ITCS-6100 Big Data for Computational Advantage

Group -18

Project Deliverable- 1

1. Team

a) Team Members

Manasa Avula - 801307493
Nikhita Sai Boyidapu - 801327682
Srikar Chamarthy - 801317299
Rachana Gullipalli - 801311637
Aravind Pabbisetty - 801274519

b) Communication Plan to include project artifact repository.

- The group meetings will be held weekly twice virtually on Google Meet. Every group member agrees to attend the meetings on time. These meetings are conducted based on the project deliverables. Further group members can also discuss and choose a time if an in-person meeting is required in some cases.
- Members agreed to respect one another and share and discuss their ideas.
- Members can share their findings with others through Emails, Group Shared Folder

Folder Link:
https://drive.google.com/drive/folders/1ObKCKcFXyW_kZzUwfqe0GMhXDjKxhRFT?usp=share_link

GitHub repository URL:

https://github.com/aravindpabbisetty/BigDataGroup18

2. Selection of data to analyze from Kaggle:

Dataset Link:
https://www.kaggle.com/datasets/jasfre/gcc-cyclistic-case-study-present-report-prompt

Dataset Size: 3 GB

The chosen dataset is of a bike sharing company based in chicago called Cyclistic. The dataset consists of information of all the rides taken using different types of bikes by various citizens recently during the period of 12/2021 to 11/2022. The data is stored in CSV files where there is an individual CSV file present for trips taken each month.

The dataset consists of attributes such as

- **ride_id:** It is the unique value assigned to each ride taken. It is used to identify a ride uniquely. Ride_id is the primary key.
- **rideable_type:** rideable_type indicates the type of bike used in the ride such as electric_bike or classic_bike.
- **started_at_date:** Date at which the ride got started.
- **started_at_time:** Time of the day the ride got started.
- **ended_at_date:** Date the ride has ended.
- **ended_at_time:** It indicates the ride ending time.
- **time_of_ride:** Total duration taken for the entire ride
- **start_station_name:** It consists of the name of the station where the ride got started.

- **end_station _name:** It consists of the name of the station where the ride has ended.
- **start_lat:** It denotes the latitude of the starting point of the ride.
- **start_lng:** It denotes the longitude of the starting point of the ride.
- **end_lat:** Latitude of the ride end point.
- **end_lng:** Longitude of the ride end point.
- **member_casual:** This attribute describes the membership of the customer such as casual_member or member.

3. Business Problem or Opportunity, Domain Knowledge

The Cyclistic Bike Share dataset provides us with valuable information about the usage of bikes in the bike-sharing industry. A potential business problem or opportunity that this dataset could be used to address is the optimization of Cyclistic's marketing strategy to attract more riders and increase revenue. There are several factors that can affect bike ridership, including weather conditions, day of the week, and time of day. By analyzing the data, we could identify patterns and trends in rider behavior and preferences, such as the most popular start and end stations, peak riding times, and trip lengths. We can also find which type of bike is mostly used, electric, casual bike or docked bike. Analyzing these factors can help us identify patterns in the dataset and understand how they impact bike usage. For example, we might find that bike usage is higher on weekdays than weekends or the usage is higher in the morning than evening etc. This kind of analysis can help bike-sharing companies to make more bikes available during those times.

To conduct our analysis, we will be using AWS S3 to store the dataset's CSV files, Amazon Quicksight to create interactive visualizations, and Amazon Sagemaker for data preparation and analysis. By leveraging these powerful AWS tools, we can efficiently extract meaningful insights from the Cyclistic Bike Share dataset to drive business growth and enhance the overall bike-sharing experience.

4. Research Objectives and Questions:

1) Which bike type is most used among the casual ,docked and electric bikes?

2) Which weekday is preferable for people to ride the bike?

3) What is the peak time during the weekday that people use the bike?

4) What types of bikes are preferred by different customer types?

5) What is the average time of ride for each bike type?

6) How does the demand for bike rides vary across different months and days of the week based on start date and time?

7) Is there a difference in the duration of rides between members and casual riders?

8) What will be the rate of customers preferring electric bikes in the next 5 years?

9) What will be the most demandable locations for the next 3 years?