Shiny App for Exploratory Data Analysis

Abhishek Umrawal

===

As course project for the Developing Data Products course of the Data Science Specialization offered by Bloomberg School of Public Health of the Johns Hopkins University through Coursera, I have created a *Shiny App for Exploratory Data Analysis*.

1. How the evaluator should reach to this Shiny App?

There are following two approaches for this:

1.1. Directly Follow the Weblink

I have deployed this App by means of *ShinyApps.io* which provides a weblink for the App, which is as follows:

https://abhishekumrawal.shinyapps.io/ShinyAppforEDA

1.2. Running the App from R Console

I have deployed this App by means of creating a GitHub Repository named as *ShinyApp* under my username *abhishekumrawal*. Everything including this documentation is shared there. The evaluator has to *simply type and run the following command on the RStudio console*:

runGitHub("ShinyAppforEDA", "abhishekumrawal")

2. How to Use the App?

2.1 App Input

The input data for the App has to be in a CSV file which is to be located in the working directory. However for trial run following datasets are already included:

Continuous Type: Normal, MilesPerGallon and PickUpTime.

Discrete Type: Poission, #ofCylinders and #ofCarburetors.

The App Input Panel asks, **Please enter name of the CSV data file without extension:** where one needs to enter the name of the data file for instance **Normal**.

The Exploratory Data Analysis has to be performed differently for Continuos and Discrete Data Types and hence the second input for the App is taken through, **Please select the Data Type:** one has to choose one from **Continuos** and **Discrete** depending upon the input Data Type.

2.2 App Execution

App execution is controlled by a Submission Button which says **Perform EDA**, as one clicks on it, the output can be seen in the Output Panel.

2.3 App Output

The App performs the **Exploratoray Data Analysis (EDA)** on the chosen data which is an approach to analyzing datasets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task. We perform the following as part of EDA.

2.3.1 Summary Statistics

In this section Basic Statistics like Mean, Median, Quartiles, Minimum and Maximum are calculated for the given data.

2.3.2 Data Visualization

In this section given data is visualized using common statistical plots.

2.3.2.1 Stem and Leaf Plot

Stem and Leaf Plot enables the user to visualize the entire data exhibiting the shape of the distribution, skewness, kurtosis, points of concentration, gaps in data etc.

2.3.2.2 Histogram/Bar Plot

If the data type is continuous then Histogram enables the user to visualize the shape of the distribution. Histogram overlaid with normal probability curve can be used to test if the given data follows a normal distribution.

If the data type is discrete then Bar Plot enables the user to visualize the frequency distribution.

2.3.2.3 Box Plot

Box Plot also enables the user to visualize the shape of the distribution and skewness. The heighth of the box represents the Inter Quartile Range (a measure of dispersion).

Box Plot enables the user also to identify the outliers (the points lying outside the whiskers endpoints) in the data.

2.3.3 Tests for Normality (if Data Type is continuous)

In this section given data (if it is continuous) is tested for normality using common statistical procedures.

2.3.3.1 Q-Q Plot

Q-Q Plot enables the user to identify if the data follows a normal distribution. It plots the sample quantiles of the data against the theoretical normal distribution quantiles.

If most of the data points fall on the 45-degree straight line in the Q-Q Plot, the data can be taken to be following normal distribution.

2.3.2.2 Shapiro-Wilk Test for Normality

Shapiro-Wilk Test is a statistical procedure to test for normality of the data.

If p-value is greater than 0.05 then the data can be taken to be following normal distribution with 95% confidence.

Conclusion: The activity patterns are observed to be different between Weekdays and Weekends. Weekend activity patterns are relatively more volatile (random) as compared to Weekday activity patterns.