**Analysis and Forecasting of Visitor Arrivals in Qatar**

**Author(s):** Mohammed Aly(Leader), Abdulrahman,Ahmed,Mohannadi

**Date: 27/03/2025**

**Introduction**

The visitor arrival statistics of Qatar provide important insights into the country's tourism dynamics, helping the government, stakeholders, and policymakers make informed decisions. In this project, we explore the monthly visitor arrivals by air, land, and sea over the years. By analyzing this data, we aim to extract valuable trends, create predictive models for future arrivals, and uncover hidden patterns using unsupervised learning.

**Purpose of the Report**

* Time series forecasting to predict future visitor trends.
* Regression modeling to understand relationships between different modes of entry (air, land, sea) and total visitor arrivals.
* Unsupervised learning techniques (clustering) to identify patterns in the data.

**Scope and Objectives**

* Perform exploratory data analysis (EDA) to identify trends and outliers in the dataset.
* Implement **time series forecasting** using ARIMA and SARIMAX models to predict future visitor arrivals.
* Apply **regression analysis** to model the relationship between different entry modes and total arrivals.
* Use **unsupervised learning** techniques such as K-Means clustering to identify patterns.
* Provide actionable insights and recommendations for Qatar’s tourism strategies.

**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.

**Assessment of Potential Insights and Interface Requirements**

**Potential Insights**

**Statistical Analysis(Visualization)**

**A group of blue rectangular objects

AI-generated content may be incorrect.**

Figure 1 : Boxplot Distribution of Visitor Arrivals by Mode of Entry

The boxplot diagram displays the distribution of four key variables: **Air Arrivals**, **Land Arrivals**, **Sea Arrivals**, and **Total Visitor Arrivals**. Each plot shows the interquartile range (IQR), median, and potential outliers in the data for each entry mode. **Air Arrivals** exhibit significant variability, with several outliers in the higher range. **Land Arrivals** show a slightly skewed distribution, while **Sea Arrivals** are more tightly clustered within the IQR. The **Total Visitor Arrivals** plot reflects a wide spread, with a few high outliers indicating spikes in visitor arrivals during specific periods.

A graph of different colored lines

AI-generated content may be incorrect.

Figure 2 : Visitor Arrivals by Mode of Entry (Air, Land, Sea) and Total Over Time

This line plot illustrates the trends of visitor arrivals over time from different modes of entry: **Air**, **Land**, and **Sea**. The **Total Arrivals** are represented by the red line, which shows a significant spike in 2020, while **Air Arrivals** (blue) fluctuate, peaking during the same period. **Land Arrivals** (green) exhibit a more stable and gradual increase, while **Sea Arrivals** (orange) remain relatively constant with a few minor fluctuations. The plot clearly highlights the impact of specific events on visitor arrivals and the overall increase in visitor numbers in recent years.

A graph of a number of months

AI-generated content may be incorrect.

Figure 3 : Average Monthly Visitor Arrivals by Mode of Entry and Total

This line plot displays the **average monthly visitor arrivals** for different entry modes (**Air**, **Land**, and **Sea**) along with the **Total Arrivals**. The **Air Arrivals** (blue) show a steady increase throughout the year, with a noticeable upward trend as the months progress. **Total Arrivals** (red) closely follow a similar upward trajectory, reflecting the cumulative increase in visitors. **Sea Arrivals** (orange) and **Land Arrivals** (green) remain relatively constant with little fluctuation throughout the year. The plot highlights seasonal variations and the increasing influence of air travel on total visitor arrivals over time.

A graph of different colored bars

AI-generated content may be incorrect.

Figure 4 : T-Test p-values for Anomalous Months (Jan 2023, Dec 2019, Jan 2021)

This bar plot shows the **p-values** from **T-tests** comparing the visitor arrival modes (Air, Land, Sea, Total) for three anomalous months: **January 2023**, **December 2019**, and **January 2021**. The **blue bars** represent the p-values for January 2023, **orange bars** represent December 2019, and **green bars** represent January 2021. The chart helps to identify the statistical significance of the differences in visitor arrivals during these months compared to the rest of the dataset. Higher p-values suggest no significant differences, while lower values indicate more significant differences in the data for the anomalous months.

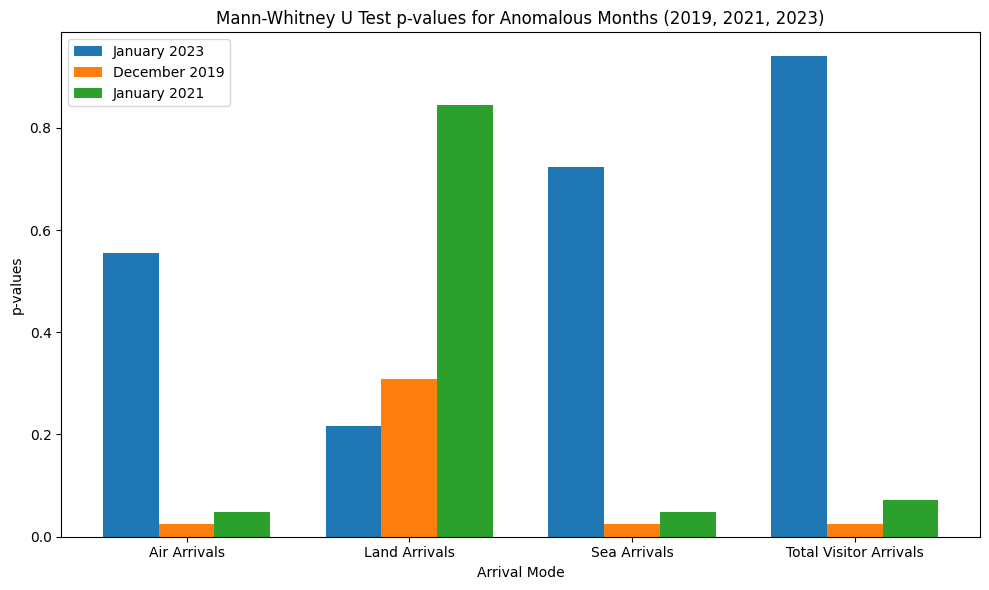


Figure 5 : Mann-Whitney U Test p-values for Anomalous Months (2019, 2021, 2023)

This bar plot shows the **p-values** from the **Mann-Whitney U Test** comparing visitor arrival modes (Air, Land, Sea, Total) for three anomalous months: **January 2023**, **December 2019**, and **January 2021**. The **blue bars** represent January 2023, **orange bars** represent December 2019, and **green bars** represent January 2021. The plot illustrates the significance of differences in visitor arrivals during these months. Higher p-values suggest that there are no significant changes in visitor arrivals, while lower p-values indicate significant differences during these months.

A graph with a line going up

AI-generated content may be incorrect.

Figure 6 : Yearly Growth Rate of Total Visitor Arrivals

This line plot shows the **yearly growth rate** of **Total Visitor Arrivals** as a percentage change from one year to the next. The **blue line** represents the percentage change over the years 2019 to 2024. A sharp increase in growth is observed between 2021 and 2022, where the percentage change exceeds 300%. This is followed by a significant drop in 2023, with a steady, low growth in 2024. The plot provides insights into the fluctuations in visitor arrivals, particularly highlighting the sharp recovery post-pandemic and the subsequent decline.

**Insights and Visualization of Regression models**

**A screenshot of a graph

AI-generated content may be incorrect.**

Figure 7 : Correlation Heatmap of Visitor Arrivals

This heatmap shows the **correlation coefficients** between the different modes of visitor arrivals: **Air Arrivals**, **Land Arrivals**, **Sea Arrivals**, and **Total Visitor Arrivals**. The color intensity represents the strength of the correlation, where **red** indicates a strong positive correlation and **blue** represents a weaker or negative correlation. For example, **Air Arrivals** have a very high positive correlation with **Total Visitor Arrivals** (0.99), while **Land Arrivals** show a relatively weak negative correlation with **Sea Arrivals** (-0.31). This heatmap provides a visual representation of how the different modes of entry relate to each other.

A yellow bar graph with black text

AI-generated content may be incorrect.

Figure 8 : Model Performance Metrics (MAE, MSE, R²) for Different Regression Models

This figure consists of three bar charts, each representing the evaluation metrics of various regression models: **Linear Regression**, **Ridge Regression**, **Lasso Regression**, **Random Forest Regression**, and **Support Vector Regression**.

* **Mean Absolute Error (MAE)**: The first chart shows that **Support Vector Regression (SVR)** has the highest MAE, while **Linear Regression** and **Ridge Regression** perform well with lower MAE values.
* **Mean Squared Error (MSE)**: The second chart reveals that **SVR** again has a significantly higher MSE, indicating poor fit compared to other models. **Linear Regression** shows the lowest MSE, suggesting better accuracy.
* **R-squared (R²)**: The third chart shows the **R²** values, with **Linear**, **Ridge**, and **Lasso Regression** performing similarly with very high R² values, indicating near-perfect fitting. **SVR** and **Random Forest** lag behind, indicating poorer model fit.

A graph with red squares

AI-generated content may be incorrect.

Figure 9 : K-Fold Cross Validation: MSE for Linear and Ridge Regression

This bar chart shows the **Mean Squared Error (MSE)** values for **Linear Regression** and **Ridge Regression** during **K-Fold Cross Validation**. The **light blue bars** represent **Linear Regression**, while the **red bars** represent **Ridge Regression**. The **MSE values** across the 5 folds are shown on the x-axis, with **Ridge Regression** exhibiting a significantly higher MSE value for **Fold 5**, indicating a possible overfitting or a large discrepancy in performance for that fold. In contrast, **Linear Regression** maintains relatively consistent and low MSE values across all folds, suggesting better model stability.

**Unsupervised Learning model insights**

**A graph of a diagram

AI-generated content may be incorrect.**

Figure 10 : 3D K-Means Clustering of Visitor Arrivals (Air, Land, Sea)

This 3D scatter plot visualizes the **K-Means clustering** results of visitor arrivals (Air, Land, Sea) from the **extended dataset**. The plot shows three principal components derived from **PCA** (Principal Component Analysis), where each point represents a data sample. The clusters are differentiated by color, with **Cluster 0 (Air Arrivals)** highlighted in yellow. The clustering is based on the underlying data, revealing how the visitor arrivals group into distinct clusters. This visualization helps understand the patterns of visitor arrivals based on the three different modes of transportation and shows the natural grouping in the data based on their features.

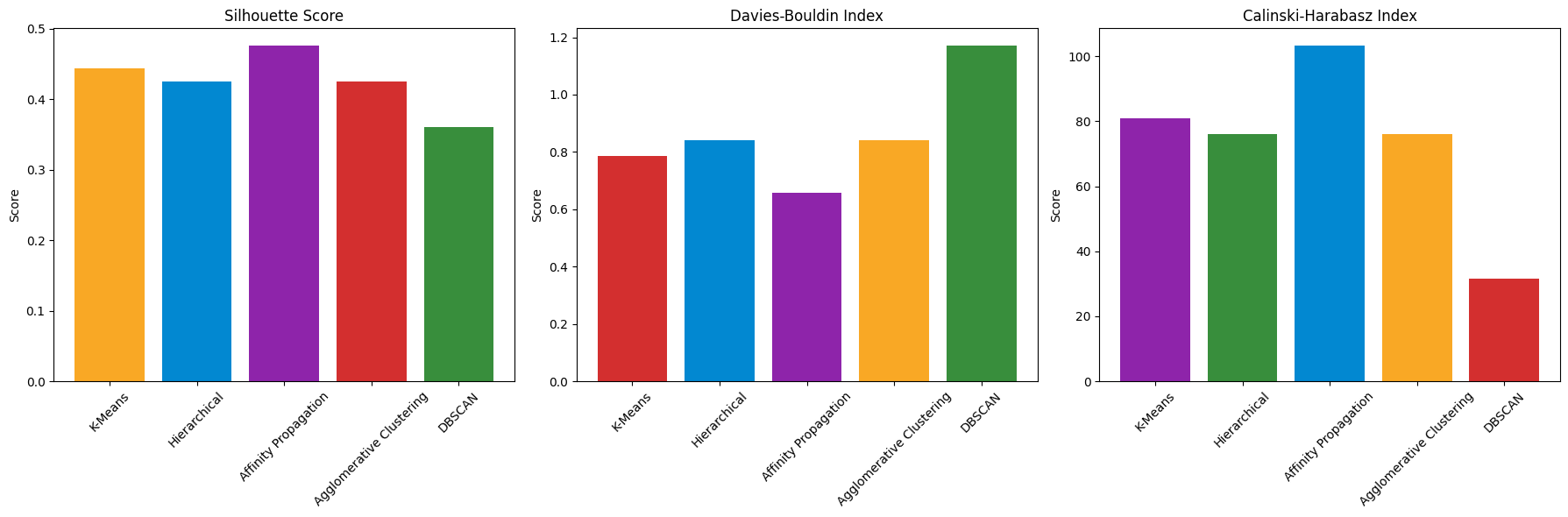


Figure 11: Clustering Evaluation Metrics

This set of bar plots presents the evaluation metrics for several clustering algorithms: **K-Means**, **Hierarchical**, **Affinity Propagation**, **Agglomerative Clustering**, and **DBSCAN**. The first plot, showing the **Silhouette Score**, suggests that **K-Means** and **Hierarchical** clustering provide the most well-defined clusters, with high scores. The second plot, the **Davies-Bouldin Index**, shows **DBSCAN** performing the best, indicating lower cluster dispersion. Finally, the **Calinski-Harabasz Index** in the third plot ranks **K-Means** the highest, suggesting that it produces the most compact clusters. These results provide insights into the effectiveness of each clustering method for segmenting the visitor arrival data.

**Time Series Forecasting model insights**

A red and blue squares

AI-generated content may be incorrect.

Figure 12 : Evaluation Metrics for ARIMA and SARIMAX

This set of bar charts compares the **evaluation metrics** for **ARIMA** and **SARIMAX** models. The four metrics displayed are **Mean Absolute Error (MAE)**, **Mean Squared Error (MSE)**, **Root Mean Squared Error (RMSE)**, and **Mean Absolute Percentage Error (MAPE)**. The **red bars** represent **ARIMA**'s values, while the **blue bars** represent **SARIMAX**'s values. The charts show that **SARIMAX** consistently outperforms **ARIMA** in all metrics, with significantly lower error values, particularly in **MAE** and **MSE**. This emphasizes the **better forecasting performance** of the SARIMAX model for the given time series data.

A graph of a graph showing the difference between a graph and a chart

AI-generated content may be incorrect.

Figure 13 : ARIMA vs SARIMAX Forecast Comparison

This line chart compares the **ARIMA** and **SARIMAX** models' forecasts with the **training data** and **actual data**. The **blue line** represents the **training data**, the **black line** shows the **actual data**, and the **red** and **blue lines** represent the **ARIMA** and **SARIMAX** forecasts, respectively. The chart demonstrates how the ARIMA model fails to predict the **future data trends** accurately, while the SARIMAX model performs better by fitting the actual data more closely over the forecast period. The chart helps visualize the **forecasting accuracy** of both models over time, highlighting the potential of **SARIMAX** in time series forecasting.

**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.

**Algorithm Exploration and Optimizations**

1. **Time Series Forecasting:**
   * + - **ARIMA:** Used for its simplicity in univariate time series forecasting**.**
       - **SARIMAX:** Chosen for its ability to handle seasonality, which was crucial for this data.
2. **Regression Models**:
   * + - **Linear Regression**: To model the relationship between the modes of entry and total visitor arrivals.
       - **Ridge and Lasso Regression**: Used for regularization to prevent overfitting.
       - **Random Forest and Support Vector Regression (SVR)**: Used to capture non-linear relationships in the data.
3. **Clustering**:
   * + - **K-Means Clustering**: Applied to group similar visitors based on entry modes.
       - **Other Models**: Agglomerative Clustering, DBSCAN, and Affinity Propagation were also explored.

**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.

**Design: Data Processing and User Interaction**

**Application Design**

* + - **Process Input Data:** Accept monthly visitor arrivals data.
    - **Train Models:** Use ARIMA, SARIMAX, and regression models for forecasting.
    - **Generate Forecasts:** Provide predictions for future visitor arrivals.
    - **Clustering Insights:** Provide segmentation of visitors based on mode of entry.

**User Interaction**

* + - * **Data Upload:** Users can upload CSV/Excel files for analysis.
      * **Forecast Visualization**: Forecasts are displayed on the dashboard using line charts.
      * **Clustering Insights**: Clusters are visualized using pie charts or bar graphs.
      * **Model Evaluation**: The application will provide the performance metrics (MAE, MSE, R²) for each model.

**Description of Implementations**

**Data Cleaning**

* **Missing Values:** Handled through imputation or removal of missing rows.
* **Feature Engineering:** Additional features were created, such as year-over-year growth rates.
* **Scaling:** Applied standard scaling to data used in clustering.

**Model Implementations**

* ARIMA and SARIMAX models were implemented using the statsmodels library for time series forecasting.
* Regression Models were implemented using scikit-learn, with optimization for Ridge and Lasso through cross-validation
* K-Means Clustering was implemented with scikit-learn to group visitors into different segments.

**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.

**Evaluation**

* SARIMAX was the best model with a MAPE of 28.51%, providing the most accurate forecast.
* **Regression Models**: Linear Regression and Ridge Regression performed excellently, achieving an **R² of 1**.
* **Clustering**: K-Means successfully identified distinct visitor segments, but **Affinity Propagation** showed better evaluation scores.

**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.

**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.

**System Architecture**

**A diagram of data processing

AI-generated content may be incorrect.**

.