

# **EXPLORATORY DATA ANALYSIS USING WEB APP**

## **A PROJECT REPORT**

*Submitted by*

<b>SHASHWAT JHA</b>	<b>(19MIM10111)</b>
<b>AMAN JAIN</b>	<b>(19MIM10064)</b>
<b>PRIYANSHU SHUKLA</b>	<b>(19MIM10043)</b>
<b>ASHRAF SHAIKH</b>	<b>(19MIM10116)</b>

*in partial fulfillment for the award of the degree  
of*

**INTEGRATED MASTERS OF TECHNOLOGY**

*In*

**COMPUTER SCIENCE AND ENGINEERING**

*Specialization in*

**Artificial Intelligence and Machine Learning**



**SCHOOL OF COMPUTING SCIENCE AND ENGINEERING**

**VIT BHOPAL UNIVERSITY**

**KOTHRI KALAN, SEHORE**

**MADHYA PRADESH - 466114**

**May 2021**

**VIT BHOPAL UNIVERSITY, KOTHRI KALAN, SEHORE  
MADHYA PRADESH – 466114**

**BONAFIDE CERTIFICATE**

Certified that this project report titled “**EXPLORATORY DATA ANALYSIS BY WEB APP**” is the bonafide work of “**SHASHWAT JHA (19MIM10111), AMAN JAIN (19MIM10064), PRIYANSHU SHUKLA (19MIM10043), ASHRAF SHAIKH (19MIM10116)**” who carried out the project work under my supervision. Certified further that to the best of my knowledge the work reported here does not form part of any other project / research work on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

**PROGRAM CHAIR**

Dr Pandimurgan Vellaisamy,  
Assistant Professor  
School of AI & ML division  
VIT BHOPAL UNIVERSITY

**PROJECT GUIDE**

Dr. Suthir. S.  
Senior Assistant Professor  
School of AI & ML division  
VIT BHOPAL UNIVERSITY

The Project Exhibition II Examination is held on \_\_\_\_\_

## **ACKNOWLEDGEMENT**

First and foremost I would like to thank the Lord Almighty for His presence and immense blessings throughout the project work.

I wish to express my heartfelt gratitude to Dr. Nageswara Guptha M, Head of the Department, School of Computer Science for much of his valuable support and encouragement in carrying out this work.

I would like to thank my internal guide Dr. Suthir S. for continually guiding and actively participating in my project, giving valuable suggestions to complete the project work.

I would like to thank all the technical and teaching staff of the School of Computer Science and Engineering, who extended directly or indirectly all support.

Last, but not the least, I am deeply indebted to my parents who have been the greatest support while I worked day and night for the project to make it a success.

## LIST OF ABBREVIATIONS

Sr No.	ABBREVIATED WORD	ABBREVIATION
1	EDA	EXPLORATORY DATA ANALYSIS
2	CSV	COMMA- SEPARATED VALUES

## LIST OF FIGURES

<b>Figure no.</b>	<b>Title</b>	<b>Page Number</b>
<b>1.</b>	<b>Pandas profiling</b>	<b>17</b>
<b>2.</b>	<b>Streamlit</b>	<b>20</b>
<b>3.</b>	<b>Performance Analysis</b>	<b>22-29</b>

**LIST OF TABLES**

<b>TABLE NO.</b>	<b>TITLE</b>	<b>PAGE NO.</b>
<b>1.</b>	<b>DATA SET</b>	<b>22</b>

## ABSTRACT

The main aim of exploratory data analysis by web app is to obtain confidence in your data to an extent where you're ready to engage a machine learning algorithm. Exploratory Data Analysis is a crucial step before we jump to machine learning or modeling our data. By doing this we 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. In every machine learning workflow, the last step is Reporting or Providing the insights to the Stake Holders/ Users and as a Data Scientist we can explain every bit of code but we need to keep in mind the audience. By completing the **EDA** we will have many plots, heat-maps, frequency distribution, graphs, correlation matrix along with the hypothesis by which any individual can understand what our data is all about and what insights we got from exploring our data set. We Are Using Pandas Profiling for executing the code and Streamlit for Making The Web App. Pandas profiling is an open source Python module with which we can quickly do an exploratory data analysis with just a few lines of code. Besides, if this is not enough to convince us to use this tool, it also generates interactive reports in web format that can be presented to any person, even if they don't know programming. Streamlit is an open source app framework specifically designed for ML engineers working with Python. It allows you to create a stunning looking application with only a few lines of code.

## TABLE OF CONTENTS

CHAPT ER NO.	TITLE	PAGE NO.
	List of Abbreviations List of figures List of Tables Abstract	
1.	<b>INTRODUCTION</b>  1.1 INTRODUCTION  1.2 MOTIVATION OF THE WORK  1.3 OUR IDEA FOR PROJECT  1.4 PROBLEM STATEMENT  1.5 OBJECTIVE OF WORK 1.6 ORGANISATION OF THESIS	



<b>2.</b>	<b>2. LITERATURE REVIEW</b>  2.1 INTRODUCTION  2.2 EXISTING ALGORITHMS  2.2.1 ALGORITHM 1  2.2.2 ALGORITHM 2  2.3 RESEARCH ISSUE/OBSERVATIONS FROM LITERATURE SURVEY	
-----------	--	--

<b>3.</b>	<b>SYSTEM ANALYSIS</b>  3.1 INTRODUCTION 3.2 DISADVANTAGES/LIMITATIONS IN THE EXISTING SYSTEM 3.3 PROPOSED WORK	
<b>4.</b>	<b>SYSTEM DESIGN AND IMPLEMENTATION</b>  4.1 MODULE 1 DESIGN AND IMPLEMENTATION 4.2 MODULE 2 DESIGN AND IMPLEMENTATION	
<b>5.</b>	<b>PERFORMANCE ANALYSIS</b>  5.1 PERFORMANCE MEASURES 5.2 PERFORMANCE ANALYSIS CODE	

<b>6.</b>	<b>FUTURE ENHANCEMENT AND CONCLUSION</b>  6.1 LIMITATIONS/CONSTRAINTS OF THE SYSTEM 6.2 FUTURE ENHANCEMENTS 6.3 CONCLUSION  <b>REFERENCES</b>	
-----------	---	--

# **1.INTRODUCTION**

## **1.1 INTRODUCTION**

In machine learning, Exploratory Data Analysis (EDA) is an approach to analyzing datasets to summarize their main characteristics, often with visual methods. EDA is used for seeing what the data can tell us before the modeling task. It is not easy to look at a column of numbers or a whole spreadsheet and determine important characteristics of the data. It may be tedious, boring, and/or overwhelming to derive insights by looking at plain numbers. Exploratory data analysis techniques have been devised as an aid in this situation.

Exploratory data analysis is generally cross-classified in two ways. First, each method is either non-graphical or graphical. And second, each method is either univariate or multivariate (usually just bivariate). Our Web App Includes Pandas Profiling and Streamlit to analyse the data and compress it for a better approach.

## **1.2 MOTIVATION OF THE WORK**

As we see today, there is a lot of data flowing all around. It may be training data or clustered data. After looking upon the current data analytics tool which takes a lot of time and man effort, so we came to conclusion of an interactive, user friendly web app for data analysis. This app uses EDA which provides faster and accurate output. As all programmes we use today for EDA are mostly paid, we strive to provide a free and feasible environment for our users.

## **1.3 OUR IDEA FOR PROJECT**

We idealize the idea of creating an exploratory data analysis based web app, as we thought of creating the data representation in graphical nature with utmost accuracy.

So that the user didn't get any issues regarding insertion of data. The most accurate output is the first priority of our web app. For the ease of visualization of our data this web app would be very useful.

## **1.4 PROBLEM STATEMENT**

Nowadays, the Charges Of Some Application Which Provides Data Analysis is Very High and Not Much Feasible For Many Users. We Wanted Our Platform To Be Free Of Cost So That As Many Users Can Be Benefitted.

## **1.5 OBJECTIVE OF WORK**

The Main Objective Of Doing Exploratory Data Analysis Through Web App Is to Help Look At Data before making any assumptions. It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, find interesting relations among the variables and data. Keeping in Mind the Increased user activity on the internet and the proliferation, we implied sophisticated tools to monitor our data.

## **1.6 ORGANISATION OF THESIS**

Chapter 1: Introduction

Chapter 2: Literature Survey

Chapter 3: Project Procedure

Chapter 4: Work done

Chapter 5: Observation

Chapter 6: Result & Conclusion

Chapter 7: Recommendation for future work

Chapter 8: References

## **2. LITERATURE REVIEW**

### **2.1 INTRODUCTION**

Exploratory data analysis is an approach of analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization 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. Exploratory data analysis was promoted by John Tukey to encourage statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments. EDA is different from initial data analysis (IDA), which focuses more narrowly on checking assumptions required for model fitting and hypothesis testing, and handling missing values and making transformations of variables as needed.

### **2.2 EXISTING ALGORITHMS**

#### **2.2.1 ALGORITHM 1**

Data Analysis with Excel is a comprehensive tutorial that provides a good insight into the latest and advanced features available in Microsoft Excel. It explains in detail how to perform various data analysis functions using the features available in MS-Excel.

#### **2.2.2 ALGORITHM 2**

R analytics (or R programming language) is a free, open-source software used for all kinds of data science, statistics, and visualization projects. R programming language is powerful, versatile, AND able to be integrated into BI platforms like Sisense, to help you get the most out of business-critical data.

## **2.3 RESEARCH ISSUE/OBSERVATIONS FROM LITERATURE SURVEY**

Reference From :- Python for Data Analysis  
Data Wrangling with Pandas, NumPy, and IPython (Second Edition)

By : -Wes McKinney

For data analysis and interactive computing and data visualization, Python will inevitably draw comparisons with other open source and commercial programming languages and tools in wide use, such as R, MATLAB, SAS, Stata, and others. In recent years, Python's improved support for libraries (such as pandas and scikit-learn) has made it a popular choice for data analysis tasks. Combined with Python's overall strength for general-purpose software engineering, it is an excellent option as a primary language for building data applications.

Reference From: - An Exploratory Data Analysis of COVID-19 in India

- By INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT)

The number of COVID-19 cases in India is increasing at a rapid pace. The National and local authorities are having a hard time to create a pattern, analyze and forecast the spread of COVID-19 in India. The main aim of this paper is to draw a statistical model for better understanding of COVID-19 spread in India by thoroughly studying the reported cases in the country till 22 April 2020. An Exploratory Data Analysis (EDA) technique is being implemented to study and analyze the reported COVID-19 cases in India. The result of the analysis divulges the impact of COVID-19 in India on daily and weekly manner, analogize India with abutting countries as well as with the countries who are badly affected and arrangement of India's Healthcare sector for such epidemic.

Keywords COVID-19, exploratory data analysis technique, India's analysis, abutting countries analysis, healthcare sector analysis.

### **3. SYSTEM ANALYSIS**

#### **3.1 INTRODUCTION**

EDA app in python using Pandas Profiling and Streamlit. With this app we will be able to upload CSV data to the app and it will automatically generate an exploratory data analysis(EDA) report with a user - friendly interface for smooth and efficient data analysis.

#### **3.2 DISADVANTAGES/LIMITATIONS IN THE EXISTING SYSTEM**

- The Web App Will Be Having Very Useful Features For The Users, Which Will Be Overcoming the Barriers Of Disadvantages created By Older, Programmes which are in-use.
- Except For R And Python, All Major Apps are Paid. So They are only Feasible apps we can prefer for Initial learning Basis of beginners.
- R Programming And Python Requires the user To Learn about Libraries Before performing any tasks, so they require prior knowledge to these programmes. If the user would not be having any prior knowledge they would face difficulties in performing Data Analyzation.
- Using App Would Make The Data Entry A Easy Job, than using existing program based softwares.
- Our Web App will be providing the easy interface , which will be user friendly, so that any type of user can use the app easily.
- In prior models there are many questions raised about the efficiency of the model using EDA. Our web App will provide as much as efficiency to fully fill the desired output of the user.



### **3.3 PROPOSED WORK**

Proposed solution is That We Are Using EDA web app. An web application in Python using Pandas Profiling and Streamlit. With this app you will be able to upload CSV data to the app and it will automatically generate an exploratory data analysis (EDA) report.

We Would Be Doing The Following Tasks :

1. Overview
  - Data Set Info
  - Variable Types
2. Variables
  - Analysation
3. Correlations
  - Like SeaBorn Heatmap
4. Missing Values
5. Sampling

## 4. SYSTEM DESIGN AND IMPLEMENTATION

### 4.1 MODULE 1 DESIGN AND IMPLEMENTATION

#### Pandas Profiling: -

Pandas profiling is an open source Python module with which we can quickly do an exploratory data analysis with just a few lines of code. Besides, if this is not enough to convince us to use this tool, it also generates interactive reports in web format that can be presented to any person, even if they don't know programming.

In short, what pandas profiling does is save us all the work of visualizing and understanding the distribution of each variable. It generates a report with all the information easily available.

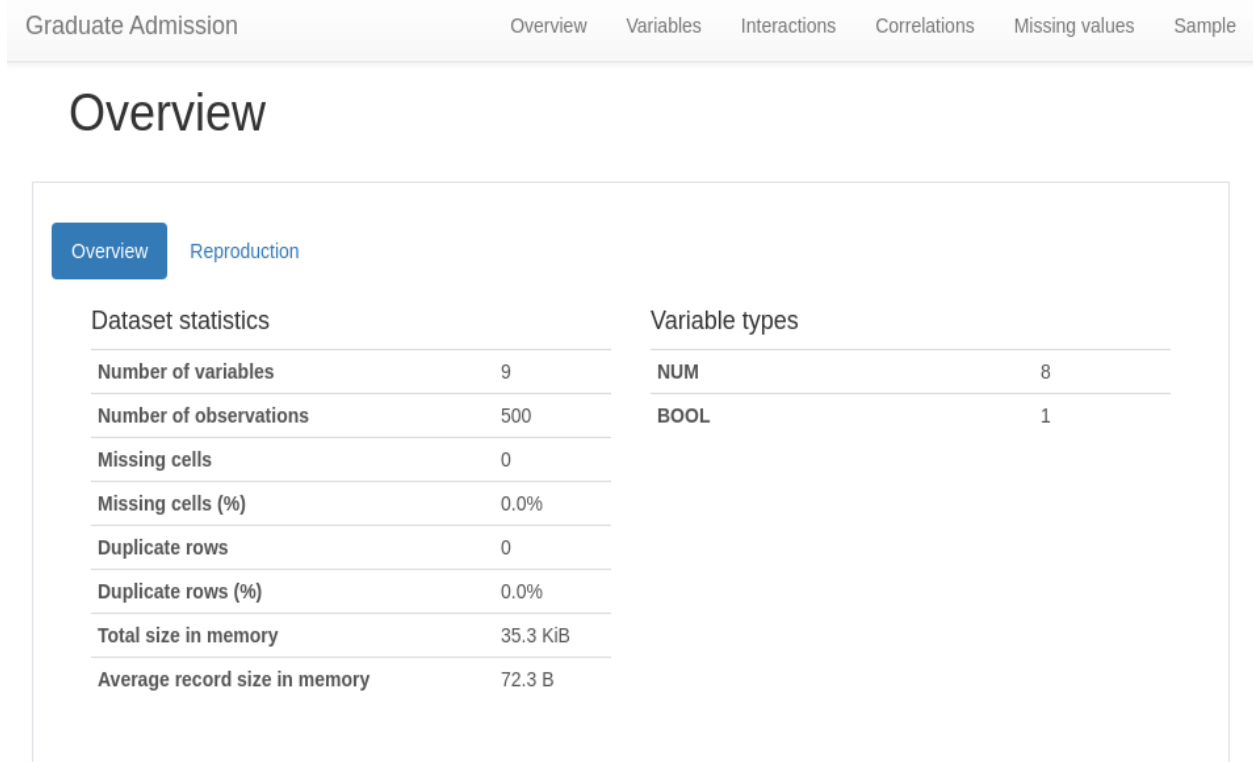


Figure :- 1

Overview	Reproduction
Reproduction	
Analysis started	2020-03-03 16:43:03.374799
Analysis finished	2020-03-03 16:43:10.440730
Version	<a href="#">pandas-profiling v2.5.0</a>
Command line	<code>pandas_profiling --config_file config.yaml [YOUR_FILE.csv]</code>
Download configuration	<a href="#">config.yaml</a>

**Figure :- 2**

### Implementation: -

First step is to install it with this command:

```
pip install pandas-profiling
```

Then we generate the report using these commands:

```
from pandas_profiling import ProfileReport
prof = ProfileReport(df)
prof.to_file(output_file='output.html')
```

## 4.2 MODULE 2 DESIGN AND IMPLEMENTATION

### Streamlit: -

- With the launch of Streamlit, developing a dashboard for your machine learning solution has been made incredibly easy.

- Streamlit is an open source app framework specifically designed for ML engineers working with Python. It allows you to create a stunning looking application with only a few lines of code.
- Streamlit lets you turn data scripts into shareable web apps in minutes, not weeks. It's all Python, open-source, and free! And once you've created an app you can use our free sharing platform to deploy, manage, and share your app with the world.
- Streamlit makes it incredibly easy to build interactive app.

### **Implementation: -**

Streamlit can easily be installed with the following command:

- `pip install streamlit`

Use the following command to see a demonstration of an application with example code:

- `streamlit hello`

Doing this will result in the following page to be opened:

Imports:

- `import pandas as pd`
- `import streamlit as st`
- `import plotly.express as px`

To run your Streamlit app:

- `streamlit run app.py`

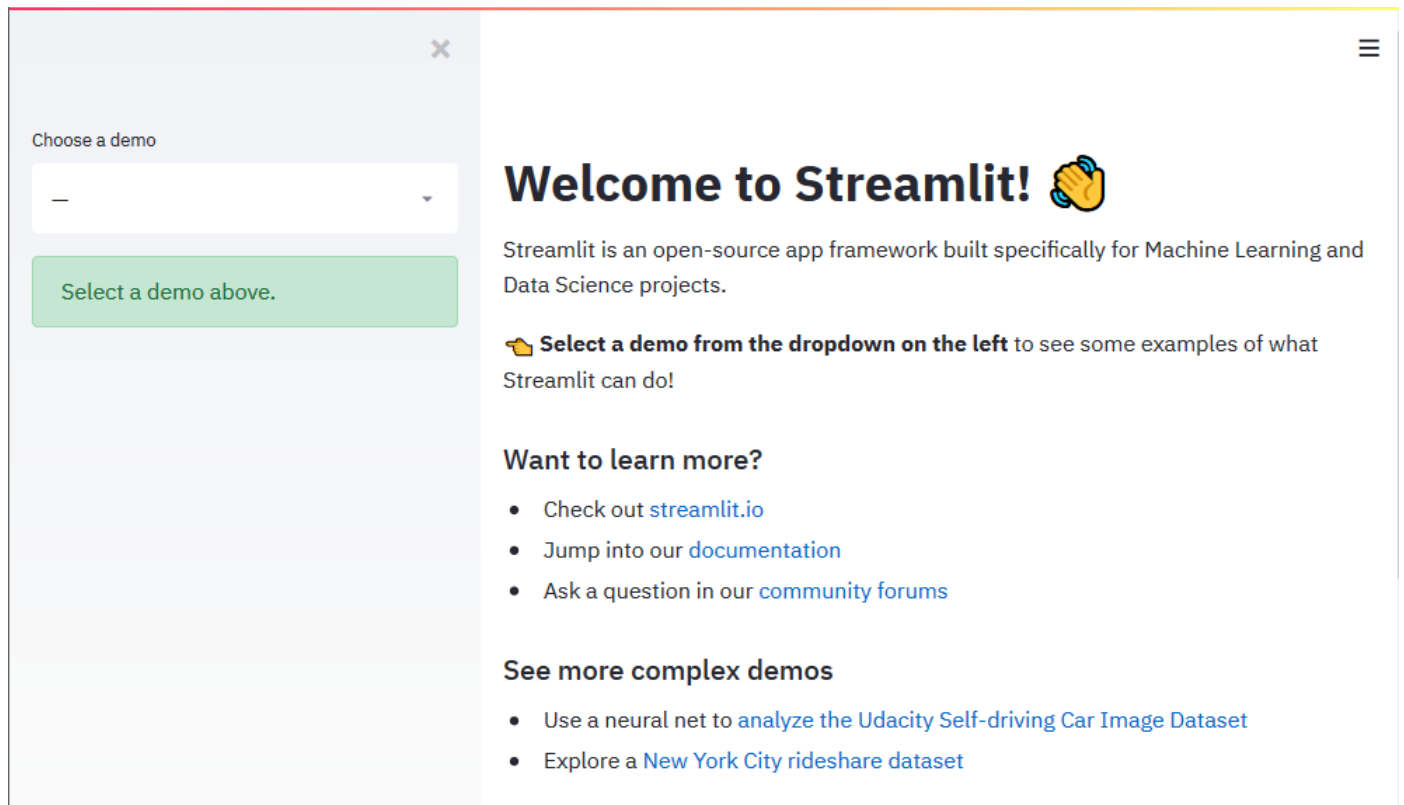


Figure :- 3

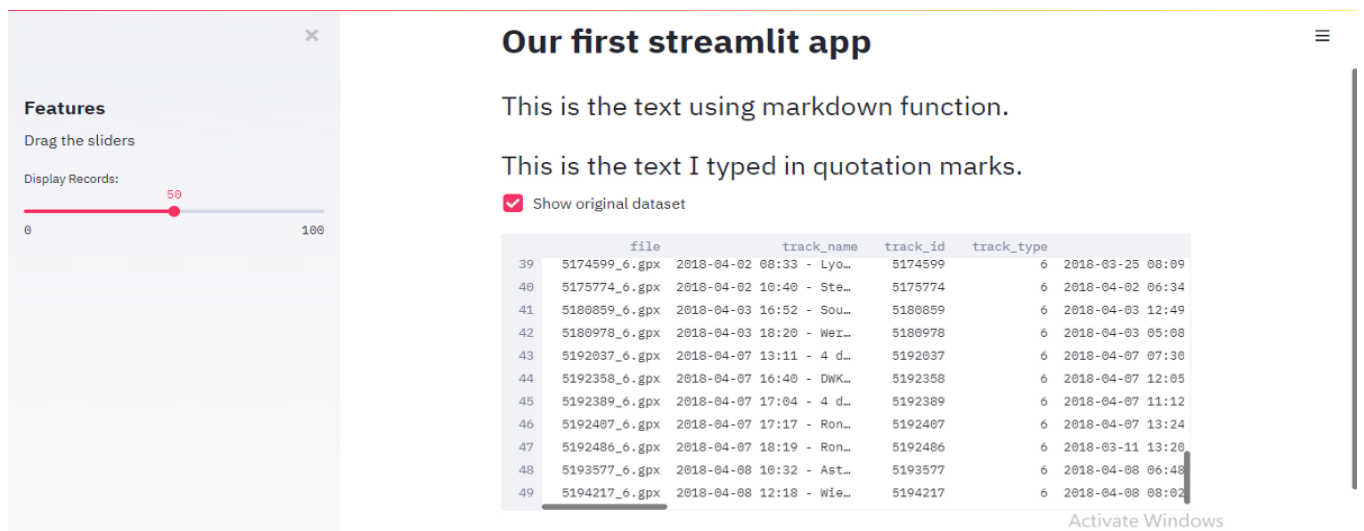
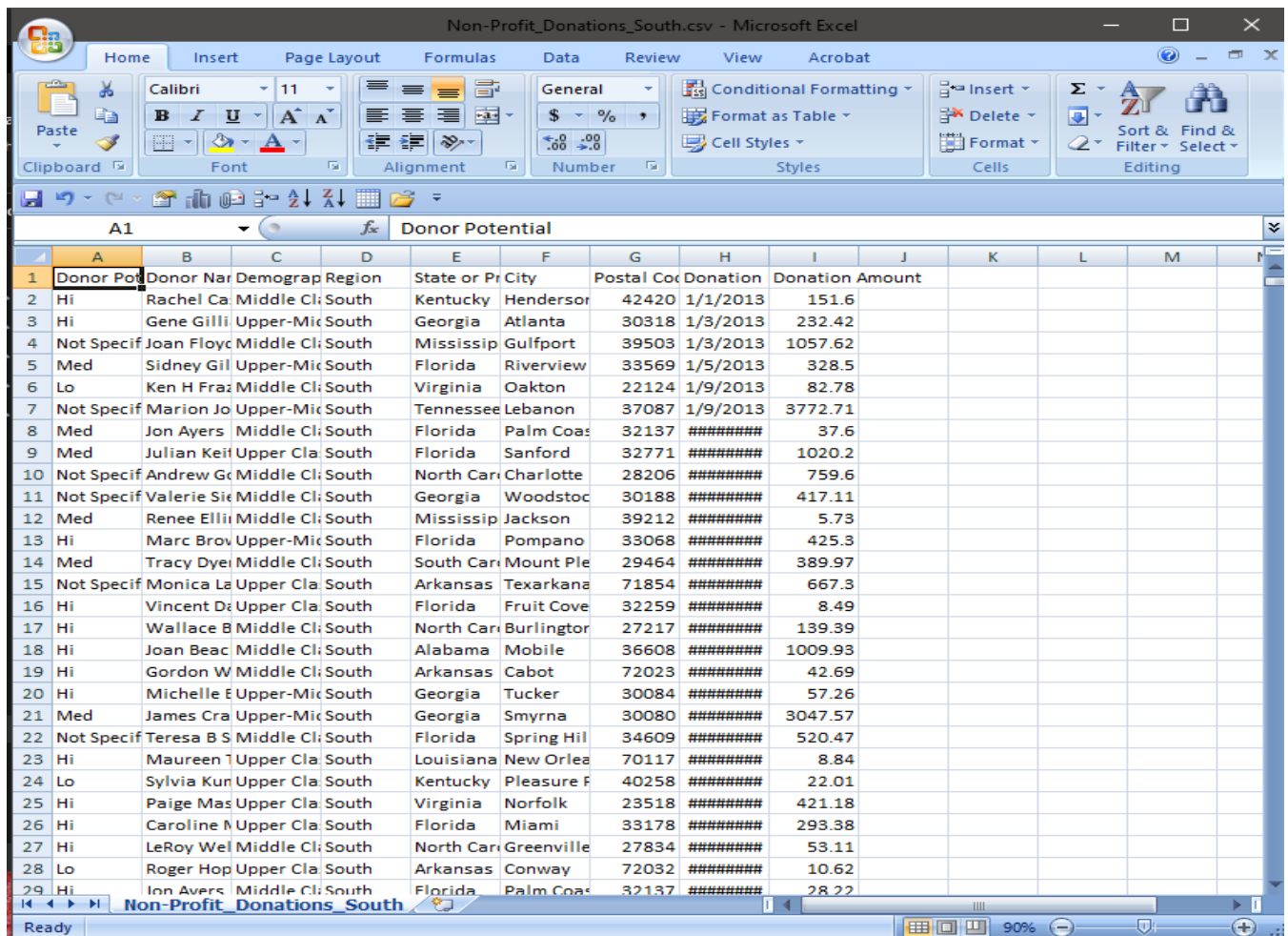


Figure :- 4

## 5. PERFORMANCE ANALYSIS

### 5.1 PERFORMANCE MEASURES



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Donor Pot	Donor Name	Demographic	Region	State or Province	City	Postal Code	Donation Date	Donation Amount				
2	Hi	Rachel Carter	Middle Class	South	Kentucky	Henderson	42420	1/1/2013	151.6				
3	Hi	Gene Gillis	Upper-Middle	South	Georgia	Atlanta	30318	1/3/2013	232.42				
4	Not Specified	Joan Floyd	Middle Class	South	Mississippi	Gulfport	39503	1/3/2013	1057.62				
5	Med	Sidney Gil	Upper-Middle	South	Florida	Riverview	33569	1/5/2013	328.5				
6	Lo	Ken H Frazier	Middle Class	South	Virginia	Oakton	22124	1/9/2013	82.78				
7	Not Specified	Marion Jones	Upper-Middle	South	Tennessee	Lebanon	37087	1/9/2013	3772.71				
8	Med	Jon Ayers	Middle Class	South	Florida	Palm Coast	32137	#####	37.6				
9	Med	Julian Keith	Upper Class	South	Florida	Sanford	32771	#####	1020.2				
10	Not Specified	Andrew G	Middle Class	South	North Carolina	Charlotte	28206	#####	759.6				
11	Not Specified	Valerie Sie	Middle Class	South	Georgia	Woodstock	30188	#####	417.11				
12	Med	Renee Ellis	Middle Class	South	Mississippi	Jackson	39212	#####	5.73				
13	Hi	Marc Brown	Upper-Middle	South	Florida	Pompano	33068	#####	425.3				
14	Med	Tracy Dye	Middle Class	South	South Carolina	Mount Pleasant	29464	#####	389.97				
15	Not Specified	Monica La	Upper Class	South	Arkansas	Texarkana	71854	#####	667.3				
16	Hi	Vincent D	Upper Class	South	Florida	Fruit Cove	32259	#####	8.49				
17	Hi	Wallace B	Middle Class	South	North Carolina	Burlington	27217	#####	139.39				
18	Hi	Joan Beach	Middle Class	South	Alabama	Mobile	36608	#####	1009.93				
19	Hi	Gordon W	Middle Class	South	Arkansas	Cabot	72023	#####	42.69				
20	Hi	Michelle E	Upper-Middle	South	Georgia	Tucker	30084	#####	57.26				
21	Med	James Cra	Upper-Middle	South	Georgia	Smyrna	30080	#####	3047.57				
22	Not Specified	Teresa B S	Middle Class	South	Florida	Spring Hill	34609	#####	520.47				
23	Hi	Maureen T	Upper Class	South	Louisiana	New Orleans	70117	#####	8.84				
24	Lo	Sylvia Kun	Upper Class	South	Kentucky	Pleasure Field	40258	#####	22.01				
25	Hi	Paige Mas	Upper Class	South	Virginia	Norfolk	23518	#####	421.18				
26	Hi	Caroline M	Upper Class	South	Florida	Miami	33178	#####	293.38				
27	Hi	LeRoy Wel	Middle Class	South	North Carolina	Greenville	27834	#####	53.11				
28	Lo	Roger Hop	Upper Class	South	Arkansas	Conway	72032	#####	10.62				
29	Hi	Jon Ayers	Middle Class	South	Florida	Palm Coast	32137	#####	28.22				

Figure :- 5

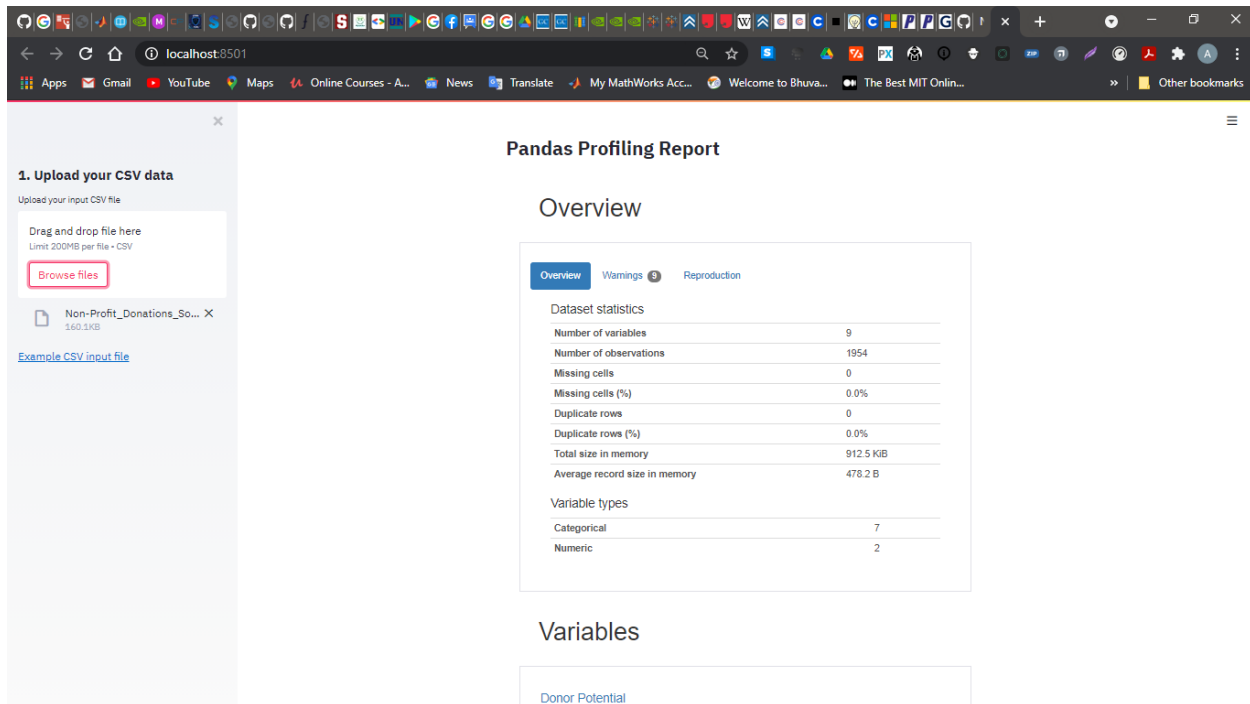


Figure :- 6

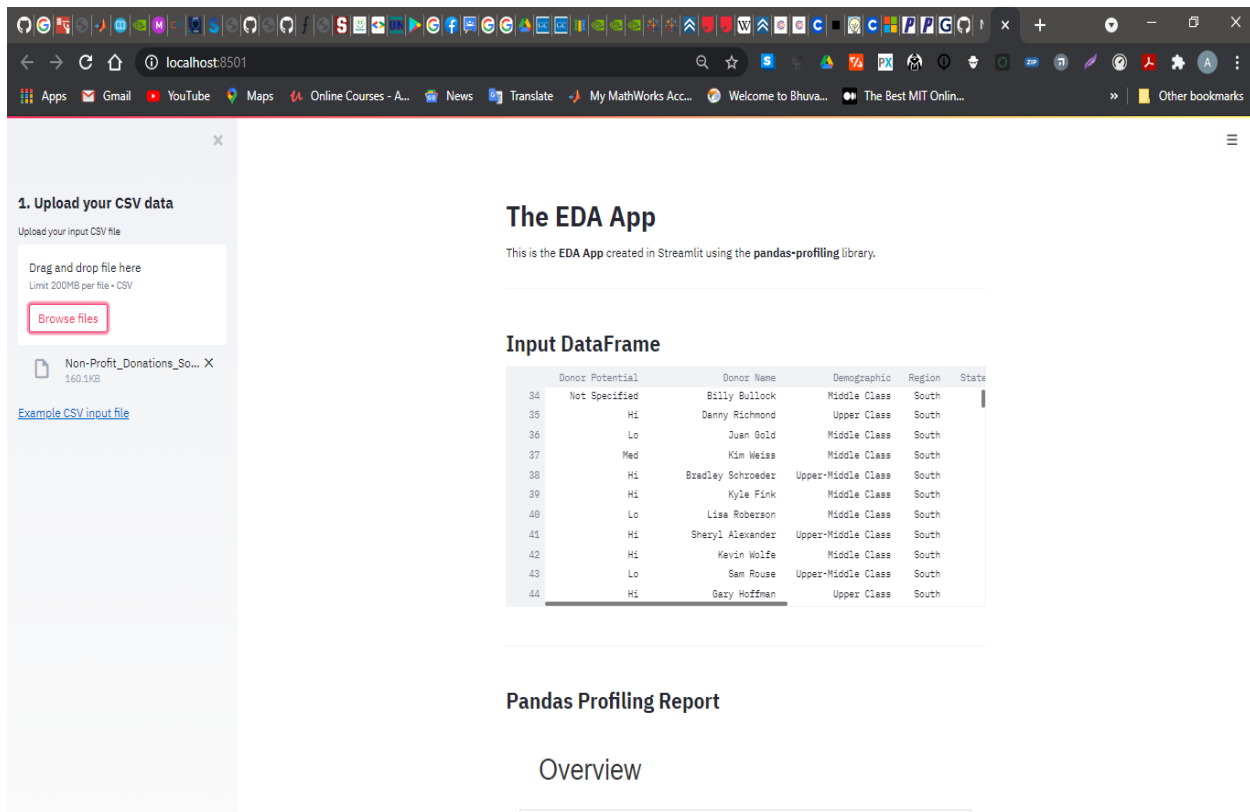


Figure :- 7

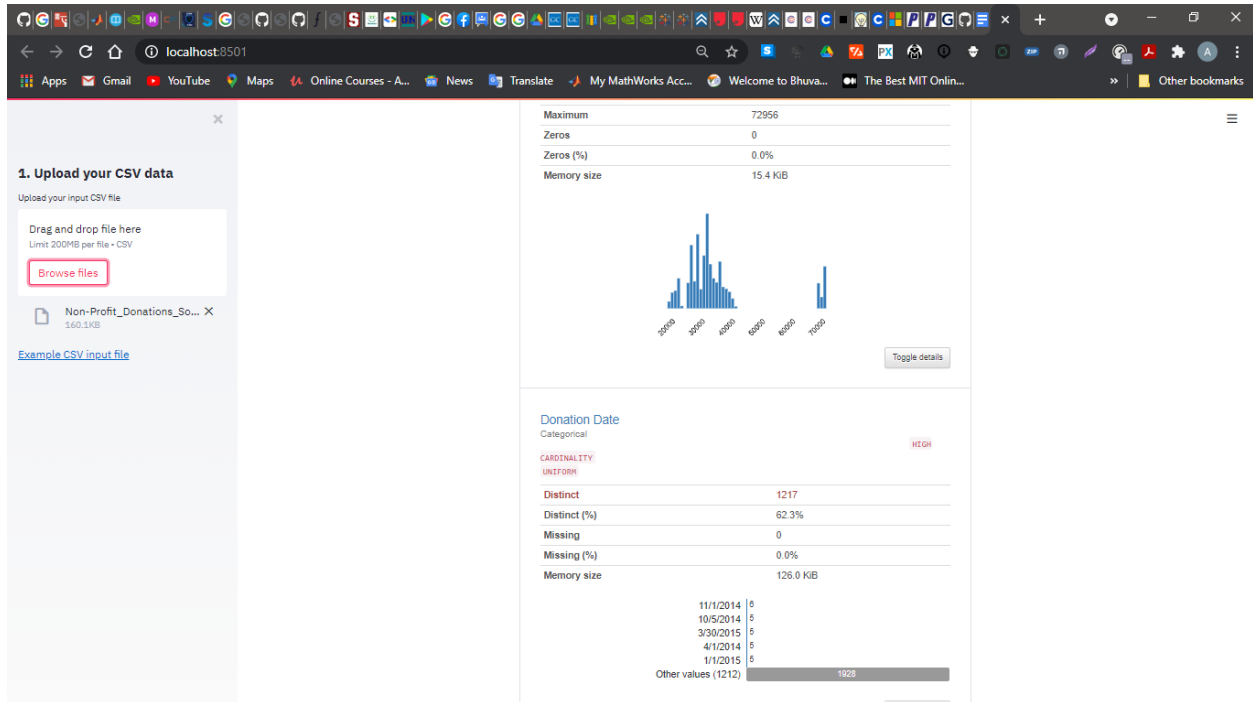


Figure : 8

## 5.2 PERFORMANCE ANALYSIS

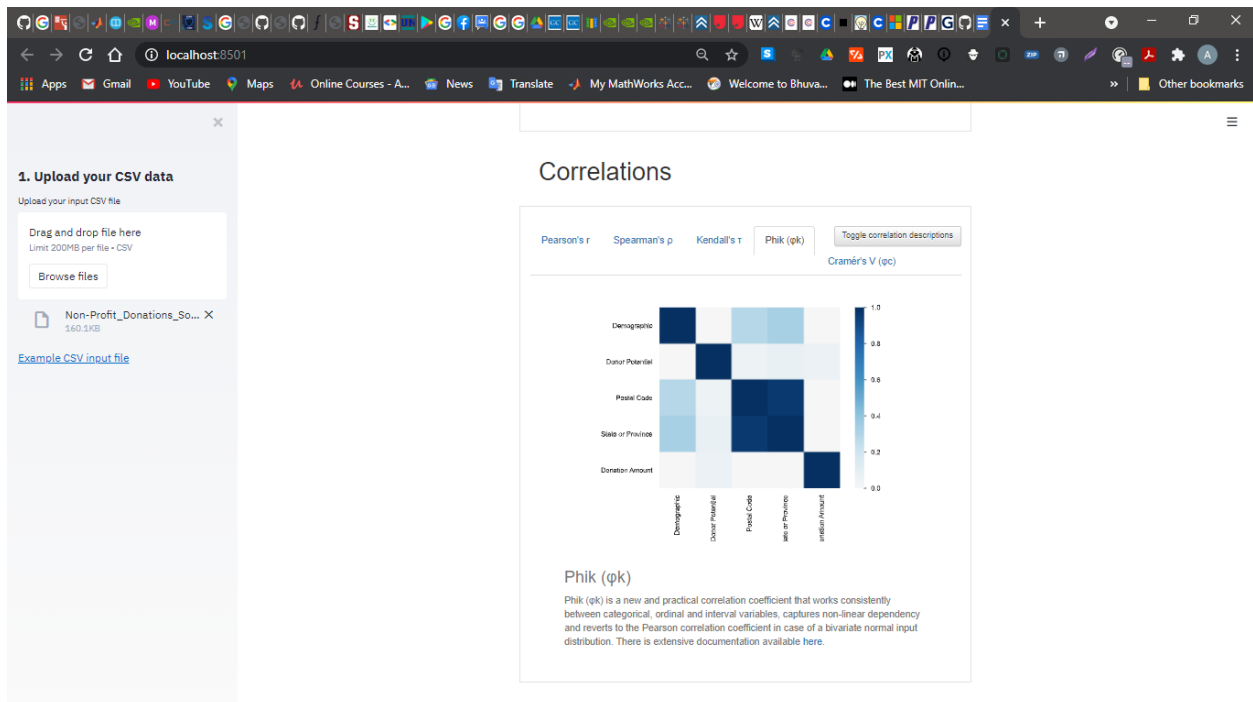
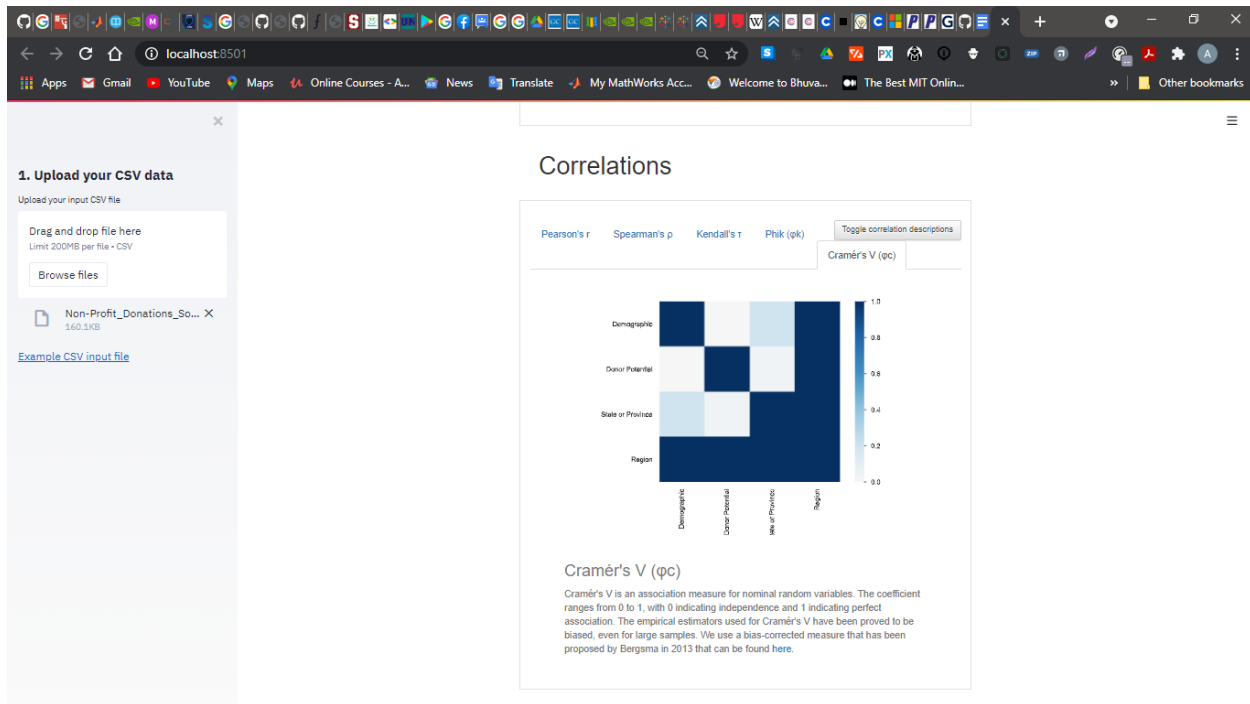
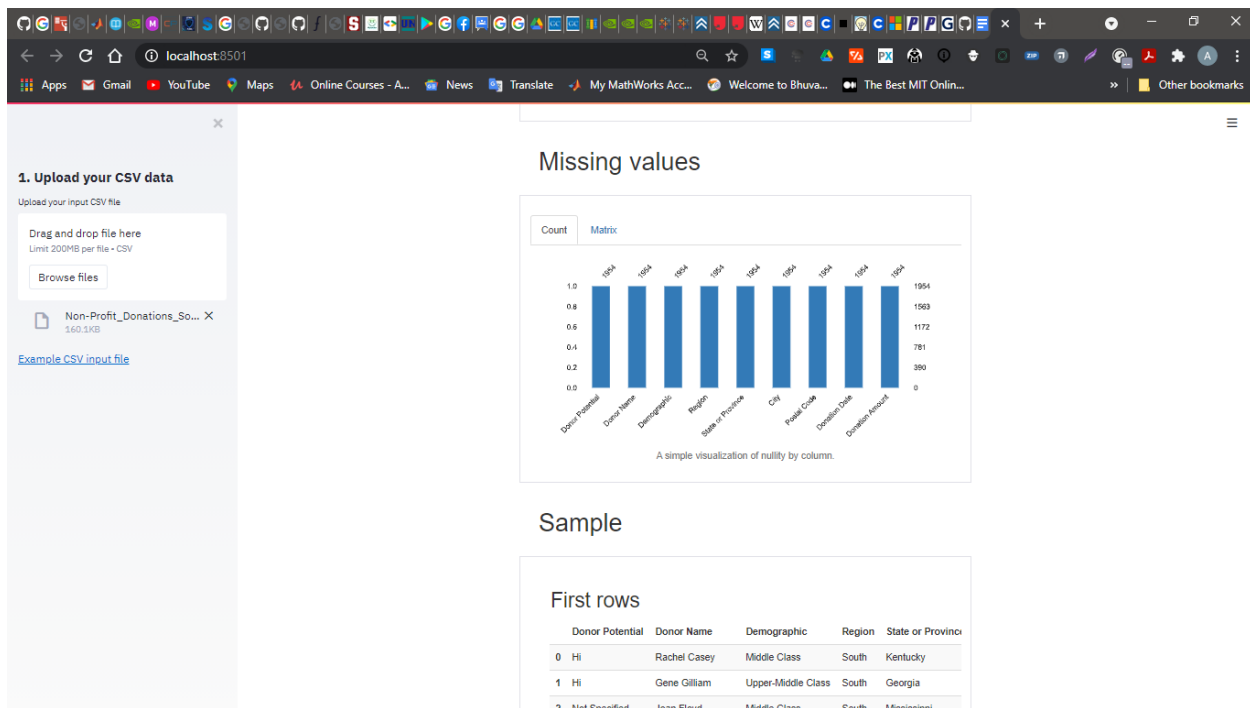


Figure :- 9

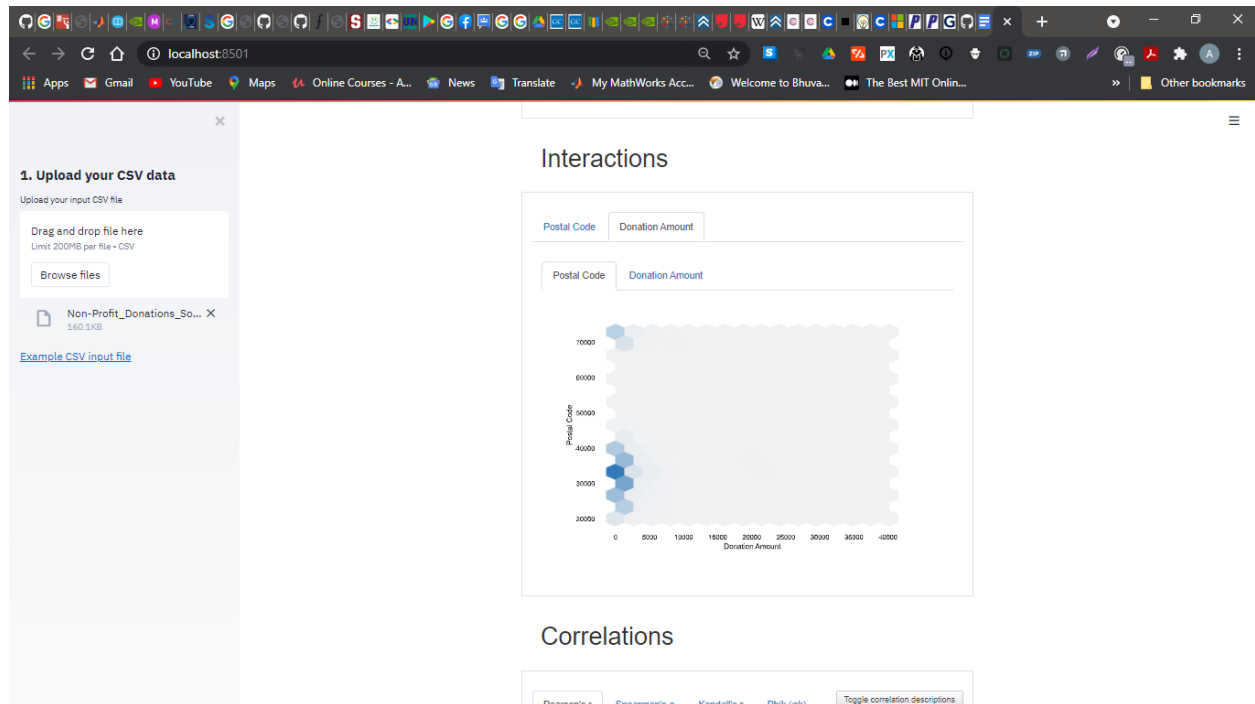




**Figure :- 10**



**Figure :- 11**



**Figure :- 12**

**CODE: -**

```
import numpy as np

import pandas as pd

import streamlit as st

from pandas_profiling import ProfileReport

from streamlit_pandas_profiling import st_profile_report


# Web App Title
```

```
st.markdown("""
```

```
# **The EDA App**
```

This is the **EDA App** created in Streamlit using the **pandas-profiling** library.

- # Upload CSV data

with st.sidebar.header('1. Upload your CSV data'):

```
    uploaded_file = st.sidebar.file_uploader("Upload your input CSV file", type=["csv"])
```

```
    st.sidebar.markdown("""
```

```
[Example                                     CSV                                     input
file](https://raw.githubusercontent.com/dataprofessor/data/master/delaney_solubility_with_descri
ptors.csv)
""")
```

```
# Pandas Profiling Report
```

```
if uploaded_file is not None:
```

```
    @st.cache
```

```
    def load_csv():
```

```
        csv = pd.read_csv(uploaded_file)
```

```
        return csv
```

```
    df = load_csv()
```

```
    pr = ProfileReport(df, explorative=True)
```

```
st.header('**Input DataFrame**')

st.write(df)

st.write('---')

st.header('**Pandas Profiling Report**')

st_profile_report(pr)
```

else:

```
st.info('Awaiting for CSV file to be uploaded.')

if st.button('Press to use Example Dataset'):

    # Example data

    @st.cache

    def load_data():

        a = pd.DataFrame(

            np.random.rand(100, 5),

            columns=['a', 'b', 'c', 'd', 'e']

        )

        return a

    df = load_data()

    pr = ProfileReport(df, explorative=True)

    st.header('**Input DataFrame**')

    st.write(df)

    st.write('---')

    st.header('**Pandas Profiling Report**')

    st_profile_report(pr)
```

## **6.FUTURE ENHANCEMENT AND CONCLUSION**

### **6.1 LIMITATIONS/CONSTRAINTS OF THE SYSTEM**

Though This web application will have a good number of Merits in providing a efficient,user-friendly platform for the users,It Has Some Demerits Which Are Given Below: -

1. Size of dataset, an user can upload.
2. The App Will Only work on the datasets available in csv or excel format.

### **6.2 FUTURE ENHANCEMENTS**

1. Input Data set size could be increased further in future.
2. This web app can be developed as a web progressive app so that it could be used in smartphones by users .

### **6.3 CONCLUSION**

The Project conclusion is Exploratory Data Analysis (EDA) and Data Visualization are powerful tools and that can highlight problems to be addressed, lead to insights, and suggest patterns in the date.

## REFERENCES

1. <https://towardsdatascience.com/exploratory-data-analysis-eda-a-practical-guide-and-template-for-structured-data-abfbf3ee3bd9>
2. Mosteller F, Tukey JW (1977) Data analysis and regression. Addison-Wesley Pub. Co., Boston.
3. Tukey J (1977) Exploratory data analysis. Pearson, London
4. Seltman HJ (2012) Experimental design and analysis. Online
5. [https://www.researchgate.net/publication/329204518\\_Exploratory\\_Daa\\_Analysis\\_EDA](https://www.researchgate.net/publication/329204518_Exploratory_Daa_Analysis_EDA)
6. Python for Data Analysis Data Wrangling with Pandas, NumPy, and IPython (Second Edition) ,Wes McKinney
7. Hands-On Exploratory Data Analysis with Python: Perform EDA techniques to understand, summarize, and investigate your data Paperback – March 27, 2020  
Suresh Kumar Mukhiya (Author), Usman Ahmed (Author)
8. Research on EDA technology and its related issues :-  
<https://ieeexplore.ieee.org/document/5541507>
9. **An Exploratory Data Analysis of COVID-19 in India :-**  
<https://www.ijert.org/an-exploratory-data-analysis-of-covid-19-in-india>
10. International Journal of Scientific & Engineering Research -IJSER

[https://ijser.org/?gclid=CjwKCAjwm7mEBhBsEiwA\\_ofTCKgCpUNXskWOuuHKeeyS5v31WXz6sqy6i4HvoD-hu5qYiPhA23TiRoCI2sOAvD\\_BwF](https://ijser.org/?gclid=CjwKCAjwm7mEBhBsEiwA_ofTCKgCpUNXskWOuuHKeeyS5v31WXz6sqy6i4HvoD-hu5qYiPhA23TiRoCI2sOAvD_BwF)