**Data Visualization with Python**

This is the Capstone Project for Course 9, *IBM Data Analyst Capstone Project*. Part of IBM's Data Analyst Professional Certificate from Coursera.

Here, we will take on the role of a Data Analyst with a global IT and Business services firm. In this role, we will be analyzing several datasets to help identify trends for emerging technologies. We have recently been hired as a Data Analyst by a global IT and business consulting services firm that is known for its expertise in IT solutions and its team of highly experienced IT consultants. To keep pace with changing technologies and remain competitive, our organization regularly analyzes data to help identify future skill requirements.

As a Data Analyst, we will be assisting with this initiative and have been tasked with collecting data from various sources and identifying trends for this year's report on emerging skills.

**Task 1**

Our first task is to collect data for the technology skills that are most in demand from various sources including job postings, blog posts, and surveys. We will begin by scraping internet websites and accessing APIs to collect data in various formats like .csv files, excel sheets, and databases.

**Task 2**

Once we've collected enough data we will take the collected data and prepare it for analysis by using data wrangling techniques like finding duplicates, removing duplicates, finding missing values, and inputting missing values.

**Task 3**

Now that the data is ready we will apply statistical techniques to analyze the data and identify insights and trends like: What are the top programming languages that are in demand? What are the top database skills that are in demand? What are the most popular IDEs? And Demographic data like gender and age distribution of developers.

**Task 4**

In the fourth task, we'll focus on choosing appropriate visualizations based on the data we want to present using charts, plots, and histograms to help reveal our findings and trends. We are going to access the Data from an SQL database and pull only the data we need into DataFrames.

**Task 5**

For task 5, we will employ Cognos/Google Looker Studio to create interactive dashboards to help analyze and present the data dynamically.

**Task 6**

For the final task, we will use our storytelling skills to provide a narrative and present the findings of our analysis.

**Table of Contents**

* [Data Description](https://github.com/rheera/IBM-data-analyst-capstone/blob/main/README.md#data-description)
* [Tools](https://github.com/rheera/IBM-data-analyst-capstone/blob/main/README.md#tools)
* [Deliverables](https://github.com/rheera/IBM-data-analyst-capstone/blob/main/README.md#deliverables)
  + [Task 1: Data Collection](https://github.com/rheera/IBM-data-analyst-capstone/blob/main/README.md#task-1-data-collection)
  + [Task 2: Data Wrangling](https://github.com/rheera/IBM-data-analyst-capstone/blob/main/README.md#task-2-data-wrangling)
  + [Task 3: Exploratory Data Analysis](https://github.com/rheera/IBM-data-analyst-capstone/blob/main/README.md#task-3-exploratory-data-analysis)
  + [Task 4: Data Visualization](https://github.com/rheera/IBM-data-analyst-capstone/blob/main/README.md#task-4-data-visualization)
  + [Task 5: Dashboard Creation](https://github.com/rheera/IBM-data-analyst-capstone/blob/main/README.md#task-5-dashboard-creation)
  + [Task 6: Presentation of Findings](https://github.com/rheera/IBM-data-analyst-capstone/blob/main/README.md#task-6-presentation-of-findings)
* [Stretch Goals](https://github.com/rheera/IBM-data-analyst-capstone/blob/main/README.md#stretch-goals)

**Data Description**

Stack Overflow, a popular website for developers, conducted an online survey of software professionals across the world. The survey data was later open sourced by Stack Overflow. The actual data set has around 90,000 responses.

The dataset we are going to use comes from the following source: <https://stackoverflow.blog/2019/04/09/the-2019-stack-overflow-developer-survey-results-are-in/> under a ODbL: Open Database License.

We will be given a subset of the original data set in this capstone project. We will explore, analyze, and visualize this dataset and present our analysis.

Note: This randomised subset contains around 1/10th of the original data set. Any conclusions we draw after analyzing this subset may not reflect the real world scenario.

|  |  |
| --- | --- |
|  |  |

**Tools**

* [python](https://www.python.org/downloads/) v3.12.2
* [pandas](https://pandas.pydata.org/?utm_medium=Exinfluencer&utm_source=Exinfluencer&utm_content=000026UJ&utm_term=10006555&utm_id=NA-SkillsNetwork-Channel-SkillsNetworkCoursesIBMML0187ENSkillsNetwork31430127-2021-01-01) for managing the data.
* [numpy](https://numpy.org/?utm_medium=Exinfluencer&utm_source=Exinfluencer&utm_content=000026UJ&utm_term=10006555&utm_id=NA-SkillsNetwork-Channel-SkillsNetworkCoursesIBMML0187ENSkillsNetwork31430127-2021-01-01) for mathematical operations.
* [seaborn](https://seaborn.pydata.org/?utm_medium=Exinfluencer&utm_source=Exinfluencer&utm_content=000026UJ&utm_term=10006555&utm_id=NA-SkillsNetwork-Channel-SkillsNetworkCoursesIBMML0187ENSkillsNetwork31430127-2021-01-01) for visualizing the data.
* [matplotlib](https://matplotlib.org/?utm_medium=Exinfluencer&utm_source=Exinfluencer&utm_content=000026UJ&utm_term=10006555&utm_id=NA-SkillsNetwork-Channel-SkillsNetworkCoursesIBMML0187ENSkillsNetwork31430127-2021-01-01) for additional plotting tools.
* [folium](https://python-visualization.github.io/folium/latest/) for geospatial data visualization such as choropleth maps.
* [plotly](https://plotly.com/python/) for interactive plotting tools.
* [Google Looker Studio](https://lookerstudio.google.com/overview) for dashboards.
* [IBM Cognos Analytics](https://www.ibm.com/products/cognos-analytics) for dashboards.

**Deliverables**

**Task 1: Data Collection**

* Collecting Data Using APIs
* Collecting Data Using Web Scraping
* Exploring Data

**Task 2: Data Wrangling**

* Finding Missing Values
* Determine Missing Values
* Finding Duplicates
* Removing Duplicates
* Normalizing Data

**Task 3: Exploratory Data Analysis**

* Distribution
* Outliers
* Correlation

**Task 4: Data Visualization**

* Visualizing Distribution of Data
* Relationship
* Composition
* Comparison

**Task 5: Dashboard Creation**

* Dashboards

**Task 6: Presentation of Findings**

* Final Presentation

**Stretch Goals**

* Create Dashboard in Google Looker or Cognos Analytics