WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Web Application for Automated Exploratory Data Analysis

Dr. Deika Rani Dhivya. K [a], *Siva Ranjini. C* [b]

[a] *Assistant Professor, Sri Krishna Arts and Science College, Coimbatore-641008, India*
[b] *Student, Sri Krishna Arts and Science College, Coimbatore-641008, India*

ABSTRACT

The data is generated everywhere, every time in this world. The data which is analyzed is used to get insights to utilize the data productively. The data need to be collected and cleaned, exploratory data analyzed, and a model must be built and deployed. The exploratory data analysis will make us to understand the features and characteristics of the collected data. The website is built to automate the EDA process with incorporated Python language, which automates common actions with a simple and elegant user interface. The data interpretation is done through rows and columns. Python is an object-oriented, interactive language with a rich set of libraries such as pandas, numpy, matplotlib, seaborn, etc. The Python code is embedded in the Streamlit library which is deployed as a web app. This open-source library which is Python based runs with the web framework of Flask and Tornado. It is built on top of Python which supports mainstream libraries such as pandas, numpy, seaborn, etc.

Keywords: Exploratory data analysis, python, pandas, streamlit

1. Introduction

The data is different types of information generally formatted in a particular manner. These data are generated in this world in numerous manners in the form of big data. The data need to be anatomized to make practicable knowledge and perceptivity from raw data and to ameliorate the growth of colorful disciplines. There comes the concept of data science and data analysis. We use data science to make insights from data. Data science is defined as a field that combines knowledge of mathematics, programming chops, sphere skills, scientific styles, algorithms, processes, and systems to prize practicable knowledge and perceptivity from both structured and unshaped data, also apply the knowledge picked from that data to a wide range of uses and disciplines. Data analytics allow businesses to understand their effectiveness and performance, and eventually helps the business make further informed opinions. For example, A company like Netflix, Youtube, and amazon uses a recommendation system in data science to lead users with applicable suggestions grounded on the choices they make. Information is also the reused data used to make opinions and take action. Reused data must meet the following benchmark for it to be of any significant use in decision-making E-commerce companies use various visualization method to understand data. The data science process contains the following steps collection, cleaning, exploratory data analysis, feature engineering, model building, and deployment. Exploratory data analysis(EDA) is an approach to epitomize the data by taking its main characteristics and fantasizing it with proper representations. It helps determine how casually to manipulate data sources to get the answers which made more accessible for data scientists to discover patterns, spot anomalies, test a thesis, or check hypotheticals. EDA focuses more hardly on checking hypotheticals needed for model fitting and thesis testing, handling missing values, and making metamorphoses of variables as demanded. EDA encompasses IDA. EDA snappily describes the data set number of rows columns, missing data, data types, and exercise. Clean spoiled data; handle missing data, invalid data types, and incorrect values. EDA fantasizes data distributions; bar maps, histograms, and box plots. Calculate and fantasize correlations(connections) between variables; heat map. Python and R are generally used tools to perform EDA. This process is automated into a web app so it would be easy to do the tasks without rendering. Python code is embedded in the Streamlit library which places a web app to do the process. This paper describes the multihued techniques of EDA which are done by just dragging and dropping datasets into our website and checkout numerous operations to understand the dataset before feature engineering in the data wisdom process.

1.1 Types of Exploratory Data Analysis

Univariate non-graphical. This is the simplest form of data analysis, where the data being anatomized consists of just one variable. Since it's a single variable, it doesn't deal with causes or connections. Univariate analysis is mostly used to explain data and identify patterns that exist within it. Univariate visual. Non-graphical styles don't give a full picture of the data. Graphical styles are thus needed. Common types of univariate plates include Stemand-splint plots, which show all data values and the shape of the distribution. Histograms are bar plots in which each stick represents the frequency (count) or proportion(count/ aggregate count) of cases for a range of values. Box plots graphically depict the five-number summary of minimum, first quartile, standard, third quartile, and outside. Multivariate nongraphical Multivariate data arises from further than one variable. Multivariate non-graphical EDA ways generally show the relationship between two or further variables of the data through cross-tabulation or statistics.

Multivariate graphical -Multivariate data uses plates to display connections between two or further sets of data. The most habituated visual is a grouped line graph or bar map with every group representing one position of one of the variables and each bar within a group representing the situations of the other variable. Other common types of multivariate plates include a scatter plot, which is used to compass data points on a vertical and a perpendicular axis to show how important one variable is affected by another. The multivariate map is a graphical representation of the connections between factors and responses. Run map, which is a line graph of data colluded over time. A bubble map is a data visualization that displays multiple circles(bubbles) in a two-dimensional plot. The heat map is a pictorial representation of data where values are depicted by shades of color.

2. Methodology

The following procedures are used to perform a complete EDA process in a web app:

Step1: Import the data

The application begins with uploading the dataset. They were supplied as the three maximum widely used file codecs are CSV, text, and Excel documents. In line with the feedback, new record formats will be brought. A dataset, or statistics set, is without a doubt a collection of data. The most effective and most not unusual layout for datasets you will find online is a spreadsheet or CSV format — an unnamed file organized as a desk of rows and columns. However some datasets can be saved in other codecs, and they don't must be just one file. The streamlit library offers a function called file_uploader which provides a drag-and-drop file option to the web app.

Step 2: Display the dataset

The head() and tail() function of the panda’s library is used to display the imported dataset in the web app. The head and tail function is embedded in the data frame function of the streamlit library. The default head() function displays the first five rows in that dataset. The tail() function displays the last five rows with respective columns.

Step 3: Sort the dataset

The if clause of the Python language is used to create a checkbox in the n streamlit library. The sort() function of the panda’s language is used to sort the dataset. This function will sort the entire dataset into ascending or descending order. This code is embedded in the write function of the streamlit library to display the sorted dataset with index or values.

Step 4: Include the Memory usage

Inside the if clause the checkbox function for memory, operation() is enrooted. This pandas library function is used to display each column's memory operation independently in bytes. The code is written inside the write() function of the streamlit library and displayed

in the web app.

Step 5: Show the unique value count

The dataset may have a huge quantity of data. It's important to know about the unique count of each column in the dataset before disemboweling it to avoid data redundancy. This function is entrenched in the streamlit library’s write function under the if clause. This count is displayed in the web app for each column.

Step 6: Datatype of Column

The dtypes() function in the panda’s library gives the datatype of the particular column. This makes the addict understand the basal datatype of which every column through EDA. The code is bedded in the streamlit library so that the checkbox is created to view the result.

Step 7: Display the information of the dataset

The info() function of the panda’s library gives information about the imported dataset similar as column details, memory operation, and description. This function is enciphered inside the streamlit library’s checkbox function so that it'll be displayed to the user through the

web app.

Step 8: Identify the missing values

The isnull() function will return the missing values or blank cells in every column of the dataset. This system is bedded in the streamlit library’s checkbox function inside the if clause to display it on the web app.

Step 9: Know the shape of the dataset

The shape() function of the pandas library will return the no. of rows and no. of columns of the imported dataset in the form of(x, y). This the function is embedded in streamlit library’s checkbox function under the if clause which is displayed in the web app.

Step 10: Understand the summary of the dataset

The describe() method will give the descriptive statistics report of the imported dataset. The summary of the dataset such as mean, quartile, percentile, count, minimum value, maximum value, and standard deviation. This function is written inside the checkbox function of the streamlit library which will be displayed through the web app to the user.

Step 11: Display the value count

The value count() function returns the no. of values present in a particular column that constantly do in the dataset. This is enciphered inside the streamlit library so that it'll be displayed in the web app.

Step 12: Declare the Correlation of the dataset

Data Correlation is used to understand the relationship between multiple variables and attributes in your dataset. Using Correlation, you can get some perceptivity similar to One or multiple attributes depending on another trait or a cause for another trait. The matplotlib library's pyplot's sub library's math show function and the seaborn library's heatmap() plot function are used to display the correlation graphically.

Step 13: Visualize the histogram

A histogram is the distribution of numeric column values. It basically creates bins in different ranges of values and plots them

where we can visualize how the values are distributed. The streamlit library st.pyplot() is called and the histogram is generated for the chosen field from the dataset.

Step 14: Put a bar chart

The graph that represents categorical values with bars with height and length proportional to the values is known as a

bar chart. The plots may be in vertical or in horizontal positions. The streamlit library st.bar_chart() is used to generate

and display bar charts in Python.

Step 13: Display the pie chart

A Pie Chart is a statistical layout in a circular form that can only show one series of data. The chart's area represents the overall

percentage of the data. The area of the pie slices reflects the proportion of the data. The ploty_chart() from the Streamlit library is used to generate the pie chart in the web app.

Step 14: Display the line Graph

The individual data points of quantitative values are connected using the line in the Line graph with a specified time interval.

The st.line_chart() function in the streamlit library is used to create and display the line_chart on the web app.

Step 15: Identify outliers through the box plot

The box plot is used to visualize the median, percentile, and quartile of the selected column. It also displays the outliers present in that particular column. The streamlit library is used to import box plots into the web app in Python.

2.2 Tools and Platforms Used to build a web app

- OS
- Windows
- Linux
- Python IDE: python 2.7.x
- Streamlit
- Python IDLE Shell 3.11.1, Anaconda

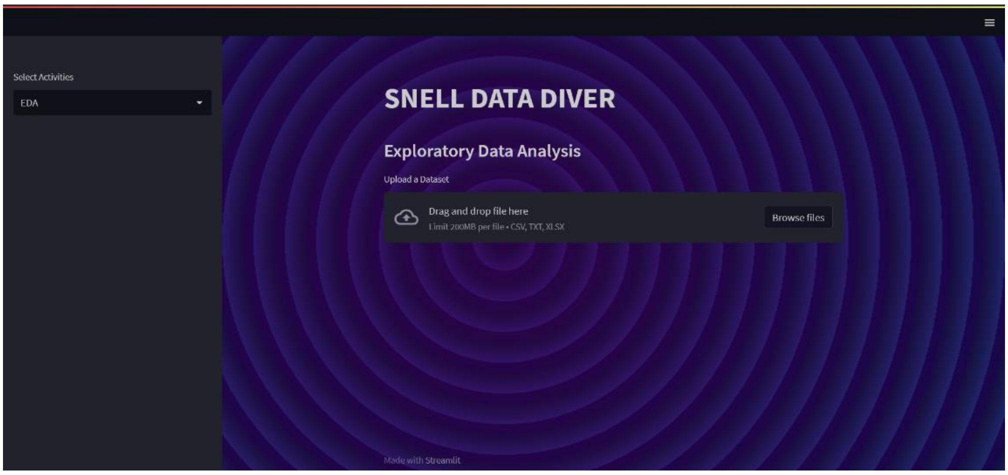
Prompt(anaconda3)

2.3 Hardware Used

- RAM:4GB
- Processor: Intel i3
- Hard Disk:500GB

3. Research and Discussion

The overall result turned out helpful to build a simple web app that automates the process of exploratory data analysis. The data is imported and further steps are done to understand the dataset and view the characteristics of every column through visualizations such as histograms, line charts, box plots, pie charts, etc.. The web app is used for time-saving and efficient usage of datasets by the user. Moreover, the plots are used for graphical EDA and the functions of the panda's library are used for non-graphical EDA.

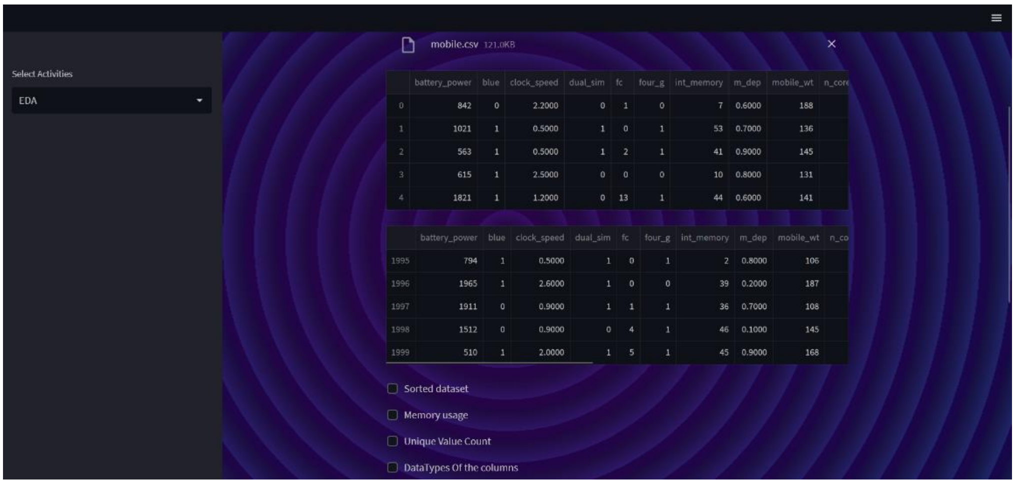


SNELL DATA DIVER

Exploratory Data Analysis

.

Fig 1: Snell data diver EDA phase



	battery_power	blue	clock_speed	dual_sim	fc	four_g	int_memory	m_dep	mobile_wt	n_core
0	842	0	2.2000	0	1	0	7	0.6000	188	
1	1021	1	0.5000	1	0	1	53	0.7000	136	
2	563	1	0.5000	1	2	1	41	0.9000	145	
3	615	1	2.5000	0	0	0	10	0.8000	131	
4	1821	1	1.2000	0	13	1	44	0.6000	141	

	battery_power	blue	clock_speed	dual_sim	fc	four_g	int_memory	m_dep	mobile_wt	n_co
1995	794	1	0.5000	1	0	1	2	0.8000	106	
1996	1965	1	2.6000	1	0	0	39	0.2000	187	
1997	1911	0	0.9000	1	1	1	36	0.7000	108	
1998	1512	0	0.9000	0	4	1	46	0.1000	145	
1999	510	1	2.0000	1	5	1	45	0.9000	168	

☐ Sorted dataset

☐ Memory usage

☐ Unique Value Count

☐ DataTypes Of the columns

Fig 2: Nongraphical functions in EDA

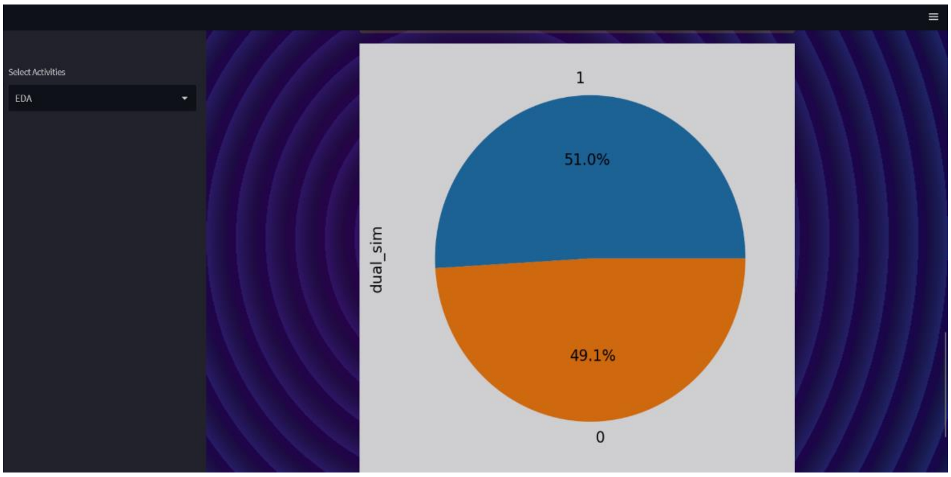
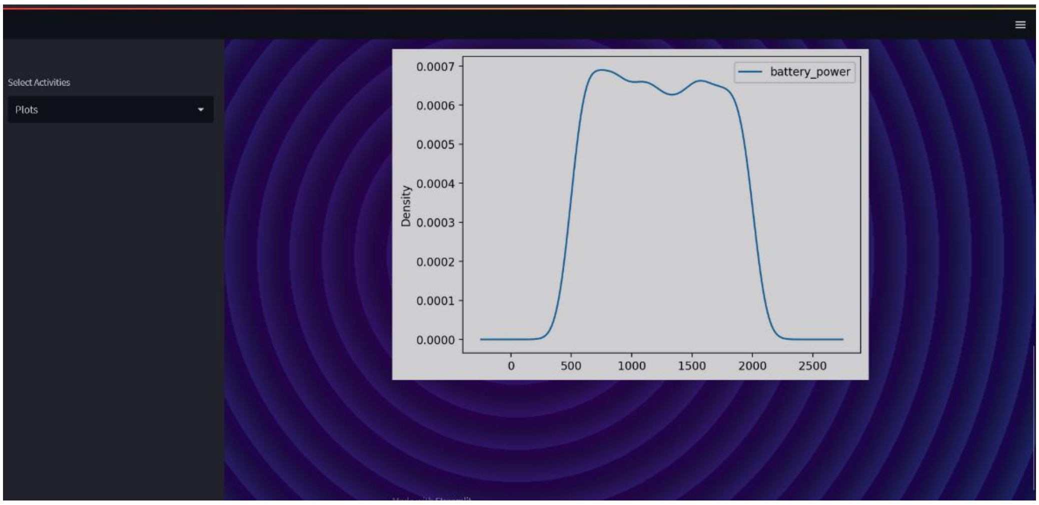


Fig 3: Graphical EDA phase of the Web app



Select Activities

Plots

Fig 4: Plots related to imported datase

4. Conclusion

We might arrive at the conclusion that exploratory data analysis is a tried-and-true approach that can assist Data Scientists in understanding the meaning of complicated datasets. You can find patterns and correlations that you would not have discovered otherwise by utilizing visualizations and other techniques. The Snell Data Diver web app has provided an efficient user interface to perform EDA without coding. It is accomplished with some functions and visualizations where the user just understands the patterns, trends, and structure of the imported datasets. This web app makes the user reduce their repetitive process in every EDA. This also helps to reduce user time and data storage from downloading various software, libraries, importing them, etc. This is a user-friendly web app that also provides data security to user datasets

References

What is Exploratory Data Analysis? | IBM

Exploratory Data Analysis (EDA): Types, Tools, Process (knowledgehut.com)

EDA - Exploratory Data Analysis: Using Python Functions | DigitalOcean

GitHub-PacktPublishing/Hands-on-ExploratoryData-Analysis-with-Python: Hands-on Exploratory Data Analysis with Python, published by Packt [5] AI Deploy - Tutorial - Deploy an interactive app for EDA and prediction using Streamlit | OVH Guide