Case Study: How Does a Bike-Share Navigate Speedy Success?

Andres Felipe Gomez Camelo – Junior Data Analyst

Curso: Google Data Analytics Capstone

Fecha: 27/08/2025

1. Introduction

Cyclistic is a bike-share company that believes its future success will depend on growing the number of annual memberships rather than relying on casual riders. In this project, we aim to evaluate this hypothesis by analyzing how these two groups use Cyclistic's services differently. The goal is to generate insights that can help stakeholders design effective strategies to increase memberships and engage casual riders through upcoming marketing campaigns.

Main question: How do annual members and casual riders use the bikes differently?

Cyclistic's customers fall into three main categories:

- Single-ride pass users
- Full-day pass users
- Annual members

The main goal of this analysis is to understand how these groups use Cyclistic's services differently and to identify patterns in their behavior. These insights will help determine strategies to convert casual riders into loyal annual members.

2. Ask (Busniess task)

The business task is to analyze patterns, correlations, and usage behaviors in the data from Cyclistic's three types of customers. The goal is to generate insights that can guide marketing strategies aimed at converting casual riders into annual members.

Wee need to keep in mind these main questions:

How do annual members and casual riders use Cyclistic bikes differently?

Why would casual riders buy Cyclistic annual memberships?

How can Cyclistic use digital media to influence casual riders to become members?

Key Stakeholders :
- Lily Moreno (Directora de Marketing)
- Equipo de marketing analytics
- Equipo ejecutivo de Cyclistic

3. Prepare (Data)

The dataset was obtained from the official Divvy public data repository (https://divvy-tripdata.s3.amazonaws.com). For this project, the most recent twelve months of data available at the time of analysis were used, covering the period from **August 2024 through July 2025**. Each month's data is provided as a compressed .zip file containing a .csv with all trips completed during that month.

The dataset was evaluated using the ROCCC framework to confirm its reliability as a source. The data is **reliable** because it is published regularly by Motivate International Inc. as part of Chicago's official bike-share system. It is **original**, since it comes directly from the company operating the program. The dataset is **comprehensive**, covering all trips during the selected twelve-month period. It is also **current**, as the files are updated monthly and the most recent twelve months were used. Finally, it is **cited**, given that it is openly available through Divvy's official public data repository, with appropriate licensing and attribution.

Type of data

The data contains detailed **transactional records** of individual bike trips, including variables such as:

- **ride_id** (unique identifier for each trip)
- **rideable_type** (bike type: classic, docked, electric)
- **started_at** and **ended_at** (timestamps of the trip)
- **start_station_id / name** and **end_station_id / name**
- **member_casual** (user type: annual member or casual rider)

These variables enable descriptive and comparative analysis of usage patterns between different customer segments.

Notes about data

- The dataset is **public** and was made available by **Motivate International Inc.**, under an open data license.
- Personally identifiable information (PII) is **not included**, ensuring data privacy and compliance.

Potential limitations include:

- Possible **inaccuracies in GPS station records** (occasionally trips may show missing or null station data).
- Seasonal effects may influence the number of trips (e.g., higher activity in summer months).

- Only **historical trip data** is available; no additional demographic details (age, gender, income) are provided.

During the initial review of the monthly CSV files we identified several data-quality issues that should be considered before analysis:

- **Date/time formatting issues:** The started_at and ended_at fields contain mixed formats across files — some records use 12-hour AM/PM notation while others include milliseconds (e.g., 2024-08-02 13:35:14.403). To ensure consistency and compatibility with SQL time types, timestamps were standardized to the 24-hour format YYYY-MM-DD HH:MM:SS (milliseconds removed) during preprocessing.
- **Missing station information:** A subset of trips has missing values in start_station_name, start_station_id, end_station_name, or end_station_id. These nulls reduce the completeness of location-based analyses and will be documented and handled during the cleaning stage (e.g., flagging or excluding affected records, or applying appropriate imputation where justified).
- **Station ID inconsistencies across months:** We observed that station identifiers are not always stable month-to-month — the same start_station_name/end_station_name can appear with different station_id values in different monthly files, likely due to system updates or station re-numbering. This requires constructing a canonical station mapping (using station names and other metadata) to normalize IDs before any multi-month aggregation or join.

These issues do not invalidate the dataset for the planned analyses but must be addressed in the Process (data cleaning) stage to ensure accurate time-series and location-level results.

4. Process (Data clean)

Process

During the processing phase, several transformations and validations were carried out to ensure that the data was in optimal condition for subsequent analysis.

Tools Used

For this stage, Microsoft Excel was mainly used (for initial cleaning, creation of derived columns, and validations), along with MySQL (for database structuring and final data loading).

Steps Performed

Separation of dates and times

The original columns contained both date and time combined.

We separated these into two distinct fields (date and time) for both the start and end of each trip, thus facilitating subsequent calculations.

Calculation of trip duration
A column named len_trip was created to represent trip duration in the hh:mm:ss format. Additionally, a column named len_trip_min was generated, transforming the duration into minutes as a numeric value, which simplifies calculations and later analysis.

Determination of the day of the week
Based on the start date, a column was created to indicate the day of the week in both numeric and text format (e.g., 1 = Sunday, 2 = Monday, etc.), enabling the identification of usage patterns by day.

Calculation of distance traveled
An approximate formula was implemented, based on the initial and final latitude/longitude coordinates, to estimate the distance traveled in degrees.

Trip validation
Conditions were established to identify valid and invalid trips:

- A trip is considered valid if it lasts more than 2 minutes or covers more than approximately 200 meters.

- Trips below these thresholds were classified as invalid, preventing bias from anomalous records or system capture errors.

Handling missing values and duplicates

- Empty final coordinate values were replaced with a neutral value (0) to preserve dataset integrity.

- Duplicate records were checked and removed, ensuring consistency.

Database structuring
A table was designed in MySQL with the appropriate data types (e.g., DATETIME for dates, DECIMAL for trip duration in minutes, VARCHAR for IDs).
Finally, the 12 monthly files were loaded into the table, consolidating all information into a single repository for queries and analysis.

Data quality verification

- Outlier cases were manually reviewed (e.g., trips crossing midnight or with unusually long durations).

- All Excel formulas used were documented, including:

    - **Date extraction:** =FECHA(AÑO(C2), MES(C2), DIA(C2))

- **Trip duration calculation:** =SI(F2<D2,(F2+1)-D2,F2-D2)
- **Conversion to minutes:** =HORA(P2)*60 + MINUTO(P2) + SEGUNDO(P2)/60
- **Day of the week:** =DIASEM(C2,1) and conversion to text
- **Distance traveled:** =RAIZ((M2-K2)^2+(N2-L2)^2)
- **Trip validation:** =SI(O(Q2>=2, Y(Q2>=0.03, S2>=0.002)), "Valid trip", "Invalid trip")

Conclusion

Through this process, the data was ensured to be clean, standardized, and ready for analysis. In addition, each step was documented, allowing for both traceability and reproducibility of the procedure in future projects.

5. Analyze (Explore and Results)

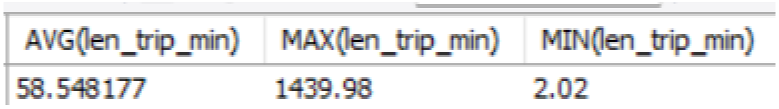
The analysis phase aimed to generate a clear understanding of trip patterns within the Cyclistic dataset, focusing on validating trips, understanding trip durations, and identifying potential anomalies. The SQL queries were designed to follow a structured exploration: first ensuring data integrity, then calculating basic descriptive statistics, and finally evaluating trip quality.

Querys

This query calculates basic descriptive statistics for the duration of bike trips, measured in minutes (len_trip_min). Specifically, it computes the **average**, **maximum**, and **minimum** trip duration for all trips that lasted more than 2 minutes. The WHERE clause ensures that extremely short trips, which are likely anomalies or system errors, are excluded from the analysis.

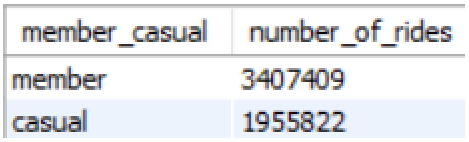
These statistics provide an initial understanding of how long riders typically use the bikes, helping to identify usage patterns and differences between casual riders and annual members.

```
SELECT
  AVG(len_trip_min),
  MAX(len_trip_min),
  MIN(len_trip_min)
FROM trips
WHERE len_trip_min >2;
```



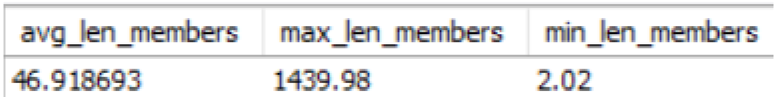
AVG(len_trip_min)	MAX(len_trip_min)	MIN(len_trip_min)
58.548177	1439.98	2.02

```
select member_casual, count(*) as number_of_rides
from trips
where len_trip_min >2
group by member_casual;
```



member_casual	number_of_rides
member	3407409
casual	1955822

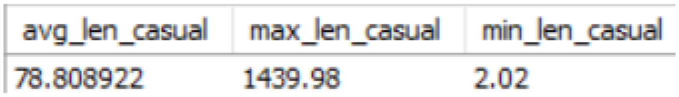
```
select avg(len_trip_min) as avg_len_members,
max(len_trip_min) as max_len_members,
min(len_trip_min) as min_len_members
from trips
where member_casual='member' and len_trip_min>2;
```



avg_len_members	max_len_members	min_len_members
46.918693	1439.98	2.02

```
select avg(len_trip_min) as avg_len_casual,
max(len_trip_min) as max_len_casual,
min(len_trip_min) as min_len_casual
from trips

where member_casual='casual' and len_trip_min>2;
```



avg_len_casual	max_len_casual	min_len_casual
78.808922	1439.98	2.02

During the analysis phase, we created several summary tables in SQL in order to better understand the patterns and behaviors within the dataset. These tables were designed to answer key business questions related to user types, time trends, and trip characteristics.

1. **Total trips by user type (member vs. casual):**
   This table was created to compare the number of trips completed by members against those completed by casual riders. The purpose was to identify usage patterns between frequent subscribers and occasional users.

   Command SQL:

   ```
   CREATE OR REPLACE VIEW vw_avg_by_type AS
   SELECT member_casual,
         AVG(len_trip_min) AS avg_duration_min,
         COUNT(*) AS total_rides,
         ROUND(STDDEV_POP(len_trip_min),2) AS sd_duration
   FROM trips
   WHERE  len_trip_min>2
   GROUP BY member_casual;
   ```

2. **Number of trips per month and year:**
   This table aggregated the trips by month and year to reveal seasonal and long-term trends. It allowed us to identify peak usage periods and assess year-over-year growth or decline in ridership.

   Command SQL:

   ```
   CREATE OR REPLACE VIEW view_monthly_trips AS
   SELECT
      member_casual,
      YEAR(start_date) AS year_date,
      MONTH(start_date) AS month_date,
      COUNT(*) AS trips,
      AVG(len_trip_min) AS avg_duration_min
   FROM trips
   WHERE len_trip_min > 2
     AND start_date IS NOT NULL
   GROUP BY member_casual, YEAR(start_date), MONTH(start_date)
   ORDER BY member_casual, YEAR(start_date), MONTH(start_date);
   ```

3. **Number of trips per hour:**
   By grouping trips by the hour of the day, we identified peak travel times. This analysis was essential to understand commuting patterns and the demand distribution throughout the day.

   Command SQL:

   ```
   CREATE OR REPLACE VIEW vw_hourly_by_type AS
   SELECT member_casual,
         HOUR(start_time) AS hour,
         COUNT(*) AS trips,
         AVG(len_trip_min) AS avg_duration_min
   FROM trips
   WHERE len_trip_min > 2
   GROUP BY member_casual, hour
   ORDER BY hour;
   ```

4. **Number of trips per day:**
   This table helped us analyze daily trends, including weekdays versus weekends. It provided insights into how rider behavior differs across the week and highlighted specific days with higher or lower activity.

   Command SQL:

   ```
   CREATE OR REPLACE VIEW vw_avg_by_type_day AS
   SELECT member_casual,
       day_of_week,
       AVG(len_trip_min) AS avg_duration_min,
       COUNT(*) AS total_rides
   FROM trips
   WHERE len_trip_min > 2
   GROUP BY member_casual, day_of_week
   ORDER BY day_of_week;
   ```

5. **Percentage of trips shorter than 5 minutes:**
   This table calculated the proportion of very short trips, which can indicate potential data anomalies, misuse of the service, or short-distance preferences. Understanding this metric is key for evaluating ride quality and operational efficiency.

   Command SQL:

   ```
   CREATE OR REPLACE VIEW vw_pct_short_trips AS
   SELECT member_casual,
       SUM(CASE WHEN len_trip_min < 5 THEN 1 ELSE 0 END) / COUNT(*) * 100 AS
   pct_short
   FROM trips
   WHERE len_trip_min >2
   GROUP BY member_casual;
   ```

Each of these tables was generated using SQL aggregation functions such as COUNT(), GROUP BY, and conditional filtering with WHERE clauses. Together, they provided the foundation for building dashboards and visualizations, helping us transform raw data into actionable insights.

6. Share (Viz and Insights)

Figure 1: Total Rides by User Type

Insight: The analysis shows that members generate a significantly higher number of rides compared to casual users. This indicates that members rely more on the service for frequent and consistent transportation needs.

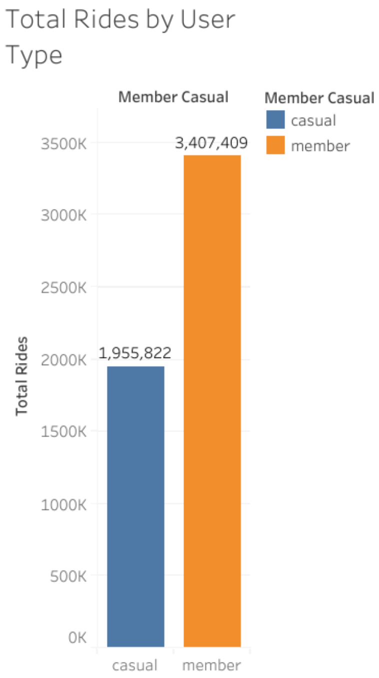
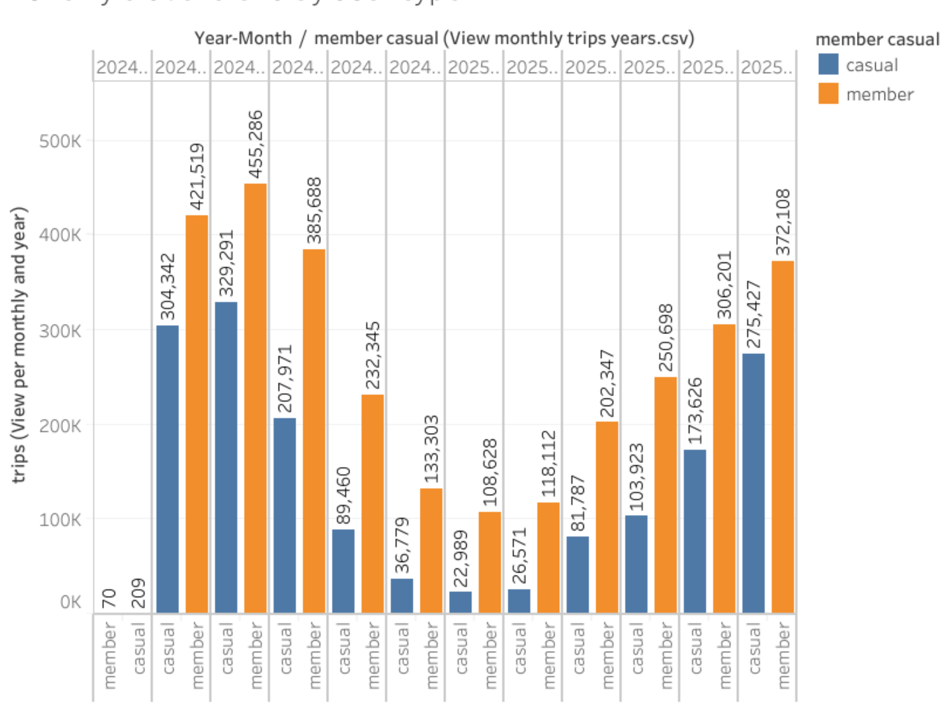
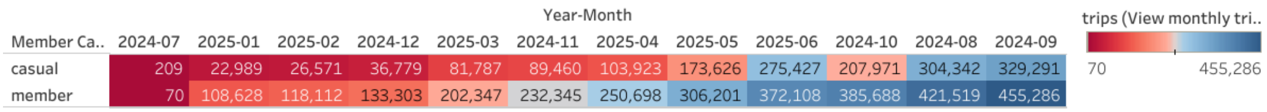


Figure 2: Number of Rides by Month-Year

Insight: The monthly distribution of rides highlights seasonal trends, with a clear increase during summer months. This suggests that weather and seasonal factors strongly influence ridership patterns, particularly for casual users.



Monthly travel trend by user type



Monthly Ride Volume Heatmap

	Year-Month											
Member Ca..	2024-07	2025-01	2025-02	2024-12	2025-03	2024-11	2025-04	2025-05	2025-06	2024-10	2024-08	2024-09
casual	209	22,989	26,571	36,779	81,787	89,460	103,923	173,626	275,427	207,971	304,342	329,291
member	70	108,628	118,112	133,303	202,347	232,345	250,698	306,201	372,108	385,688	421,519	455,286

Figure 3: Number of Rides by Hour of the Day

Insight: The hourly distribution reveals that members mostly ride during commuting hours (morning and late afternoon), while casual users show higher activity in midday and evening hours, likely linked to leisure and recreational purposes.

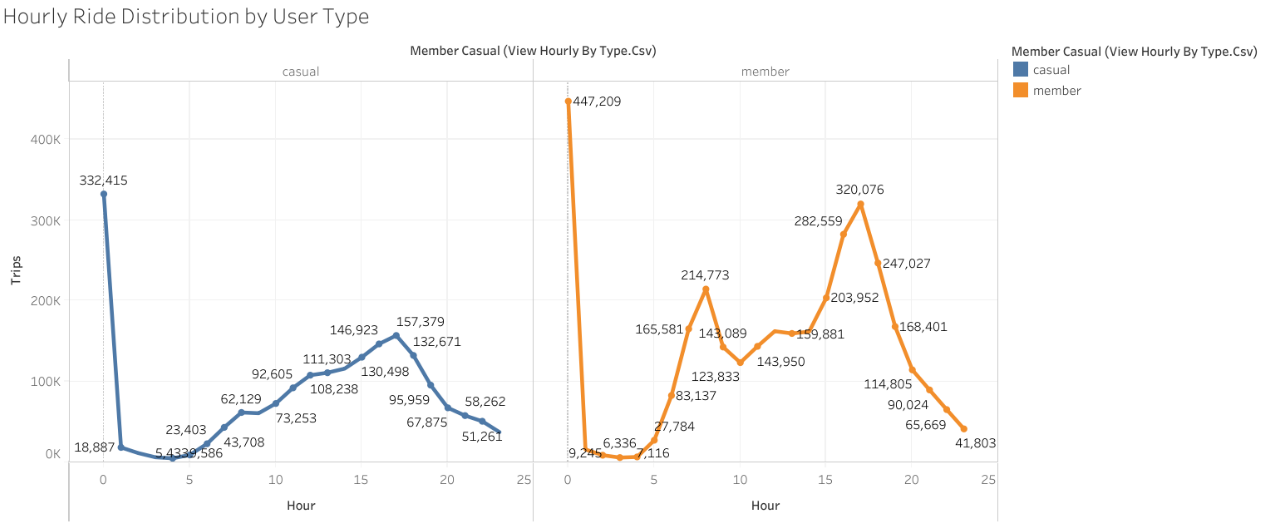


Figure 4: Number of Rides by Day of the Week

Insight: Members maintain a consistent pattern throughout the workweek, while casual users exhibit a noticeable peak during weekends. This reinforces the idea that members use the service for commuting, while casual users associate it more with leisure.

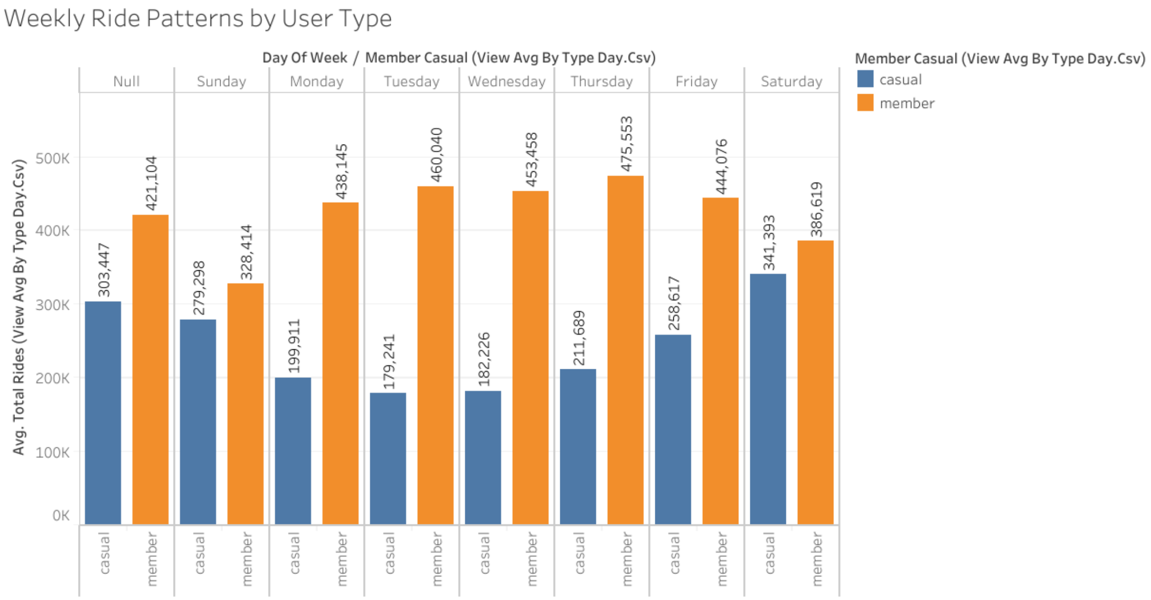
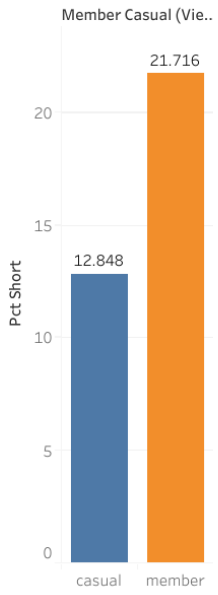


Figure 5: Percentage of Rides Shorter than 5 Minutes

Insight: A relatively small percentage of trips last less than 5 minutes, which may suggest user testing, errors in ride usage, or very short-distance trips. Understanding these cases is important to identify potential system inefficiencies or misuse.

Percentage of Short Rides (<5 min) by User Type



7. Act (Recomendations)

Recommendations

Based on the analysis, the following strategic recommendations are proposed to help Cyclistic convert casual riders into annual members and optimize marketing initiatives:

1. **Leverage peak seasonal demand (August–September 2024)**

   - Insight: Casual riders recorded the highest volume of trips during August and September 2024, representing a key period of demand.

   - Recommendation: Cyclistic should launch targeted membership promotions during these months (e.g., discounted first-month membership or bundled weekend offers). Capturing riders at their peak usage moment increases the likelihood of conversion.

2. **Capitalize on weekend and Saturday patterns**

- Insight: Casual riders are significantly more active on Saturdays compared to weekdays. This suggests leisure and tourism-driven usage.
- Recommendation: Introduce "Weekend Plus" incentives for members (e.g., extra ride credits or bonus minutes on Saturdays). By positioning memberships as not only practical but also rewarding for weekend activities, casual users may see more value in upgrading.

3. **Promote membership for longer rides**
    - Insight: Only 12.81% of casual trips lasted less than 5 minutes, meaning the majority of trips were longer. This indicates that casual riders are not just testing the service but actively using it, possibly as tourists or for leisure.
    - Recommendation: Offer discounts or bonus minutes for rides over 10 minutes as part of the membership plan. This directly appeals to casual riders who typically engage in longer trips, highlighting the cost-effectiveness of becoming a member.

4. **Explore late-night riding behaviors**
    - Insight: An unusual spike in casual rides was observed around midnight. While this may be linked to nightlife or tourism, the behavior requires further exploration.
    - Recommendation: Cyclistic could test targeted digital campaigns (e.g., ads on social platforms used late at night) to capture this audience. However, this pattern should be validated further before committing significant resources.

8. Final Reflection

Through this project, I gained a deeper understanding of the critical role of exploratory data analysis (EDA) as a first step to assess the structure, integrity, and limitations of the dataset. I strengthened my skills in data cleaning by applying different practices in Excel and ensuring that the data was formatted appropriately for integration into a database management system such as MySQL. This process enhanced my proficiency in writing SQL queries to extract, transform, and interpret relevant insights.

Additionally, I expanded my experience with Tableau by designing dashboards and visualizations that effectively communicate findings in a clear and actionable way. Beyond the technical aspects, this project allowed me to put into practice the concepts and methodologies covered in the Google Data Analytics course, reinforcing my readiness to apply them in real-world scenarios.

In terms of limitations, the analysis was constrained by the scope and quality of the available dataset. With more time or access to additional data sources, I would focus on conducting a more robust statistical analysis, exploring predictive modeling, and enriching the dashboards with deeper context to drive stronger decision-making insights.

Overall, this project not only solidified my technical skills but also prepared me to approach data challenges with a structured, industry-relevant mindset.