**R PRACTICE REPORT**

**PROBABILITY AND STATISTICS ASSIGNMENT**

**WEEK 2**

**MODULE 2**

By : **HARSHIT GAUR**

MASTER OF PROFESSIONAL STUDIES IN ANALYTICS

ALY 6010 : PROBABILITY THEORY AND INTRODUCTORY STATISTICS

NOVEMBER 14, 2021

To : **PROF. AMIN KARIMPOUR**

# **ABSTRACT**

**Data analytics** is a discipline focused on extracting insights from data. It comprises of the processes, tools and techniques of data analysis and management, including the collection, organization, and storage of data. The chief aim of data analytics is to apply statistical analysis and technologies on data in order to find trends and solve problems.

**Sales analytics** is the practice of generating insights from the sales data, trends, and metrics and utilise them to set targets and forecast future sales performance. The best practice for sales analytics is to closely tie all activities to determine revenue outcomes and set objectives for your sales team.

Analysis should focus on improvement and developing a strategy for improving your sales performance in both the short-terms and long-terms.

It provides insights about the top performing and underperforming products/services, the problems in sale-and-market opportunities, sales forecasting, and sales activities that generate revenue.

# **INTRODUCTION**

It is strongly encouraged to find and choose a data set in an area where one is more interested and is personally motivated to explore about.

Initially, I decided to go with a dataset that will help me hone the skills that I have learned so far in this course and go even beyond it by acquiring more knowledge and broadening my prowess of the analytical skills and R programming language. Then, I came across a dataset which not only interested me very much but also gave me various ideas to implement using that data set.   
  
 I grew up watching my all-time favourite TV Show (Anime) - Pokémon and the dataset which I chose is about the different types of Pokémon present along with their attributes. The dataset contains various numerical and categorical data. It has 1,045 data points with 11 features related to the primary types, attack attributes, defence attributes, hit points, special attributes and more. I decided to put my soft spot for the series to better use and up-skill myself in the analytical, visualization, and programmatical aspects of the domain.

The features available in the data set are -

|  |  |  |
| --- | --- | --- |
| **S. No.** | **Feature** | **Dictionary** |
| 1. | Name | Name of the Pokémon |
| 2. | Name2 | Secondary name of the Pokémon |
| 3. | Primary.Type | Primary type |
| 4. | Secondary.Type | Secondary type to which Pokémon belongs to |
| 5. | Attack | Attack attribute |
| 6. | Defense | Defence Quality |
| 7. | HP | Hit Points |
| 8. | Sp.Attack | Special Attack attribute |
| 9. | Sp.Defense | Special Defense attribute |
| 10. | Speed | Speed of the Pokémon |
| 11. | Total | Total Qualities (Summation) |

*Table 1: Features of the data set with their dictionary.*

The data set was obtained from the below URL and will be referred in the bibliography as well :

*The World of Pokemons*. (2021, September 29). Kaggle. https://www.kaggle.com/hamdallak/the-world-of-pokemons

From the structure of the data set, the features, their types, and values can be determined.

|  |
| --- |
| *Figure 1: Structure of the data set.* |

Some of the data points from the summary of the data set are present below.

|  |
| --- |
| *Figure 1: Structure of the data set.* |

# **DATA PRE-PROCESSING AND CLEANING**

The first blush of the data set portrayed the data as not clean. After an **Initial Data Analysis (IDA)** process assisted with some graph visualizations as well, I found out that the data set requires pre-processing and cleaning in it.

Data cleaning is an important & necessary factor in the data analysis process. As the famous quote says - *"Garbage In, Garbage Out."*

I proceeded with the now needed & necessary step of data cleaning using RStudio application on this data set to wrangle and eliminate all the garbage values which consisted of :

* Missing values
* Duplicate values

Some of the actions performed for data cleaning are :

* 1. The categorical features in the data set were checked for inconsistencies in them. If there were some found, we would need to normalize them using *if..elseif..else* condition with data manipulation.

|  |
| --- |
| *Figure 2: Inconsistencies Check on the feature 'Primary Type' in the data set.* |

* 1. We checked the data set for NA and NULL values in its features using the functions which can be referred from the below snapshot.

|  |
| --- |
| *Figure 3: Checking records with NA, NULL values from the data set.* |

* 1. There were many data points with missing or blank values in the features 'Name2' and 'Secondary.Type'. Removing these data points from the data set made no sense as it would have lost us a majority of the data. Instead, the missing or empty values were replaced by the word "Unknown" using a function ***GSUB()***.

|  |
| --- |
| *Figure 4: Eliminating records with NA, missing values from the data set.* |

* 1. The data set was checked for duplicate values in the combination of 'Name' and 'Name2' because of their uniqueness when combined. We found some duplicate values from the output below where "TRUE" s written.

|  |
| --- |
| Text  Description automatically generated  *Figure 5: Check for Duplication in the combination of features from the data set.* |

* 1. The data set was checked for duplicate values in the combination of 'Name' and 'Name2' because of their uniqueness when combined. We found some duplicate values from the output below where "TRUE" is written.

|  |
| --- |
| *Figure 6: Retrieving the duplicate records from the data set.* |

* 1. The duplicate values found in the above step were eliminated from the data set using **'filter'** function of the ***dplyr*** library.

|  |
| --- |
| *Figure 7: Eliminating the duplicate records from the data set.* |

* 1. After performing the previous steps, the data set was omitted for NA values if present using the below function **'na.omit'**.

|  |
| --- |
| *Figure 8: Removing the NA, missing values from the data set.* |

* 1. The data set do not contain any kind of unbalanced features.

# **EXPLORATORY DATA ANALYSIS**

The descriptive statistics of the features of the data set can be summarised to calculate the statistics -

|  |
| --- |
| *Figure 9: Summary of the data set.* |

We used the **'DESCRIBE()**' function from the package **'PSYCH'** to find out the descriptive statistics of features of the data set. The following observations can be made using the statistics found in summary of the data set -

1. The **mean** of the attribute *Attack* is around **80.47** with a **standard deviation** of **32.41** and **quartiles value (Lower Quartile - 55, Higher Quartile - 100).**   
   From the observations, we can calculate that the data points around **maximum value (190)** can be *outliers* to the feature.
2. The **mean** of the attribute *Defense* is around **74.66** with a **standard deviation** of **31.24** and **quartiles value (Lower Quartile - 50, Higher Quartile - 90).**   
   From the observations, we can calculate that the data points around **maximum value (250)** can be *outliers* to the feature.
3. The **mean** of the attribute *HP (Hit Point)* is around **70.07** with a **standard deviation** of **26.67** and **quartiles value (Lower Quartile - 50, Higher Quartile - 82).**   
   From the observations, we can calculate that the data points around **maximum value (255)** can be *outliers* to the feature.
4. The **mean** of the attribute *Sp.Attack (Special Attack)* is around **73.02** with a **standard deviation** of **32.73** and **quartiles value (Lower Quartile - 50, Higher Quartile - 95).**   
   From the observations, we can calculate that the data points around **maximum value (194)** can be *outliers* to the feature.
5. The **mean** of the attribute *Sp.Defense (Special Defense)* is around **72.29** with a **standard deviation** of **28.07** and **quartiles value (Lower Quartile - 50, Higher Quartile - 90).**   
   We can calculate that data points around **maximum value (250)** can be *outliers* in it.
6. The **mean** of the attribute *Speed* is around **68.80** with a **standard deviation** of **30.21** and **quartiles value (Lower Quartile - 45, Higher Quartile - 90).**   
   From the observations, we can calculate that the data points around **maximum value (200)** can be *outliers* to the feature.
7. The **mean** of the attribute *Total (Sum of all attributes)* is around **439.32** with a **standard deviation** of **121.97** and **quartiles value (Lower Quartile - 330, Higher Quartile - 515).**   
   From the observations, we can calculate that the data points around **maximum value (1125)** can be *outliers* to the feature.

# NORMALITY

The normality of all the features of the data set was checked in order to understand the type of distribution -

|  |
| --- |
| *Figure 10: Density Plot of features of the data set.* |
| *Figure 11: Shapiro-Wilks Test of the data set.* |
| *Figure 12: Q-Q Plots of the numeric features for Normality.* |

The density plots, Shapiro-Wilks Test, and the Quantiles-Quantiles plots of the features gave out the below results about the data set -

1. The density graphs of all the features show that they can be considered as normally distributed.
2. **Shapiro-Wilks Test** was used to verify the normality of a feature of the data set. But, since the p-value was less than 5%, we could verify the normality of the feature.
3. The features *Attack, Sp.Attack, Speed* have the same line for Sample and Theoretical Quantiles. They can be considered to be normally distributed.
4. The other features will need to be treated to make them normally distributed.

# DESCRIPTIVE STATISTICS

After the process of data pre-processing and cleaning, I have calculated the descriptive statistics of all the features of the data set to check the difference between the previous set and this new set of descriptive statistics. The function **DESCRIBE()** of the package **{PSYCH}**was used to calculate them. I have included *Quantiles* and *Inter-Quartile Range* columns as well.

|  |
| --- |
| *Figure 12: Describe the data set for descriptive statistics..* |
| *Figure 13: Descriptive Statistics after cleaning the data set.* |

1. We have eliminated some duplicate records from the data set. This changed the descriptive statistics values.
2. The **mean** of the attribute *Attack* is around **80.49** instead of **80.47** (earlier) with changes in the *standard deviation.*
3. The **mean** of the attribute *Defense* is around **74.70** instead of **74.66** (earlier) with a slight change of 0.02 in **standard .**
4. Few other statistics have been changed which can be figured out from the table above.

# DESCRIPTIVE STATISTICS (by Grouping)

|  |
| --- |
| *Figure 14: Describe the data set with GROUPINGS (based on features)* |
| *Table  Description automatically generated*  *Figure 15: Descriptive Statistics by GROUPS (based on Primary Attack feature)* |
| *A screenshot of a computer  Description automatically generated with low confidence*  *Figure 16: Descriptive Statistics by GROUPS (based on Primary Attack & Secondary Attack features)* |

1. The descriptive statistics of the data set was calculated after applying **GROUPINGs** based on *Primary Type* and *Secondary Type*.
2. There are 18 groups for which these descriptive statistics were calculated.
3. The mean of the *Attack* feature from the groups above shows that :

|  |  |  |
| --- | --- | --- |
| **GROUP** | **MEAN (*ATTACK*)** | **RANKING (attack based)** |
| BUG | 71.07 | 3RD |
| DARK | 81.93 | 2ND |
| DRAGON | 107.02 | 1ST |

1. The standard deviation of the *HP (Hit Power)* feature from the groups above shows that :

|  |  |  |
| --- | --- | --- |
| **GROUP** | **S.D. (*Hit Power*)** | **RANKING (HP based)** |
| BUG | 17.34 | 3RD |
| DARK | 31.63 | 2ND |
| DRAGON | 37.13 | 1ST |

1. Similar type of observations can be made from the statistics table above.

**Primary Type vs Summation of Attributes of Pokémon**

The below plot can help in visualising the summation of all the attributes of Pokémons grouped according to their Primary Type (Species)

|  |
| --- |
| *Figure 17: Box Plot of Bakery Items vs Promotion Types* |

1. From the grouped line graph, we can figure out that Pokémons of the *primary type (Species)* - **WATER** in the data set have the **cumulative largest abilities** in terms of all the attributes (*Attack, Defense, Hit Point, Special Attack, Special Defense, Speed*) amongst all the other species.
2. Pokémons of the *primary type (Species)* - **FLYING** in the data set have the **cumulative least abilities** in terms of all the attributes (*Attack, Defense, Hit Point, Special Attack, Special Defense, Speed*) amongst all the other species.
3. Pokémons of the *primary type (Species)* - **NORMAL and PSYCHIC** in the data set have the **2nd highest and 3rd cumulative highest abilities respectively** in terms of all the attributes (*Attack, Defense, Hit Point, Special Attack, Special Defense, Speed*) amongst all the other species. *They are very near to each other in terms of abilities.*
4. The depiction also shows that almost all the Pokémon Types (Species) have all of their abilities in accordance with each other amongst themselves.

**Scatterplot Matrix between Abilities of Pokémon Species :**

We can plot, visualize, and figure out the correlations (Positive, Negative, No correlation) between the abilities (features) of Pokémons using the Scatterplot Matrix. This matrix infographic provides the correlations of each abilities in accordance with all the other abilities in order to determine the patterns in them.

|  |
| --- |
| *Figure 18: Scatter Plot Matrix of abilities (attributes)* |

1. We can figure out that Attack is positively correlated to Defense in our data set. The famous statement also holds true in our data set that - *Attack is the best defense.*It seems that those Pokémons who are better at attacking usually develop defensive techniques as well and usually do rely on attacks for defense as well.
2. But, the same cannot be told about the inverse correlation between Defense and Attack. Those Pokémons who are defensive in nature usually do not develop very high attacking prowess.
3. The *Attack, Hit Point* prowess or abilities of Pokémons are also positively correlated to each other for the attacking style Pokémons.
4. We can figure out the types of correlations between all the abilities (Attributes) of the Pokémons from this scatterplot matrix. The types or correlations are :
   1. No Correlation
   2. Positive Correlation, and Negative Correlations

**Scatter & Jitter Plot of Primary Type of Pokémon vs Attack**

We have plotted the Scatter chart and Jitter plot of *Primary Attack* of the Pokémons varied with the *Attack* attribute which represents the distribution of the attack feature of all the Pokémons in the data set.

|  |
| --- |
| *Figure 19: Scatter Plot of Primary Type vs Attack* |
| *Figure 20: Jittered Scatter Plot of Primary Type vs Attack* |

1. From the scatter plot, we can see the distribution of the attack attribute for all the types of Pokémons. The typ**e PSYCHIC** has a largest pool of Pokémons with very high attack attribute.
2. Using the jitter plot, we have added noise to the scattering of points for better visualization and understanding of the data patterns.
   1. **WATER** type Pokémons are **very highly concentrated** in the range of **50 to 110** in terms of their attack feature.
   2. **FAIRY** type Pokémons are **very sparsely concentrated** in terms of their attack feature and seems to contain some outliers as well. We will find out about the outliers in the later plots.

**Scatter Plot of Attack vs Defense (GROUPED by 3 Primary Types)**

We have grouped 3 primary types namely **DRAGON, PSYCHIC, WATER** in order to graph a scatter plot amongst them based on *Attack vs Defense* attributes. We have also plotted lines to represent the distribution in this scatter plot using **ABLINE()** and **GEOM\_SMOOTH()** functions.

|  |
| --- |
| *Figure 21: Scatter Plot of Attack vs Defense (Grouped by 3 Primary Types - Dragon, Psychic, Water)* |

1. From the scatter plot, we can see the distribution as well as the correlations between the *Attack* and *Defense* attributes of the Pokémons of type **DRAGON, PSYCHIC, WATER**.
2. The correlation seems to be **positive** for all the **3 types up to 125 attack** attribute.
3. **After 125, the correlation becomes negative for PSYCHIC type Pokémons.** This means that PSYCHIC type Pokémons tend to lose their defense abilities when they acquire more attacking prowess.

**Scatter & Jitter Plot of Primary Type of Pokémon vs Attack (Defense varied)**

We have plotted the Scatter chart and Jitter plot of *Primary Attack* of the Pokémons vs *Attack* attribute varied with the *Defense* attribute which represents the distribution of the attack feature of all the Pokémons in the data set.

|  |
| --- |
| *Figure 22: Scatter Plot of Primary Type vs Attack (varied with Defense)* |
| *Figure 23: Jittered Scatter Plot of Primary Type vs Attack (varied with Defense)* |

1. From both the scatter plot and jittered scatter plot, we can figure out that **majority of the Pokémons** have **defensive abilities** in the range of **0 - 150**.
2. There are only **few Pokémons** of each type who **possess very high defensive abilities** around **250**. They are known as *Defensive type Pokémons*.

**Box & Jittered Box Plot of Primary Type of Pokémon vs Attack**

We have plotted the Box plot and Jittered Box plot of *Primary Attack* of the Pokémons vs *Attack* attribute varied with the *Defense* attribute which represents the distribution of the attack feature of all the Pokémons in the data set. It also provides the **OUTLIERS** in the data set.

|  |
| --- |
| *Figure 24: Box Plot of Primary Type vs Attack* |
| *Figure 25: Jittered Box Plot of Primary Type vs Attack* |

1. Both the Box plot and jittered box plot depict the median and quantiles (Lower and Upper) along with the Inter-Quartile Range at both extremes for all the Pokémons types.
2. **OUTLIERS** have been found out in both of the plots for all Pokémons types. These outliers have been highlighted in bigger size and bold in font.
3. The data set contains :
   1. Outlier in BUG type Pokémons at around 190.
   2. Outlier in DARK type Pokémons at around 150.
   3. Outliers in FAIRY type Pokémons at around 135 and 170.
   4. The STEEL type Pokémons have Outliers in the other extreme as other types do. The outlier lie somewhere around 25 for this type of Pokémons.
   5. WATER type Pokémons have multiple outliers at around 150.

# **THREE LINE TABLE FORMAT**

The 3 line table format is a particular format which is universally used and approved by APA. In this format, the horizontal line should appear before the column headings, after the column headings, and after the last record in the table.

The column headings should appear between the first and second lines of your table, while the row headings should appear on the left hand side. As with the title of your table, your column and row headings should be descriptive, clear, and brief. Do NOT, for example, use “Question 1,” Question 2,” “Question 3,” etc. for your headings. Please, refer to the below screenshot for proper understanding of the 3 line table APA format.

Table

Description automatically generated

# **CONCLUSION**

The dataset of 'The World of Pokémons' has provided various insights about the types, abilities of the Pokémons and the patterns between them as well. We performed initial data analysis, exploratory data analysis, calculated various statistics, plotted several visualisation graphs in order to understand the analysis properly. The below points can be inferred from the analysis :

* Since, the Pokémons with the least occurrences are abnormally low in the data set and we do not have any other information, we can *put up an anticipation that FLYING type Pokémons are very rare in the world of Pokémon*s**. This can only be answered properly if more data or information gets available related to it.**
* It can be concluded that the **most powerful species** in the Pokémon world is **DRAGON**. The total mean value of abilities is around **536**.*In most cases, dragon species Pokémons would come out as winners in the fights.*
* The typ**e PSYCHIC** has a largest pool of Pokémons with very high attack attribute.
* **WATER** type Pokémons are **very highly concentrated** in the range of **50 to 110** in terms of their attack feature.
* **After 125, the correlation becomes negative for PSYCHIC type Pokémons.** This means that PSYCHIC type Pokémons tend to lose their defense abilities when they acquire more attacking prowess.
* It can be concluded that the **least powerful species** in the Pokémon world is **BUG**. The total mean value of abilities is around **384**.  
  *In most cases, bug species Pokémons would come out as losers in the fights.*

# **BIBLIOGRAPHY**

1. *The World of Pokemons*. (2021, September 29). Kaggle. https://www.kaggle.com/hamdallak/the-world-of-pokemons
2. *Guide To Data Cleaning: Definition, Benefits, Components, And How To Clean Your Data*. (2021). Tableau. https://www.tableau.com/learn/articles/what-is-data-cleaning
3. Burns, E. (2021, January 16). *Data Cleaning in R Made Simple - Towards Data Science*. Medium. https://towardsdatascience.com/data-cleaning-in-r-made-simple-1b77303b0b17
4. *Find duplicated rows (based on 2 columns) in Data Frame in R*. (2011, August 8). Stack Overflow. https://stackoverflow.com/questions/6986657/find-duplicated-rows-based-on-2-columns-in-data-frame-in-r
5. *Normality Test in R - Easy Guides - Wiki - STHDA*. (2021). STHDA. http://www.sthda.com/english/wiki/normality-test-in-r
6. Nguyen, C. (2021, September 29). *Guide To Data Visualization With ggplot2 - Towards Data Science*. Medium. https://towardsdatascience.com/guide-to-data-visualization-with-ggplot2-in-a-hour-634c7e3bc9dd
7. *A Grammar of Data Manipulation*. (2021). Dplyr. https://dplyr.tidyverse.org/
8. *ggplot2 - Essentials - Easy Guides - Wiki - STHDA*. (2021). GGPLOT2 | STHDA. http://www.sthda.com/english/wiki/ggplot2-essentials
9. *Home - RDocumentation*. (2021). Functions in R - Documentation. https://www.rdocumentation.org/
10. Holtz, Y. (2021). *The R Graph Gallery – Help and inspiration for R charts*. The R Graph Gallery. https://www.r-graph-gallery.com/index.html
11. S. (2021d). *R Basics*. Smoothing! http://statseducation.com/Introduction-to-R/modules/graphics/smoothing/
12. S. (2021e, June 19). *R Ggplot2 Jitter*. Tutorial Gateway. https://www.tutorialgateway.org/r-ggplot2-jitter/
13. Moran, M. (2021, June 23). *Maintaining APA Format for Tables*. Statistics Solutions. https://www.statisticssolutions.com/maintaining-apa-format-for-tables/

# **APPENDIX**

#---------------------- Week\_2\_Module\_2\_R-Script ----------------------#

print("Author : Harshit Gaur")

print("Week 2 Assignment - Module 2 R Pratice")

# Importing the packages.

listOfPackages <- c(

"dplyr", "tidyr", "plyr", "tidyverse", "RColorBrewer", "plotrix", "scales", "ggplot2",

"data.table", "reshape", "gridExtra", "vtable", "moments", "ggpubr", "psych", "GGally"

)

for (package in listOfPackages) {

if (package %in% rownames(installed.packages()) == FALSE)

{ install.packages(package) }

# Importing the package.

library(package, character.only = TRUE)

}

# STEP 2: Import data set

# Note: Change the working directory as per the file's location.

setwd("/Users/HarshitGaur/Documents/Northeastern University/MPS Analytics/ALY 6010/Class 2/R Practice Assignment/")

pokemon\_dataset <- read.csv("pokemons dataset.csv", header = TRUE)

# Display the data set.

View(pokemon\_dataset)

# Print the structure of 'pokemons dataset.csv' data set

str(pokemon\_dataset)

# Print the summary of 'pokemons dataset.csv' data set

summary(pokemon\_dataset)

st(pokemon\_dataset)

# Describe the summary of the data set by providing Descriptive Statistics.

View(describe(pokemon\_dataset, skew = FALSE, quant = c(0.25, 0.75), IQR = TRUE))

#------------------- Data Cleaning -------------------#

# To check inconsistencies in the 'Primary Type - Character' feature of the data set.

unique(pokemon\_dataset$Primary.Type)

# To check NA, NULL values in the data set.

sum(is.na(pokemon\_dataset))

sum(is.null(pokemon\_dataset))

# Replacing Empty Values in the features with the word 'Unknown'

pokemon\_dataset$Name2 <- gsub('^$', 'NoName', pokemon\_dataset$Name2)

pokemon\_dataset$Secondary.type <- gsub('^$', 'NoType', pokemon\_dataset$Secondary.type)

# Check the duplicate values in a combination of 2 features.

duplicated(pokemon\_dataset[,1:2])

# Retrieving the duplicated records from the data set.

pokemon\_dataset[which(duplicated(pokemon\_dataset[,1:2])),]

# Eliminating the duplicated records using the indexes provided by the above step.

pokemon\_dataset <- pokemon\_dataset %>% filter( !row\_number() %in% 44)

# Removing 'NA, Missing Values' from the data set.

pokemon\_dataset <- na.omit(pokemon\_dataset)

#------------------- Exploratory Data Analysis -------------------#

# Check Normality using Density Graphs of all the univariates.

normality\_attack <- ggdensity(pokemon\_dataset$Attack, main = "Density plot of Attack", xlab = "Attack", fill = "#ffa514")

normality\_defense <- ggdensity(pokemon\_dataset$Defense, main = "Density plot of Defense", xlab = "Defense", fill = "#edf759")

normality\_hp <- ggdensity(pokemon\_dataset$HP, main = "Density plot of Hit Points", xlab = "Hit Points", fill = "#baf54c")

normality\_spAttack <- ggdensity(pokemon\_dataset$Sp.Attack, main = "Density plot of Special Attack", xlab = "Special Attack", fill = "#4cf5bd")

normality\_spDefense <- ggdensity(pokemon\_dataset$Sp.Defense, main = "Density plot of Special Defense", xlab = "Special Defense", fill = "#40c7f7")

normality\_speed <- ggdensity(pokemon\_dataset$Speed, main = "Density plot of Speed", xlab = "Speed", fill = "#d27afa")

normality\_total <- ggdensity(pokemon\_dataset$Total, main = "Density plot of Total Attributes", xlab = "Total Attributes", fill = "#eabdff")

grid.arrange(normality\_attack,normality\_defense,normality\_hp,normality\_spAttack,normality\_spDefense,normality\_speed,normality\_total)

# Check Normality using Shapiro-Wilks Test

shapiro.test(pokemon\_dataset$Attack)

# Check Normality using Q-Q Plot of all the numeric features.

# Function to plot graph

qq\_plot <- function(numeric\_feature, mainTitle) {

qqnorm(numeric\_feature, pch = 5, frame = TRUE, main = mainTitle)

qqline(numeric\_feature, col = "#ffa514", lwd = 2)

}

# Changing Plot Matrix Size to 3x3.

par(mfrow = c(3,3))

# Check Normality using Q-Q Plot of 'Attack' Feature.

qq\_plot(pokemon\_dataset$Attack, "Attack")

# Check Normality using Q-Q Plot of 'Defense' Feature.

qq\_plot(pokemon\_dataset$Defense, "Defense")

# Check Normality using Q-Q Plot of 'HP' Feature.

qq\_plot(pokemon\_dataset$HP, "Horse Power")

# Check Normality using Q-Q Plot of 'Special Attack' Feature.

qq\_plot(pokemon\_dataset$Sp.Attack, "Special Attack")

# Check Normality using Q-Q Plot of 'Special Defense' Feature.

qq\_plot(pokemon\_dataset$Sp.Defense, "Special Defense")

# Check Normality using Q-Q Plot of 'Speed' Feature.

qq\_plot(pokemon\_dataset$Speed, "Speed")

# Check Normality using Q-Q Plot of 'Total Attributes' Feature.

qq\_plot(pokemon\_dataset$Total, "Total Attributes")

# Resetting Plot Matrix Size to 1x1.

par(mfrow = c(1,1))

# Check Skewness of the features

skewness(pokemon\_dataset$Attack)

skewness(pokemon\_dataset$Defense)

skewness(pokemon\_dataset$HP)

skewness(pokemon\_dataset$Sp.Attack)

skewness(pokemon\_dataset$Sp.Defense)

skewness(pokemon\_dataset$Speed)

skewness(pokemon\_dataset$Total)

# Check Skewness of the features

kurtosis(pokemon\_dataset$Attack)

kurtosis(pokemon\_dataset$Defense)

kurtosis(pokemon\_dataset$HP)

kurtosis(pokemon\_dataset$Sp.Attack)

kurtosis(pokemon\_dataset$Sp.Defense)

kurtosis(pokemon\_dataset$Speed)

kurtosis(pokemon\_dataset$Total)

# Describe the summary of the data set by providing Descriptive Statistics.

View(describe(pokemon\_dataset, skew = FALSE, quant = c(0.25, 0.75), IQR = TRUE))

# Describe the summary of the data set by grouping

describeBy(pokemon\_dataset, group = pokemon\_dataset$Primary.Type)

describeBy(pokemon\_dataset, group = pokemon\_dataset$Secondary.type)

# Plot an Grouped Line graph for attack using 'ggplot'

# ----------- Plot 1: Primary Type vs Summation of Attributes ----------- #

# Summation of 'Attributes' according to Primary Type feature

summationTable <- pokemon\_dataset %>% group\_by(pokemon\_dataset$Primary.Type) %>% dplyr::summarise(

sum\_attack = sum(Attack),

sum\_defense = sum(Defense),

sum\_hp = sum(HP),

sum\_spAttack = sum(Sp.Attack),

sum\_spDefense = sum(Sp.Defense),

sum\_speed = sum(Speed),

sum\_total = sum(Total),

)

# Converting the summation Table into a Data Frame.

summationTable <- data.frame((summationTable))

par(mar = c(4, 10, 8, 4) + 0.1)

ggplot(summationTable, aes(x = pokemon\_dataset.Primary.Type, group = 1)) +

geom\_line(aes(y = sum\_attack, color="Attack"), linetype="dotted", size=1) +

geom\_line(aes(y = sum\_defense, color="Defense"), linetype="longdash", size=1) +

geom\_line(aes(y = sum\_hp, color="HP"), linetype="F1", size=1) +

geom\_line(aes(y = sum\_spAttack, color="Special Attack"), linetype="F1", size=1) +

geom\_line(aes(y = sum\_spDefense, color="Special Attack"), linetype="solid", size=1) +

geom\_line(aes(y = sum\_speed, color="Speed"), linetype="twodash", size=1) +

scale\_color\_manual(

name = "Legends",

values = c("Attack" = "red", "Defense" = "blue", "HP" = "yellow", "Special Attack" = "orange", "Special Defense" = "Darkgreen", "Speed" = "magenta")) +

theme(

panel.background = element\_rect("#c9f5e4"),

axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust=1, size = 12)

) +

labs(

title = "Plot 2: Primary Type vs Summation of Attributes",

x = "Primary Type",

y = "Summation of Attributes"

) +

theme(plot.title = element\_text(hjust = 0.5, colour = "#7F3D17", face = "bold"))

# Plot an Scatter Plot Matrix graph for 'Attributes' using 'pairs'

# ----------- Plot 2: Scatter Plot Matrix between Abilities (Attributes) ----------- #

pairs(~Attack + Defense + HP + Sp.Attack + Sp.Defense + Speed, pokemon\_dataset, col=c("YELLOW", "RED"), pch = 5)

# Plot an Scatter Chart graph for 'Primary Type vs Attack' using 'ggplot'

# ----------- Plot 3: Scatter Plot of Primary Type vs Attack ----------- #

par(mar = c(2,4,2,4))

ggplot(data = pokemon\_dataset, aes(x = Primary.Type, y = Attack, color = Primary.Type)) +

geom\_point() +

labs(title = 'Scatter Plot of Primary Type vs Attack', x = 'Primary Type', y = 'Attack Attribute') +

theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust=1, size = 10))

# Plot an Jitter Chart graph for 'Primary Type vs Attack' using 'ggplot'

# ----------- Plot 4: Jittered Scatter Plot of Primary Type vs Attack ----------- #

par(mar = c(2,4,2,4))

ggplot(data = pokemon\_dataset, aes(x = Primary.Type, y = Attack, color = Primary.Type)) +

geom\_jitter(position = position\_jitterdodge()) +

labs(title = 'Jitter Plot of Primary Type vs Attack', x = 'Primary Type', y = 'Attack Attribute') +

theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust=1, size = 10))

# Plot an Scatter Chart graph for 'Attack vs Defense (Grouped in 3 Types)' using 'ggplot'

# ----------- Plot 5: Scatter Plot of Attack vs Defense (Grouped in 3 Types) ----------- #

dgw\_group <- filter(pokemon\_dataset, Primary.Type %in% c("DRAGON", "WATER", "PSYCHIC"))

par(mar = c(2,4,2,4))

ggplot(data = dgw\_group, aes(x = Attack, y = Defense, color = Primary.Type)) +

geom\_point() +

labs(title = 'Scatter Plot of Attack vs Defense', x = 'Attack Attribute', y = 'Defense Attribute') +

theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust=1, size = 10)) +

geom\_smooth(se = FALSE)

# Plot an Scatter Chart graph for 'Primary Type vs Attack varied with Defense' using 'ggplot'

# ----------- Plot 6: Scatter Plot of Primary Type vs Attack ----------- #

par(mar = c(2,4,2,4))

ggplot(data = pokemon\_dataset, aes(x = Primary.Type, y = Attack, color = Defense)) +

geom\_point() +

labs(title = 'Scatter Plot of Primary Type vs Attack varied with Defense', x = 'Primary Type', y = 'Attack Attribute') +

theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust=1, size = 10))

# Plot an Jitter Chart graph for 'Primary Type vs Attack varied with Defense' using 'ggplot'

# ----------- Plot 7: Jitter Plot of Primary Type vs Attack ----------- #

par(mar = c(2,4,2,4))

ggplot(data = pokemon\_dataset, aes(x = Primary.Type, y = Attack, color = Defense)) +

geom\_jitter(position = position\_jitterdodge()) +

labs(title = 'Jitter Plot of Primary Type vs Attack varied with Defense', x = 'Primary Type', y = 'Attack Attribute') +

theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust=1, size = 10))

# Plot an Box Plot for 'Primary Type vs Attack' using 'ggplot'

# ----------- Plot 8: Box Plot of Primary Type vs Attack ----------- #

par(mar = c(2,4,2,4))

ggplot(data = pokemon\_dataset, aes(x = Primary.Type, y = Attack, color = Primary.Type)) +

geom\_boxplot(outlier.size = 3.5) +

labs(title = 'Box Plot of Primary Type vs Attack varied with Defense', x = 'Primary Type', y = 'Attack Attribute') +

theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust=1, size = 10))

# Plot an Jitter Chart graph for 'Primary Type vs Attack varied with Defense' using 'ggplot'

# ----------- Plot 9: Jittered Box Plot of Primary Type vs Attack ----------- #

par(mar = c(2,4,2,4))

ggplot(data = pokemon\_dataset, aes(x = Primary.Type, y = Attack, color = Primary.Type)) +

geom\_boxplot(outlier.size = 3.5, outlier.stroke = 2) +

geom\_jitter(position = position\_jitterdodge()) +

labs(title = 'Jitter Plot of Primary Type vs Attack', x = 'Primary Type', y = 'Attack Attribute') +

theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust=1, size = 10))