**THE WORLDWIDE LANDSCAPE OF DATA SCIENCE: DEMAND & DEMOGRAPHICS**

**Team Members:**

KEERTHI REDDY PAPAIAHGARI – 11772193

LAKSHMI DEEPIKA PAGADALA – 11696381

THARUN KURAVADI SATHISH BABU – 11659825

THARUN RAMULA – 11706360

**Introduction:**

In our project, we're exploring how population demographics influence the demand for data scientists and their salaries on a global scale. As data science becomes increasingly important across industries, understanding the relationship between population size and data science trends can provide valuable insights. With our team's expertise in data analysis and visualization, we aim to uncover patterns that can inform hiring practices, policy decisions, and career choices in the field of data science.

**Goals and Objectives:**

Our project is motivated by the need to understand the impact of population demographics on the demand for data scientists and their salary structures. By examining this relationship, we aim to provide actionable insights for various stakeholders. Our study holds significance for companies seeking to optimize their data science workforce, policymakers shaping workforce policies, and individuals considering careers in data science. Through our objectives, we aim to explore correlations, analyse variations, and identify local trends to deepen our understanding of this dynamic landscape.

**Features:**

We leverage data from Kaggle, including 'World Population by Country' and 'Data Science Salaries,' to drive our analysis. Using Python for data cleaning and D3.js for visualization, we create visually appealing and interactive representations of our findings. Additionally, tools like Tableau complement our analysis by uncovering correlations between datasets. Through interactive features such as scatter plots, users can engage with the data and gain insights into the complex relationship between population demographics and data science salaries.

**Related Work (Background):**

In this section, we contextualize the project within existing research and studies on the field of data science and its global demand. Previous work has acknowledged the exponential growth of data science and its pivotal role in various industries. Researchers have explored factors contributing to this growth, such as advancements in technology, the enhancement of big data, and the increasing emphasis on data\*driven decision\*making. However, while studies have recognized the rising demand for data scientists, few have specifically examined how population demographics influence this demand and its correlation with salary structures. By reviewing, the project aims to build upon existing knowledge and contribute to a deeper understanding of the dynamics shaping the global data science landscape.

**Dataset:**

This section outlines the primary datasets utilized in the project and provides key details about their contents. The 'World Population by Country' dataset sourced from Kaggle offers comprehensive information on the population sizes of various nations, providing a foundational understanding of global demographics. Conversely, the 'Data Science Salaries' dataset contains insights into the salaries of data science professionals worldwide, including details such as job titles, experience levels, and salary currencies. By combining these datasets, the project aims to uncover correlations between population size and data science salaries, facilitating a nuanced analysis of how global population demographics impact the demand for data science skills and compensation structures.

**Detail Design of Methods:**

1. Data Collection:

* Selection of Kaggle as the data source: Kaggle is chosen due to its reputation as a reliable platform for high\*quality datasets. It offers a diverse range of datasets contributed by the community and provides easy access to data relevant to various research topics.
* Identification and acquisition of relevant datasets: The 'World Population by Country' dataset is selected to provide comprehensive information on global population demographics. This dataset includes data for the year 2023,2022, allowing for current insights into population sizes worldwide. The 'Data Science Salaries' dataset is chosen to understand the salary structures of data science professionals globally. It contains detailed information on job titles, experience levels, and salary currencies, providing valuable insights into the compensation trends in the field of data science.

1. Data Cleaning and Manipulation:

* Handling missing values:

Missing data points are identified and addressed through imputation techniques and removal, depending on the extent and nature of the missingness. Imputation methods such as mean or median imputation been used to fill in missing values where appropriate.

* Standardization of data formats:

Data formats across the datasets are standardized to ensure consistency and compatibility for analysis. This may involve converting currencies to a common unit or standardizing date formats.

* Merging datasets:

The datasets are merged based on common identifiers, such as country names or codes, to create a unified dataset for analysis. This allows for a comprehensive examination of population demographics and data science salaries.

**Analysis:**

In the analysis phase of the project, the focus is on examining the relationship between population demographics and the demand for data science skills, as well as its impact on salary structures. This involves several key steps:

1. Exploratory Data Analysis (EDA): Initially, the team conducted exploratory data analysis to understand the distributions, trends, and relationships within the datasets. This includes summarizing key statistics, visualizing data distributions, and identifying potential outliers or anomalies.
2. Correlation Analysis: Statistical techniques are employed to analyse the correlations between population size and various factors related to data science, such as the number of data science job offerings, average salaries, and employment trends. This helped to identify significant associations between population demographics and the demand for data science skills.
3. Regional and Local Trends: The analysis also explores regional and local trends in data science demand and salary structures. This involves comparing data across different countries, regions, or even cities to identify variations and underlying factors driving these differences.
4. Causal Inference: Additionally, efforts are made to infer potential causal relationships between population demographics and data science metrics. This involves hypothesis testing or advanced statistical modelling to ascertain the extent to which population size influences data science demand and compensation.
5. Interpretation: Finally, the analysis phase concludes with interpreting the findings to derive actionable insights. This includes synthesizing the results of statistical analyses and visualizations to provide a coherent narrative of how population demographics impact the global data science landscape.

**Implementation:**

1. Data Cleaning and Manipulation:

* Utilization of Python with the Pandas module: Python is chosen as the programming language for data manipulation due to its versatility and extensive libraries for data analysis. The Pandas module provides powerful tools for data cleaning, transformation, and manipulation, streamlining the process.
* Execution of cleaning procedures: cleaning procedures, such as handling missing values and standardizing data formats, are carried out systematically to ensure the integrity and reliability of the data for subsequent analysis.

1. Statistical Analysis:

* Execution of statistical methods: Statistical analyses are performed using Python libraries such as NumPy and SciPy to explore relationships between population demographics and data science salaries. These analyses involve calculating descriptive statistics to summarize the data and inferential statistics to test hypotheses and identify significant associations.
* Generation of descriptive statistics: Descriptive statistics, including measures of central tendency and dispersion, are generated to provide insights into the distribution and variability of data science salaries across different population demographics.

1. Visualization:

* Creation of interactive visualizations using D3.js, Tableau: D3.js & tableau employed to create interactive and dynamic visualizations that enhance the presentation of findings. Interactive elements such as scatter plots are utilized to allow users to explore the data and gain deeper insights into correlations and trends.
* Implementation of visualization techniques: Visualizations such as scatter plots are used to visualize relationships between population size and data science salaries, while bar graphs with sized markers represent average salaries in different countries. These visualizations enable users to interact with the data and uncover patterns that may not be apparent through traditional static charts.

1. Validation:

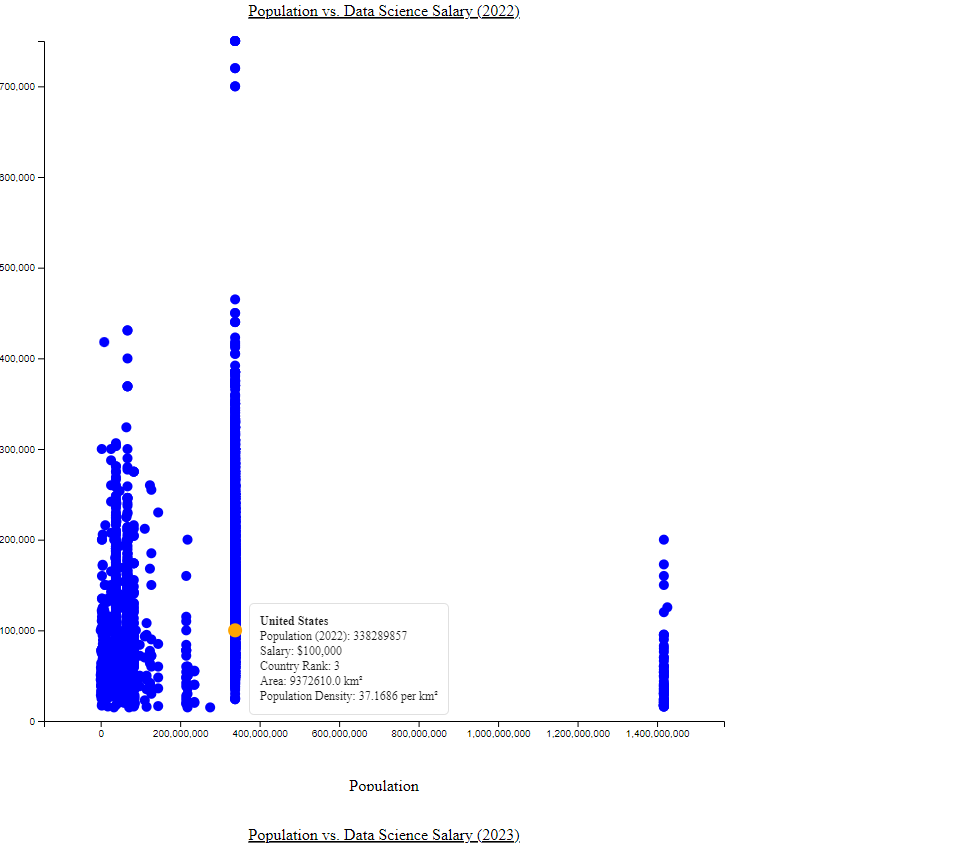
* Regular checks for data accuracy: Regular checks are conducted throughout the implementation process to validate the accuracy and integrity of the analysis. These checks help ensure that any anomalies or errors in the data are identified and addressed promptly.
* Verification of results: Results obtained from statistical analyses and visualizations are verified to ensure their reliability and consistency. Validation techniques such as cross validation employed to assess the robustness of statistical models and ensure that findings are generalizable.

1. Collaboration:

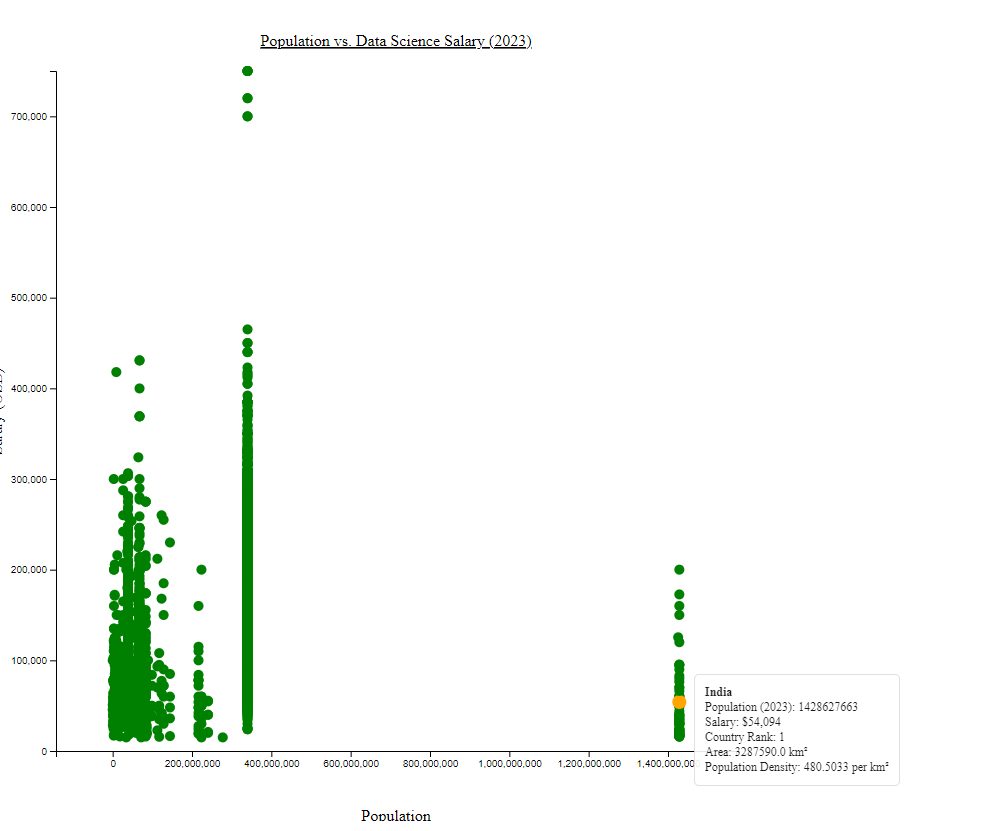
* Team collaboration for seamless progress: Collaboration among team members is crucial for effective project management and utilization of resources. By Regular communication and coordination ensured that tasks are completed efficiently and deadlines are met.
* Communication and coordination: Open communication channels and regular meetings facilitated the sharing of progress updates, addressing challenges or concerns, and fostering a collaborative working environment conducive to project success.

**Preliminary Results:**

Preliminary findings suggest a nuanced relationship between population size, demand for data science skills, and salary structures. Initial visualizations indicate potential correlations between certain demographic factors and data science job offerings and compensation.

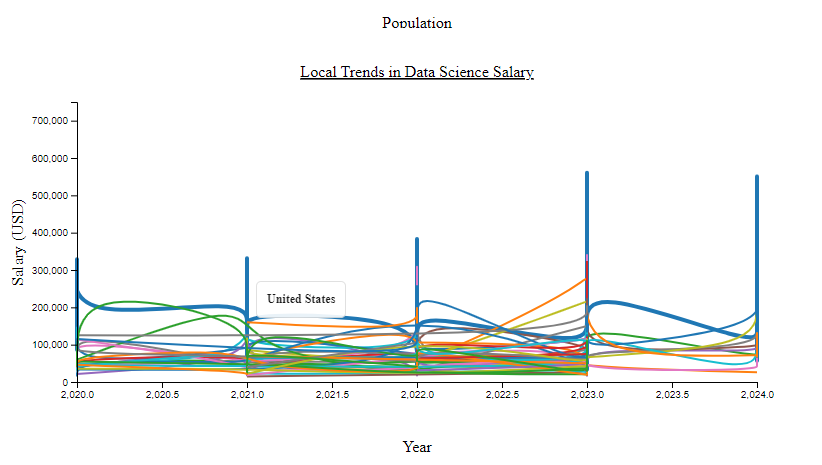
D3.js visualization:  
Two scatter plots are created where one for population data in 2022. The x-axis represents population size, while the y-axis represents the number of data scientists. Each point represents a country. On hovering it displays features of particular countries' population, salary, country Rank, Area and Population density.

Population vs. Data Science Salary (2023) in D3.js  
The data point for India is highlighted on the graph. It shows that in India (population of 1,428,627,663 in 2023), the average data science salary is $54,094. This is the highest data science salary listed although it isn't necessarily the peak salary.

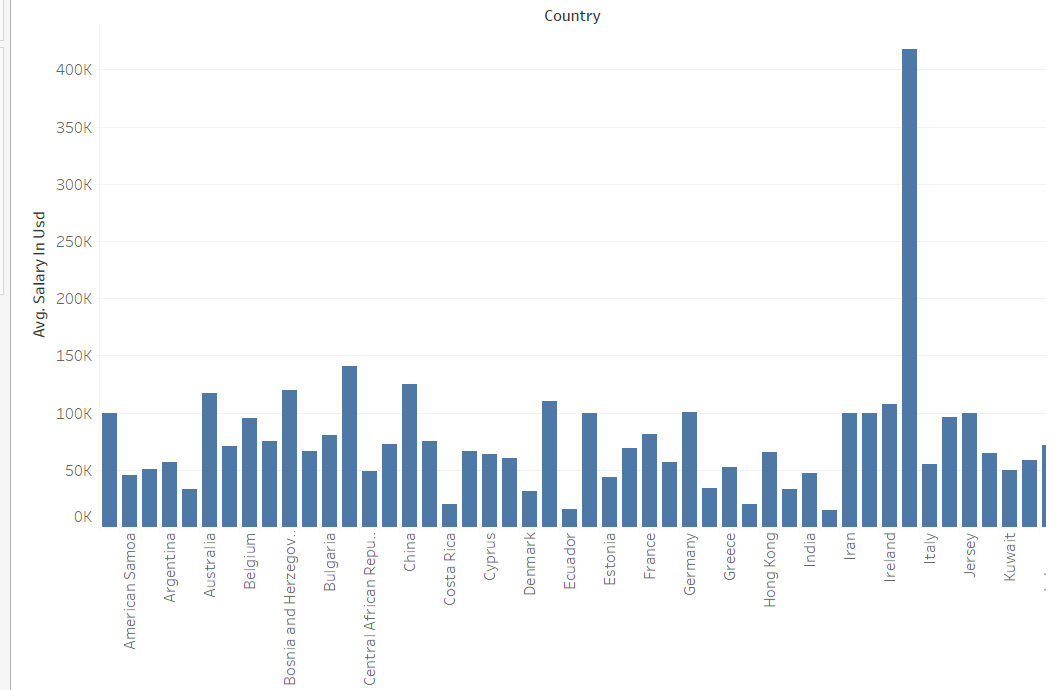


**Local trends in data science salary**

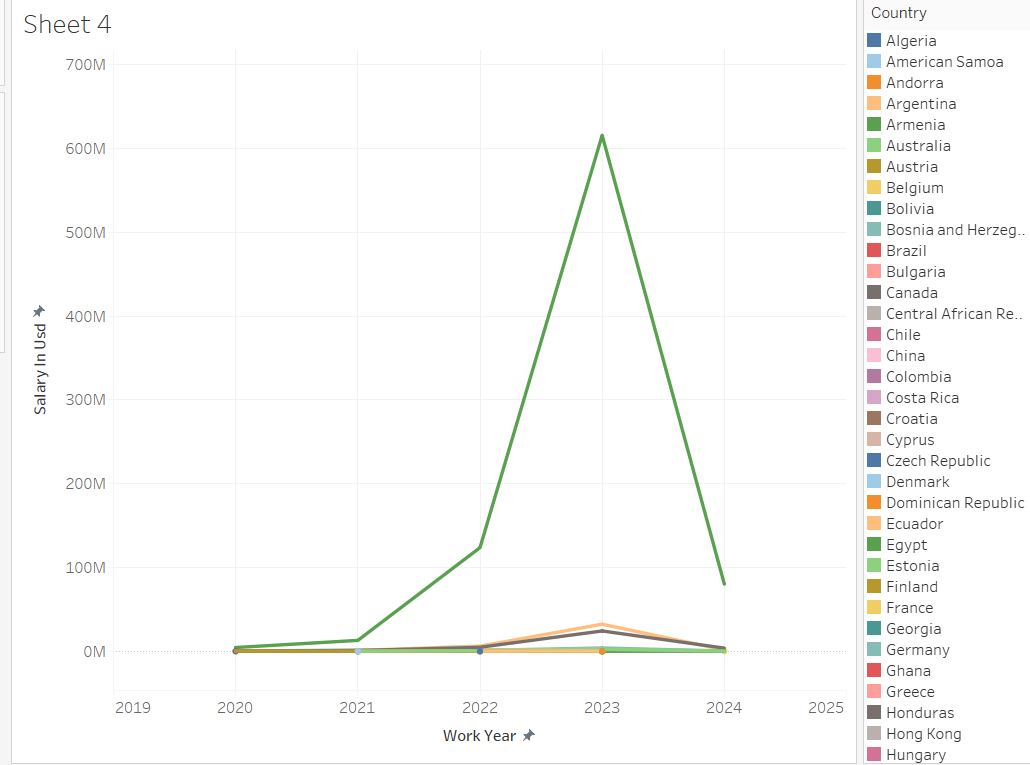
It shows a relationship between two variables - data science salary (likely average) and population. Each data point likely represents a country. The highlighted point emphasizes a specific country, probably here in plot its United States. It shows United States's population on the horizontal axis and its corresponding average data science salary on the vertical axis.



**Average data science salary by country**  
This plot visualizes the average data science salary across different countries. A bar plot is generated. The x-axis displays the countries, and the y-axis shows the average salary in USD. The bars represent the average salary for each country, and the plot is rotated for better readability.



**Local trends in data science salary using Tableau**  
This plot examines the local trends in data science demand and salary structures over the years for each country. A line plot is created. The x-axis represents the years, the y-axis shows the average salary in USD, and each line represents a country. Different colors are used to distinguish between countries, and a legend is added to identify each country’s trend over time.

****

**Project Management:**

Implementation Status Report:

**Contributions:**

1. **THARUN RAMULA**
   * **Responsibility:** Data Collection and Preparation, Documentation and Presentation
   * **Contribution:** Collected and loaded datasets, performed data cleaning and merged datasets based on common attributes. Prepared project documentation, including workflow diagrams and implementation status reports, and contributed to the creation of the PowerPoint presentation summarizing the project findings.
2. **KEERTHI REDDY PAPIAHGARI**
   * **Responsibility:** Data Analysis and Visualization in code
   * **Contribution:** Investigated the relationship between population size and the number of data scientists employed, analyzed variations in data science salaries across countries, and created visualizations using Python with Matplotlib and Seaborn.
3. **THARUN KURAVADI SATHISH BABU**
   * **Responsibility:** Data Analysis and Visualization in tableau
   * **Contribution:** Implemented Tableau visualization to understand the relation between population in countries and their salaries, and also relationship between work year and its salary from the datasets.
4. **LAKSHMI DEEPIKA PAGADALA**
   * **Responsibility:** Data Analysis and Visualization in d3.js
   * **Contribution:** Implemented interactive features using D3.js for visualizations, integrated visualizations into a webpage for presentation, and ensured the smooth deployment of the project.

**Concerns:**

* Limited availability of certain data points in the datasets may impact the depth of analysis.
* For Ensuring consistency and accuracy in data cleaning and manipulation processes.

**References:**

* <https://www.prb.org/resources/human-population/>
* <https://researchfdi.com/resources/articles/the-effects-of-globalization-on-economic-development/>
* <https://www.kaggle.com/datasets/rajkumarpandey02/2023-world-population-by-country>
* <https://www.kaggle.com/datasets/sazidthe1/data-science-salaries>

GITHUB LINK: <https://github.com/ks-tharun-14/Scientific-Data-Visualization-Project>