MINI PROJECT ON DATA VISUALIZATION REPORT

In Partial Fulfillment For The Award Of Degree Of

BACHELOR OF TECHNOLOGY

IN

INFORMATION SCIENCE AND TECHNOLOGY

Under The Guidance Of

Mr. DEEAPK SAKKARI

Professor

Department Of Engineering

PRESIDENCY UNIVERSITY

November 27, 2021

Itgalpura, Rajanukunte, Yelahanka, Bengaluru 560064

Website: www.presidencyuniversity.in

TEAM MEMBERS:

- ➢NARESH RAO SS – IST0052
- ➢NAVYA DS – IST0053
- ➢NOOR AYESHA – IST0055
- ➢VAMSHIKA – IST0086
- ➢KAVYA – IST0091

Table of Contents

Certificate ……………………………………………………...……3

Acknowledgement ……………………………………………………..4

Data Visualization & Application ……………………….………………………...……5

Data-Set Explanation & Requirements……………………………………….……...……6

Libraries Used ……...……………………………………….…..….7

Technique Used……………………………………….……….....10

Reference …………………………………………………….……….11

CERTIFICATE

Certified that the Initiation Report on Data Visualization Course entitled "MARATHON DATASET" carried out by VII Semester student of Department of Computer Engineering, Presidency University, Bangalore in partial fulfilment of the requirements for the award of degree of Bachelor of Technology in Computer Engineering. The report has been approved as it satisfies the academic requirements in respect of University Learning Course prescribed for the said Degree.

Mr. DEEPAK SAKKARI (SOE)

Presidency University

ACKNOWLEDGE

For completing this Data Visualization dataset on Marathon Report, I have received support and guidance from many people to whom I would like to place on record my deep gratitude.

Firstly, I cordially thank Presidency University in platforming such intellectual works and supporting us in all means for our successful completion of the Data Visualization.

I would like to express our sincere gratitude and indebtedness to our coordinators Mr. Deepak Sakkari, Department of Management, Presidency University for their valuable guidance.

I am thankful to Dr: K G Mohan, Head of the Department, Department of Computer Engineering, Presidency University for his moral support, advice and encouragement provided to us.

Mr. DEEPAK SAKKARI (SOE)

Presidency University

DATA VISUALIZATION

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

In the world of Big Data, data visualization tools and technologies are essential to analyze massive amounts of information and make data-driven decisions

APPLICATIONS IN MAJOR SECTORS

- **E- Commerce**

  Data visualization is a brilliant method of data representation. It helps you interpret customers, so it provides valuable insights to improve your sales strategy and your customer relationships, enrich shoppers' buying experience, and feed their needs

- **Education**

  Users may visually engage with data, answer questions quickly, make more accurate, data-informed decisions, and share their results with others using intuitive, interactive dashboards. The ability to monitor students' progress throughout the semester, allowing advisers to act quickly with outreach to failing students. When they provide end users access to interactive, self-service analytic visualizations as well as ad hoc visual data discovery and exploration, they make quick insights accessible to everyone – even those with little prior experience with analytics.

- **Finance**

  For exploring/explaining data of linked customers, understanding consumer behavior, having a clear flow of information, the efficiency of decision making, and so on, data visualization tools are becoming a requirement for financial sectors. For associated organizations and businesses, data visualization aids in the creation of patterns, which aids in better investment strategy. For improved business prospects, data visualization emphasizes the most recent trends

- **Health Care**

A dashboard that visualises a patient's history might aid a current or new doctor in comprehending a patient's health. It might give faster care facilities based on illness in the event of an emergency. Instead, then shifting through hundreds of pages of information, data visualisation may assist in finding trends. Health care is a time-consuming procedure, and the majority of it is spent evaluating prior reports. By boosting response time, data visualisation provides a superior selling point. It gives matrices that make analysis easier, resulting in a faster reaction time.

- **Sales**

Data scientists generally create visualisations for their personal use or to communicate information to a small group of people. Visualization libraries for the specified programming languages and tools are used to create the visual representations. Open-source programming languages, such as Python, and proprietary tools built for complicate data analysis are commonly used by data scientists and academics. These data scientists and researchers use data visualisation to better comprehend data sets and spot patterns and trends that might otherwise go undiscovered.

DATASETS EXPLANATION

Here we have used **Marathon** dataset for following project. As we know that the Marathon consists of 42km of continuous running.

And it has many sub-data in it such as *No of kilometer, pace per minute per kilometer and per hour, gender, age, name of participants, location, country, finish timing of race*.

By using these details, we are going to visualize the data through various libraires and techniques.

This is going to be much simpler to understand the data rather than completing looking into the csv precisely know as excel sheet.

We will be making use of different plotting technique.

Hence by this we can come to a conclusion the positions of each participant in the marathon race and time taken by them and to whom 3 prizes has been awarded.

REQURIMENTS

- Intel 5 processor PC / Laptop.
- Internet connectivity.
- Google Colab.
- Kaggle dataset.
- Good Understanding Knowledge on Data Visualization Concept.

LIBRARIES

➢ import numpy as np

NumPy is an open-source numerical Python library.

NumPy contains a multi-dimensional array and matrix data structures.

It can be utilised to perform a number of mathematical operations on arrays such as trigonometric, statistical, and algebraic routines. Therefore, the library contains a large number of mathematical, algebraic, and transformation functions.

NumPy is an extension of Numeric and Numarray.

Numpy also contains random number generators.

NumPy is a wrapper around a library implemented in C.

Pandas objects rely heavily on NumPy objects. Essentially, Pandas extends Numpy.

➢ import matplotlib.pyplot as plt

matplotlib.pyplot is a collection of command style functions that make matplotlib work like MATLAB.

Each pyplot function makes some change to a figure: e.g., creates a figure, creates a plotting area in a figure, plots some lines in a plotting area, decorates the plot with labels, etc.

In matplotlib.pyplot various states are preserved across function calls, so that it keeps track of things like the current figure and plotting area, and the plotting functions are directed to the current axes.

- import pandas as pd

Pandas (all lowercase) is a popular Python-based data analysis toolkit which can be imported using import pandas as pd.

It presents a diverse range of utilities, ranging from parsing multiple file formats to converting an entire data table into a NumPy matrix array.

This makes pandas a trusted ally in data science and machine learning. Similar to NumPy, pandas deal primarily with data in 1-D and 2-D arrays; however, pandas handle the two differently.

- import seaborn as sns

Seaborn is a library for making statistical graphics in Python. It builds on top of matplotlib and integrates closely with pandas data structures.

Seaborn helps you explore and understand your data. Its plotting functions operate on data frames and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plots.

Its dataset-oriented, declarative API lets you focus on what the different elements of your plots mean, rather than on the details of how to draw them.

- import plotly.express as px

Plotly express is the easy-to-use, high-level interface to Plotly, which operates on a variety of types of data and produces easy-to-figure styles.

Plotly Express provides functions to visualize a variety of types of data. Most functions such as px.bar or px.scatter expect to operate on column-oriented data of the type you might store in a Pandas DataFrame (in either "long" or "wide" format, see below).

px.imshow operates on matrix-like data you might store in
a numpy or xarray array and functions like px.choropleth and

px.choropleth_mapbox can operate on geographic data of the kind you might store in a GeoPandas GeoDataFrame.

This page details how to provide column-oriented data to most Plotly Express functions.

- import plotly.graph_objects as go

The figures created, manipulated and rendered by the plotly Python library are represented by tree-like data structures which are automatically serialized to JSON for rendering by the Plotly.js JavaScript library.

These trees are composed of named nodes called "attributes", with their structure defined by the Plotly.js figure schema, which is available
in machine readable form.

The plotly.graph_objects module (typically imported as go) contains
an automatically generated hierarchy of Python Classes which represent non-leaf nodes in this figure schema. The term "graph objects" refers to instances of these classes.

The primary classes defined in the plotly.graph_objects module are Figure and an ipywidgets-compatible variant called FigureWidget, which both represent entire figures.

Instances of these classes have many convenience methods for Pythonically manipulating their attributes
(e.g. .update_layout() or .add_trace(), which all accept “magic underscore” nation) as well as renderingthem (e.g. .show()) and exporting them to various formats
(e.g. .to_json() or .write_image() or .write_html()).

- %matplotlib inline

%matplotlib inline sets the backend of matplotlib to the 'inline' backend:

With this backend, the output of plotting commands is displayed inline within frontends like the Jupyter notebook, directly below the code cell that produced it.

The resulting plots will then also be stored in the notebook document.

TECHNIQUES USED

- **Histograms**

A histogram is basically used to represent data provided in a form of some groups. It is accurate method for the graphical representation of numerical data distribution. It is a type of bar plot where X-axis represents the bin ranges while Y-axis gives information about frequency.

- **Scatter Plot**

Scatter plots are used to observe relationship between variables and uses dots to represent the relationship between them. The scatter() method in the matplotlib library is used to draw a scatter plot. Scatter plots are widely used to represent relationship among variables and how change In one affects the other.

- **Time series:**

Time series show a particular behavior over time with high probability that it will follow the same in the future. It is important to remove seasonality. If we do not remove it, it will create problems in future prediction. There are many ways to detect seasonality.

- **Plotly:**

Plotly allows users to import, copy and paste, or stream data to be analyzed and visualized.
For analysis and styling graphs, Plotly offers a Python sandbox (NumPy supported), DataGrid, and GUI. Python scripts can be saved, shared, and collaboratively edited in Plotly.

- **Seaborn:**

  Seaborn is an open-source Python library built on top of matplotlib. It is used for data visualization and exploratory data analysis. Seaborn works easily with data frames and the Pandas library. The graphs created can also be customized easily.

- **Regression:**

  The term regression is used when you try to find the relationship between variables. In Machine Learning, and in statistical modeling, that relationship is used to predict the outcome of future events.

- **Word Cloud:**

  Word Cloud is a data visualization technique used for representing text data in which the size of each word indicates its frequency or importance. Significant textual data points can be highlighted using a word cloud.

REFERENCE

Google Colab Link:

https://colab.research.google.com/drive/1HPOvVwkNeP-ZEd8zW3UwTlZJU1b1f85s#scrollTo=gJfDxxCA8Jkm