**Dataset Analysis on qatar-monthly-statistics-visitor-arrivals-by-mode**

**Write your team and member name here**

**Introduction :**

In this project, we analyze the visitor arrivals to Qatar, focusing on the three major modes of entry: Air, Land, and Sea. The dataset used for this analysis is provided by the government of Qatar, offering monthly statistics on the number of visitors arriving through each mode. The primary objective of this analysis is to gain valuable insights into the behavior of visitors entering Qatar, and to understand the trends and seasonality of visitor arrivals across different time periods.

The dataset includes detailed information about visitor arrivals from 2010 onwards, providing a comprehensive overview of tourism patterns in Qatar. The project explores various aspects of this data, including the identification of seasonal trends, the impact of global events such as the COVID-19 pandemic, and the effectiveness of different modes of entry in attracting tourists.

Through this report, we will employ **data preprocessing** techniques to clean the data, **exploratory data analysis** (EDA) to visualize trends and patterns, and apply **time series forecasting** and **clustering models** to predict future visitor arrivals and identify underlying patterns. Our analysis will provide actionable insights for policymakers and businesses in Qatar to understand the dynamics of tourism and plan for future growth and challenges.

By using advanced data analysis techniques, this report aims to provide a detailed view of Qatar's tourism industry and contribute to the development of more effective tourism policies and strategies.

This analysis could be particularly beneficial for understanding the recovery of the tourism sector after global disruptions such as the COVID-19 pandemic, and for identifying opportunities to strengthen Qatar's position as a global tourist destination.

Through this approach, we hope to generate insights that not only forecast future visitor arrivals but also identify significant factors affecting these trends, thereby assisting decision-makers in better planning and strategizing for Qatar's tourism industry.

**Data Collection :**

For this project, we utilized the Qatar Monthly Statistics on Visitor Arrivals dataset, which was sourced from the official open data platform of the Qatari government. The dataset is publicly available on the following link: [Qatar Monthly Statistics - Visitor Arrivals by Mode of Entry](https://www.data.gov.qa/explore/dataset/qatar-monthly-statistics-visitor-arrivals-by-mode-of-entery/table/?sort=month). This dataset provides comprehensive records of visitor arrivals to Qatar from January 2010 to the present, categorized by entry mode (Air, Land, and Sea) and total visitor arrivals each month. The dataset was chosen for its relevance, completeness, and the level of detail it provides on the tourism sector in Qatar.

The data was originally in the form of a table that includes the following key columns:

* **Month**: The month for which the data is recorded.
* **Air**: The number of visitors arriving via air.
* **Land**: The number of visitors arriving via land.
* **Sea**: The number of visitors arriving via sea.
* **Total Visitor Arrivals**: The total number of visitors arriving to Qatar by any mode of transport.

We collected the data directly from the online government database, and this dataset serves as the basis for all analyses performed in this report. The data was in an Excel format, and after downloading, it was pre-processed and cleaned for further analysis.

**Discussion of the Open Data Considered**

Several open data sources were reviewed, but the dataset on monthly visitor arrivals by entry mode was selected due to its direct relevance to Qatar's tourism sector. The dataset offers granular monthly statistics, which are critical for identifying seasonal trends and forecasting future arrivals. The primary considerations for selecting this dataset were:

* **Relevance**: The data provided detailed information on visitor arrivals, directly related to the project’s goals of trend analysis and forecasting.
* **Quality and Consistency**: The dataset was relatively clean, with only minor missing values, and covered a substantial period to analyze trends and patterns effectively.
* **Time Series Nature**: The dataset's structure was ideal for time series forecasting, allowing for robust predictions of future arrivals.
* **Lack of Granularity**: Some datasets lacked the monthly breakdown, which was necessary for time series forecasting.
* **Data Completeness**: Some datasets contained missing data that couldn't be easily handled for forecasting.

**Rationale for Dataset Selection**

The chosen dataset was selected due to its high granularity and availability for multiple years. It also provided detailed information on the mode of arrival, which was useful for identifying different patterns of visitor behavior and predicting future trends. The dataset was ideal for time series forecasting and regression analysis, which were key aspects of this project.

**Methodology:**

The analysis was carried out using a systematic approach, focusing on data cleaning, exploratory data analysis (EDA), and advanced machine learning techniques. Below are the main methodologies used throughout this project:

1. **Data Preprocessing**:
   * The dataset was cleaned to remove any missing values, duplicates, or irrelevant information. We ensured that the data was in a suitable format for further analysis, converting date-related columns into **datetime** format to facilitate time-based analysis.
   * Columns containing extra spaces or unnecessary characters (like the extra spaces in column names) were corrected to make them more accessible for analysis.
2. **Exploratory Data Analysis (EDA)**:
   * Various **visualizations** were generated to explore patterns, distributions, and trends in the dataset. This included histograms to understand the distribution of arrivals for each mode of transport, boxplots to identify outliers, and trend lines to observe seasonal patterns.
   * We also conducted **correlation analysis** to check the relationships between different entry modes (Air, Land, Sea) and the total number of arrivals, identifying how the modes influence each other.
3. **Time Series Forecasting**:
   * Given the time-based nature of the data, **time series forecasting models** such as **ARIMA** and **SARIMAX** were applied to predict future visitor arrivals. These models were chosen to handle the temporal aspect of the data and to account for seasonality and trends that affect visitor numbers.
   * **SARIMAX** was specifically used because it accounts for seasonal effects, which are important when analyzing visitor arrivals, as they often fluctuate based on the time of year.
4. **Supervised Learning**:
   * **Linear Regression**, **Ridge Regression**, **Lasso Regression**, and **Random Forest Regression** models were employed to predict the total number of visitor arrivals based on the number of visitors arriving by air, land, and sea.
   * These models were trained on historical data, and various **evaluation metrics** such as **MAE**, **MSE**, and **R²** were used to assess model performance.
5. **Unsupervised Learning**:
   * **K-Means Clustering** and **DBSCAN** were used to identify patterns in the data, particularly to segment the visitors based on their entry mode (Air, Land, Sea) and to discover any hidden groups that exhibit similar behavior.
   * **Principal Component Analysis (PCA)** was employed for dimensionality reduction to make the clustering results more interpretable and visualizable.
6. **Model Evaluation**:
   * For the time series models (ARIMA and SARIMAX), metrics such as **MAE**, **MSE**, **RMSE**, and **MAPE** were used to evaluate the forecasting accuracy.
   * For clustering, we used **Silhouette Score**, **Davies-Bouldin Index**, and **Calinski-Harabasz Index** to evaluate the quality of the clusters formed by different unsupervised learning models.

By utilizing these methods, the analysis aimed not only to provide insights into historical trends but also to make predictions and classify the visitor arrival patterns to assist with future planning and tourism strategy development in Qatar. The insights derived from this analysis can be used by policymakers, businesses, and tourism authorities to adjust strategies based on changing patterns in visitor arrivals.

**Data Analysis:**

The data analysis for the visitor arrivals dataset from Qatar reveals insights into visitor trends and the distribution of arrivals by different modes. The following analysis includes three visualizations that were generated using the dataset:

A graph with purple lines and numbers

AI-generated content may be incorrect.

Figure 1 : Total Visitor Arrivals Trend Over Time

The line graph shown here (Figure 1) illustrates the trend in total visitor arrivals from 2018 to 2024. The data indicates several significant peaks, with the highest peaks occurring during 2019 and 2024. There are also noticeable dips in some periods, especially in 2020, which likely corresponds to the global impacts of the COVID-19 pandemic. This analysis highlights how visitor arrivals have fluctuated over time, with a general decline after the peaks in 2019 and 2020.

A graph with purple rectangles

AI-generated content may be incorrect.

Figure 2 : Total Visitor Arrivals Per Year

Figure 2 presents a bar chart showing the total number of visitors per year from 2018 to 2024. It clearly shows that 2018 and 2019 experienced the highest number of arrivals, with a significant drop in 2020 due to the pandemic. The years 2021 to 2024 reflect a steady recovery in the number of visitors, although the numbers have not reached the pre-pandemic levels seen in 2018 and 2019.

A pie chart with numbers and text

AI-generated content may be incorrect.

Figure 3 : Distribution of Visitor Arrivals by Mode (Air, Land, Sea)

The pie chart in Figure 3 represents the distribution of visitor arrivals by different modes of transportation: Air, Land, and Sea. It is clear that the majority of visitors arrive by air, comprising 81% of the total arrivals. Land arrivals account for 9.4%, and Sea arrivals make up 9.6% of the total. This distribution gives insight into the preferred modes of travel for visitors to Qatar, with air travel dominating the arrivals.

These insights provide a clear picture of the patterns in visitor arrivals to Qatar, with air travel being the dominant mode of entry. The trends over time also highlight how external factors such as the pandemic have affected visitor numbers, with a noticeable decline in 2020 followed by a gradual recovery.

**Algorithms:**

The main objective of the analysis was to examine the trends and patterns of visitor arrivals in Qatar across different modes of entry (air, land, and sea) and to understand the fluctuations over time. Several algorithms and techniques were applied to achieve this goal:

1. **Data Preprocessing and Cleaning**: The data collected from the Qatar Monthly Statistics dataset was first cleaned and preprocessed to ensure it was usable. Missing values were identified, and data types were corrected for analysis. A critical step in data preprocessing was ensuring the 'Month' and 'Year' columns were extracted from the datetime field, allowing for proper time-series analysis and aggregation by year.

2. **Time-Series Analysis**: To analyze the trends in visitor arrivals over time, we applied time-series analysis using **line graphs** to visualize the number of visitors across different months and years. The goal was to capture seasonality, trends, and irregularities in the data. This allowed us to detect peaks and valleys in the arrival numbers, such as the sharp drop during the COVID-19 pandemic and the subsequent recovery.

3. **Statistical Analysis and Distribution**: We used **descriptive statistics** to explore the distribution of visitor arrivals by mode of entry. The **pie chart** showed that air arrivals dominate, followed by land and sea arrivals. Statistical methods such as calculating the mean, median, and standard deviation were applied to understand the central tendencies and dispersion in visitor arrivals for each mode.

**Optimizations**

* + - **Hyperparameter Tuning:** Parameters for ARIMA and SARIMAX models were optimized to improve forecasting accuracy.
    - **Feature Scaling**: StandardScaler was applied to ensure that regression and clustering models performed optimally.
    - **Cross-Validation**: Used to evaluate the performance and generalization ability of models like Ridge and Linear Regression.

4. **Visualization Algorithms**: Data visualization was a significant component of the project. **Matplotlib** and **Seaborn** libraries were used to generate:

* Line plots for trends over time,
* Bar charts for yearly comparisons,
* Pie charts to visualize the distribution of visitor arrivals by mode of entry. These visualizations helped communicate the insights derived from the dataset effectively.
* **Overall Evaluation**

The project successfully achieved the objective of analyzing and forecasting visitor arrivals. The combination of time series forecasting, regression modeling, and unsupervised learning provided a comprehensive understanding of visitor trends. The forecasting models, especially SARIMAX, offered accurate predictions, and the clustering model provided insights into visitor behavior.

**Application Design :**

The application design was centered on creating an interactive system to allow users to explore the trends and distribution of visitor arrivals data in Qatar. The system was designed to include the following components:

1. User Interface: The user interface was built using Streamlit, a Python framework that allows the rapid development of web applications. The interface allows users to:
   * Select the mode of arrival (Air, Land, or Sea),
   * View time-series trends for total visitor arrivals across various months and years,
   * Analyze the annual distribution of visitor arrivals.
2. Functionality:
   * Data Selection: The user can select different time periods (e.g., by month or by year) to view visitor arrivals.
   * Visualizations: Upon user input, the system dynamically generates appropriate charts, including:
     + Total Visitor Arrivals Trend: Line charts showing the trend of total visitor arrivals across time.
     + Yearly Distribution: Bar charts that show total arrivals per year.
     + Mode of Entry Distribution: Pie charts depicting the breakdown of visitors based on how they arrived (Air, Land, Sea).
   * The system updates in real-time as users adjust the inputs.
3. Data Integration: The dataset was integrated into the application from the link provided, which was extracted into a Pandas DataFrame. The user interface dynamically loads and processes the dataset, ensuring that the visualizations reflect the most up-to-date data available.

**Implementation**

The implementation phase involved writing the code to:

1. Load and Clean Data: The dataset was loaded into a Pandas DataFrame. The 'Month' column was split to extract the year and month values, and data preprocessing steps like filling or removing missing values were carried out.
2. Generating Visualizations: Using Matplotlib and Seaborn, various types of visualizations were created:
   * Line Charts: Used to visualize trends over time.
   * Bar Charts: Used for displaying the total visitor arrivals by year.
   * Pie Charts: Used for showing the distribution of visitor arrivals by mode.
3. Error Handling and Optimization: The implementation also included error handling for edge cases, such as invalid inputs (e.g., when a user tries to input a non-numeric value for any column, the system asks for correct input). The applicationwas optimized to load data efficiently, ensuring a smooth user experience.

**Interface Requirements**

* + - **Data Input Interface:** Ability to upload visitor arrival data in CSV or Excel format.
    - **Dashboard:** Display visualizations of visitor arrivals, forecasts, and clustering insights.
    - **Forecasting Panel:** Section to view predictions for future visitor arrivals.
    - **Clustering Insights:** Visualizations of clustering results, such as pie charts or bar graphs, showing the distribution of visitors by mode of entry.

**Testing**

* + **Unit Testing:** Ensured that data processing steps, such as handling missing values and scaling, were functioning correctly.
  + **Model Evaluation:** Used MAE, MSE, and R² to assess model accuracy.
  + **Cross-Validation:** Applied k-fold cross-validation to assess model robustness.
  + **User Interface Testing:** Tested the interface for user-friendliness and ease of interaction.

**System Architecture**

**A diagram of data processing

AI-generated content may be incorrect.**

**Libraries**

* + 1. **pandas -** Used for data manipulation and analysis.
    2. **matplotlib -** Used for creating static, animated, and interactive visualizations.
    3. **scipy -** Used for scientific and technical computing (e.g., for statistical functions like zscore).
    4. **seaborn -** Used for making statistical graphics, often used in conjunction with matplotlib.
    5. **sklearn -** A machine learning library (used for regression, clustering, and other ML tasks).
    6. **statsmodels -** Used for statistical modeling and time series analysis (like ARIMA, SARIMAX).
    7. **numpy** - Used for numerical computing and working with arrays.

**Conclusion:**

The combination of time-series analysis, statistical methods, and visual analytics allowed us to gain significant insights from the visitor arrivals data in Qatar. The system was designed to provide users with a simple yet interactive interface, making it easy to explore the trends and distributions of visitor arrivals across various modes of entry. The use of Streamlit for the front-end and Matplotlib/Seaborn for the backend analysis provided an intuitive and effective platform for this data-driven application. The final implementation successfully demonstrates how data can be processed, analyzed, and visualized for informed decision-making.

**Reference**

*Gitbook*. (n.d.). Retrieved from https://git-scm.com/book/en/v2.

Havill, Jessen;. (2016). *Discovering Computer Science: Interdisciplinary Problems, Principles, and*

*Python Programming.* Boca Raton, Florida: CRC Press.

Kopoc, D. (2019). *Classic Computer Science Problems in Python.* Manning Publications.

*Matplotlib*. (n.d.). Retrieved from https://matplotlib.org/stable/index.html

*Network*. (n.d.). Retrieved from https://networkx.org/documentation/stable/tutorial.html

Pual, Ores, et al;. ( 2017). *Practical Programming: An Introduction to Computer Science Using*

*Python 3.6.* Ralcigh, North Carolina: The Progmatic Bookshelf.