Data Science EST

ggplot2

ggplot2 (***grammar of graphics***) is a data visualization package for the R programming language. It provides a powerful and flexible framework for creating complex and aesthetically pleasing visualizations by layering different components such as data, aesthetics, geometries, and statistical transformations.

```
Loading ggplot2
library(ggplot2)
```

Introduction

ggplot2 builds plots using a *layered* approach, the various layers are:

1. Data: The dataset to be visualized.
2. Aesthetics: Mappings between data variables and visual properties (e.g., x and y axes, colors, sizes).
3. Geometries: The visual representation of the data (e.g., points, lines, bars).

Basic Plot

To create a basic scatter plot using ***ggplot2*** , you can use the following code:

```
Basic Scatter Plot
ggplot(data = mtcars, aes(x = wt, y = mpg)) +
  geom_point()
```

This code creates a scatter plot of the `mtcars` dataset, mapping the weight (`wt`) to the x-axis and miles per gallon (`mpg`) to the y-axis.

`geom_point()` adds points to the plot, representing individual data points.

The other common geometries include:

1. `geom_line()`: For line plots.
2. `geom_bar()`: For bar charts.
3. `geom_histogram()`: For histograms.
4. `geom_boxplot()`: For box plots.

Customization

Its possible to add various customizations to the plots, such as titles, labels, themes, and colors. For example:

```
Customized Scatter Plot
ggplot(data = mtcars, aes(x = wt, y = mpg, color = factor(cyl))) +
  geom_point(size = 3) +
  labs(title = "Scatter Plot of MPG vs Weight",
       x = "Weight (1000 lbs)",
       y = "Miles per Gallon",
       color = "Cylinders") +
  theme_minimal()
```

This code adds color to the points based on the number of cylinders (`cyl`), customizes the point size, and adds labels and a minimal theme to the plot.

```
Grouping by Colors
color = factor(cyl)
```

This maps the `cyl` variable to different colors in the plot, allowing for easy differentiation between groups based on the number of cylinders in the cars.

For example, assume the `cyl` variable has three unique values: 4, 6, and 8. The points representing cars with 4 cylinders will be colored differently from those with 6 or 8 cylinders, making it

easier to visualize how the number of cylinders relates to the weight and miles per gallon of the cars in the dataset.