Exploratory Data Analysis

Submitted in partial fulfillment of the requirements of the degree of

T.E in Computer Engineering

By

Roll no. 05 Tushar Avhad

Roll no. 24 Prathamesh Kadam

Roll no. 44 Bhavesh Patil

Roll no. 72 Mahendra Valvi

Supervisor:

Prof. Kanchan Doke

(Department of Computer Engineering)

Bharati Vidyapeeth College of Engineering, Navi Mumbai

Table of Contents

Chapter – One

Introduction

1.1 Abstract

The main aim of the project is to perform analysis easily, conveniently with reusable code. All this can be done by just our project "Exploratory Data Analysis". Exploratory data analysis (EDA) is sometimes suggested as a hypothesis identification approach. It is often used as such in problem solving and consists of the analysis of observational data, often collected without well-defined hypotheses, with the purpose of finding clues that could inspire ideas and hypotheses. This project includes BoxPlot, ScatterPlot, Histogram, Heatmap and Bar Plot of a given dataset. Project should have emphasis on integration, experiential learning, and real-world problem solving and hence we choose the project by considering all these factors mentioned above. During the third semester, all the group members worked on the Project Planning which included preparing report, presentation and Project Proposal.

Keywords: EDA (Exploratory Data Analysis), CSV (Comma-Separated Values), Heatmap (Map used to check correlation between two attributes), pandas, Matplotlib, seaborn, NumPy.

1.2 Proposed Problem

Exploratory data analysis (EDA) is sometimes suggested as a hypothesis identification approach. It is often used as such in problem solving and consists of the analysis of observational data, often collected with-out well-defined hypotheses, with the purpose of finding clues that could inspire ideas and hypotheses. It is a good practice to understand the data first and try to gather as many insights from it. Exploratory Data Analysis is a crucial step before you jump to machine learning or modeling your data. By doing this you can get to know whether the selected features are good enough to model, are all the features required, are there any correlations based on which we can either go back to the Data Preprocessing step or move on to modeling.

Once **EDA** is complete and insights are drawn, its feature can be used for supervised and unsupervised machine learning modeling. By completing the EDA you will have many plots, heat-maps, frequency distribution, graphs, correlation matrix along with the hypothesis by which any individual can understand what your data is all about and what insights you got from exploring your data set. Data scientists can use exploratory analysis to ensure the results they produce are valid and applicable to any desired business outcomes and goals. EDA can help answer questions about standard deviations, categorical variables, and confidence intervals. Once EDA is complete and insights are drawn, its features can then be used for more sophisticated data analysis or modeling, including machine learning.

EDA is all about making sense of data in hand, before getting them dirty with it. By completing the EDA, you will have many plots, heat-maps, frequency distribution, graphs, correlation matrix along with the hypothesis by which any individual can understand what your data is all about and what insights you got from exploring your data set.

1.3 Definition and explanation

Exploratory Data Analysis (EDA) is an approach to analyze the data using visual techniques. It isused to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations. Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions.

EDA is primarily used to see what data can reveal beyond the formal modeling or hypothesis testing task and provides a provides a better understanding of data set variables and the relationships between them. It can also help determine if the statistical techniques you are considering for data analysis are appropriate. Originally developed by American mathematician John Tukey in the 1970s, EDA techniques continue to be a widely used method in the data discovery process today.

1.4 Aim and Scope

The main aim of the project is to perform analysis easily, conveniently with reusable code. All this can be done by just our project "Exploratory Data Analysis". This project includes BoxPlot, ScatterPlot, Histogram, Heatmap and BarPlot of a given dataset.

In the days of Tukey-style EDA, the analyst was typically well aware of how the data they were analyzing was generated. However, now as organizations generate vast numbers of datasets internally as well as acquire third-party data, the analyst is typically far removed from the data generation process. If the data is not what you think it is, then your results could be poorly affected, or worse, misinterpreted and acted on.

One example of a way data generation can be misinterpreted and cause problems is when data is provided at the user level but is actually generated at a higher level of granularity (such as for the company, location, age group the observation is a part of). This situation results in data being the same for otherwise disparate users within a group. Insight implies detecting and uncovering underlying structure in the data. Such underlying structure may not be encapsulated in the list of items above; such items serve as the specific targets of an analysis, but the real insight and "feel" for a data set comes as the analyst judiciously probes and explores the various subtleties of the data. The "feel" for the data comes almost exclusively from the application of various graphical techniques, the collection of which serves as the window into the essence of the data.

Chapter – Two

Literature Survey

2.1 Published Papers

Author	Title	Published at	Abstract
Kabita Sahoo, Abhaya Kumar Samal, Jitendra Pramanik, Subhendu Kumar Pani	Exploratory Data Analysis using Python	October 2019 International Journal of Innovative Technology and Exploring Engineering	This paper gives information about Exploratory data analysis. This paper also gives us knowledge regarding python. it is object oriented interpreted and interactive programming language. it is open source with rich sets of libraries like pandas, Matplotlib, seaborn etc.
Matthieu Komorowski, Justin D Salciccioli, Dominic C Marshall, Yves Crutain	Exploratory Data Analysis	September 2016 In book: Secondary Analysis of Electronic Health Records	From this paper, the reader will learn about the most common tools available for exploring a dataset, which is essential in order to gain a good understanding of the features and potential issues of a dataset, as well as helping in hypothesis generation.
Chong Ho Yu	Exploratory data analysis in the context of data mining and resampling	June 2010 International Journal of Psychological Research	EDA is introduced in the context of data mining and resampling with an emphasis on three goals : cluster detection, variable selection, and pattern recognition. TwoStep clustering, classification trees, and neural networks, which are powerful techniques to accomplish the preceding goals, respectively, are illustrated with c Concrete examples

Frederick Hartwig, Brian E. Dearing	Exploratory Data Analysis	1979 SAGE University Paper	An introduction to the underlying principles, central concepts, and basic techniques for conducting and understanding exploratory data analysis - with numerous social science examples.
Suresh Kumar Mukhiya, Usman Ahmed	Hands-On Exploratory Data Analysis with Python	27 March 2020 by Packt Publishing	Discover techniques to summarize the characteristics of your data using Pyplot, NumPy, SciPy, and pandas.

2.2 Study Of Existing Systems

Type Of EDA Technique	Description
Univariate Graphical EDA	Univariate GEDA provides statistical summary for each field in the raw data set or the summary only on one variable.
Bivariate Graphical EDA	Bivariate GEDA is accomplished to understand the connections between each variable in the dataset and the target variable of interest or using two variables and finding connection among them.
Multivariate Graphical EDA	Multivariate GEDA is accomplished to understand the connections between different fields in the dataset or finding the connections between more than two variables.
Univariate Non-graphical EDA	Univariate Non-graphical EDA is the simplest form of data analysis as during this we use just one variable to research the info. The standard goal of univariate non-graphical EDA is to know the underlying sample distribution/ data and make observations about the population
Multivariate Non-graphical EDA	Multivariate non-graphical EDA technique is usually wont to show the connection between two or more variables within the sort of either cross-tabulation or statistics.

Chapter – Three

Methodology

3.1 Technologies Used

- **Java:** Python is a high-level, general-purpose and a very popular programming language. Python programming language (latest Python 3) is being used in web development, Machine Learning applications, along with all cutting-edge technology in the Software Industry.
- **Pandas:** Pandas is an open-source library that is made mainly for working with relational or labeled data both easily and intuitively. It provides various data structures and operations for manipulating numerical data and time series. This library is built on top of the NumPy library. Pandas is fast and it has high performance & productivity for users.
- **Matplotlib:** Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like tkinter, wxPython, Qt, or GTK.
- **Seaborn:** Seaborn is a Python data visualization library based on matplotlib. It is mostly used for statistical plotting in Python. It provides a high-level interface for drawing attractive and informative statistical graphics.
- **Streamlit:** Streamlit is an open-source python framework for building web apps for Machine Learning and Data Science. We can instantly develop web apps and deploy them easily using Streamlit. Streamlit allows you to write an app the same way you write python code. Streamlit makes it seamless to work on the interactive loop of coding and viewing results in the web app.

3.2 Functionalities

- **Data Exploration:** It is the first stage of data analysis. It tells about the size of the data. We can find the missing value of data. We can find the possible relationship among data. Data visualization is done by the use of tabular data and understanding the characteristics.
- **Data Managing:** It is process of detecting the corrupt data, removing the irrelevant parts

of the data and replacing the correct data. The actual process of data cleaning is to remove the error and validating the data. Data can be cross checked to remove the error. Issue can be resolved by validating the data.(Fig 1.)

- **Model Building:** We use the statistical model or machine learning model to describe the variable and working of the variable. Model can be supervised or unsupervised model. We can use classification, regression model to get the output. We can visualize the result by the use of model. After that we have to evaluate the model.
- **Present Result:** We can visualize large amount of complex data by the use of chart, graph and tables. Human brain can process information using chart, graphs. It is an easy way to convey the concept. It can identify the area which needs improvement. It can clarifythe factor very well.

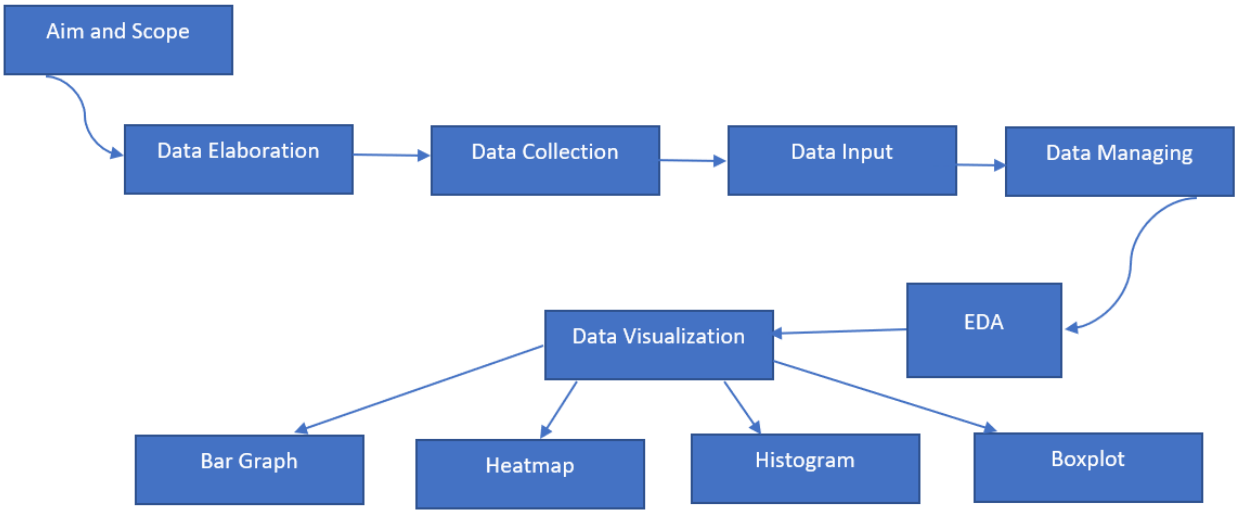

Fig 1. Architecture of System

3.3 Methodology / Project Workflow

- ***Steps for Exploratory Data Analysis:***

  Step 1: Import all the libraries needed.
  Step 2: Read the CSV file with read_csv().
  Step 3: Display dataset with head() and tail().
  Step 4: Check for duplicate values.
  Step 5: Remove Duplicate Values.
  Step 6: Check for NULL Values.
  Step 7: Remove NULL values.
  Step 8: Detect Outliers.
  Step 9: Plot BoxPlot.
  Step 10: Plot Histogram.
  Step 11: Plot Heatmap.
  Step 12: Plot BarGraph.

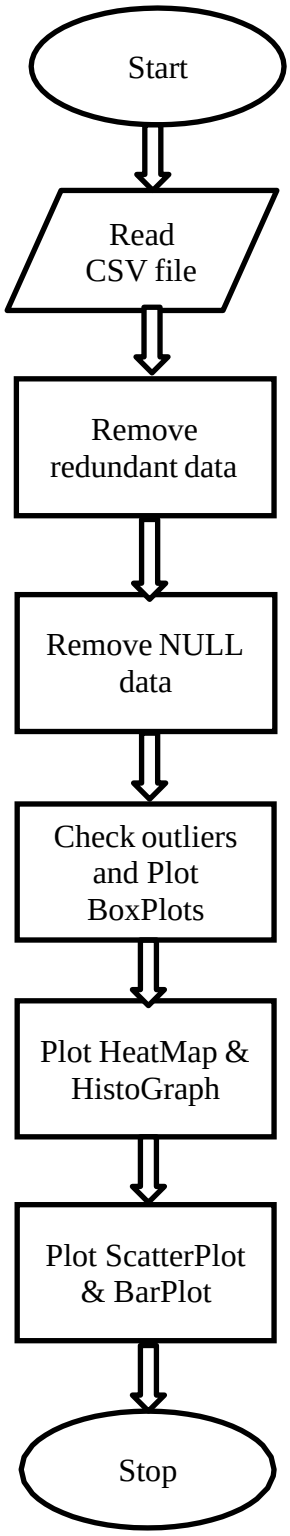

Fig 2. Flowchart for performing Exploratory Data Analysis

Chapter – Four

Results and Discussion

4.1 Current Outcomes

When a team or an individual is handling a project, it must be perfect, right from the initial time phase to the final product without any fault. Implementation of any project means the correct execution of that particular work, which later proves to be a useful and successful tool. After the planning and structuring are completed, project implementation is the most crucial phase. It is where the plan is put into action, and all the strategies are carried out. The thought-out structure is executed in this phase, and the results are analyzed later. Taking into account all these points, we hereby look into the two major topics under this section. First and foremost is the Technology stack and the second one is the Output.

The primary goal of Exploratory Data Analysis is to assist in the analysis of data prior to making any assumptions. It can help with the detection of obvious errors, a better comprehension of data patterns, the detection of outliers or unexpected events, and the discovery of interesting correlations between variables. Data scientists can employ exploratory analysis to ensure that the results they produce are accurate and acceptable for any desired business outcomes and goals.

Chapter – Five

Conclusion

5.1 Conclusion

Exploratory data analysis comes in handy whenever a data scientist needs to gain new insights into a massive quantity of data sets. In this aspect, EDA can be beneficial for fields such as research and development, engineering, and data science. Hence, In today's age, with access to advanced computing power along with the support of modern analytics. EDA can be a stimulating and engaging experience for researchers or data scientists to explore unexpected value in a massive quantity of complex data sets.

We can also implement prediction and plot prediction graphs; we can do multiple things using EDA. This is a very simple example like execution, we can create good marketing and business strategies from analysis like EDA and many more. we have explained the detail about explorative data analysis. We have used the language python programming language for implementation. We have used jupyter note book for detail analysis. We have implemented different library packages of python. We got the required result taking different parameter. In future we will use more data sets andother functions to get the clear idea related to exploratory data analysis.

5.2 Scope for Future Development

Business Intelligence and analytics tools will continue to focus on usability and increasing natural language that enables business users to extract data and create reports without needing to understand the underlying algorithms. Not only will this increase efficiencies and create further adoption throughout companies, but it will also help alleviate some of the problems created by the data scientist shortage. EDA also assists stakeholders by ensuring that they are asking the appropriate questions. Standard deviations, categorical variables, and confidence intervals can all be answered with EDA

In future we plan to expand our application by adding functionalities for downloading images we get from the analysis. We also plan to provide to provide a full-fledged report on the analysis of the dataset. This will assist users or companies to have a detailed summary of analysis in a pdf format for future use

Chapter – Six

Appendices

6.1 Screenshots

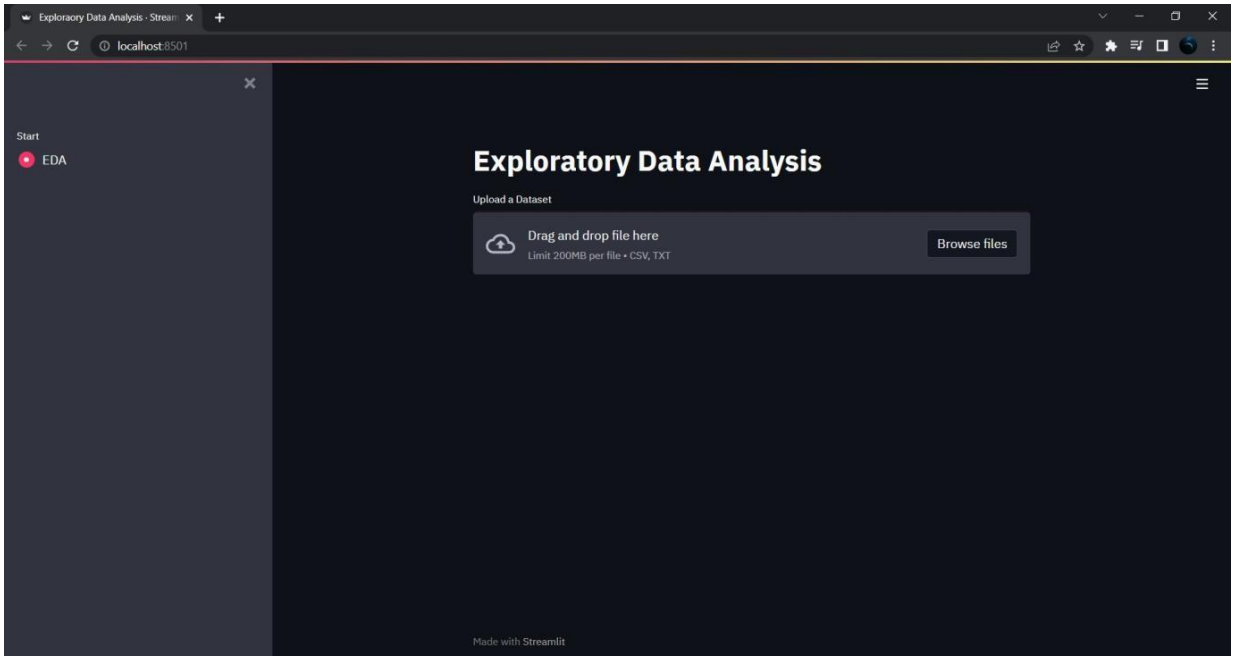

Fig 3 *Starting page where the user is asked to upload the dataset (csv format)*

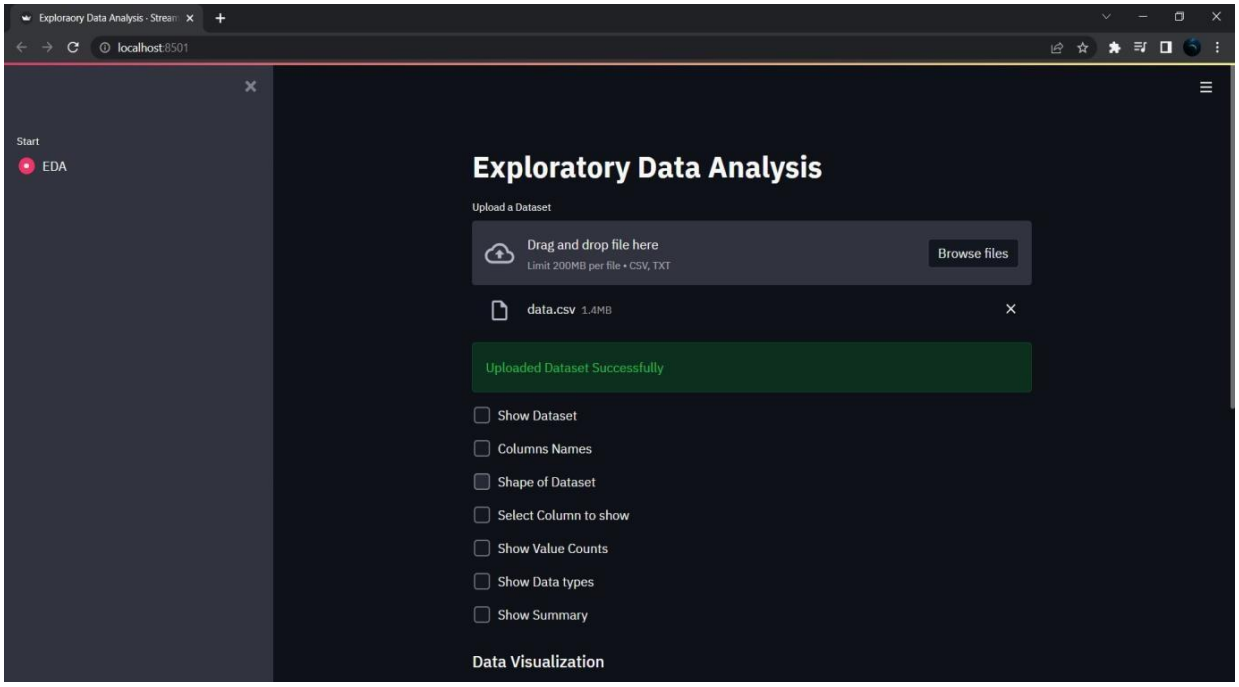

Fig 4 *After uploading the dataset, options for viewing the dataset are shown*

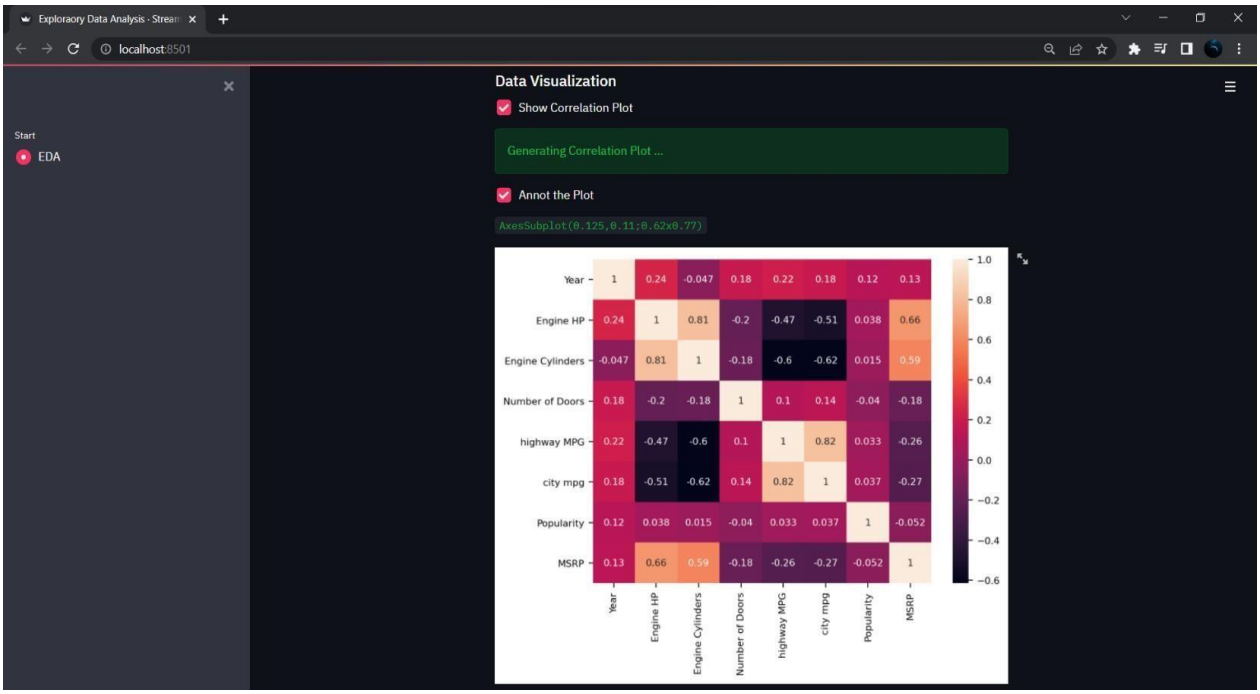

Fig. 5 *When you click on show correlation plot*

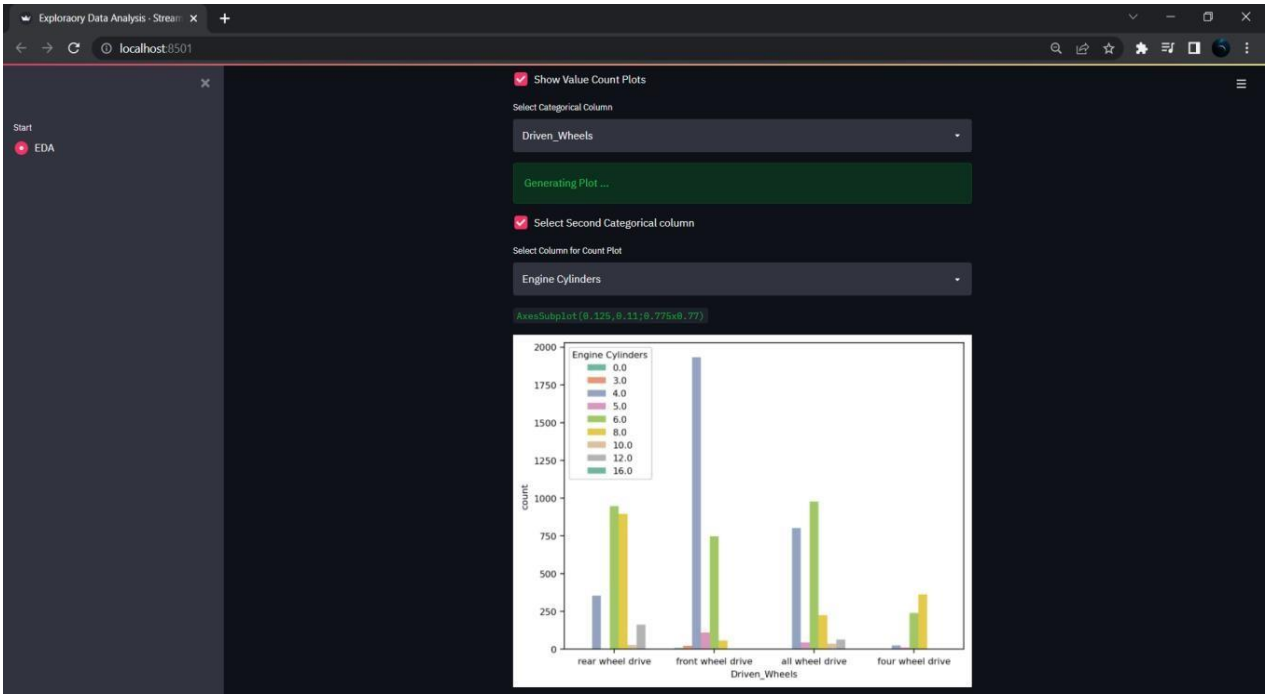

Fig. 6 *Value count plot for column "Driven_Wheels"*

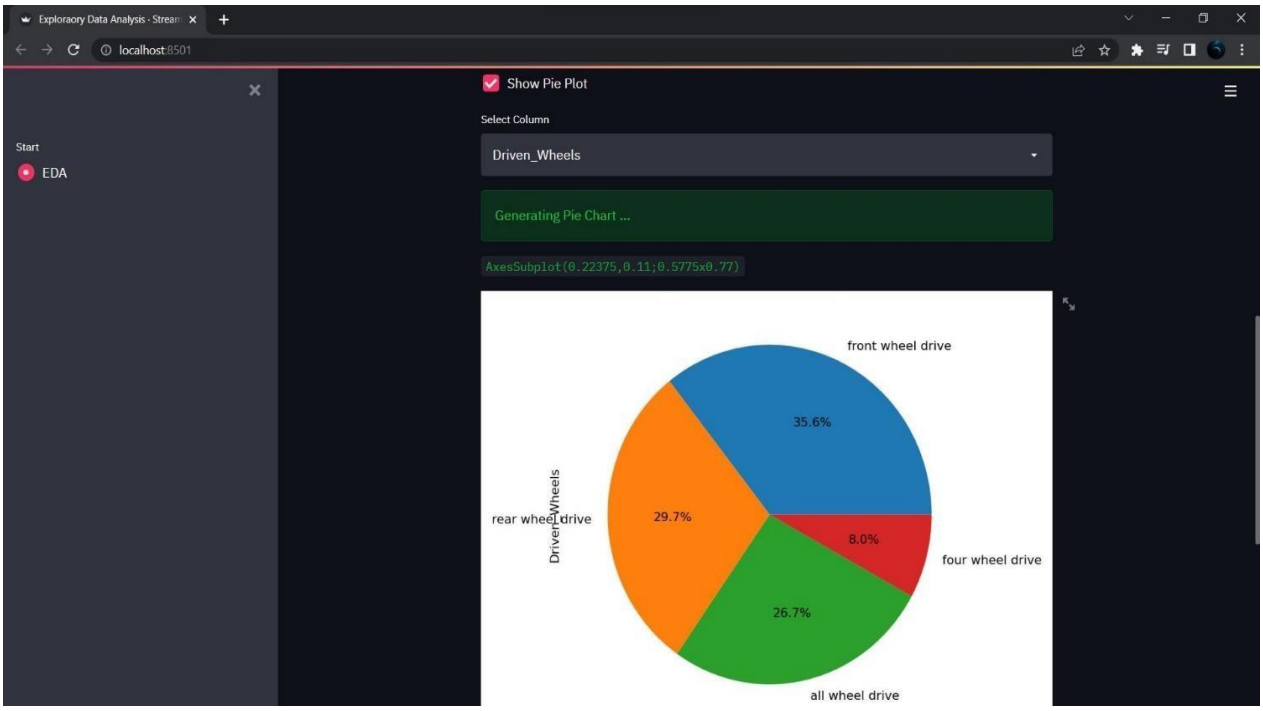

Fig 7 *When you click on pie plot for column "Driven_Wheels"*

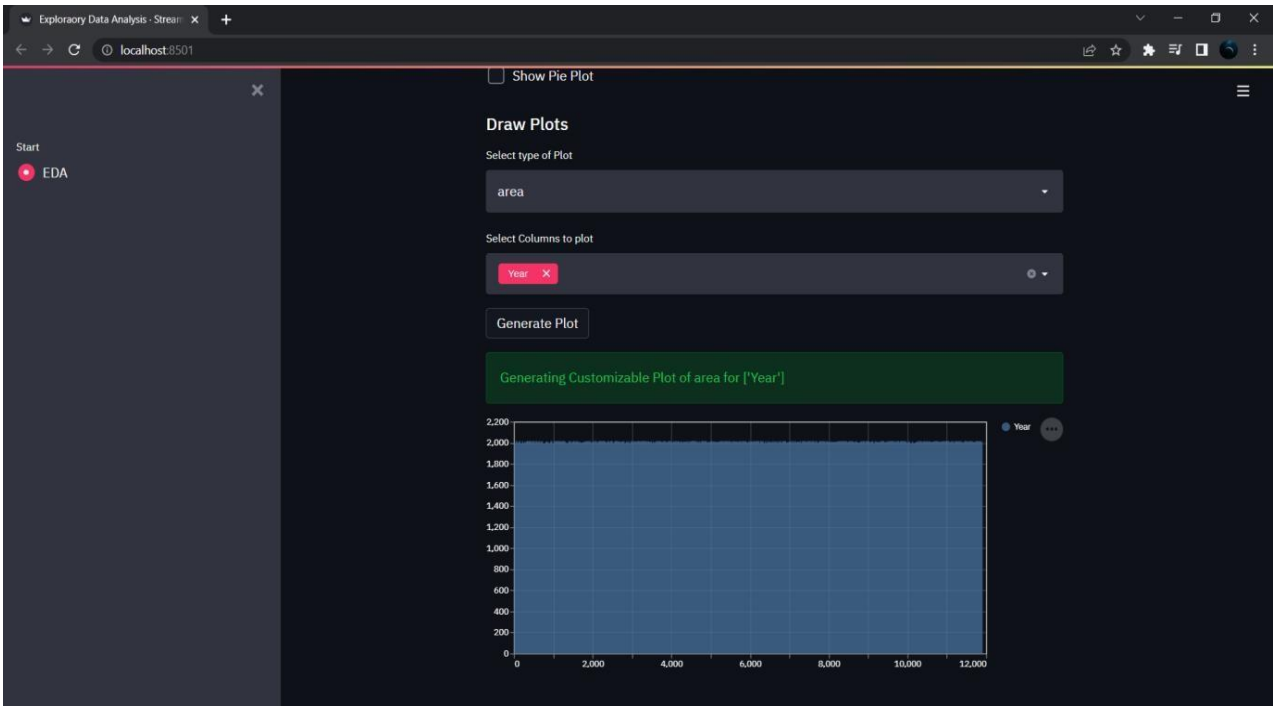

Fig 8 *Area plot for column "Year"*

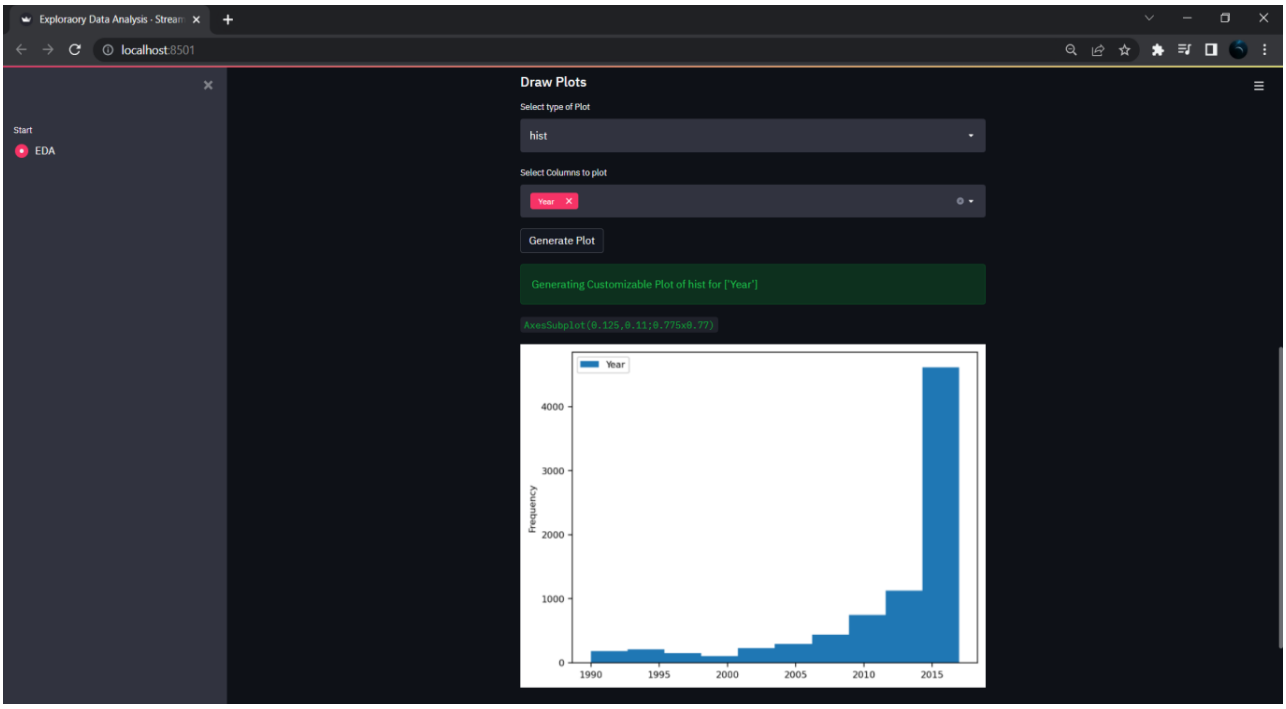

Fig 9 *Histogram for column "Year"*

Chapter – Seven

References

IEEE Publications

[1] Role of Exploratory Data Analysis in Data Science. https://ieeexplore.ieee.org/document/9488986

[2] https://ieeexplore.ieee.org/document/9225621

[3] Exploratory data analysis https://ieeexplore.ieee.org/document/1163294

[4] Wendy L. Martinez, Angel R. Martinez and Jeffrey L. Solka. (2018), "Exploratory Data Analysis with MATLAB®" from "https://www.researchgate.net/publication/254392773_Exploratory_Data_Analysis_with_MATLAB_Second_Edition_by_Wendy_L_Martinez_Angel_R_Martinez_Jeffrey_L_Solka"

[5] Babangida Ibrahim Babura, Mohd Bakri Adam, Muhammad Sani, Usman Waziri and Felix Yakubu Eguda (2020),"Construction and Applications of Stairboxplot for Exploratory Data" from "https://www.researchgate.net/publication/351006278_Construction_and_Applications_of_Stairboxplot_for_Exploratory_Data_Analysis"

[6] Kabita Sahoo, Abhaya Kumar Samal, Jitendra Pramanik, Subhendu Kumar Pani. (2019), "Exploratory Data Analysis using Python" from https://www.researchgate.net/publication/341121348_Exploratory_Data_Analysis_using_Python

[7] Kai Puolamäki, Emilia Oikarinen, Bo Kang, Jefrey Lijffijt, Tijl De Bie. (2017), "Interactive Visual Data Exploration with Subjective Feedback: An Information-Theoretic Approach" from https://arxiv.org/abs/1710.08167

[8] https://www.ijitee.org/wp-content/uploads/papers/v8i12/L35911081219.pdf

[9] Jean-Daniel Fekete, Romain Primet. (2016), "Progressive Analytics: A Computation Paradigm for Exploratory Data Analysis" from https://arxiv.org/abs/1607.05162

Websites:

[1] https://en.wikipedia.org/wiki/Exploratory_data_analysis

[2] https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15

[3] https://www.kaggle.com/CooperUnion/cardataset

[4] https://www.researchgate.net/publication/260146543_Exploratory_data_analysis_with_MATLAB

[5] https://www.ibm.com/in-en/cloud/learn/exploratory-data-analysis

[6] https://www.upgrad.com/blog/exploratory-data-analysis-and-its-importance-to-your-business/

[7] https://www.clicdata.com/blog/future-of-data-analytics