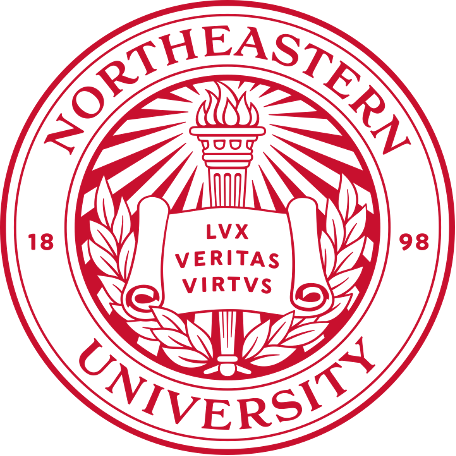
****

**Module 4 Final Project**

**Group 4: Chengjun jiang, Jinsheng Wei, Junchen Yi**

**College Professional Studies in Analytics, Northeastern University**

**ALY\_6110: Big Data and Data Management**

**Dr. Donhoffner**

**Submission Date: 09/06/2024**

**Introduction**

This report aims to analyze the demographics and web session records of Airbnb users to predict the destination country of a new user's first booking (Airbnb new user bookings, n.d.). We focus on two primary questions:

1. User Profile Analysis:

* What are the demographics of Airbnb users?
* Are the users predominantly male or female?
* What age groups do they belong to?

2. Time Series Traffic Analysis:

* Which months are the peak seasons for Airbnb usage?
* Which days of the week see the most travel activity?
* During which times of the day is Airbnb used the most?

3. User Behavior pattern Analysis:

* What are the operational behavior patterns of users when using Airbnb?
* Which operations and device types are associated with high usage hours?

**Data Overview**

The dataset provided includes user demographics, web session records, and summary statistics. The target is to predict one of the 12 possible outcomes for the destination country: 'US', 'FR', 'CA', 'GB', 'ES', 'IT', 'PT', 'NL', 'DE', 'AU', 'NDF' (no destination found), and 'other'.

**User Profile Analysis**

1. Where are the main users of Airbnb from?

The distribution of user destinations shows that the majority of users have bookings in the US (47,286) and many have no destination found (NDF) (55,041). Other significant destinations include 'other' countries (7,329). This indicates that most users either book within the US or do not have a specific destination.

2. Are the users male or female?

The gender distribution indicates that 40.59% of users are male, 46.13% are female, and 13.28% are unspecified or unknown. This suggests that female users slightly outnumber male users.

3. What age group do they belong to?

The age distribution reveals that most users fall into the 30-39 age group (1,610,233), followed by the 20-29 age group (939,536) and the 40-49 age group (887,388). This shows that the primary user demographic is adults aged 30-39.

**Time Series Traffic Analysis**

1. Which months are peak seasons?

The analysis of account creation by month shows peak activity in May (14,784) and June (15,482). Similarly, first booking activity peaks in May (7,785) and June (7,935). Therefore, May and June are identified as peak seasons for Airbnb usage.

2. Which days of the week have the most travel?

The data indicates that account creation is highest on Mondays (20,060), followed by Wednesdays (19,529) and Tuesdays (18,809). First bookings are also highest on Wednesdays (10,742), followed by Tuesdays (10,708) and Thursdays (10,413). Hence, the most active days are from Tuesday to Thursday.

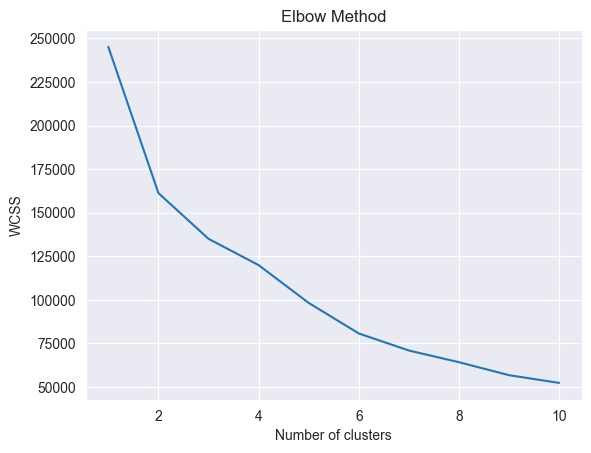
3. During which time of the day is Airbnb used the most?

The provided dashboard does not include a specific analysis of activity by time of day. Further data analysis is required to determine peak usage times within a day.

**User Behavior Pattern Analysis**

Based on the clustering analysis results and visualization charts, we can identify the user's operation behavior patterns and draw the following conclusions:

We classified the users into three clusters, Cluster 0, Cluster 1 and Cluster 2. These clusters represent different user behavior patterns (Optimove, 2024).



According to the Elbow Method graph, the optimal number of clusters is 3 or 4, as the decrease in WCSS slows down significantly at these points, indicating that further increase in the number of clusters does not improve the clustering effect much. This indicates that dividing customers into 3 or 4 clusters can effectively reflect different customer behavior patterns. Based on our dataset, we chose to categorize users into three clusters.

1. Cluster 0:

* Users with high values for total\_actions and unique\_actions indicate that these users perform a large number and variety of operations on the platform.
* Their total\_secs\_elapsed is also higher, indicating that they spend more time on the platform.
* The average\_secs\_per\_action is relatively low, indicating that these users perform operations more frequently and more quickly.

2. Cluster 1:

* Users have moderate values for total\_actions and unique\_actions, indicating that these users perform a moderate amount of varied operations on the platform.
* Their total\_secs\_elapsed is moderate, indicating that they spend a moderate amount of time on the platform.
* The average\_secs\_per\_action is also moderate, indicating that these users have a balanced frequency and duration of operations.

3. Cluster 2:

* Users have lower values for total\_actions and unique\_actions, indicating that these users have fewer and less varied operations on the platform.
* Their total\_secs\_elapsed is also lower, indicating that they spend less time on the platform.
* Higher values on average\_secs\_per\_action indicate that these users operate less frequently but spend more time per action.

Through cluster analysis, we can identify three main patterns of user behavior:

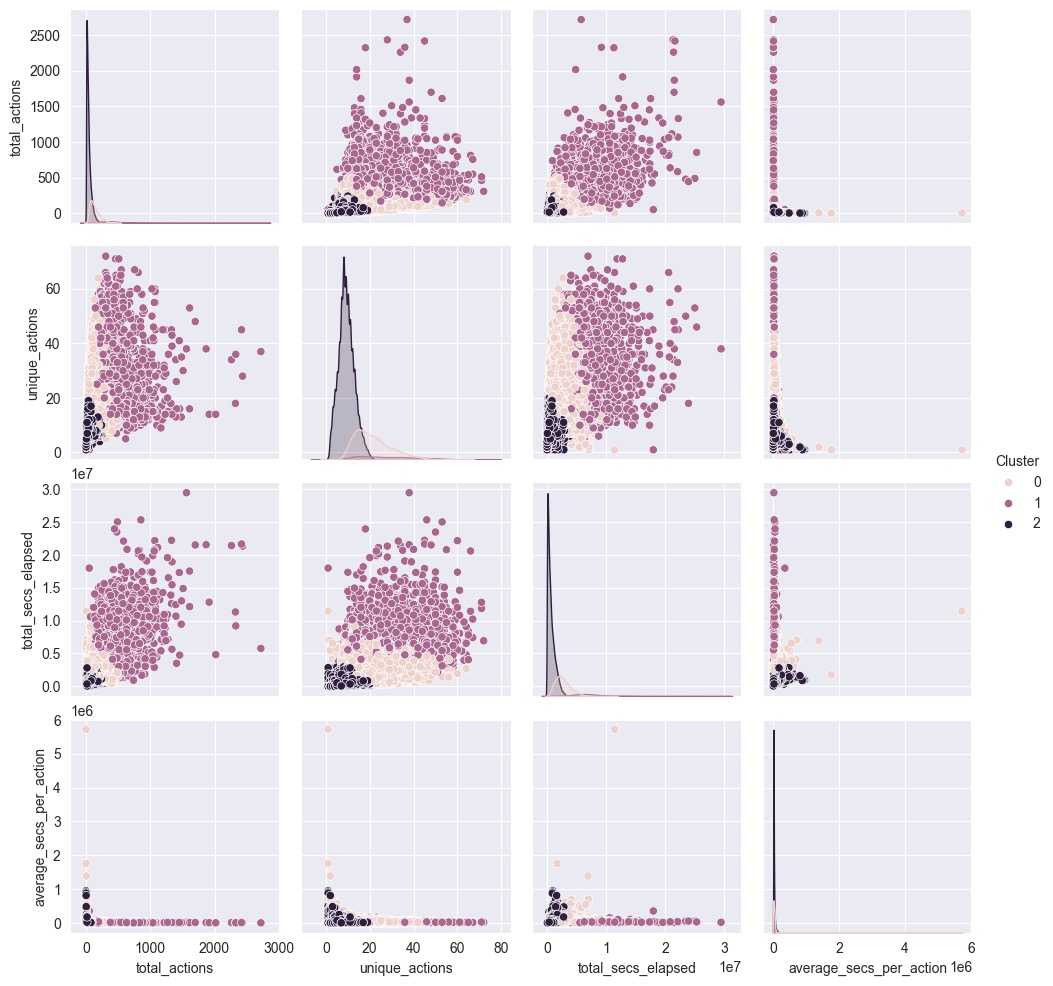
1. High-frequency and efficient users (Cluster 0): these users perform a large number and variety of operations on the platform, with a shorter but more frequent time per operation. They spend more total time on the platform and may be heavy or specialized users of the platform.

2. Moderate Frequency Users (Cluster 1): These users perform a moderate number and variety of actions on the platform, with a balanced duration and frequency of each action. They spend a moderate amount of time on the platform and may be average users.

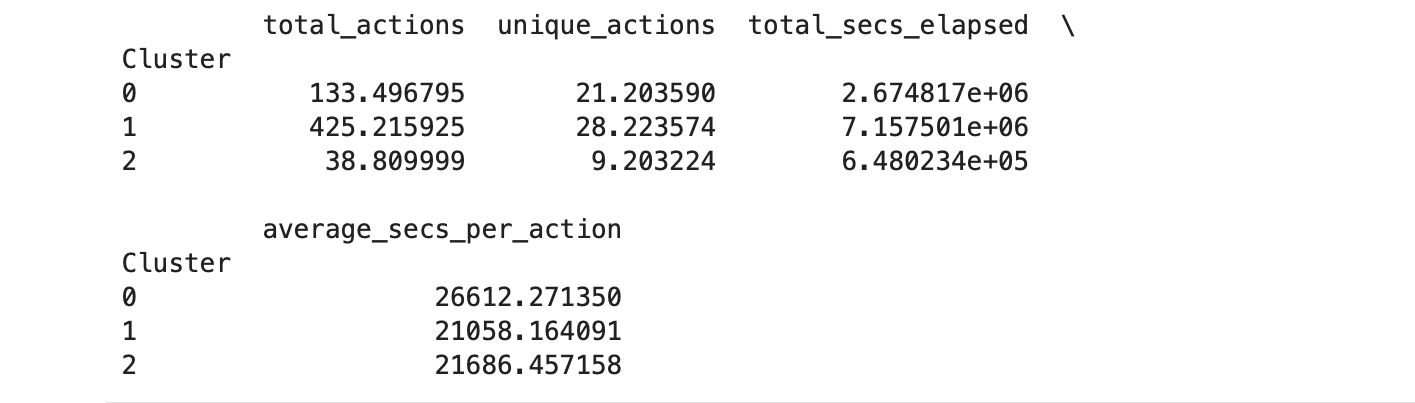
3. Low Frequency and Low Efficiency Users (Cluster 2): These users have a low number and variety of operations on the platform, and each operation is longer but less frequent. They spend less total time on the platform and may be new or occasional users.

All in all, these behavioral patterns help us understand how different types of users interact with the platform and provide data support for optimizing user experience. For example, more advanced features can be introduced for high-frequency and efficient users, while more guidance and incentives can be provided for low-frequency and inefficient users to increase their activity and satisfaction.

**Clustering Visualization**



**Mean of Each Cluster**



**Dashboard Presentation**

To support the analysis and provide a clear visual representation of the data, a dashboard has been created using Tableau (Tableau, n.d.). This dashboard includes graphical elements that highlight key insights from the data:

User Age Distribution: Displays the age groups of Airbnb users.

Distribution of User Destination: Shows the distribution of booking destinations.

Gender Characteristics of Users: Pie chart illustrating the gender distribution.

Distribution of Users by Language: Bar chart showing the language preferences of users.

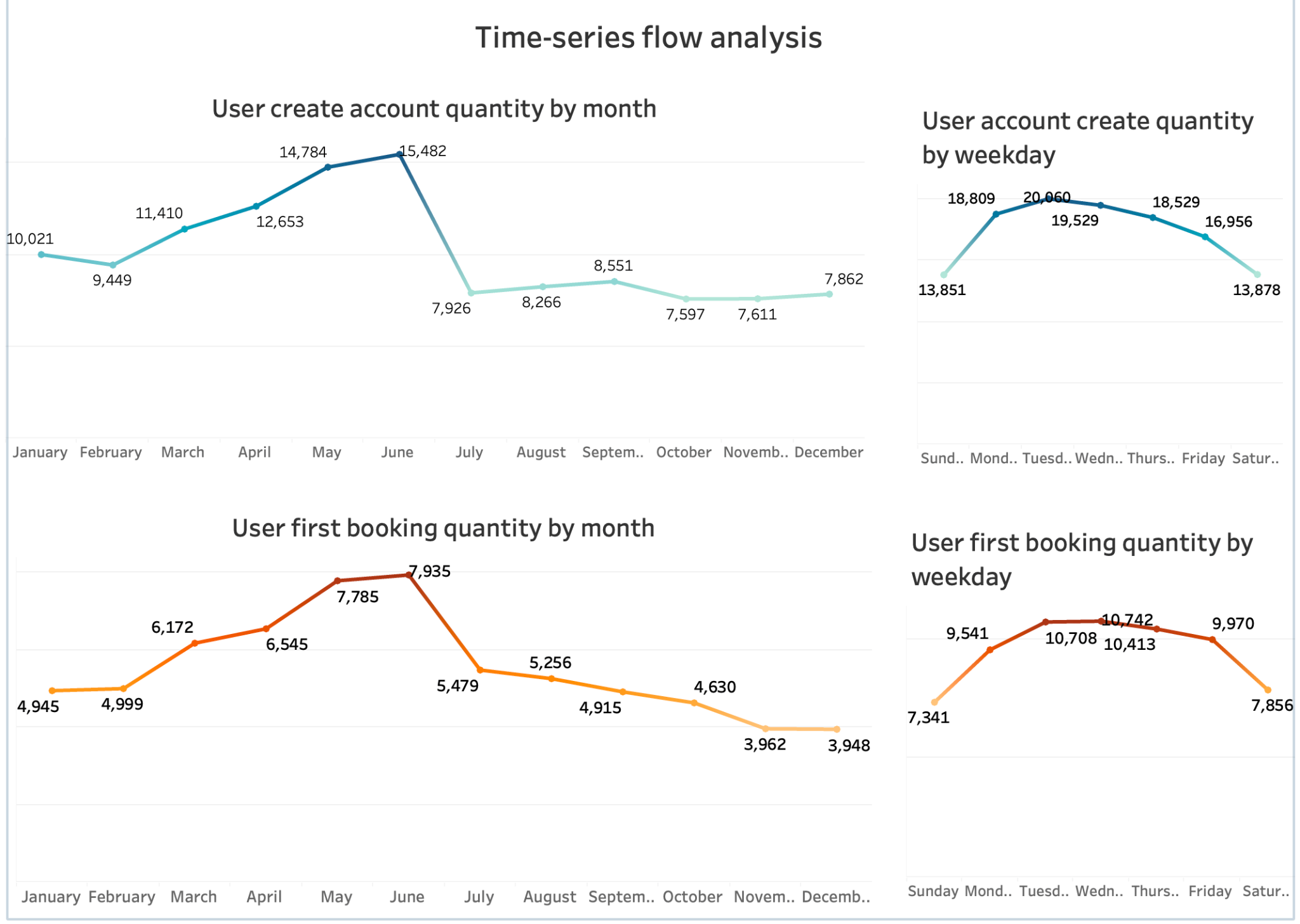
User Account Creation by Month: Line graph depicting the number of accounts created each month.

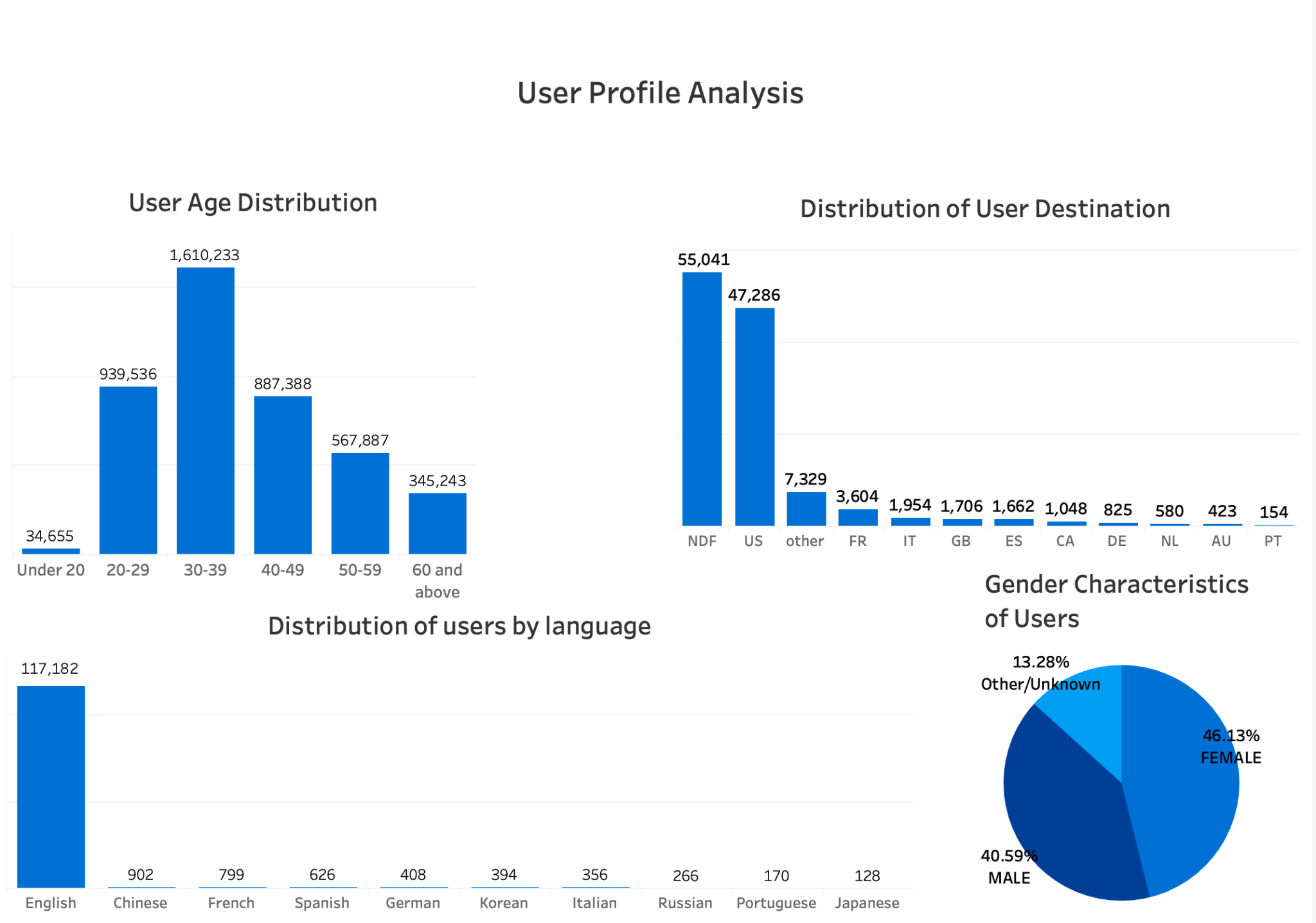
User First Booking by Month: Line graph showing the first bookings by month.

User Account Creation by Weekday: Line graph illustrating account creation by day of the week.

User First Booking by Weekday: Line graph depicting first bookings by day of the week.

The dashboard effectively communicates the analysis results, making it easier to understand the patterns, clustering, and correlations in the data.





**Conclusion**

The analysis provides valuable insights into the demographics and usage patterns of Airbnb users. The majority of users are from the US, with a significant proportion not specifying a destination. Female users slightly outnumber males, and the primary age group is 30-39 years. May and June are peak months for Airbnb activity, and the most active days are from Tuesday to Thursday. Further analysis is needed to understand usage patterns within a day.

This report, along with the associated dashboard created using Tableau, helps visualize these findings and supports the interpretation of the patterns, clustering, and correlations found in the data (Kriebel, 2023).

**Reference**

*Airbnb new user bookings*. Kaggle. (n.d.). <https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings/data>

*Dashboards*. Tableau. (n.d.). <https://help.tableau.com/current/pro/desktop/en-us/dashboards.htm>

Kriebel, A. (2023, June 20). *How to use a Tableau Dashboard: Tableau Tutorial for beginners*. YouTube. <https://www.youtube.com/watch?v=s_uDUmIUeWA>

*Customer segmentation via Cluster Analysis*. Optimove. (2024, March 14). <https://www.optimove.com/resources/learning-center/customer-segmentation-via-cluster-analysis>