INTEL UNNATI PROJECT REPORT

PS-12

Knowledge Representation and Insight Generation from Structured Datasets

Project Name

Data Analysis App

Created By-

Name - Aditya Srivastava

E-Mail - adi.official.2990@gmail.com

Project GitHub

Deployed App

INTRODUCTION

Problem Statement:-

In the age of big data, organizations in diverse sectors are producing vast quantities of data daily. When properly processed and analyzed, this data can yield valuable insights that greatly enhance decision-making. However, the challenge is in effectively representing this knowledge and extracting actionable insights.

Our objective is to develop an AI-based solution to address this challenge. Given a structured dataset, our solution will process and analyze the data, represent the contained knowledge effectively, and generate meaningful insights.

Objectives:-

- Develop an AI-based solution for handling large datasets.
- Process and analyze structured datasets effectively.
- Represent the knowledge contained within the data.
- Generate valuable insights from the data.
- Enhance decision-making through data-driven insights.
- Overcome challenges in knowledge representation and insight extraction.

SOLUTION BRIEF

- **Overview:**

  This project involves the development of an AI-based data analysis application using Streamlit, a popular open-source app framework.This application aims to simplify the data analysis process, making it accessible and efficient for users to extract meaningful insights from their data.

- **Key Features:**
  - **Data Upload and Display:** Users can upload their dataset (CSV format), which the app displays along with a basic summary.
  - **Missing Value Handling:** The app provides multiple methods for handling missing values, including mean, median, mode, dropping columns, and using KNN Imputer
  - **Outlier Detection:** Users can detect outliers using IQR and Z-Score methods in their datasets.
  - **Data Visualization:** Users can visualize their data through various plots, including bar plots for categorical data and scatter, line, and bar plots for numerical data.
  - **Regression Plots:** Users can plot regression plots to see the general trend between any two features.
  - **Clustering:** The app includes functionality for clustering data points, allowing users to identify patterns and group similar data points.

- **Technologies Used:**
  - **Python Libraries:** NumPy, pandas, Streamlit, Matplotlib, Seaborn, Scikit-learn

- **Data Preprocessing:** Includes functions to convert boolean and categorical columns to numeric, and to handle missing values using KNN Imputer.
- **Visualization:** Utilizes Seaborn and Matplotlib for creating a variety of data visualizations.

DATASET DESCRIPTION

Dataset: Better Life Index 2024 (Source)

- This dataset, used for testing the data analysis application, provides comprehensive information on various quality of life indicators for 38 countries. The dataset includes 26 columns, each representing different aspects of life that contribute to overall well-being.

- This dataset serves as a robust example for testing the capabilities of the data analysis application, demonstrating its ability to preprocess, visualize, and extract meaningful insights from real-world data.

- This dataset can also be found in the dataset folder of the GitHub repository

- **Summary:**
  - **Missing Values:** Some columns contain missing values, which can be handled using various imputation methods provided in the application.
  - **Data Types:** The dataset includes numerical (int64, float64) and categorical (object) data types.

- **Descriptive Statistics:** Basic statistical details such as mean, median, standard deviation, min, and max values are provided for numerical columns.

METHODOLOGY

Overview:-

The methodology for this project involves several key steps to ensure effective data preprocessing, visualization, and analysis using an AI-based approach. This structured approach allows users to interact with the data through a user-friendly interface, providing valuable insights and enhancing decision-making.

Steps:

1. **Data Upload and Initial Display**
   - Users upload their datasets through the application interface.
   - The uploaded dataset is displayed with a basic summary, including the number of features, samples, and columns.
2. **Data Preprocessing**
   - **Conversion of Boolean and Categorical Data:**
     - Boolean columns are converted to numeric values (0 and 1) for consistency.
     - Categorical columns are converted to numeric values using `LabelEncoder`.
   - **Handling Missing Values:**
     - The application provides multiple methods for handling missing values:
       - **Fill with Mean:** Replace missing values with the mean of the column.

        - **Fill with Median:** Replace missing values with the median of the column.
        - **Fill with Mode:** Replace missing values with the mode of the column.
        - **Drop Column:** Remove columns with missing values.
        - **KNN Imputer:** Use K-Nearest Neighbors imputation to fill missing values based on the similarity of data points.
    - Users select the method for handling missing values through the interface, and the dataset is updated accordingly.
3. **Data Visualization**
    - Users can select columns for visualization from a dropdown menu.
    - **Graph Types:**
        - **Categorical Columns:** Bar plots are created with categorical columns on the y-axis and numerical columns on the x-axis.
        - **Numerical Columns:** Users can select from scatter plots, line plots, and bar plots, with both axes representing numerical data.
    - The application uses `Seaborn` and `Matplotlib` to generate visualizations, which are then displayed within the interface.
4. **Clustering Analysis**
    - Users can perform clustering on the dataset to identify patterns and group similar data points.
    - **KMeans Clustering:**
        - Users select the number of clusters and the column to be used for clustering.
        - The application applies KMeans clustering and visualizes the clusters using scatter plots, highlighting the different groups formed.

Tools and Technologies Used:

- **Programming Language:** Python
- **Libraries:**
  - **Data Processing and Analysis:**
    - `pandas`: For data manipulation and analysis.
    - `numpy`: For numerical computations.
  - **Machine Learning:**
    - `scikit-learn`: For Label Encoding, KNN Imputer, and KMeans clustering.
  - **Visualization:**
    - `Matplotlib`: For creating basic plots and figures.
    - `Seaborn`: For creating advanced, aesthetically pleasing visualizations.
  - **Web Framework:**
    - `Streamlit`: For building an interactive web application that allows users to upload datasets, handle missing values, visualize data, and perform clustering.

.

RESULTS AND DISCUSSION

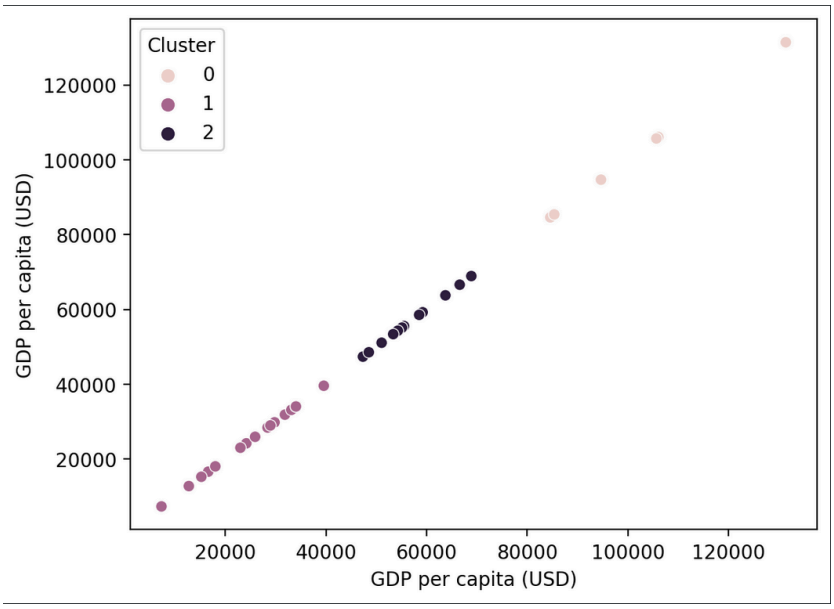
- This section includes the various insights generated in the Better Life Index Dataset 2024 using the Data Analysis App. All the graphs are generated by the Data Analysis App.

Insights

- Insight 1:-
  - The countries can be divided into 3 groups on the based on their GDP into:-
    - High GDP per capita Group (> 80,000$)
    - Middle GDP per capita Group(50,000$ to 80,000$)
    - Low GDP per capita Group (20,000$ to 50,000$)

  - Algorithm Used - K-Means Clustering

  - Graph Used-



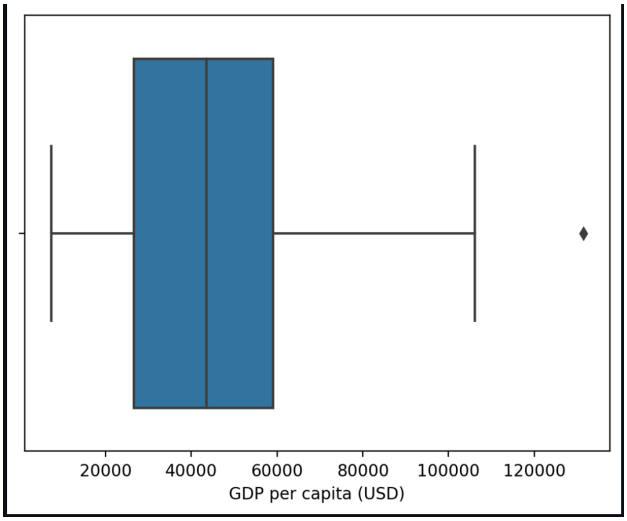
- Insight 2:-
  - Luxembourg is an outlier with the highest GDP of 131,384 USD
  - Algorithms Used - Outlier detection using Interquartile Range.
  - Application Snippet:-

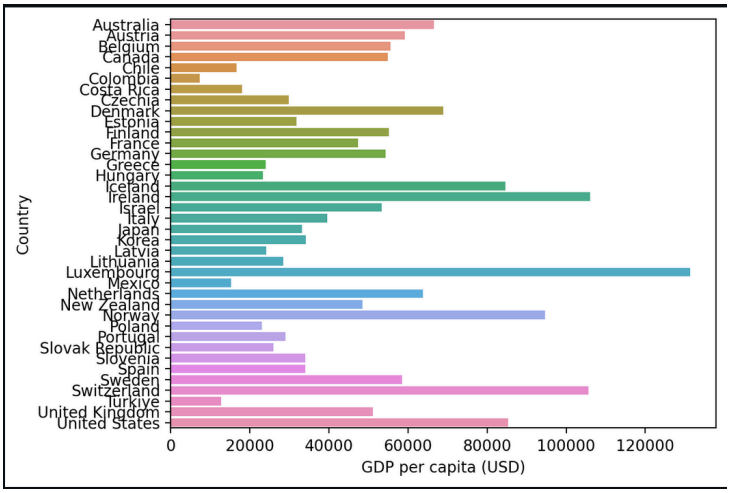
Outliers

The outliers in the dataset based on GDP per capita (USD) using IQR method are:

	Country	GDP per capita (USD)	Dwellings without basic facilities	Housing expenditure	Rooms p
23	Luxembourg	131,384	0.1	20.7	

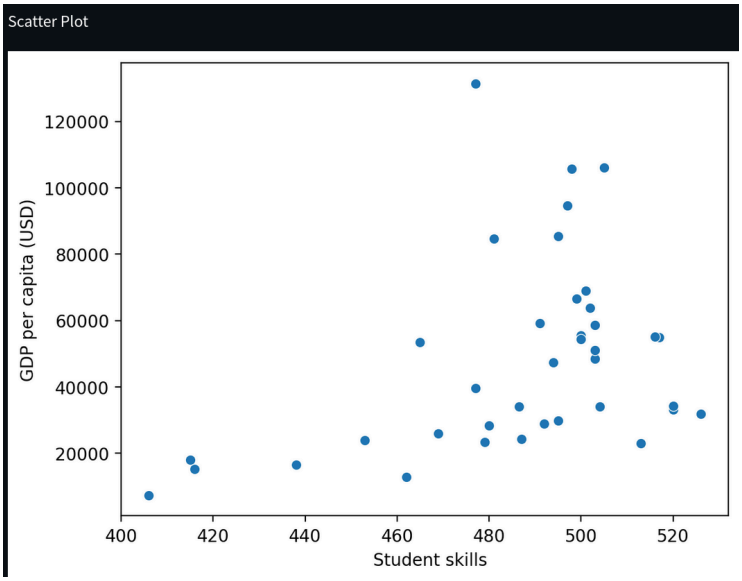
  - Graphs Used:-

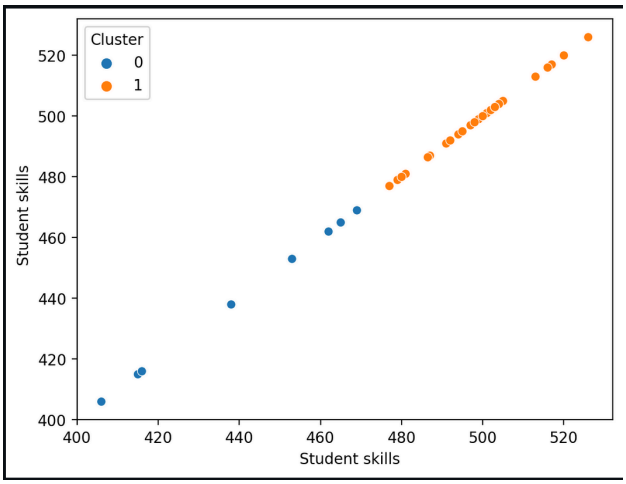




- Insight 3:-
    - There is no relation between the GDP of a country and the skills that can be attained by a student. Students of a lot of countries with GDP less than 50,000 USD have shown similar skill levels to the relatively higher GDP countries. A probable reason for such a trend could be the easy access to the internet in developing countries.

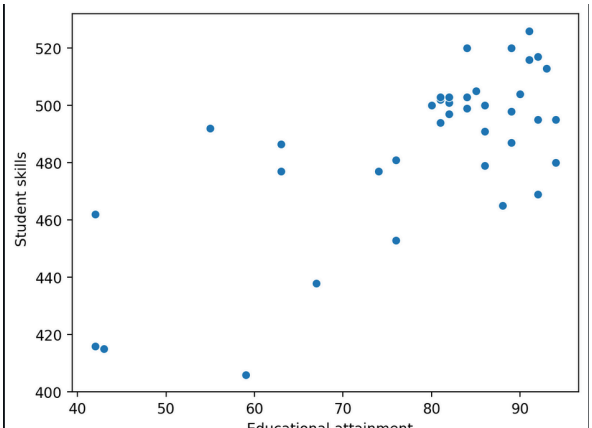
    - Graph Used:-





- Insight 4:-
    - Countries where the education attainment percentage is high, The students are found to be more skilled than the countries where the education attainment percentage is relatively low

    - Graph Used:-



- Insight 5:-
    - There is somewhat of a linear relationship between employment rate and homicide rate. In general, higher the employment rate and lower the homicide rate. Columbia and Mexico seem to have extremely high homicide rates.

    - Algorithm Used - Z-Score Outlier detection.

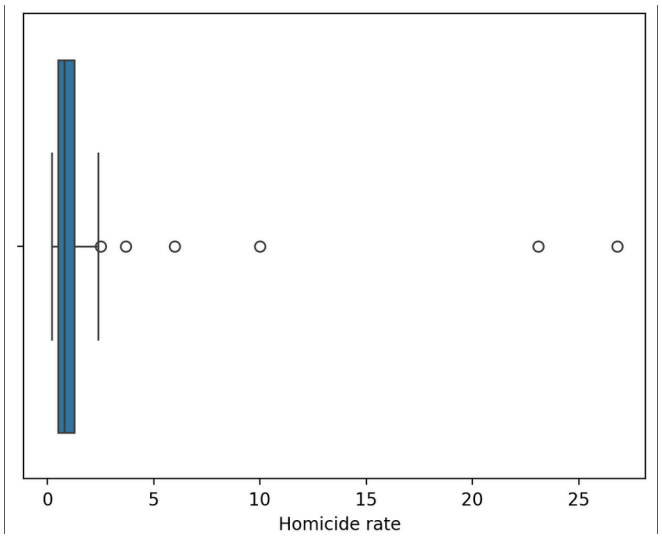
    - Application Snippet:-

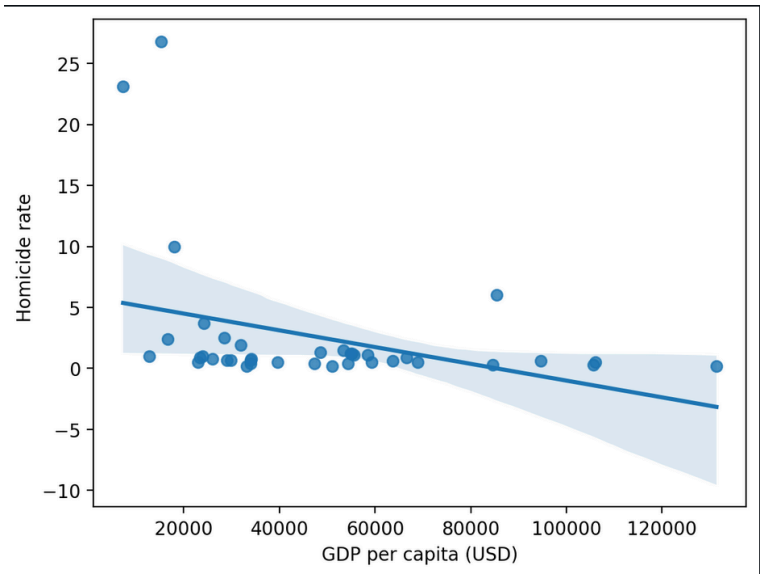
Outliers

The outliers in the dataset based on Homicide rate using z-score method are:

	Country	GDP per capita (USD)	Dwellings without basic facilities	Housing expenditure	Rooms per p
5	Colombia	7,327	12.3	20.4943	
24	Mexico	15,249	25.9	17.8	

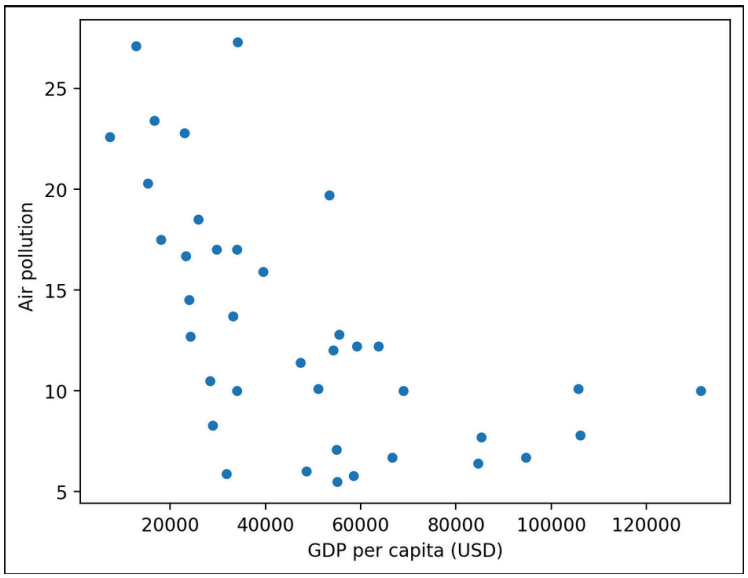
    - Graphs Used :-

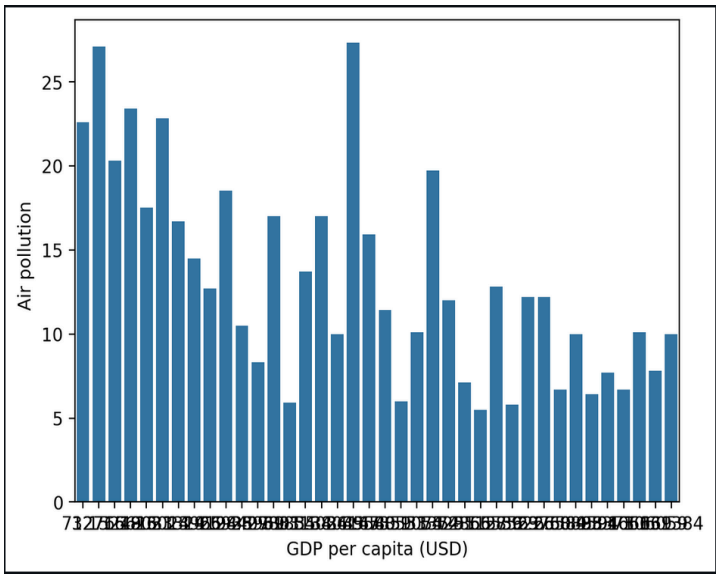




- Insight 6:-
    - Higher GDP countries have cleaner air when compared to countries with relatively lower GDP

    - Graphs Used:-





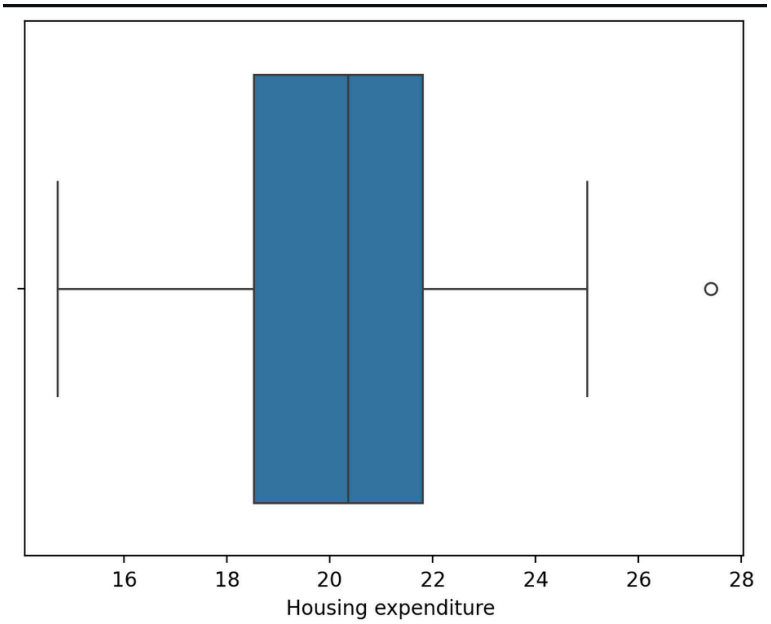
- Insight 7:-
    - The Slovak Republic is a country that spends the most of all the other countries on its housing.
    - Algorithm Used - IQR Outlier Detection Method
    - Application Snippet:-

Outliers

The outliers in the dataset based on Housing expenditure using IQR method are:

	Country	GDP per capita (USD)	Dwellings without basic facilities	Housing expenditure	Room
30	Slovak Republic	25,935	1.5	27.4	

- Graph Used:-



- Insight 8:-
  - Columbia, Mexico, Costa Rica and Turkiye have the worst work life balances where employees where more than 25% of the working population work for extremely long hours.
  - Algorithm Used - IQR Outlier Detection.

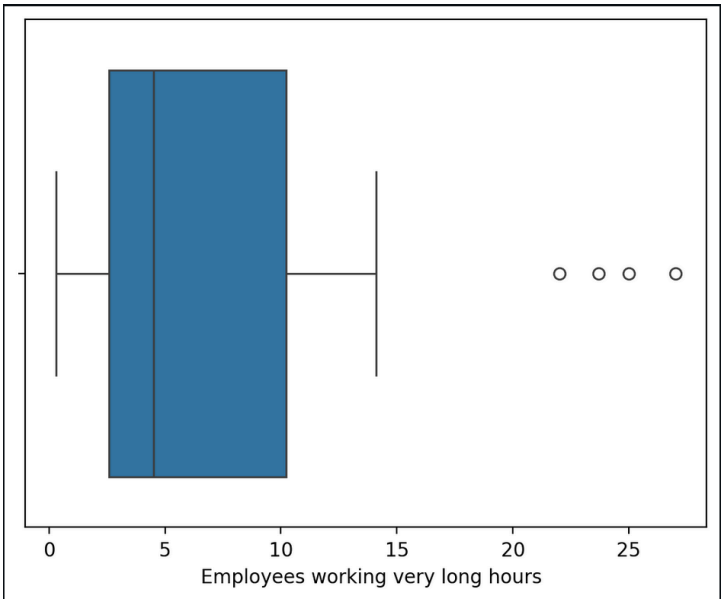
  - Application Snippet:-

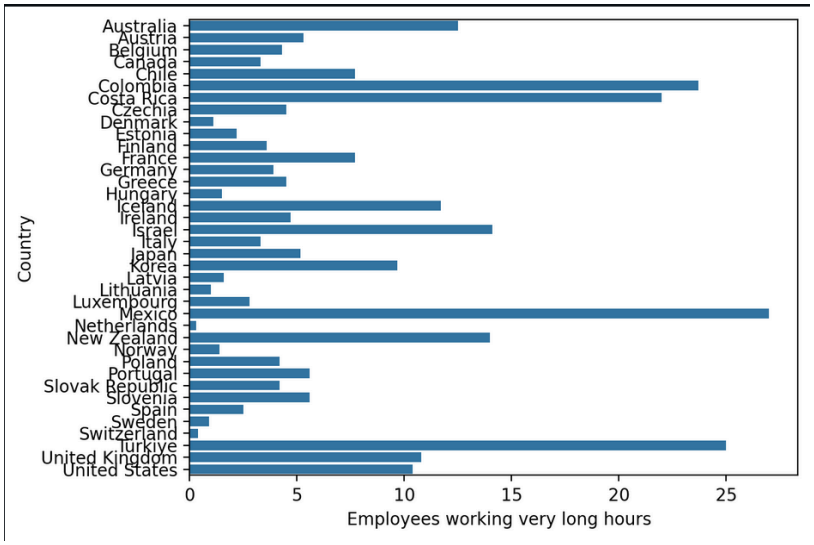
Outliers

The outliers in the dataset based on Employees working very long hours using IQR method are:

	Country	GDP per capita (USD)	Dwellings without basic facilities	Housing expenditure	Rooms per
5	Colombia	7,327	12.3	18.3667	
6	Costa Rica	18,031	2.3	17	
24	Mexico	15,249	25.9	17.8	
35	Türkiye	12,765	4.9	18.9	

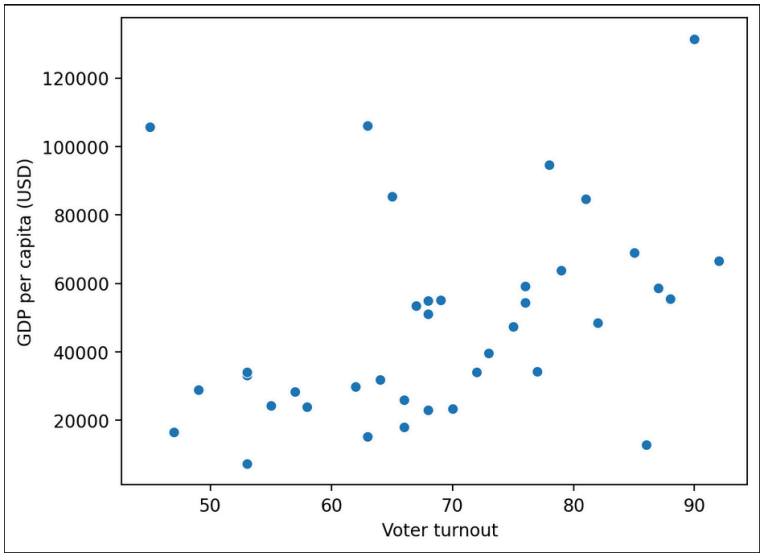
- Graphs Used:-

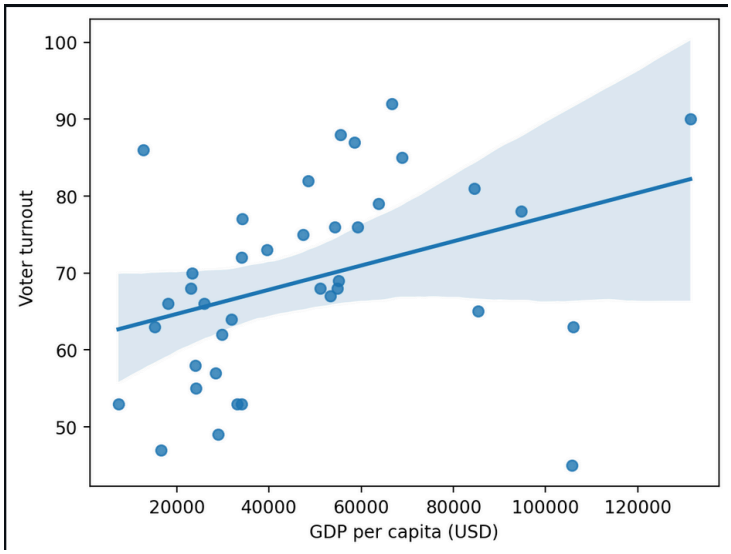




- Insight 9:-
  - Higher GDP countries seem to have a higher voter turnout. That is, the general population takes far more interest in the country's political scene than in relatively richer countries.

  - Graph Used:-





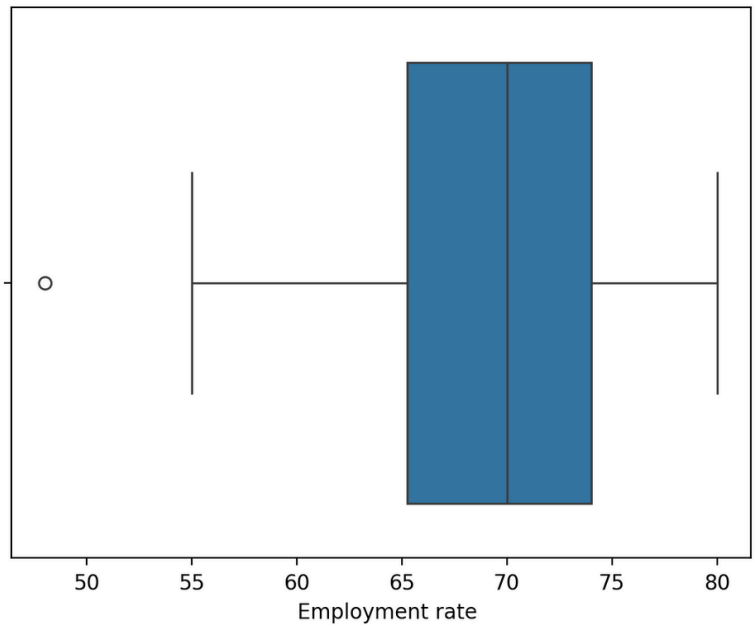
- Insight 10:-
    - Turkiye seems to have the lowest employment rate among all the other countries.
    - Algorithm Used- IQR Outlier Detection
    - Application Snippet:-

Outliers

The outliers in the dataset based on Employment rate using IQR method are:

	Country	GDP per capita (USD)	Dwellings without basic facilities	Housing expenditure	Rooms per pe
35	Türkiye	12,765	4.9	18.9	

    - Graph Used:-



FUTURE WORK

- **Natural Language Processing (NLP) Capabilities**:
  - **Text Analysis**: Introduce functionalities for text analysis, including sentiment analysis, topic modeling, and keyword extraction.
  - **Automated Insights**: Use NLP to generate automated insights and summaries based on the data.

- **Comprehensive Documentation and Tutorials**:
  - **In-App Tutorials**: Offer guided tutorials and walkthroughs to help users understand and utilize the app's features effectively.
  - **Detailed Documentation**: Provide extensive documentation and examples for each feature to assist users in their analysis tasks.

- **Enhanced User Experience**:
  - **User Authentication and Profiles**: Implement user authentication to save user settings, datasets, and analysis results for future sessions.
  - **Collaboration Features**: Add features for collaborative analysis, allowing multiple users to work on the same dataset simultaneously.

CONCLUSION

The Data Analysis App is a versatile and user-friendly tool designed to streamline the process of data analysis for users of all skill levels. By providing an intuitive interface and a suite of powerful features, the app allows users to upload datasets, handle missing values, detect and visualize outliers, and perform clustering analysis with ease.

Key features include:

- **Data Handling**: Effortlessly manage missing values using various methods such as mean, median, mode, and KNN Imputer.
- **Outlier Detection**: Identify outliers in your data using Z-score and IQR methods.
- **Data Visualization**: Create insightful visualizations, including scatter plots, line plots, bar graphs, and regression plots, to better understand the relationships within your data.
- **Clustering**: Perform clustering to segment your data into meaningful groups.

This app is an essential tool for anyone looking to perform comprehensive data analysis efficiently and effectively. Whether you are a novice or an experienced data analyst, the Data Analysis App provides the necessary tools to gain deeper insights from your data, making it an invaluable addition to your analytical toolkit.