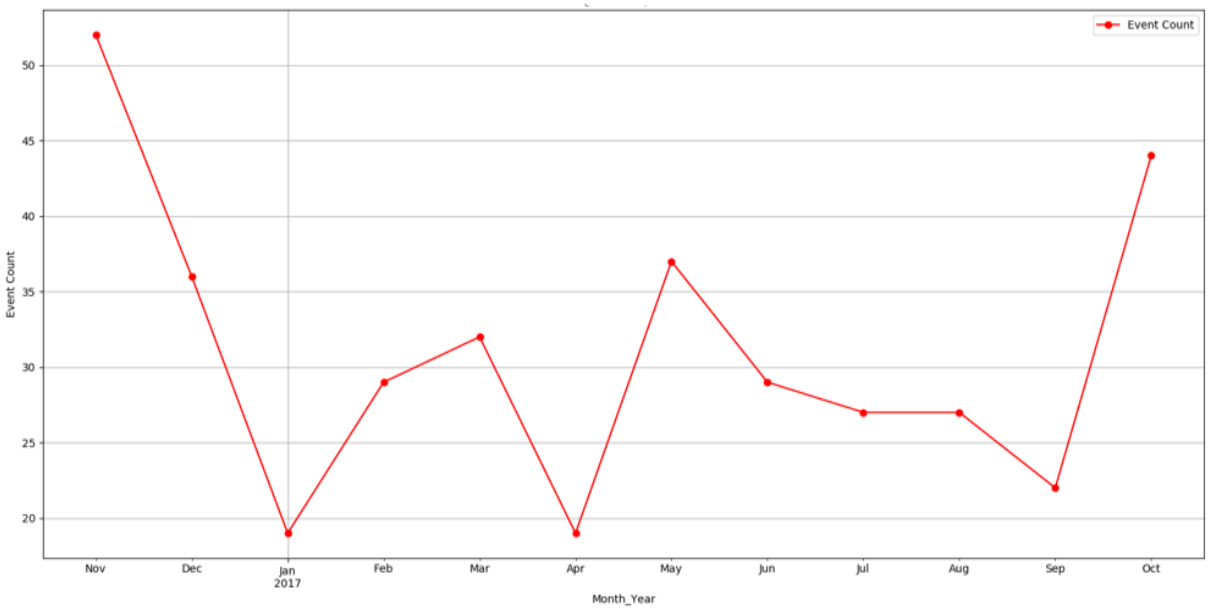
R PROGRAMMING

POKÉMON DATASET ANALYSIS

COMPLETED BY:
Ravi Sharma -11021210017

COURSE : B.Tech - CSE (DS AI)
BATCH : 2021 - 2025



Introduction

R, a versatile and powerful programming language for data analysis, will be our guiding tool. The ggplot2 library within R allows us to transform raw data into compelling visualizations, bringing clarity and insights to the vast Pokémon universe. With R and ggplot2, we can unveil the hidden treasures within the Pokémon dataset.

The Pokémon franchise, is one of the most successful franchise till date having a warm place in the heart of every 2000's kid, created by Satoshi Tajiki and Ken Sugimori has captured the hearts of millions worldwide since its inception in 1996. Pokémon, short for "Pocket Monsters," is a multimedia franchise encompassing video games, trading card games, animated series, movies, merchandise, and more. One of the most captivating aspects of Pokémon is the diverse world of fictional creatures, each with unique abilities, types, and characteristics.

This project signifies the fusion of two seemingly divergent worlds — the imaginative realm of Pokémon and the analogical world of data science. It is not merely an exploration of data.

In this project, we embark on an exciting journey that bridges the enchanting world of Pokémon with the power of data analysis and visualization using the R programming language. We delve into the rich "Pokémon dataset" to explore and extract meaningful insights, revealing hidden patterns and trends within the Pokémon universe.

In our quest, we seek to uncover the mysteries hidden within the vast expanse of Pokémon data. From the elemental strengths, attacks and weaknesses of Pokémon types whether it be legendary or non-legendary to the intricacies of battle statistics and evolutionary paths, our analysis will shed light on the dynamics that govern this fantastical realm.

Pokémon Dataset

The Pokémon dataset is a comprehensive collection of data that allows us to explore and analyze the world of Pokémon in a structured and systematic manner. It serves as a valuable resource for understanding the characteristics and attributes of these fictional creatures.

This project aims to extract meaningful insights from the Pokémon dataset by employing data analysis and visualization techniques. By exploring this rich dataset, we hope to gain a deeper understanding of the characteristics, trends, and relationships that define the world of Pokémon.

Let's understand the structure and scale of the dataset -

Structure and Size:
The dataset is ingeniously organized as a data frame or table, comprising rows and columns that neatly encapsulate vital information. Within its confines, we find a repository of knowledge encompassing a staggering array of 801 unique Pokémon species.

Attributes and Columns:
Each row within the dataset serves as a portal to an individual Pokémon species, offering a glimpse into its distinctive characteristics and in-game statistics. Through a diverse array of columns, we uncover a wealth of details, ranging from elemental types and evolutionary paths to combat prowess and rarity distributions.

This dataset, with its meticulous curation and expansive scope, provides a fertile ground for exploration and discovery. With each analysis, we inch closer to unraveling the mysteries and magic that define the captivating world of Pokémon.

The dataset is in the form of csv file available in the GitHub Repository.

Attributes:

- **Name:** The English name of the Pokémon.
- **Japanese Name:** The original Japanese name of the Pokémon.
- **Pokedex Number:** The entry number of the Pokémon in the National Pokédex.
- **Percentage Male:** The percentage of the species that are male.
- **Classification:** The classification of the Pokémon as described by the Sun and Moon Pokédex.
- **Height (m):** Height of the Pokémon in meters.
- **Weight (kg):** The weight of the Pokémon in kilograms.
- **Capture Rate:** Capture rate of the Pokémon.
- **Base Egg Steps:** The number of steps required to hatch an egg of the Pokémon.
- **Experience Growth:** The experience growth of the Pokémon.
- **Base Happiness:** Base happiness of the Pokémon.
- **Type1:** Primary type of the Pokémon.
- **Type2:** Secondary type of the Pokémon.

Combat Attributes:

- **Against [Type]:** Eighteen features that denote the amount of damage taken against an attack of a particular type.
- **HP:** The base HP of the Pokémon.
- **Attack:** The base Attack of the Pokémon.
- **Defense:** The base Defense of the Pokémon.
- **Sp. Attack:** The base Special Attack of the Pokémon.
- **Sp. Defense:** The base Special Defense of the Pokémon.
- **Speed:** The base Speed of the Pokémon.

Additional Information:

- **Generation:** The numbered generation in which the Pokémon was first introduced.
- **Is Legendary:** Denotes if the Pokémon is legendary.

Importing the Dataset

Once the packages are installed and the dataset is imported, we're equipped to embark on our exploration of the Pokémon dataset. Here's a brief overview of the steps we've taken and what comes next:

- Installation of Packages: We've installed two essential packages, ggplot2 and tidyverse, using the install.packages() function. These packages provide powerful tools for data visualization and manipulation in R.

- Library Loading: After installation, we load the packages into our R environment using the library() function. This makes their functions and capabilities available for use in our scripts.

- Dataset Import: With the packages in place, we read the Pokémon dataset from a CSV file using the read_csv() function from the tidyverse package. The dataset is stored in an object named pokemon_data.

- Previewing the Dataset: To get a glimpse of the structure and content of the dataset, we use the head() function to display the first few rows and columns.

- Data Exploration: Now that the dataset is loaded, we're ready to dive deeper into its contents. We can explore various attributes, conduct statistical analyses, and create visualizations to uncover patterns and insights within the Pokémon data.

With the groundwork laid and the dataset at our fingertips, we're poised to unleash the power of R and embark on our journey of exploration and discovery within the Pokémon universe. Let's dive in and see what secrets await us!

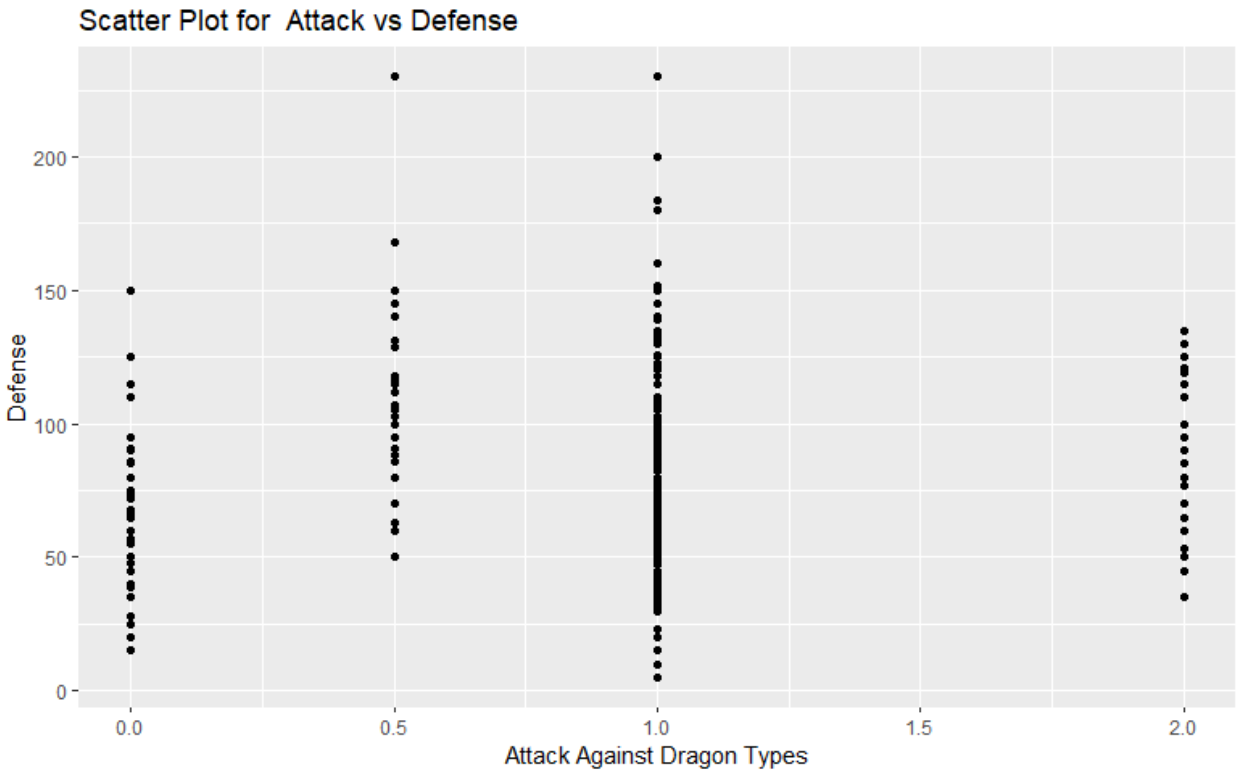
PLOT 1:

Scattered Plot

A scatter plot is also called a scatter chart, scattergram, or scatter plot, XY graph. The scatter diagram graphs numerical data pairs, with one variable on each axis, show their relationship.

#plot1

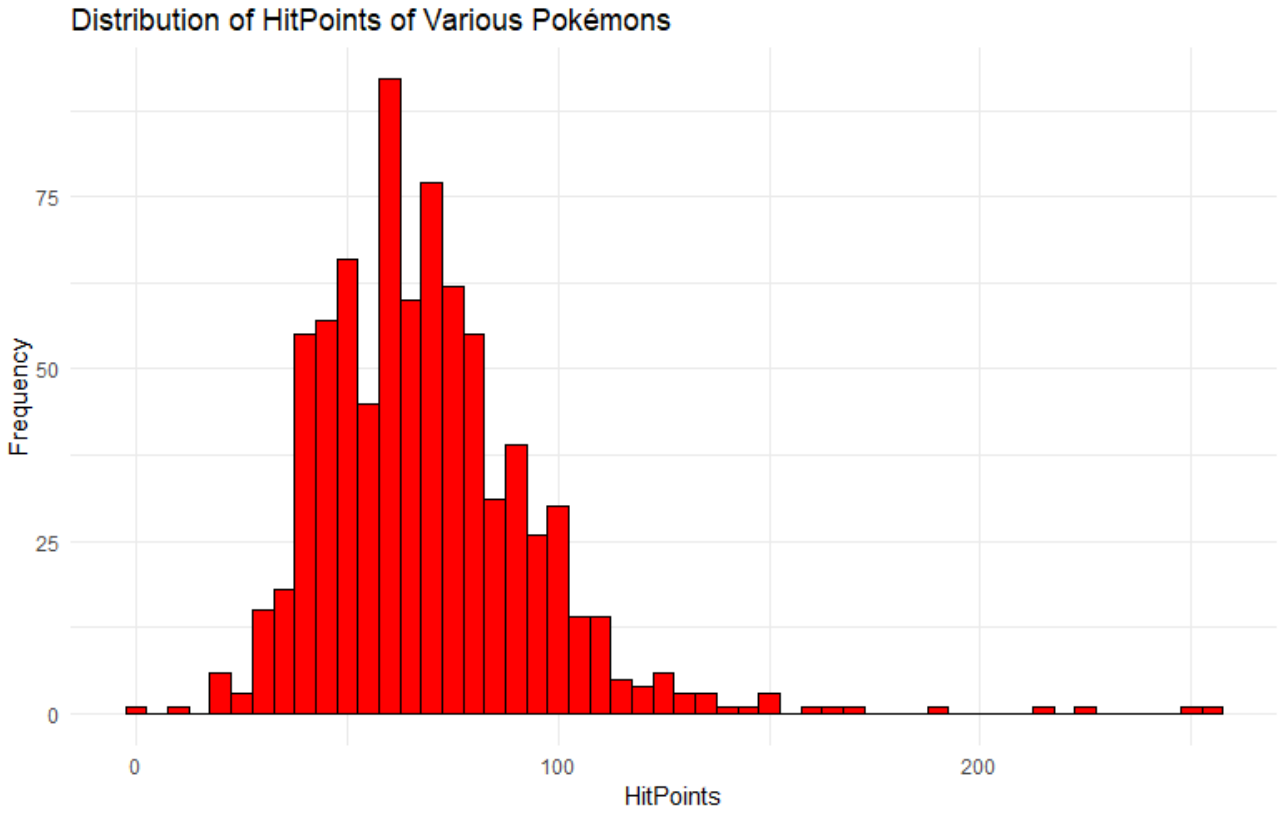
```
ggplot(data = pokemon_data, aes(x = against_dragon, y = defense)) +
 geom_point() +
 labs(title = "Scatter Plot for  Attack vs Defense", x = "Attack Against Dragon Types", y = "Defense")
```



PLOT 2:

Histogram Plot

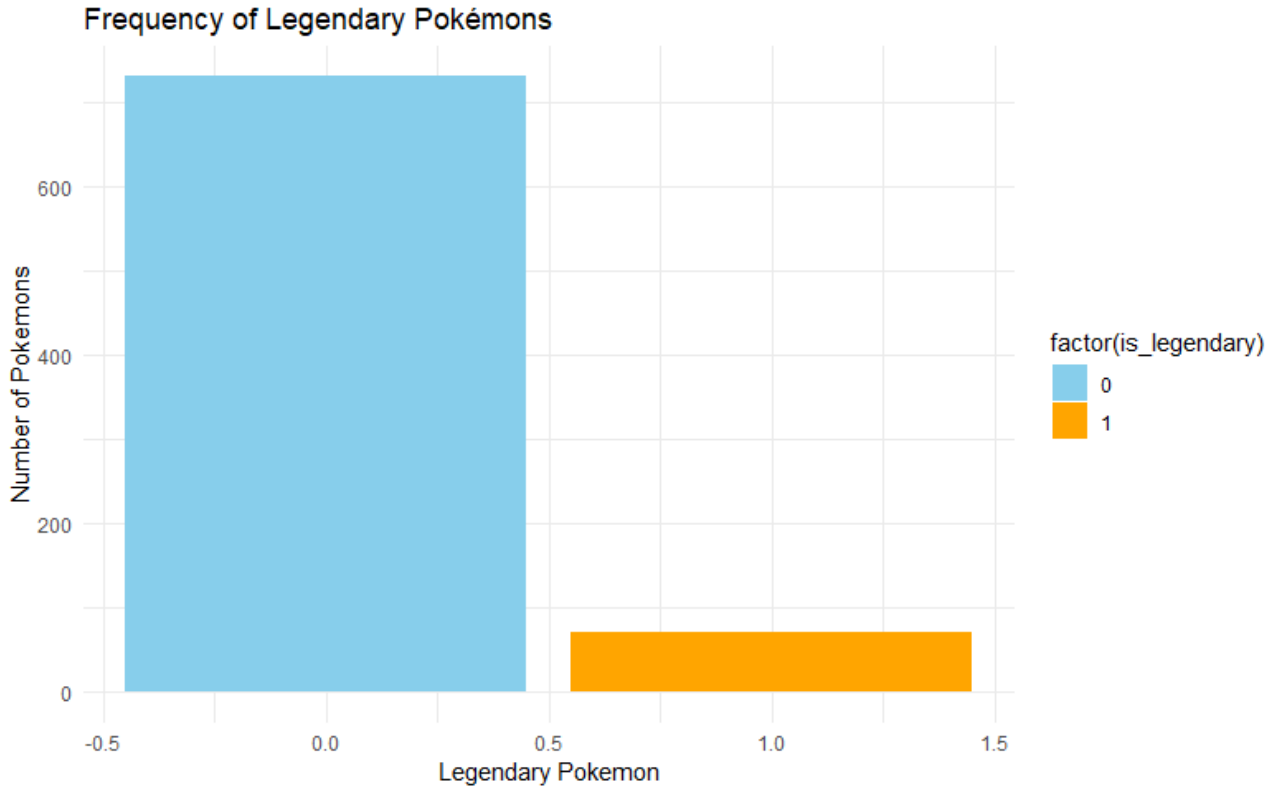
```
#plot2
ggplot(data = pokemon_data) +
  geom_histogram(mapping = aes(x = hp),
                 binwidth = 5,  # Set the bin width
                 fill = "Red",  # Set fill color
                 color = "black"  # Set border color
  ) +
  labs(title = "Distribution of HitPoints of Various Pokémons",
       x = "HitPoints",
       y = "Frequency",
  )
```



PLOT 3:

Vertical Bar Plot

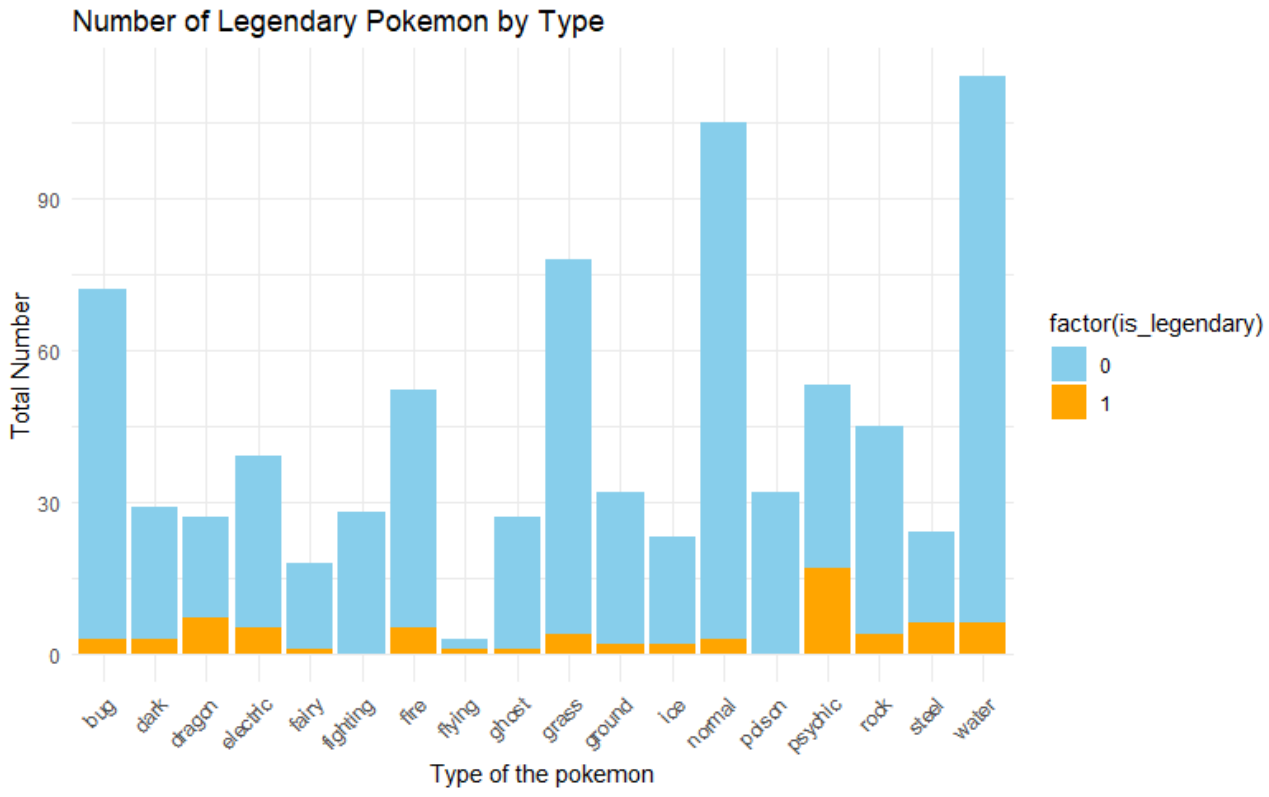
```
#plot3
ggplot(data = pokemon_data) +
  geom_bar(mapping = aes(x = is_legendary, fill = factor(is_legendary))) +
  labs(title = "Frequency of Legendary Pokémons",
       x = "Legendary Pokemon",
       y = "Number of Pokemons"
  ) +
  scale_fill_manual(values = c("skyblue", "orange")) +
  theme_minimal()
```



PLOT 4:

Grouped Bar Plot

```
#plot4
ggplot(data = pokemon_data) +
  geom_bar(mapping = aes(x = type1, fill = factor(is_legendary)))
+
  labs(title = "Number of Legendary Pokemon by Type",
       x = "Type of the pokemon",
       y = "Total Number"
  ) +
  scale_fill_manual(values = c("skyblue", "orange")) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

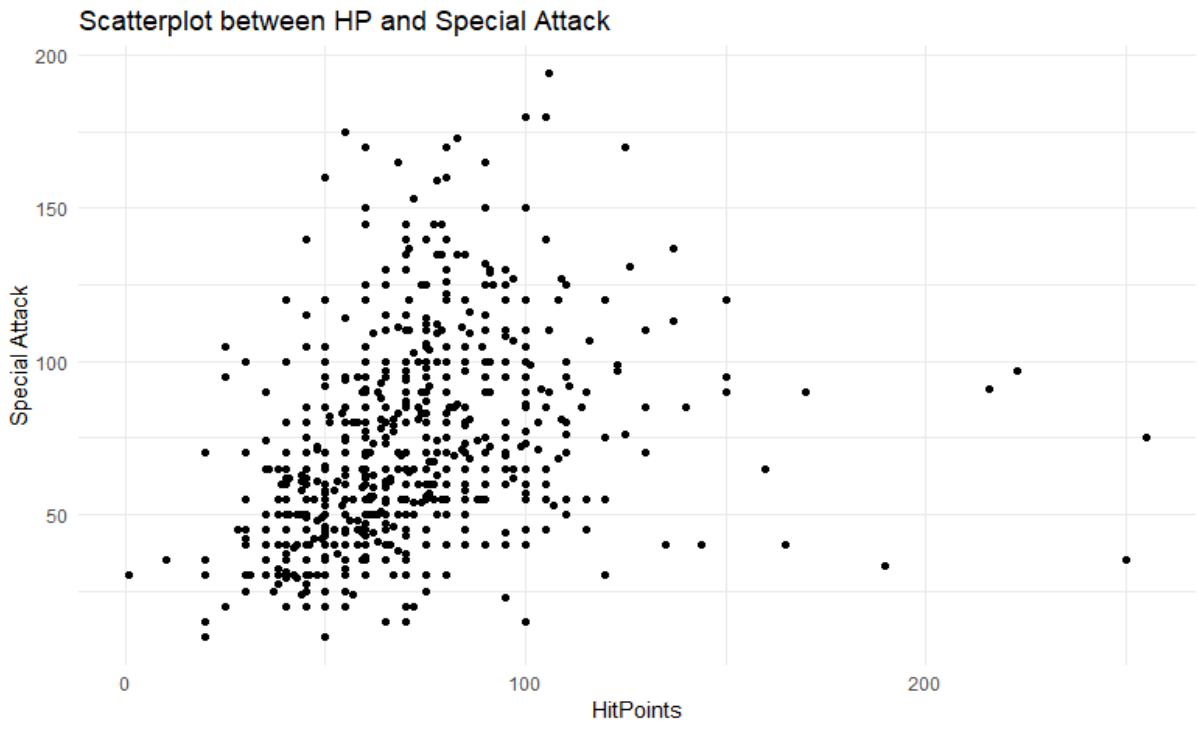


PLOT 5:

Scattered Plot

A scatter plot is a graphical representation of individual data points in a two-dimensional space, where each point corresponds to the values of two variables.

```
#plot5
ggplot(data = Pokemon) +
  geom_point(mapping = aes(x = hp, y = sp_attack)) +
  labs(
    title = "Scatterplot between HP and Special Attack",
    x = "HitPoints",
    y = "Special Attack"
  )
```



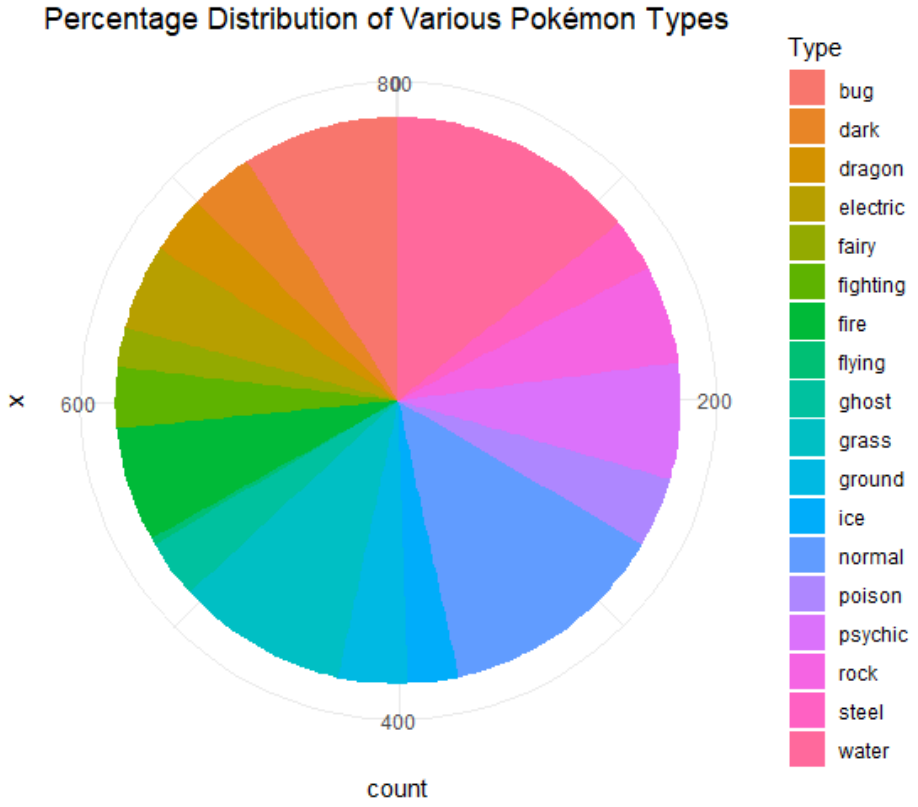
PLOT 6:

Rose Diagram

A pie chart is a type of graph representing data in a circular form, with each slice of the circle representing a fraction or proportionate part of the whole.

#plot6

```
ggplot(data = Pokemon) +
  geom_bar(mapping = aes(x = "", fill = type1)) +
  coord_polar(theta = "y") +
  labs(title = "Percentage Distribution of Various Pokémon Types",
fill = "Type") +
  theme_minimal()
```



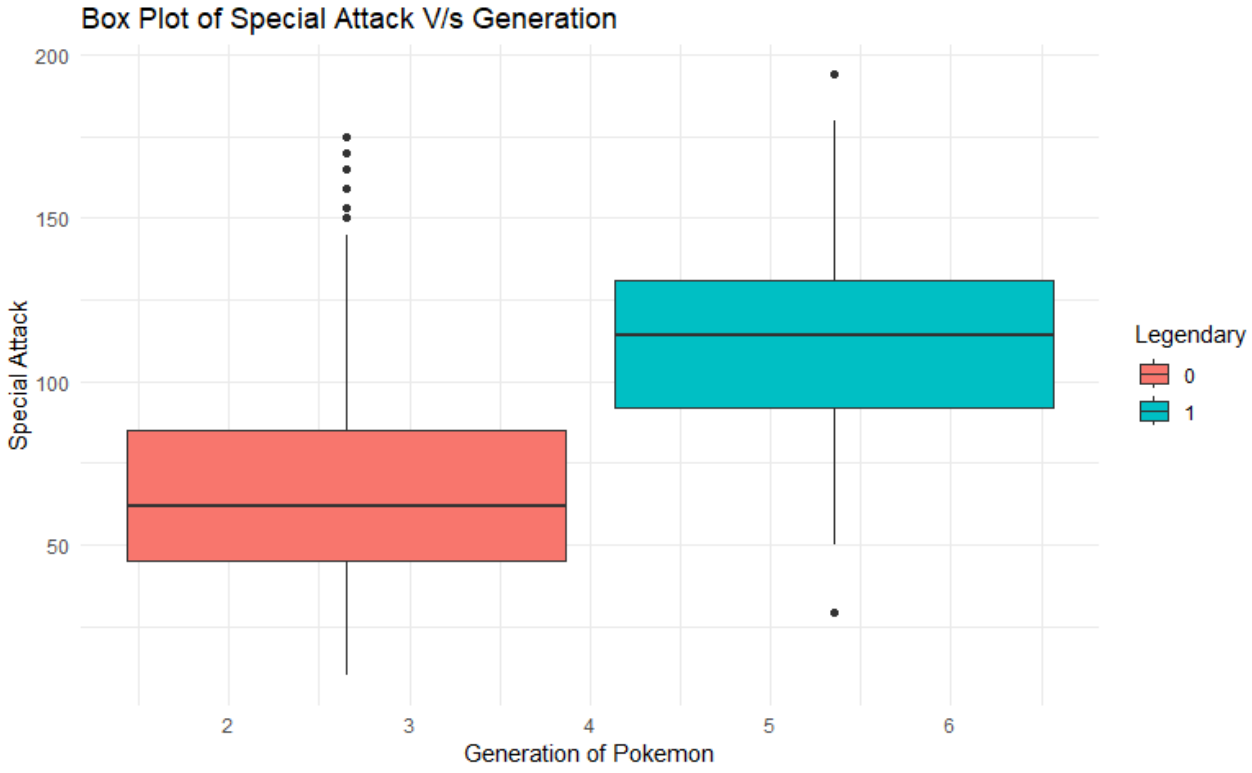
PLOT 7:

Vertical Box Plot

A box and whisker plot is a special type of graph that is used to show groups of number data and how they are spread.

#plot7

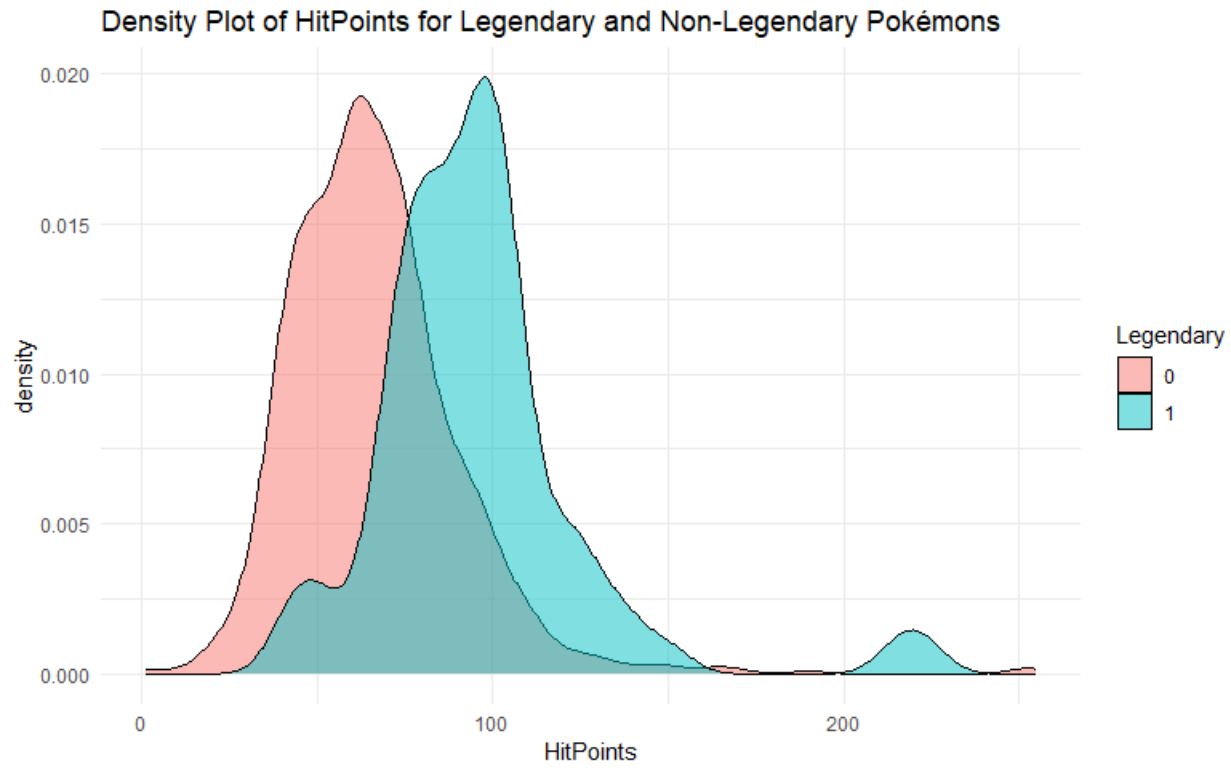
```
Pokemon$is_legendary <- as.factor(Pokemon$is_legendary)
ggplot(data = Pokemon) +
  geom_boxplot(mapping = aes(x = generation, y = sp_attack, fill = is_legendary)) +
  labs(title = "Box Plot of Special Attack V/s Generation ",
       x = "Generation of Pokemon",
       y = "Special Attack",
       fill = "Legendary")
```



PLOT 8:

Density Plots

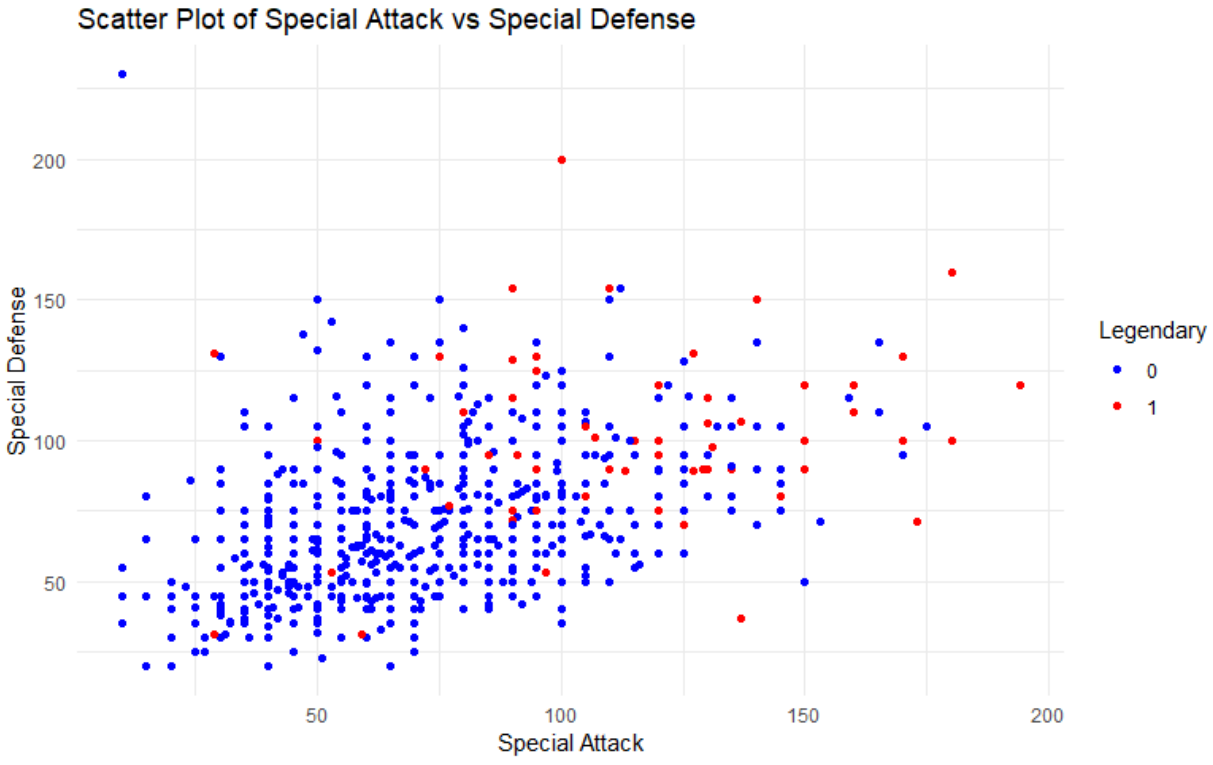
```
#plot8
ggplot(data = pokemon_data) +
  geom_density(mapping = aes(x = hp, fill = factor(is_legendary)), alpha = 0.5) +
  labs(
    title = "Density Plot of HitPoints for Legendary and Non-Legendary Pokémons",
    x = "HitPoints",
    fill = "Legendary"
  ) +
  theme_minimal()
```



PLOT 9:

Clustered Scatter Plot

```
#plot9
 ggplot(data = pokemon_data) +
  geom_point(mapping = aes(x = sp_attack, y = sp_defense, color =
 factor(is_legendary))) +
  labs(
    title = "Scatter Plot of Special Attack vs Special Defense",
    x = "Special Attack",
    y = "Special Defense",
    color = "Legendary"
  ) +
  scale_color_manual(values = c("blue", "red"))
```

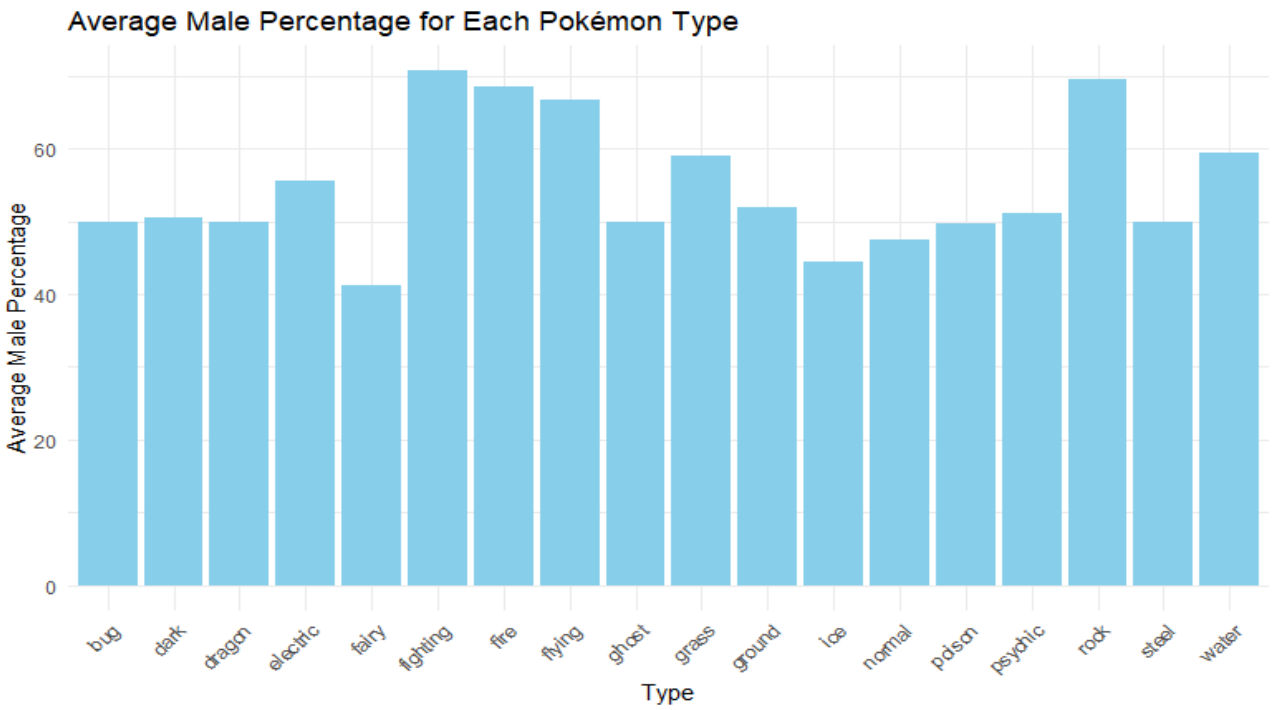


PLOT 10:

Vertical Bar Plots

#plot10

```
average_male_percentage <- pokemon_data %>%
  group_by(type1)    %>%        summarise(average_male_percentage=
mean(percentage_male, na.rm = TRUE))
ggplot(data = average_male_percentage) +
  geom_bar(mapping=aes(x=type1,y= average_male_percentage), stat
= "identity", fill = "skyblue") +
  labs(title = "Average Male Percentage for Each Pokémon Type",
       x = "Type",
       y = "Average Male Percentage") +
theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

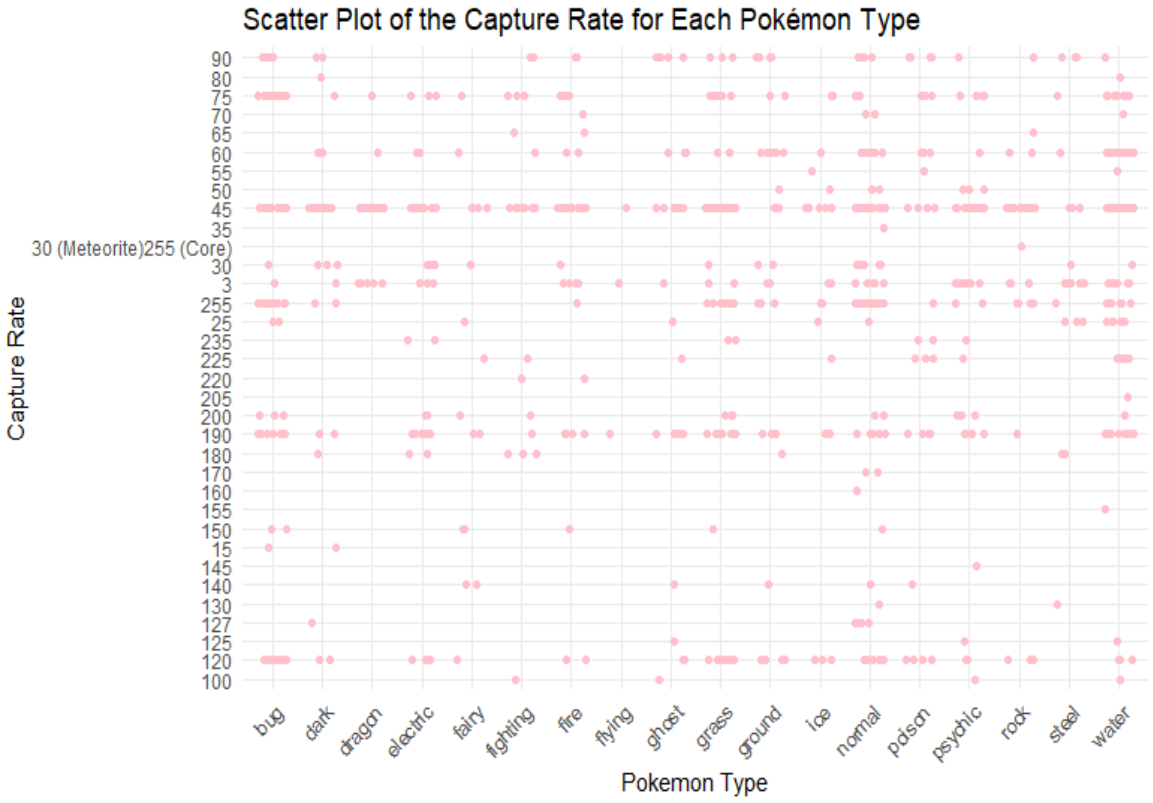


PLOT 11:

Scattered Plots

#plot11

```
ggplot(data = pokemon_data, aes(x = type1, y = capture_rate)) +
  geom_point(position = position_jitter(width = 0.3, height = 0), color
= "pink") +
  labs(title = "Scatter Plot of the Capture Rate for Each Pokémon Type",
       x = "Pokemon Type",
       y = "Capture Rate") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



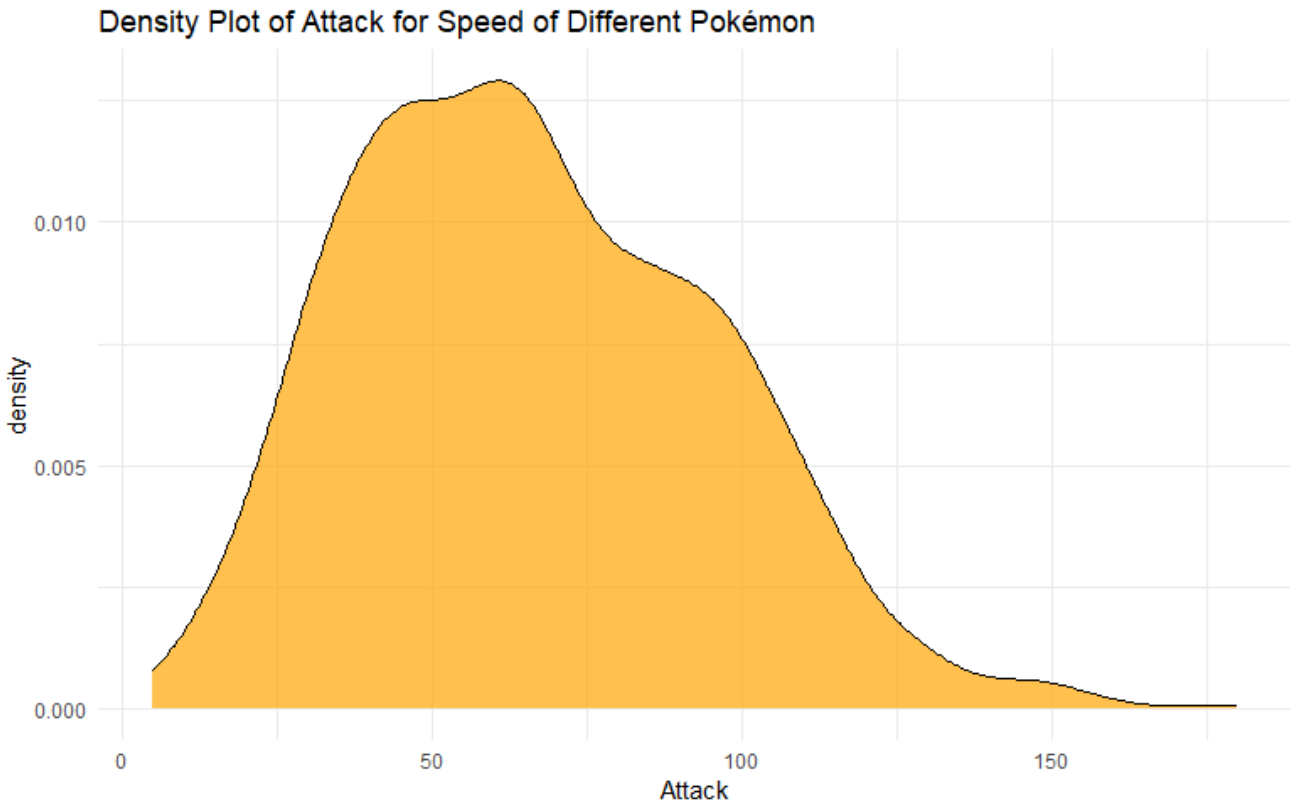
PLOT 12:

Kernel Density Plot

A kernel density plot is a non-parametric way to estimate the probability density function of a continuous random variable.

#plot12

```
ggplot(data = pokemon_data) +
  geom_density(mapping = aes(x = speed), fill = "orange", alpha = 0.7) +
  labs(title = "Density Plot of Attack for Fire-type Pokémon", x = "Attack") +
  theme_minimal()
```

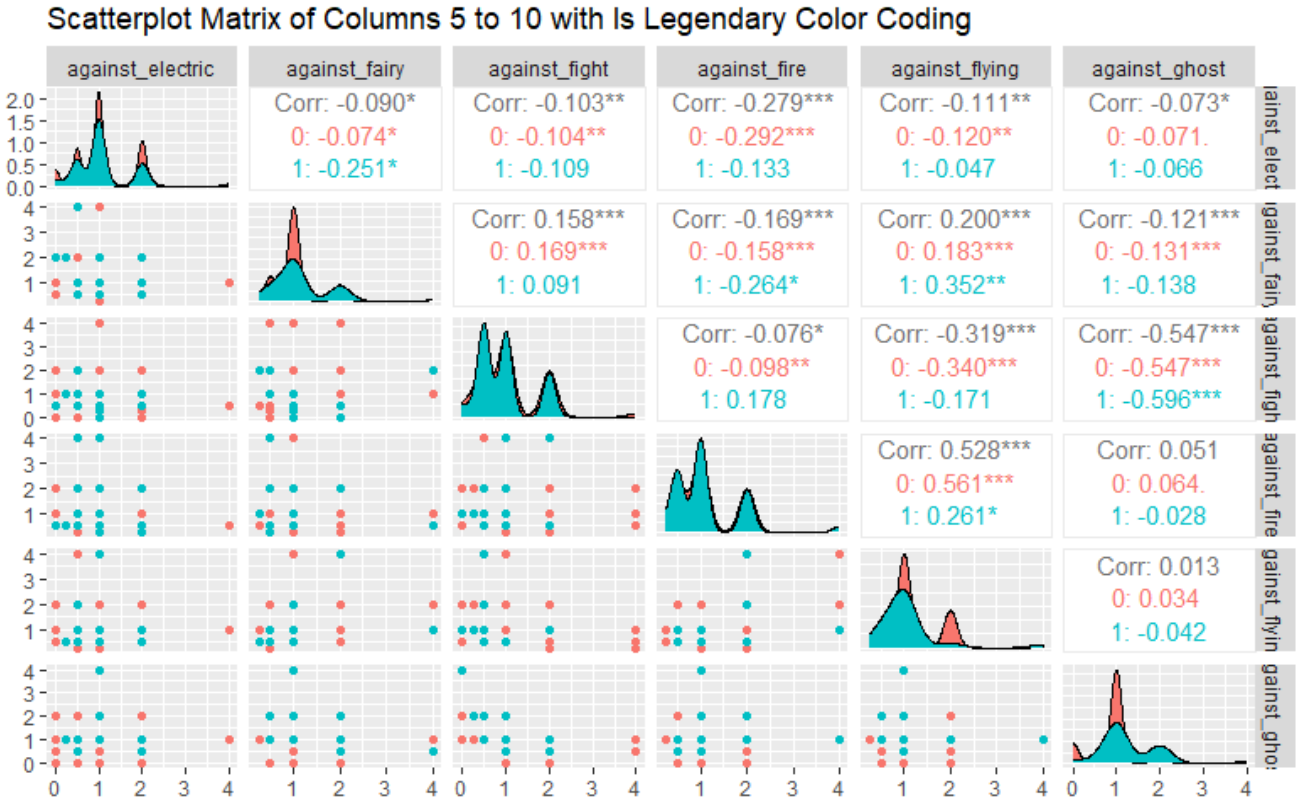


PLOT 13:

Scatterplot Matrix

#plot13

```
library(GGally)
data(pokemon_data)
pokemon_data$is_legendary <- factor(pokemon_data$is_legendary)
ggpairs(pokemon_data, columns = 5:10, ggplot2::aes(color = is_legendary)) +
  ggtitle("Scatterplot Matrix of Columns 5 to 10 with Is Legendary Color Coding")
```



PLOT 14:

Line Graph

A line graph, also known as a line chart or line plot, is a type of data visualization that represents data points over a continuous interval or time span by connecting them with straight lines.

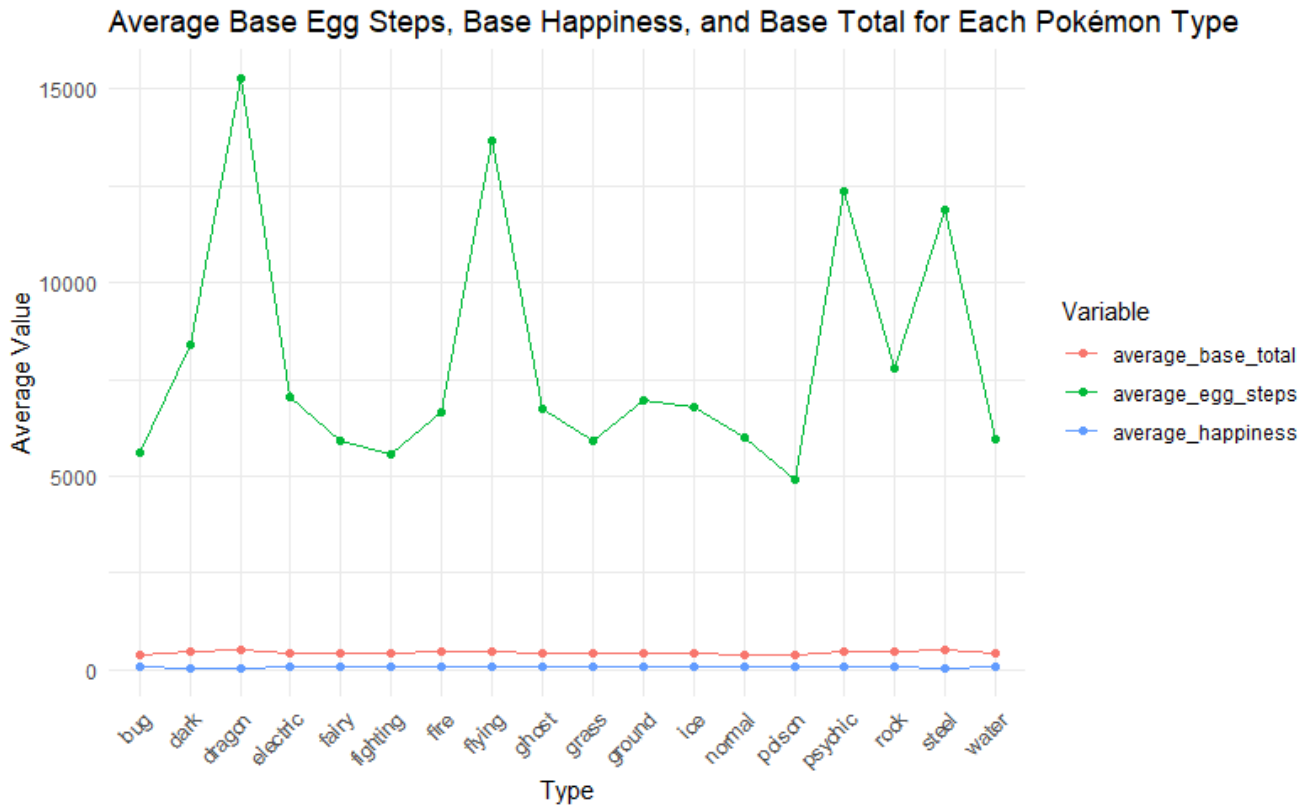
It is particularly useful for illustrating trends, patterns, or relationships in data. Each data point on the graph corresponds to a specific value on the y-axis, and the position of the points is determined by their corresponding values on the x-axis.

```
#plot14
library(ggplot2)
library(tidyverse)
average_values <- pokemon_data %>%
  group_by(type1) %>%
  summarise(
    average_egg_steps = mean(base_egg_steps, na.rm = TRUE),
    average_happiness = mean(base_happiness, na.rm = TRUE),
    average_base_total = mean(base_total, na.rm = TRUE)
  )
average_values_long <- average_values %>%
  pivot_longer(cols = c(average_egg_steps, average_happiness,
average_base_total),

           names_to = "Variable", values_to = "Value")
ggplot(data = average_values_long, aes
(x = type1, y = Value, color = Variable, group = Variable))
```

```
+

geom_line() +
 geom_point()
+
 Labs
(title = "Average Base Egg Steps, Base Happiness, and Base Total
for Each Pokémon Type",
      x = "Type",
      y = "Average Value")
 +
 theme_minimal()
+
 theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

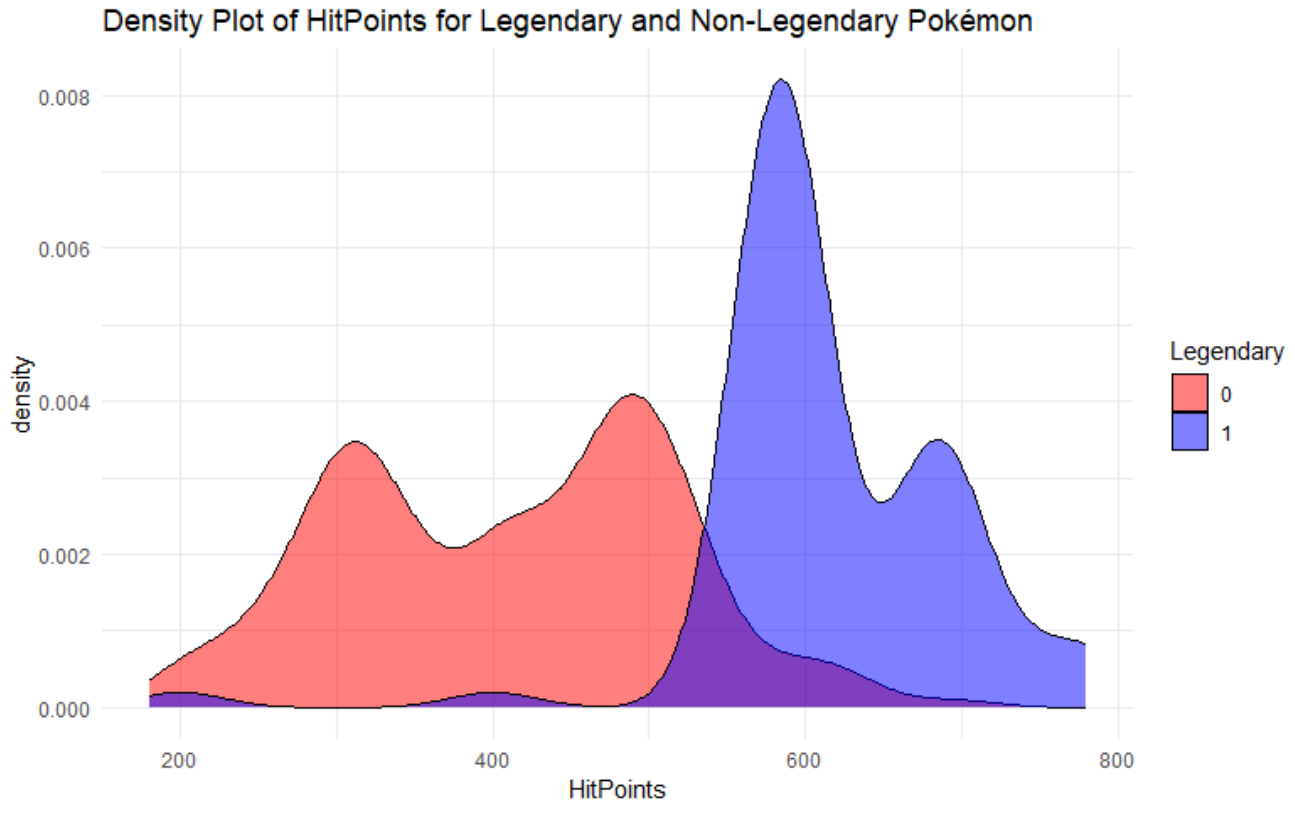


PLOT 15:

Density Plot

#plot15

```
ggplot(data = pokemon_data) +
  geom_density(mapping = aes(x = base_total, fill = factor(is_legendary)), alpha = 0.5, color = "black") +
  scale_fill_manual(values = c("red", "blue")) +  # Set fill colors to red and blue
  labs(
    title = "Density Plot of HitPoints for Legendary and Non-Legendary Pokémon",
    x = "HitPoints",
    fill = "Legendary"  )
```

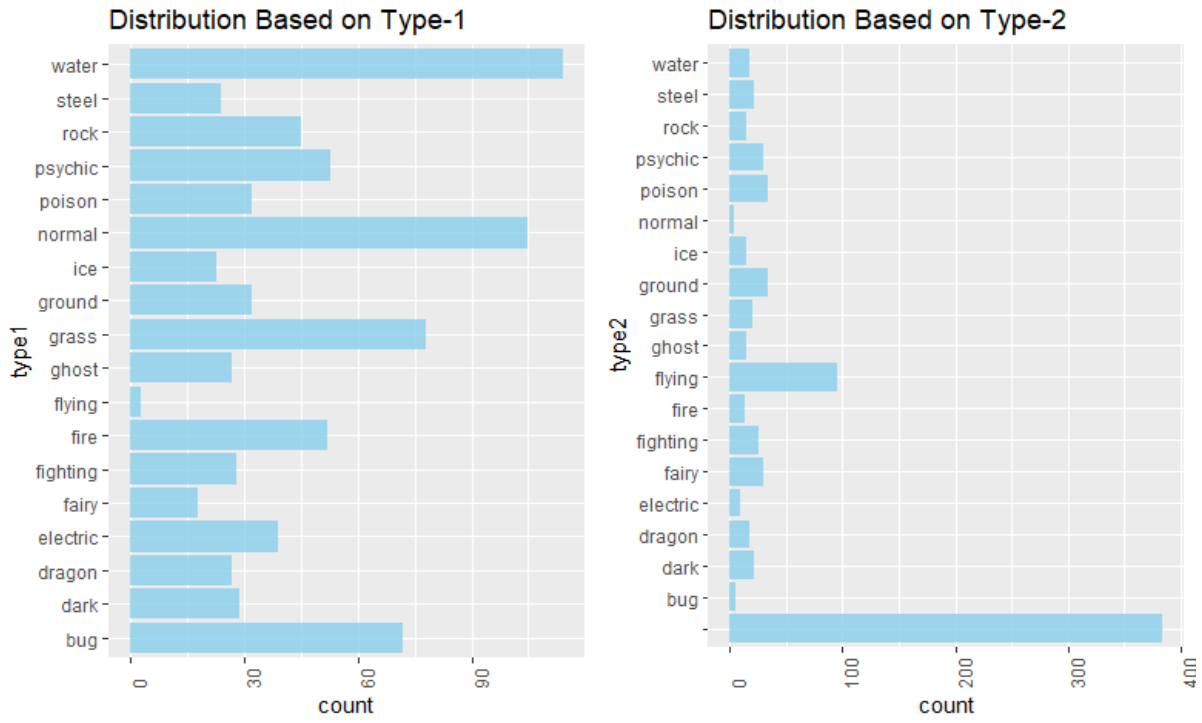


PLOT 16:

Two Horizontal Bar Plots

#plot16

```
type_1_poke <- ggplot(data = pokemon_data, aes(x = type1)) +
  geom_bar(fill = "skyblue", alpha = 0.8, stat = "count") +
  theme(axis.text.x = element_text(angle = 90, hjust = 0)) +
  ggtitle("Distribution Based on Type-1") +
  coord_flip()
type_2_poke <- ggplot(data = pokemon_data, aes(x = type2)) +
  geom_bar(fill = "skyblue", alpha = 0.8, stat = "count") +
  theme(axis.text.x = element_text(angle = 90, hjust = 0)) +
  ggtitle("Distribution Based on Type-2") +
  coord_flip()
grid.arrange(type_1_poke, type_2_poke, ncol = 2)
```

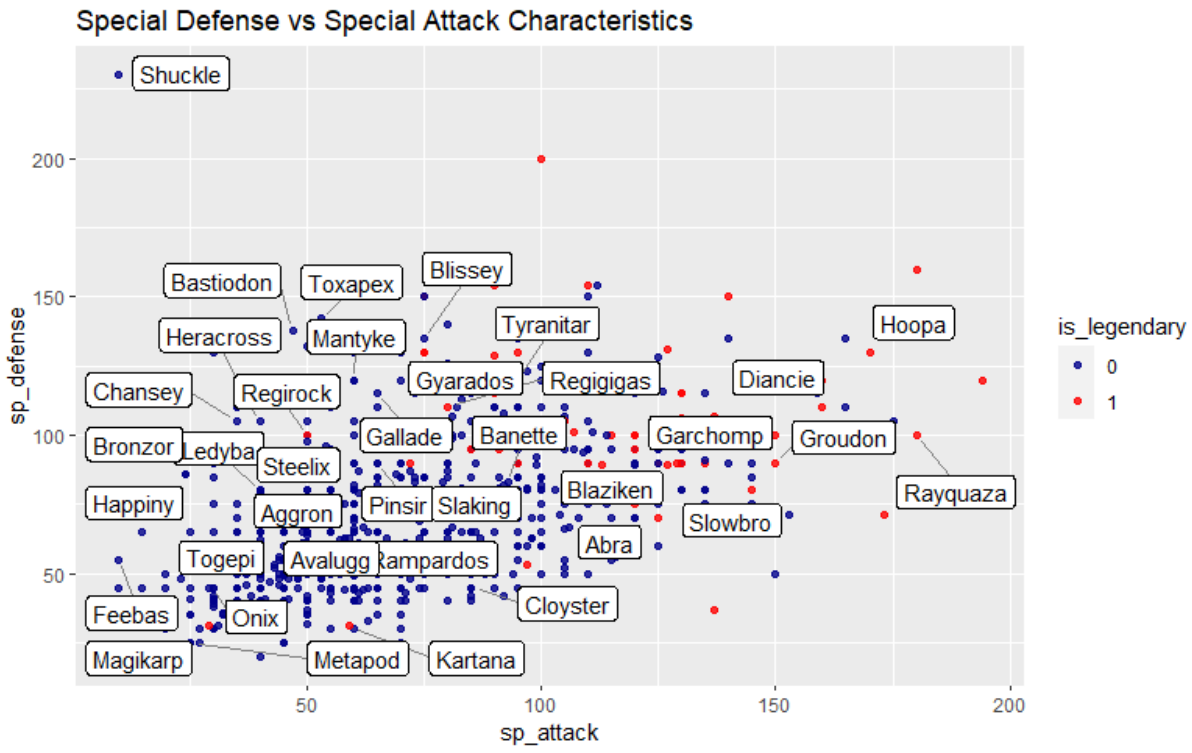


PLOT 17:

Scattered Plot

#plot17

```
pokemon_data$is_legendary<-
as.factor(pokemon_data$is_legendary)
ggplot(data = pokemon_data, aes(sp_attack, sp_defense)) +
      geom_point(aes(color = is_legendary), alpha = 0.8) +
 scale_color_manual(values = c("darkblue", "red")) +
  ggtitle("Special Defense vs Special Attack Characteristics") +
  ggrepel::geom_label_repel(
   data = subset(pokemon_data, attack > 150 | defense > 150 | attack
< 25),   aes(label = name),box.padding = 0.35,point.padding = 0.5,
segment.color = 'grey50')
```



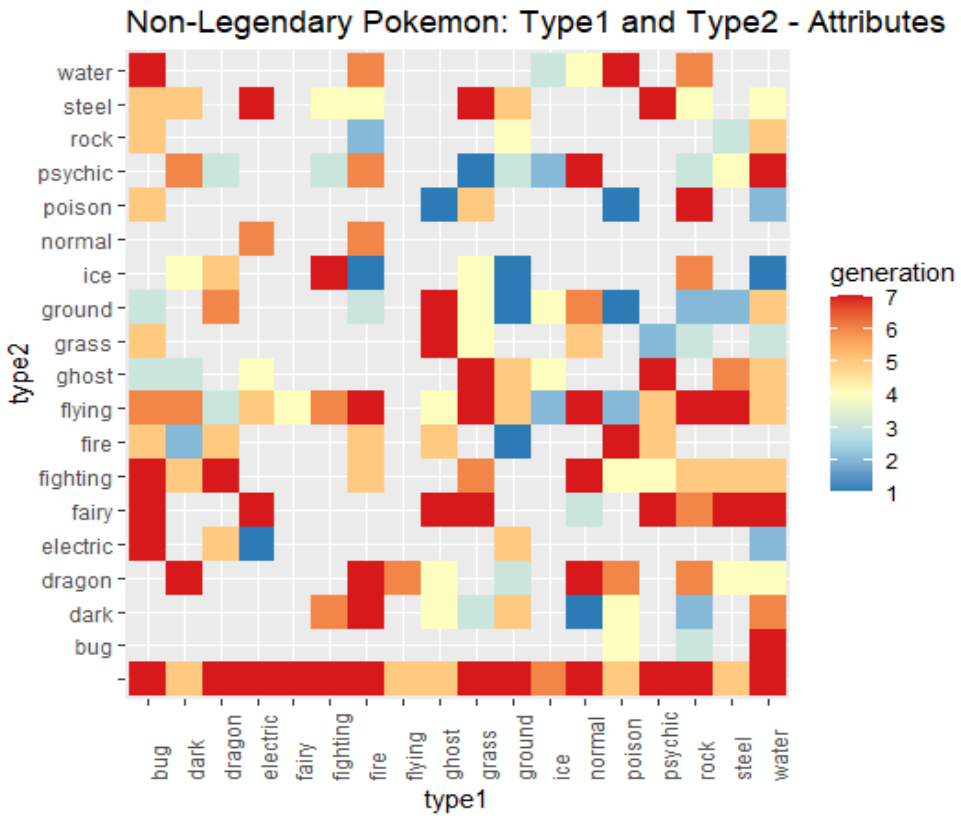
PLOT 18:

Heatmap Plot

#plot18

```
hm.palette <- colorRampPalette(rev(brewer.pal(5, 'RdYlBu')), space='Lab')

ggplot(data = pokemon_data, aes(x = type1, y = type2)) +
  geom_tile(aes(fill = generation)) +
  ggtitle("Non-Legendary Pokemon: Type1 and Type2 - Attributes") +
  scale_fill_gradientn(colors = hm.palette(100)) +
  theme(axis.text.x = element_text(angle = 90, hjust = 0)) +
  coord_equal()
```



PLOT 19:

Horizontal Bar

A horizontal bar chart is a graphical representation of data that uses horizontal bars to display the values of different categories. In a horizontal bar chart, the length of each bar is proportional to the value it represents, and the bars are arranged horizontally along the y-axis. This type of chart is useful for comparing the magnitudes of different categories or groups.

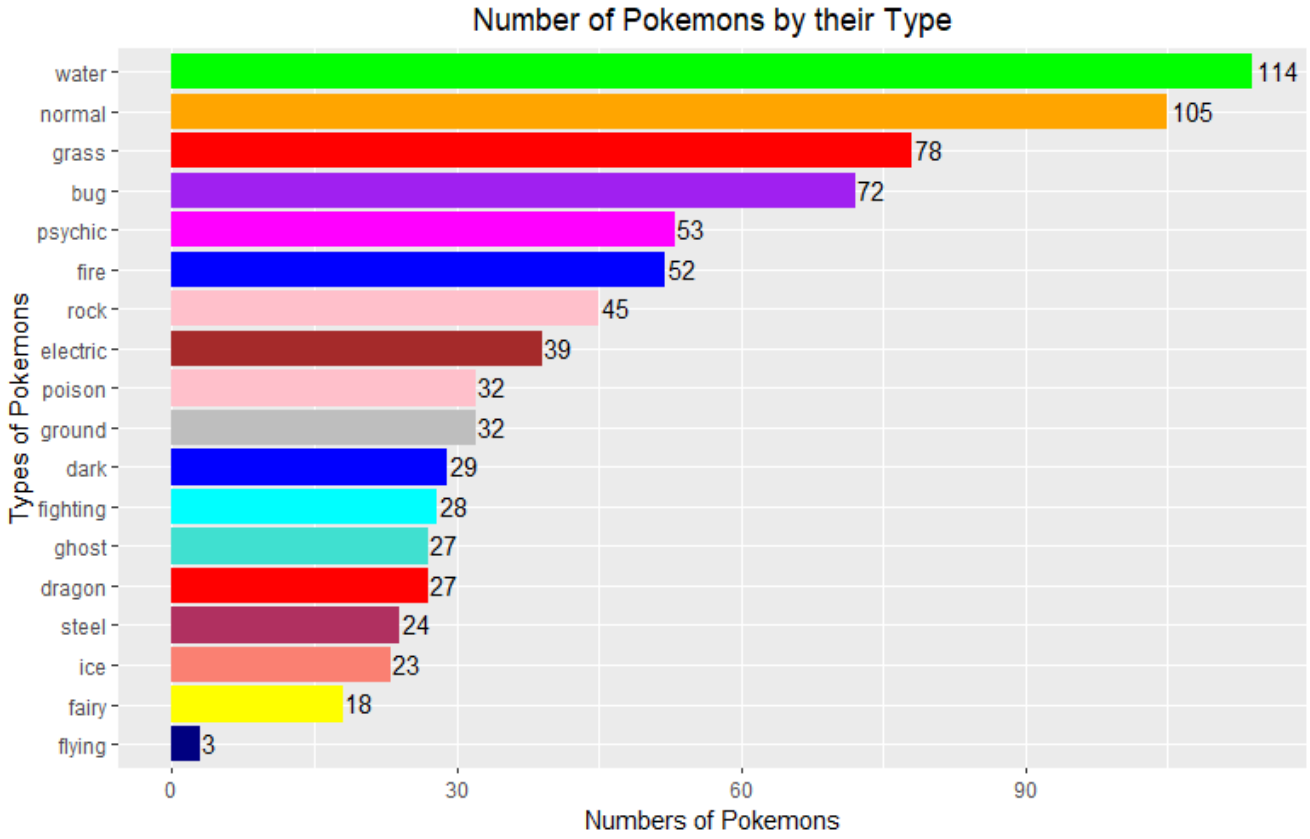
```
#plot19
library(ggplot2)
library(dplyr)
unique_types <- unique(pokemon$type1)
colors <- c("red", "blue", "green", "purple", "orange", "pink", "brown", "grey", "yellow", "cyan", "magenta", "pink", "turquoise", "salmon", "red", "blue", "maroon", "navy")

if (length(colors) != length(unique_types)) {
  stop("Number of colors must match the number of unique types.")
}

pokemon %>%
  group_by(type1) %>%
  summarise(number = n()) %>%
  ggplot(aes(x = reorder(type1, number), y = number, fill = type1))

+
```

```
  geom_bar(stat = 'identity') +
  labs(x = "Types of Pokemons", y = "Numbers of Pokemons", title
= "Number of Pokemons by their Type")
+
  theme(plot.title = element_text(hjust = 0.5))
+
  theme(legend.position = "none")
 +
  coord_flip()
+
  geom_text(aes(label = number), hjust = -0.1, vjust = 0.4)
 +
  scale_fill_manual(values = setNames(colors, unique_types))
```



PLOT 20:

Heatmap Plot

A heatmap is a graphical representation of data where values in a matrix are represented as colors. It is an effective way to visualize complex data sets and identify patterns or trends. In a heatmap, each cell of the matrix is typically assigned a color based on the numerical value it represents.

Importing the Dataset

In the context of plotting points on a heatmap, each point contributes to the density of points within its vicinity. Areas with a higher density of points appear darker on the heatmap, while areas with fewer points appear lighter.

The color intensity or hue reflects the magnitude of the value, making it easy to discern patterns and variations in the data.

#plot20

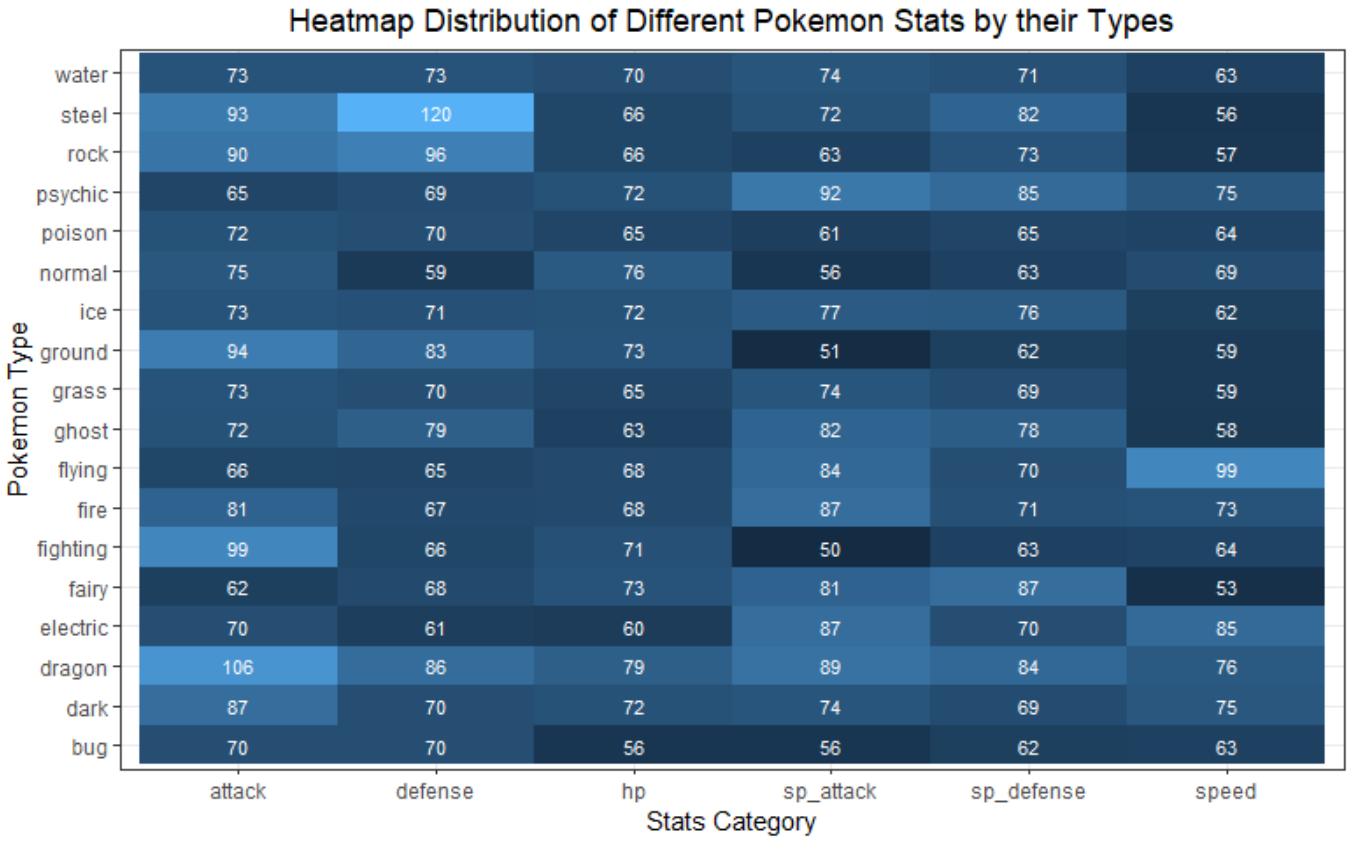
```
pokemon_data %>%
  gather(., key, value, hp, speed, defense, attack, sp_attack,
sp_defense) %>%
  group_by
 (., type1, key) %>%
  summarise(., Stat = as.integer(mean(value))) %>%
  ggplot(., aes(y = type1, x = key))
+
  geom_tile(aes(fill = Stat))
 +
  theme_bw()
+
theme(legend.position = 'bottom')
```

```
+
 geom_text(aes(label = Stat), color = 'white', size = 3)

+

labs(x = 'Stats Category', y = 'Pokemon Type', title = 'Heatmap Distribution of Different Pokemon Stats by their Types')
+
 theme(plot.title = element_text(hjust = 0.5))
+
 theme(legend.position = "none")
```

In this heatmap I have plotted the the various statistics of different pokemon types that were present in the dataset , The various types include “bug”,”Poison”,”Flying”,”Water”,”Grass” etc



Conclusion

In this analysis, we delved into the Pokémon dataset using R, employing a diverse set of graphical representations to gain insights into the characteristics of these iconic creatures. The dataset, containing information on various Pokémon species and their attributes, provided a rich source for exploration.

- Distribution of Pokémon Types:
  Our initial exploration involved visualizing the distribution of Pokémon types. Bar charts and pie charts effectively conveyed the prevalence of different types, shedding light on the diversity within the Pokémon world.

- Statistical Analysis of Pokémon Stats:
  Through box plots and violin plots, we examined the statistical distribution of key attributes such as HP, Attack, Defense, Special Attack, Special Defense, and Speed. These visualizations not only highlighted the central tendencies but also revealed the presence of outliers, contributing to a comprehensive understanding of Pokémon stat distributions.

- Correlation Analysis:
  Utilizing scatter plots and correlation matrices, we explored potential relationships between different pairs of stats. This analysis allowed us to identify positive and negative correlations, uncovering insights into how certain stats may influence each other.

- Legendary Pokémon Analysis:
  A dedicated set of visualizations focused on legendary Pokémon, providing a closer look at how these rare creatures differ in terms of stats compared to their non-legendary counterparts. Box plots and bar charts unveiled distinct patterns in legendary Pokémon attributes.

This analysis serves as a testament to the power of data visualization in extracting actionable insights from complex datasets.