...
 $ SolarRadiation     : num  0 0 0 0 0 0 0 0 0.01 0.23 ...
 $ Rainfall           : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Snowfall           : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Seasons            : chr  "Winter" "Winter" "Winter" "Winter" ...
 $ Holiday            : chr  "No Holiday" "No Holiday" "No Holiday" "No Holiday" ...
 $ Functioning.Day    : chr  "Yes" "Yes" "Yes" "Yes" ...
```

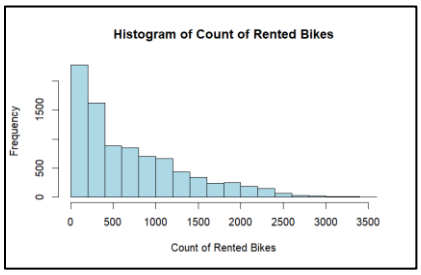
2. Summary Statistics

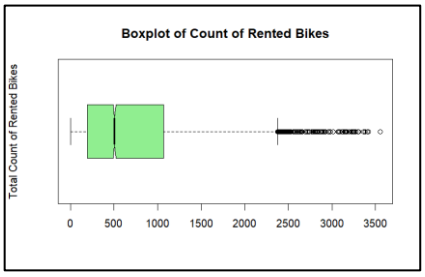
Numerical variable chosen: Rented Bike Count

Metric	Value
Mean	704.6021
Median	504.5
Standard Deviation	644.9975
Minimum	0
Maximum	3556

- The mean being significantly higher than the median, suggests a **right-skewed distribution**.
- The large standard deviation compared to the mean indicates high variability in count of rented bikes. The range from 0 to 3556 suggests varied demand patterns, which could be influenced by factors like weather, time of day, or day of the week.

3. Distribution Visualization

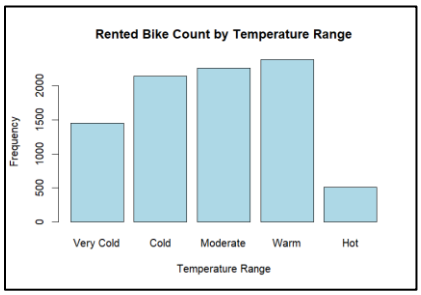




- The distribution of the count of rented bikes is **right-skewed**, as the majority of the data is concentrated on the lower end, and the tail stretches towards higher values.
- The boxplot confirms the right-skewed nature of the data since the whisker on the upper end (right side) is longer. There are outliers on the higher end, representing days with unusually high bike rentals (above 3000).

4. Categorical Variable Analysis

Categorical variable chosen: Temperature Range



- Bike rentals are highest in moderate and warm conditions, indicating that mild and pleasant weather drives demand.
- Hot and very cold conditions lead to reduced bike rentals, possibly due to the physical discomfort of biking in extreme weather.

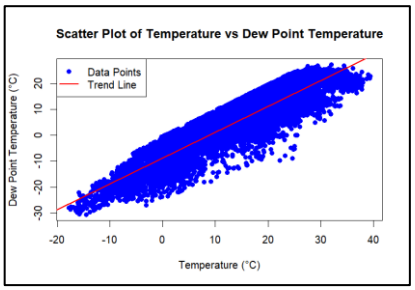
Multivariate Analysis

5. Correlation Analysis

Variables taken: Temperature and Dew Point Temperature

- **Correlation coefficient**: 0.9127982
- The correlation coefficient is **positive** and is **> 0.9**, implying a **highly strong positive correlation**.
- Dew point temperature generally increases as air temperature increases. This is because warmer air can hold more moisture, raising the dew point when the air contains significant water vapor.

6. Scatter Plot Visualization



- The scatterplot displays a **strong positive trend**, implying that as temperature increases dew point temperature also tends to increase, confirming the positive correlation observed previously.
- This relationship is expected since higher temperatures can hold more moisture, increasing the dew point.

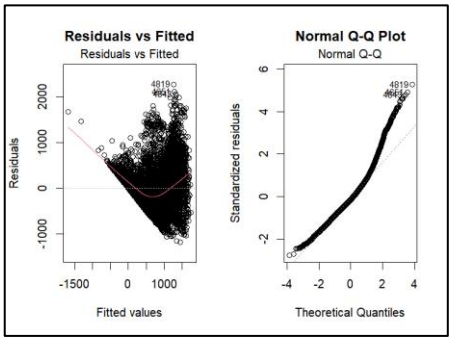
7. Multiple Regression

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         -293.1022    41.4503  -7.071 1.65e-12 ***
Hour                  27.3962     0.7328  37.385  < 2e-16 ***
Temperature           26.4276     0.8696  30.390  < 2e-16 ***
Humidity              -8.2589     0.3054 -27.043  < 2e-16 ***
WindSpeed             19.2092     5.0722   3.787 0.000153 ***
SolarRadiation       -83.8503     7.2551 -11.557  < 2e-16 ***
Rainfall             -60.1243     4.2402 -14.180  < 2e-16 ***
Snowfall              29.8882    11.1747   2.675 0.007495 **
SeasonsSpring       -140.4130    13.5229 -10.383  < 2e-16 ***
SeasonsSummer       -148.9070    17.1237  -8.696  < 2e-16 ***
SeasonsWinter       -371.3057    19.3879 -19.151  < 2e-16 ***
HolidayNo Holiday    116.5347    21.6099   5.393 7.12e-08 ***
Functioning.DayYes   929.3409    26.6494  34.873  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 433 on 8747 degrees of freedom
Multiple R-squared:  0.5499,    Adjusted R-squared:  0.5493
F-statistic: 890.6 on 12 and 8747 DF,  p-value: < 2.2e-16
```

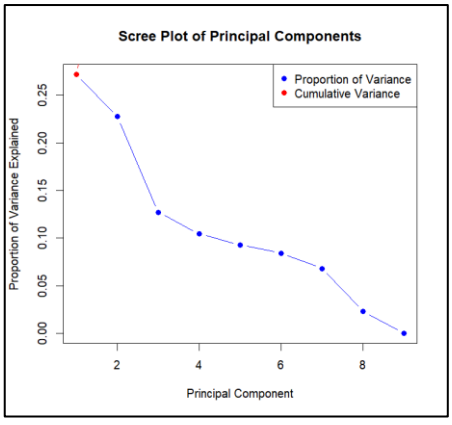
- All the variables available were used to fit the model, and the ones which had the least p-value were chosen as the predictors.
- R squared of **0.5499** indicates that about **54.99%** of the variability in the dependent variable is explained by the model.
- The moderately low R-squared indicates the presence of non-linearity in the dataset, and hence a non-linear model/component should be incorporated for better accuracy.

8. Model Diagnostics



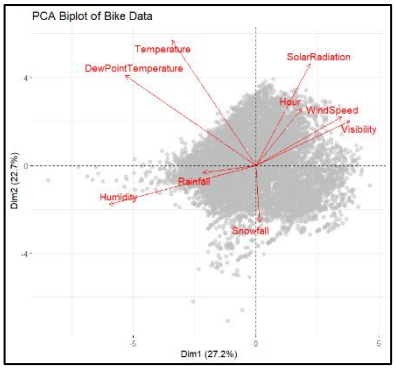
- The residuals are not randomly scattered around the horizontal line. There is a clear pattern, with increased spread at higher fitted values, implying **heteroscedasticity**.
- Residuals deviate from the diagonal line at both tails, suggesting that the residuals are **not normally distributed**.

9. Principal Component Analysis



- The first two principal components explain almost **50%** of the total variance.
- The curve has an elbow around the 3rd component.
- The remaining components explain a much smaller portion of the variance.
- Based on the elbow plot, 3 components can be selected because they capture the largest changes in variance.

10. PCA Interpretation



- PC1 - High negative loadings of **humidity** and **rainfall** indicate that these variables contribute significantly to PC1. **Visibility** and **Wind Speed** have moderate contributions, also aligning with this dimension.
- PC2 - **Temperature**, **Solar Radiation** and **Hour** have strong positive loadings, suggesting that warmer, sunnier times of the day are associated with higher bike rentals. Strong negative loading of snowfall reflects its inverse relationship with rented bike count.
- Patterns & Groupings - Bike rentals are positively associated with higher temperatures, solar radiation, and specific times of the day(hour), indicating favorable conditions for biking, and negatively affected by factors like snowfall, rainfall, and humidity, which likely discourage outdoor activities.

Conclusion

- Univariate analysis of the Bike Dataset reveals patterns and trends in bike rentals. The distribution of rented bike count is right skewed, as observed from the histogram and the boxplot. It is also confirmed from the summary statistics by the fact that mean is significantly greater than the median. A bar plot against temperature revealed that bike rentals were high in moderate and warm conditions, and dropped in extreme conditions.
- Multivariate analysis revealed strong positive correlations between Temperature and Dew Point Temperature. The regression model was able to explain 55% of the variance in bike rentals, suggesting the need to incorporate non-linearity along with linear components for better results.

Dataset 4: Hitters Data

Univariate Analysis

1. Data Overview

There is a total of 322 observations with 20 variables.

```
> str(Hitters)
'data.frame':   322 obs. of  20 variables:
 $ AtBat    : int  293 315 479 496 321 594 185 298 323 401 ...
 $ Hits     : int  66 81 130 141 87 169 37 73 81 92 ...
 $ HmRun    : int  1 7 18 20 10 4 1 0 6 17 ...
 $ Runs     : int  30 24 66 65 39 74 23 24 26 49 ...
 $ RBI      : int  29 38 72 78 42 51 8 24 32 66 ...
 $ Walks    : int  14 39 76 37 30 35 21 7 8 65 ...
 $ Years    : int  1 14 3 11 2 11 2 3 2 13 ...
 $ CAtBat   : int  293 3449 1624 5628 396 4408 214 509 341 5206 ...
 $ CHits    : int  66 835 457 1575 101 1133 42 108 86 1332 ...
 $ CHmRun   : int  1 69 63 225 12 19 1 0 6 253 ...
 $ CRuns    : int  30 321 224 828 48 501 30 41 32 784 ...
 $ CRBI     : int  29 414 266 838 46 336 9 37 34 890 ...
 $ CWalks   : int  14 375 263 354 33 194 24 12 8 866 ...
 $ League   : Factor w/ 2 levels "A","N": 1 2 1 2 2 1 2 1 2 1 ...
 $ Division : Factor w/ 2 levels "E","W": 1 2 2 1 1 2 1 2 2 1 ...
 $ PutOuts  : int  446 632 880 200 805 282 76 121 143 0 ...
 $ Assists  : int  33 43 82 11 40 421 127 283 290 0 ...
 $ Errors   : int  20 10 14 3 4 25 7 9 19 0 ...
 $ Salary   : num  NA 475 480 500 91.5 750 70 100 75 1100 ...
 $ NewLeague: Factor w/ 2 levels "A","N": 1 2 1 2 2 1 1 1 2 1 ...
```

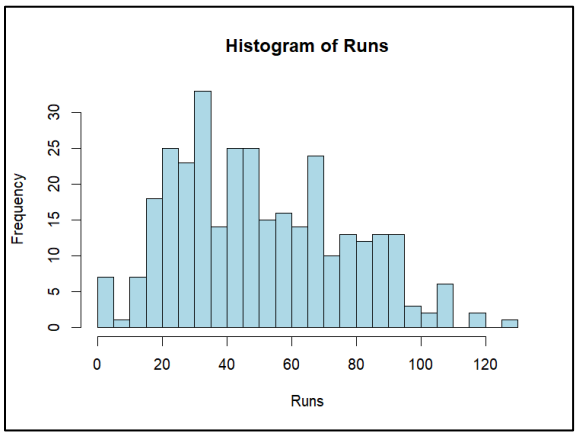
2. Summary Statistics

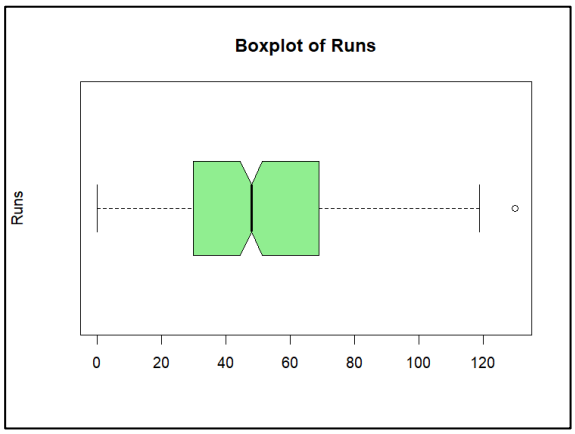
Numerical variable chosen: Runs

Metric	Value
Mean	50.90994
Median	48
Standard Deviation	26.0241
Minimum	0
Maximum	130
Skewness	0.41384
Kurtosis	2.471173

- The mean is quite close to the median, which suggests that the dataset might be normal. The skewness (close to 0) and kurtosis (close to 3) suggest that the Runs variable is approximately normal, though it is slightly skewed to the right with lighter tails.
- There is some variability in runs, with a few high-scoring players pulling the mean slightly upward.

3. Distribution Visualization

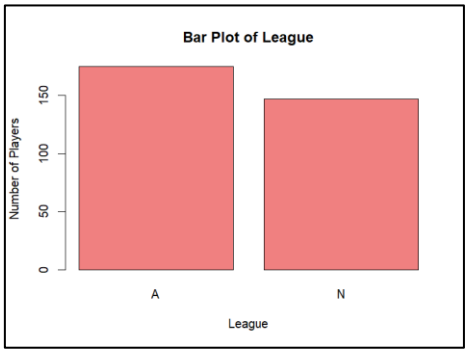




- The distribution of the runs is normal with a very slight skewness towards the right with lighter tails. Most players scored runs between 20 and 80.
- There are fewer players scoring extremely high runs (above 100).
- One point outside the whisker on the right side (above 120 runs) represents an outlier. This suggests that a few players have significantly higher run counts than the rest of the dataset.

4. Categorical Variable Analysis

Categorical variable chosen: League



- The dataset has a reasonably balanced representation of players across both leagues, which is helpful for general comparisons.
- However, the slight overrepresentation of the American League might introduce a bias if not accounted for in league-based analyses.

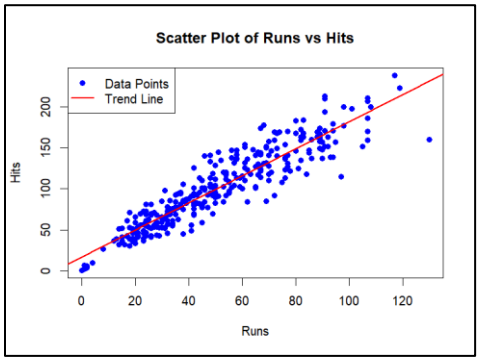
Multivariate Analysis

5. Correlation Analysis

Variables taken: Runs and Hits

- **Correlation coefficient**: 0.9221872
- The correlation coefficient is **positive** and is **> 0.9**, implying a **highly strong positive correlation**.
- In baseball, hits directly contribute to a team's ability to score runs. This high correlation reflects the logical connection between these two metrics, as more successful hits give players opportunities to reach bases and eventually score runs.

6. Scatter Plot Visualization



- The scatterplot displays a strong positive trend, implying that as the number of hits increases, runs also tend to increase, confirming the positive correlation observed previously.
- There are a few points that deviate slightly from the trend line (e.g., players with relatively more hits but fewer runs). These could indicate players who perform well in hitting but whose hits do not consistently translate into runs due to other factors.

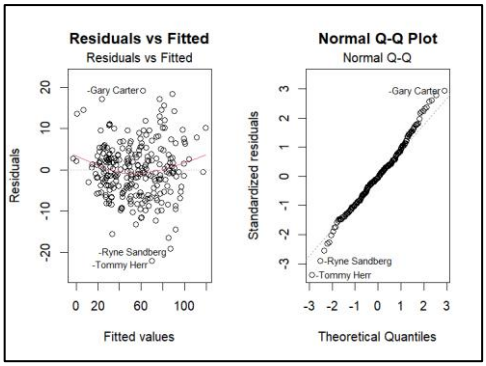
7. Multiple Regression

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.985382   1.261666  -1.574 0.116835
Hits         0.445709   0.021032  21.192  < 2e-16 ***
HmRun        0.899343   0.118427   7.594 6.11e-13 ***
RBI         -0.206392   0.053031  -3.892 0.000128 ***
Walks        0.262658   0.032879   7.989 4.95e-14 ***
CAtBat       0.007412   0.002319   3.195 0.001575 **
CHits       -0.094230   0.012094  -7.792 1.75e-13 ***
CHmRun      -0.177616   0.032522  -5.461 1.14e-07 ***
CRuns        0.133887   0.013001  10.298  < 2e-16 ***
CRBI         0.060029   0.014225   4.220 3.42e-05 ***
CWalks      -0.033573   0.006354  -5.284 2.75e-07 ***
Assists     -0.006386   0.003483  -1.833 0.067936 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.751 on 251 degrees of freedom
Multiple R-squared:  0.9331,    Adjusted R-squared:  0.9301
F-statistic:   318 on 11 and 251 DF,  p-value: < 2.2e-16
```

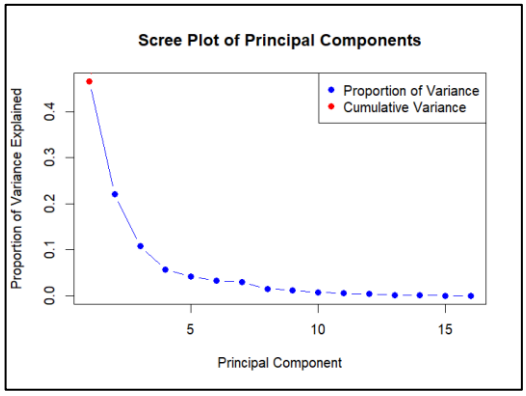
- Initially, a linear regression model is fitted to predict Runs using all available predictors. Stepwise regression is then applied to refine the model by selecting a subset of.
- Significant predictors with p-values ≤ 0.05 are identified. A new reduced model is then fit using only the significant predictors, resulting in a simplified and interpretable model that focuses on the most influential variables for predicting Runs.

8. Model Diagnostics



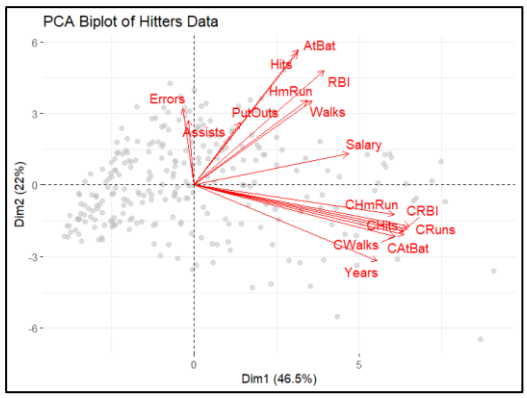
- The residuals are randomly scattered around the horizontal line at 0. So, while the linearity assumption is reasonable, there may be slight signs of **heteroscedasticity** (some spread in residuals increases with fitted values).
- Residuals do not deviate much from the diagonal line at both tails, suggesting that the residuals are **largely normally distributed**, with minor deviations at the tails.

9. Principal Component Analysis



- The first two principal components explain almost 70% of the total variance.
- The curve has an elbow around the 4th/5th component.
- The remaining components explain a much smaller portion of the variance.
- Based on the elbow plot, 4/5 components can be selected because they capture the largest changes in variance.

10. PCA Interpretation



- PC1 - High positive loadings of **CAtBat, CHits, CRuns, CRBI, CWalks, Years** and **Salary** indicate that these variables contribute significantly to PC1. They cluster together and are strongly correlated to each other.
- PC2 - **Errors** and **Assists** have strong positive loadings along PC2. **AtBat** and **Hits** have moderate contributions, but not as significant as the defensive stats.
- Patterns & Groupings - PC1 primarily captures a player's cumulative career performance and their associated salary. Players with high career stats tend to have higher salaries and longer careers. PC2 captures aspects of defensive performance and variability in game actions.

Conclusion

- Univariate analysis of the Hitters data reveals patterns and trends in Major League Baseball Data. The distribution of runs is approximately normal, as observed from the histogram and the boxplot. It is also confirmed from the summary statistics by the fact that mean is quite close to the median. A bar plot reveals one league being slightly over-represented than the other.
- Multivariate analysis revealed strong positive correlations between Runs and Hits. The regression model was able to explain 93% of the variance in Runs, suggesting that there is a very strong linear relationship and the model is adequately able to capture the patterns and structure of the data.