Data mining Report

Student	ID
Mashael Alburaidi	442200107
Lina	442200438
Ruba	442204800
Jana	442200472

Supervised by<HANAN ALTAMIMI>

Introduction:

Our dataset pertaining to the Las Vegas Strip was sourced from http://archive.ics.uci.edu/
This website is regarded as a reliable and valuable resource for discovering diverse datasets, making it an excellent choice for researchers and data enthusiasts seeking high-quality data for various purposes.

Description of data:

The dataset signifies the quantity of reviews submitted through the TripAdvisor application for hotels in Las Vegas. This data was gathered and generously donated on July 22, 2017, and it essentially represents individuals who have taken the time to share their feedback and experiences regarding any hotel in Las Vegas using the app.

Description of attributes:

1. User Country: Represents the country of origin of the reviewer, indicating where they are from.

2. Number of Reviews: Reflects the total count of reviews written by the reviewer on TripAdvisor, providing insight into their reviewing activity.

3. Number of Hotel Reviews: Indicates the total number of reviews written by the reviewer specifically for hotels on TripAdvisor, highlighting their hotel-related reviewing history.

4. Helpful Votes: Counts the total number of people who have found the reviewer's review helpful and voted accordingly, gauging the perceived value of their feedback.

5. Score: Represents the numeric rating given by the reviewer to the hotel they stayed at, typically on a scale from 1 to 5, with 1 being the lowest and 5 being the highest.

6. Period of Stay: Denotes the month in which the reviewer stayed at the hotel, discretized to allow reviewers to specify the month of their stay within a given period.

7. Traveler Type: Categorizes the type of traveler or group the reviewer was part of during their stay, with options including Couples, Friends, Business travelers, Families, and Solo travelers.

8. Pool: A binary attribute indicating whether the hotel being reviewed has a pool or not

9. Gym: A binary attribute representing the presence of a gym facility at the hotel or not

10. Tennis Court: A binary attribute indicating whether the hotel provides tennis facilities or not

11 .Spa :A binary attribute indicating the presence of a spa at the hotel or not

12. Casino: A binary attribute representing whether the hotel has a casino or not

13. Free Internet: A binary attribute indicating if the hotel offers free internet access.

14. Hotel Name: Represents the name of the hotel being reviewed.

15. Hotel Stars: Reflects the classification of the hotel, often on a scale from 1 to 5 stars, with 1 star indicating basic accommodations and 5 stars representing luxurious facilities.

16. Number of Rooms: Specifies the total number of rooms available at the hotel.

17. User Continent: Represents the continent from which the reviewer originates.

18. Member Years: Indicates the number of years that the reviewer has been a member of TripAdvisor.

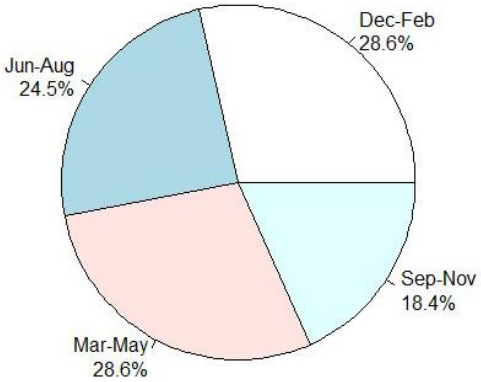
19. Review Month: Denotes the specific month in which the reviewer wrote the review.

20. Review Weekday: Represents the day of the week when the reviewer wrote the review.

These attributes collectively provide a comprehensive view of reviewer demographics, reviewing behavior, opinions about hotels, and various hotel features and amenities, along with temporal and geographical context for the reviews.

Plotting Methods:

[Figure 1: Pie chart represents period of stay]



Pie chart :

We took a sample of 50 respondents in the ' Period of Stay ' category, and the results are represented in a pie chart. The breakdown of Period of Stay is as follows:

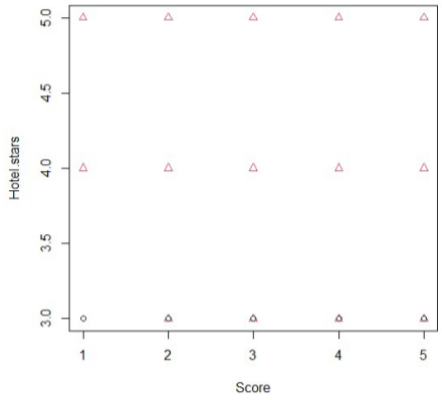
In this sample, the majority was staying from Dec-Feb

- Dec-Feb: 14 respondents (28.6 %)
- Sep-Nov: 9 respondents (18.4%)
- Mar-May: 14 respondents (28.6 %)
- Jun-Aug-May: 12 respondents (24.4 %)

In conclusion we observed distinct trends in reviewers choices of stay periods. Approximately 28.6% of reviewers favored staying in Las Vegas during the winter months, specifically from December to February. This preference for winter occupancy may be influenced by the appeal of holiday vacations, milder weather compared to the scorching summer,comprising 28.6% of reviewers, opted to visit during the

springtime, from March to May, possibly drawn by the pleasant weather, outdoor activities, and spring events in Las Vegas. Conversely, about 24.4% of reviewers chose the period from June to August, highlighting a preference for the vibrant summer atmosphere and outdoor pool experiences. Lastly, 18.4% of reviewers selected the fall months from September to November, indicating an attraction to the autumn ambiance and potential seasonal events.

[Figure 2: Scatter plot represents score of hotels]



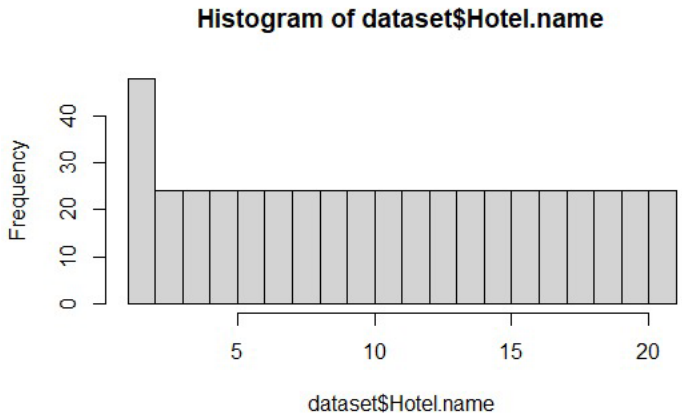
scatter plot:

Analyzing a scatter plot that compares hotels based on the number of stars and their scores reveals fascinating insights into the hospitality industry. The y-axis of the scatter plot represents the number of stars awarded to each hotel, ranging from one to five stars. The x-axis displays the hotels' scores or ratings, reflecting their overall quality and guest satisfaction. In this visual representation, we observe a clear relationship between a hotel's star rating and the likelihood of it having a swimming pool, with triangles denoting the presence of a swimming pool and circles indicating its absence.

One striking observation is that as the number of stars increases, so does the probability of a hotel featuring a swimming pool. This trend aligns

with the expectations of many travelers who associate higher star ratings with enhanced amenities. Hotels with four or five stars predominantly feature triangles, indicating a strong correlation between luxury and pool availability. As the star rating decreases to three and below, we start to see more circles, suggesting that budget and lower-rated hotels are less likely to have swimming pools.

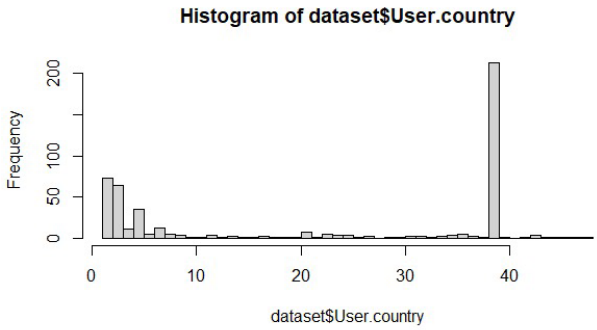
[Figure 3: Histogram represents the frequency of hotels name]



Histogram:

After analyzing this histogram of hotel visitors, with the x-axis representing hotel names and the y-axis indicating the frequency of visitors, this graphical analysis underscores the diverse popularity levels among hotels. "Hotel 1" takes the lead in terms of visitor frequency, while the remaining hotels show a comparable level. By understanding what sets "Hotel 1" apart and considering strategies for the other hotels to distinguish themselves, the industry can work toward optimizing its appeal to a broader range of visitors and achieving a more balanced distribution of guests across its various hotel

[Figure 4: histogram represents the frequency Nationality of hotel visitors]



Analyzing a histogram depicting hotel visitors in Las Vegas, where the x-axis symbolizes different countries represented by numbers and the y-axis represents the count of repeat visitors from each country

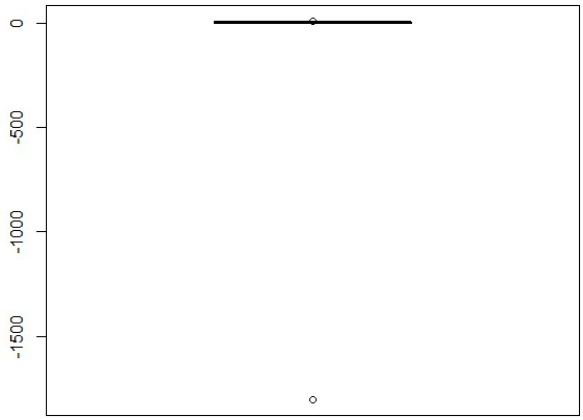
Notably, the most frequently repeated number is 39, corresponding to the USA. This observation underscores the significance of the domestic market, emphasizing that American visitors form a substantial portion of Las Vegas' hospitality clientele. The next most prevalent numbers are 1, 2, and 4, signifying Saudi Arabia, the United Kingdom, and India, respectively. This data indicates a considerable influx of visitors from .these countries

Hotels in Las Vegas can strategically leverage this information to enhance their guest experience and capture the attention of repeat visitors from these countries. Recognizing the cultural diversity and preferences of guests from Saudi Arabia, the United Kingdom, and India, hotels can tailor their services, amenities, and marketing efforts to better accommodate and appeal to these specific demographics. This may include offering cuisine, entertainment, and activities that resonate with the cultural .backgrounds of these guests

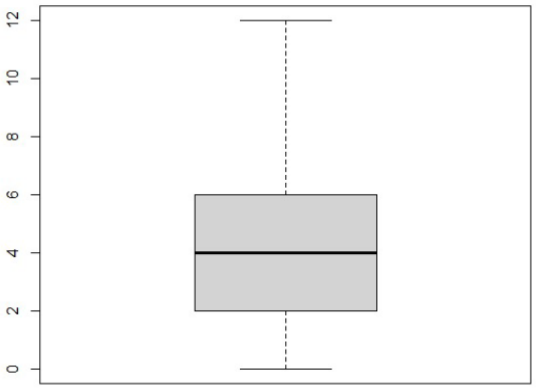
Preprocessing:

Outliers

[Figure 5: boxplot before removing outliers]



[Figure 6: boxplot after removing outliers]



Analyzing a boxplot (1) of hotel member years, with outliers represented by circles, reveals significant insights into the distribution of this data. In this case, there are two distinct outliers, one with a value of 13 years and another at an implausible -1806 years. These outliers have a notable .impact on the distribution, skewing the overall perception of the data

To obtain a more accurate representation of the typical member years duration, it is often considered necessary to deal with these outliers. One common approach is to remove the rows of these outliers. By doing so

boxplot (2), you can better understand the central tendency and spread of the majority of the data, which may provide a more realistic view of visitor membership duration in the hotels