Google Data Analytics Capstone Project: Cyclistic Case Study

Scenario

I am a junior data analyst working on the marketing analyst team at Cyclistic. The director of marketing believes the company's future success depends on maximizing the number of annual memberships. Therefore, my team wants to understand how casual riders and annual members use Cyclistic bikes differently. From these insights, my team will design a new marketing strategy to convert casual riders into annual members. But first, Cyclistic executives must approve our recommendations, so they must be backed up with compelling data insights and professional data visualizations.

About the company

In 2016, Cyclistic launched a successful bike-share offering. Since then, the program has grown to a fleet of 5,824 bicycles that are geo tracked and locked into a network of 692 stations across Chicago. The bikes can be unlocked from one station and returned to any other station in the system anytime.

Until now, Cyclistic's marketing strategy relied on building general awareness and appealing to broad consumer segments. One approach that helped make these things possible was the flexibility of its pricing plans: single-ride passes, full-day passes, and annual memberships. Customers who purchase single-ride or full-day passes are referred to as casual riders. Customers who purchase annual memberships are Cyclistic members.

Cyclistic's finance analysts have concluded that annual members are much more profitable than casual riders. Although the pricing flexibility helps Cyclistic attract more customers, Moreno (the director of marketing and my manager) believes that maximizing the number of annual members will be key to future growth. Rather than creating a marketing campaign that targets all-new customers, Moreno believes there is a very good chance to convert casual riders into members. She notes that casual riders are already aware of the Cyclistic program and have chosen Cyclistic for their mobility needs.

Moreno has set a clear goal: Design marketing strategies aimed at converting casual riders into annual members. To do that, however, the marketing analyst team needs to better understand how annual members and casual riders differ, why casual riders would buy a membership, and how digital media could affect their marketing tactics. Moreno and her team are interested in analysing the Cyclistic historical bike trip data to identify trends.

<u>Ask Stage:</u>

Guiding questions:

- What is the problem am I trying to solve?
- ➢ I am trying to solve how do annual members and casual riders use Cyclistic bikes differently.

- How can my insights drive business decisions?
- ➢ By understanding the user behaviour, marketing team can craft the marketing strategies to convert the casual riders to annual members.

Key tasks:

- Identify the business task
  - Design marketing strategies aimed at converting casual riders into annual members.

- Consider key stakeholders:
  - Lily Moreno: The director of marketing and my manager.
  - Cyclistic marketing analytics team: A team of data analysts who are responsible for collecting, analyzing, and reporting data that helps guide Cyclistic marketing strategy.
  - Cyclistic executive team: The notoriously detail-oriented executive team will decide whether to approve the recommended marketing program.

Deliverables:

- A clear statement of the Business Task
  - **“**Design marketing strategies aimed at converting casual riders into annual members.”

Prepare Stage:

Guiding questions:

- Where is my data located?
  - The data is located here. I have used the past 12 months (Jul 2023 – Jun 2024) rides data which is stored across 12 csv files.

- How is the data organized?
  - Data is organized by month and year in csv file formats.

- Are there issues with bias or credibility in this data? Does my data ROCCC?
  - The data used in this analysis is ROCC safe:
    - Reliable
    - Original
    - Comprehensive
    - Current
    - Cited

- How am I addressing licensing, privacy, security, and accessibility?
  - The data has been made available by Motivate International Inc. under this license. Data-privacy issues prohibit me from using riders’ personally identifiable information. This means that I won’t be able to connect pass purchases to credit card numbers to determine if casual riders live in the Cyclistic service area or if they have purchased multiple single passes.

Key tasks:

- Download data and store it appropriately.
  - I have downloaded the 12 csv files and stored in a folder with proper naming conventions. The files are being stored with proper naming conventions.

- Identify how it’s organized.
- ➢ The data has been organised by month and year in csv file formats.
  Each dataset has been organised across 13 columns.

- Determine the credibility of the data.
- ➢ The dataset has a total of 933350 null values in start_station_name and start_station_id column.
  980959 null values in start_station_id and end_station_id.
  7936 null values in end_lat and end_lng column. All though the number of null values are quite significant in station name columns, there are very few null values in the location coordination columns. Through that, station names can be mapped to the null values for most of the null values in station name columns.

Deliverables:

- A description of all data sources used:
- ➢ 12 Months historical ride data of Cyclistic bike share company span across Jul 2023 to Jun 2024.
  The dataset can be downloaded from this link.
  The data has been made available by Motivate International Inc. under this license.
  Each dataset has 13 columns:

Columns	Description	Data Types
ride_id	Unique ride id for each ride	object
rideable_type	3 unique bike types are in use: Classic Bike, Electric Bike and Docked Bike	object
started_at	Start time of trip store in yyyy-mm-dd h:m:s format	object
ended_at	End time of trip store in yyyy-mm-dd h:m:s format	object
start_station_name	Ride starting station name	object
start_station_id	Ride starting station id	object
end_station_name	Ride end station name	object
end_station_id	Ride end station id	object
start_lat	Ride starting station latitude	Float64
start_lng	Ride starting station longitude	float64
end_lat	Ride end station latitude	float64
end_lng	Ride end station longitude	float64
member_casual	2 unique user types: Casual and Member	object

Process Stage:

Guiding questions:

- What tools am I choosing and why?
- ➢ I am using Jupyter Notebook and Python programming language to process, analyse and visualise the data for this project.
  Reason behind to use these, the dataset is very large, consisting over 5.7 million of rows combined. This amount of large data is not possible to handle in spreadsheet applications.

- Have I ensured my data's integrity?
- ➢ Yes, I have ensured my data's integrity by:
  - ▪ Inspecting the date format across all the datasets.
  - ▪ Ensuring not a single data record is missed out while combining the datasets.
  - ▪ Checking if the data types of data fields across the datasets are of same type.

- What steps have I taken to ensure that my data is clean?
- ➢ I have created user defined function to clean the data with python.
- ➢ I have removed the leading and trailing spaces from column headers.

- How can I verify that my data is clean and ready to analyse?
- ➢ I have used various panda methods to check the followings if the data is clean:
  - ▪ Extra spaces and characters
  - ▪ Mismatched data types
  - ▪ Inconsistent date formats
  - ▪ Null data
  - ▪ Business Logic

- Have I documented the cleaning process so I can review and share those results?
- ➢ Yes, I have created a change log file to record the cleaning and manipulation process.

Key tasks:

- Check the data for errors.
- ➢ I have checked the data for errors.

- Choose my tools.
- ➢ I have chosen Python programming language and Jupyter Notebook.

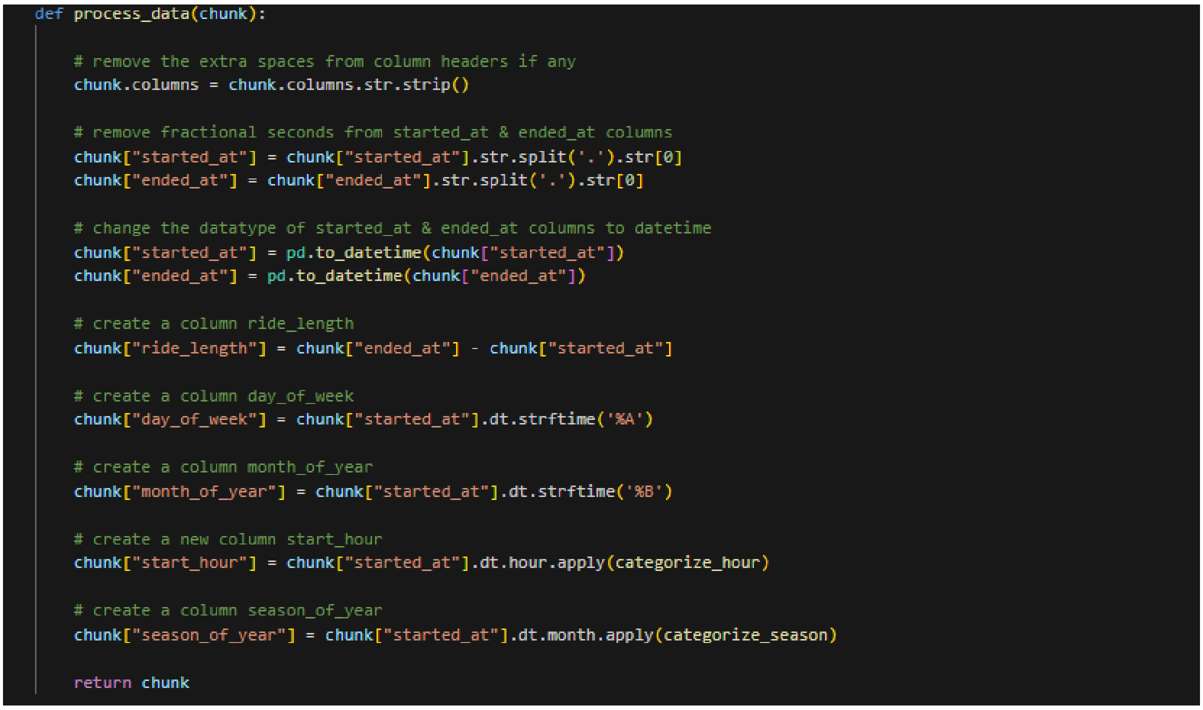
- Transform the data so I can work with it effectively.
- ➢ I have transformed the data by cleaning, converting the data types, creating required additional columns for analysis.

Deliverables:

- Documentation of any cleaning or manipulation of data.
- Changelog:

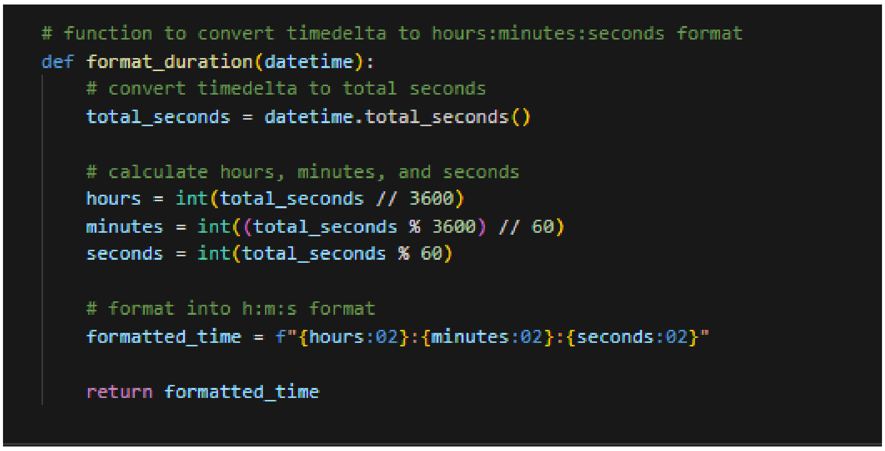
```
# Changelog
This file contains the notable changes to the Case_Study_1 Cyclistic Project

## New
  - Added new columns to the dataset ["ride_length", "day_of_week",
"month_of_year", "start_hour", "season_of_year"]
  - Added new column "ride_len_min" for ease of calculation of average ride
duration

## Changes
  - Removal of possible leading and trailing extra spaces from column header
  - Removal of fractional seconds from Date Time columns ["started_at",
"ended_at"]
  - Changed data type of date time columns from object to datetime format
["started_at",   "ended_at"]
  - Applied formatting to ride_length column to convert days to hours
```

- Created a user-defined function for data cleaning and data manipulation on the dataframes:

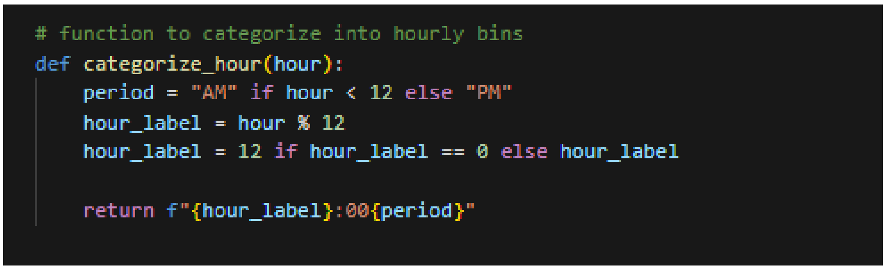


```python
def process_data(chunk):

    # remove the extra spaces from column headers if any
    chunk.columns = chunk.columns.str.strip()

    # remove fractional seconds from started_at & ended_at columns
    chunk["started_at"] = chunk["started_at"].str.split('.').str[0]
    chunk["ended_at"] = chunk["ended_at"].str.split('.').str[0]

    # change the datatype of started_at & ended_at columns to datetime
    chunk["started_at"] = pd.to_datetime(chunk["started_at"])
    chunk["ended_at"] = pd.to_datetime(chunk["ended_at"])

    # create a column ride_length
    chunk["ride_length"] = chunk["ended_at"] - chunk["started_at"]

    # create a column day_of_week
    chunk["day_of_week"] = chunk["started_at"].dt.strftime('%A')

    # create a column month_of_year
    chunk["month_of_year"] = chunk["started_at"].dt.strftime('%B')

    # create a new column start_hour
    chunk["start_hour"] = chunk["started_at"].dt.hour.apply(categorize_hour)

    # create a column season_of_year
    chunk["season_of_year"] = chunk["started_at"].dt.month.apply(categorize_season)

    return chunk
```

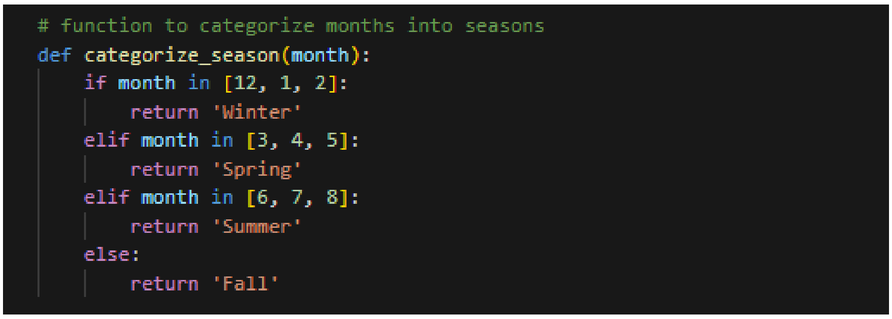
- Created user-defined functions for formatting:



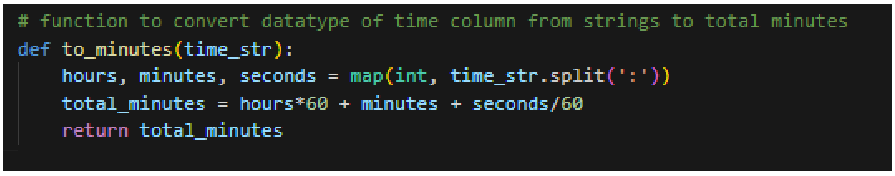
```python
# function to convert timedelta to hours:minutes:seconds format
def format_duration(datetime):
    # convert timedelta to total seconds
    total_seconds = datetime.total_seconds()

    # calculate hours, minutes, and seconds
    hours = int(total_seconds // 3600)
    minutes = int((total_seconds % 3600) // 60)
    seconds = int(total_seconds % 60)

    # format into h:m:s format
    formatted_time = f"{hours:02}:{minutes:02}:{seconds:02}"

    return formatted_time
```



```
# function to categorize into hourly bins
def categorize_hour(hour):
    period = "AM" if hour < 12 else "PM"
    hour_label = hour % 12
    hour_label = 12 if hour_label == 0 else hour_label

    return f"{hour_label}:00{period}"
```



```
# function to categorize months into seasons
def categorize_season(month):
    if month in [12, 1, 2]:
        return 'Winter'
    elif month in [3, 4, 5]:
        return 'Spring'
    elif month in [6, 7, 8]:
        return 'Summer'
    else:
        return 'Fall'
```



```
# function to convert datatype of time column from strings to total minutes
def to_minutes(time_str):
    hours, minutes, seconds = map(int, time_str.split(':'))
    total_minutes = hours*60 + minutes + seconds/60
    return total_minutes
```

Analysis Stage:

Guiding questions:

- How should I organize my data to perform analysis on it?
- ➢ I should sort and filter out the data for each metrics.

- Has my data been properly formatted?
- ➢ Yes, the data has been properly formatted by creating and using user defined functions.

- What surprises did I discover in the data?
- ➢ I've found out that docked bikes, that have been used for highest durations, are rented on Wednesday, not on weekends.

- What trends or relationships did I find in the data?
- ➢ I have found out that:
    - ▪ Annual members use Cyclistic Bikes for work related purposes
    - ▪ Casual users use Cyclistic Bikes for leisure activities.

- How will these insights help answer my business questions?
- ➢ These insights will help to understand the Cyclistic riders behaviour based on that the marketing strategies can be crafted.

Key tasks:

- Aggregate my data so it’s useful and accessible.
- ➢ I have aggregated the data to perform analysis by:
  - Casual Riders
  - Annual Member Riders
  - Months and grouping them into seasons
  - Hours of the day

- Organize and format my data.
- ➢ I have applied specific formats to datetime columns
- ➢ I have filtered out the data based on:
  - Rider types
  - Rideable bike types
  - Seasons of the year

- Perform calculations.
- ➢ I have calculated the followings:
  - Average ride duration of different user types
  - Total number of rides of different user types
  - Popular location to start and end a ride by different user types

- Identify trends and relationships.
- ➢ I have found out several trends such as which type of rider type is renting out bikes at specific hours, rideable bike type preference of each user category, hourly, weekly and seasonal trends of riders etc.

Deliverables:

- A summary of my analysis.
- ➢ 35.7% of total users of Cyclistic company are casual riders, rest are annual members. While casual riders are in less number than the annual members, Casual Riders ride approximately for 2x time the annual members on average.
- ➢ Annual members mostly prefer Classic Bike while casual riders prefer electric bikes over classic bikes. Docked bikes are only being used by casual riders.
- ➢ Casual riders tend to use the Cyclistic bikes for leisure activities while annual members use the bikes for day-to-day essential activities.

<u>Share Stage:</u>

Guiding questions:

- Was I able to answer the question of how annual members and casual riders use Cyclistic bikes differently?
- ➢ Yes, I can answer to the question of how annual members and casual riders use Cyclistic bikes differently.

- What story does my data tell?

➢ The data reveals distinct differences in the behaviour and preferences between casual riders and annual members.

- How do my findings relate to my original question?

➢ The findings of this analysis are directly related to the original question.

- Who is my audience? What is the best way to communicate with them?

➢ My audience is consisting of:

  - Lily Moreno: The director of marketing and my manager.
  - Cyclistic marketing analytics team: A team of data analysts who are responsible for collecting, analyzing, and reporting data that helps guide Cyclistic marketing strategy.
  - Cyclistic executive team: The notoriously detail-oriented executive team will decide whether to approve the recommended marketing program.

➢ The best way to communicate with them is through presentation with proper data visualisations.

- Can data visualization help me share my findings?

➢ Yes, data visualization will help me to share my findings to the audience effectively.

- Is my presentation accessible to my audience?

➢ The presentation will be uploaded to the google drive through which my audience will access the presentation.

Key tasks:

- Determine the best way to share my findings.

➢ The best way to share findings is through presentation with proper data visualisations.

- Create effective data visualizations.

➢ I have created effective presentations with the help of Python libraries such as Matplotlib, Seaborn, Folium.

- Present my findings.

➢ I have presented my findings in a ppt presentation.

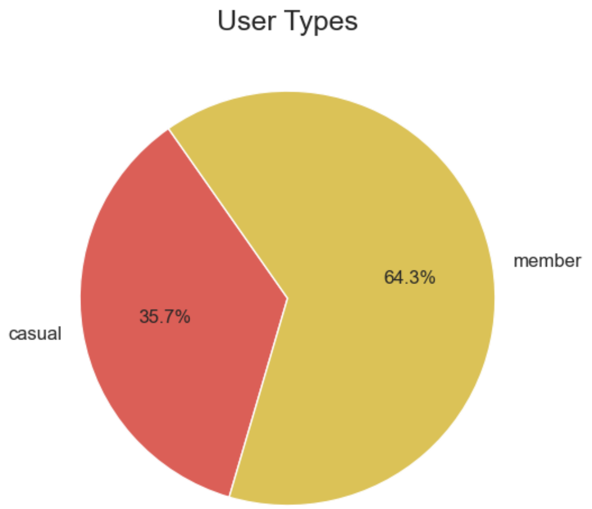
- Ensure my work is accessible.

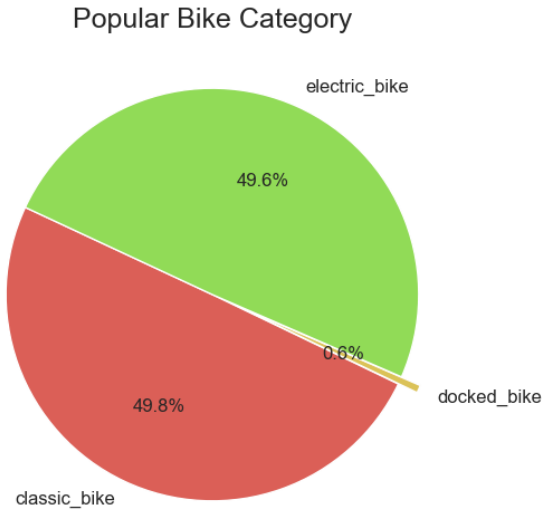
➢ I have ensured my work is accessible by uploading it to my GitHub profile.

Deliverables:

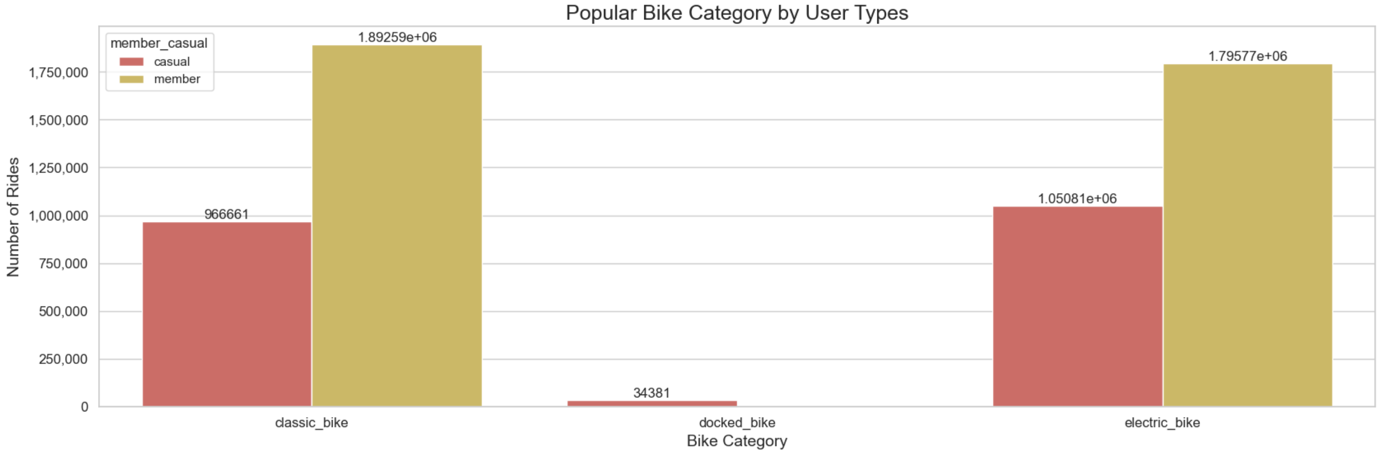
- Supporting visualizations and key findings.

User Demographics:





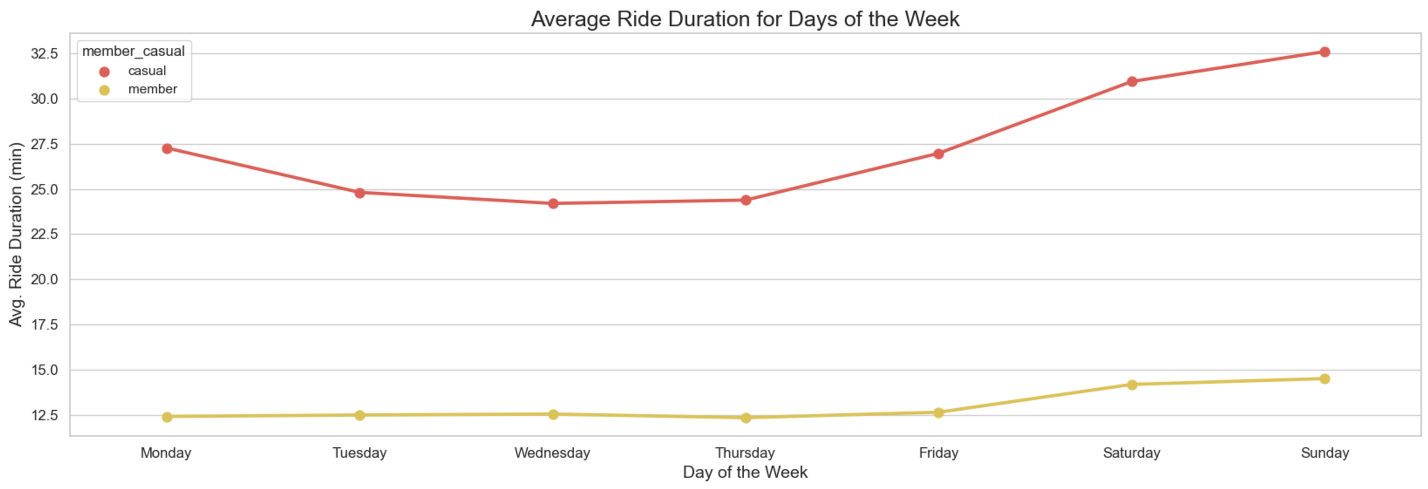
Bike Preferences:

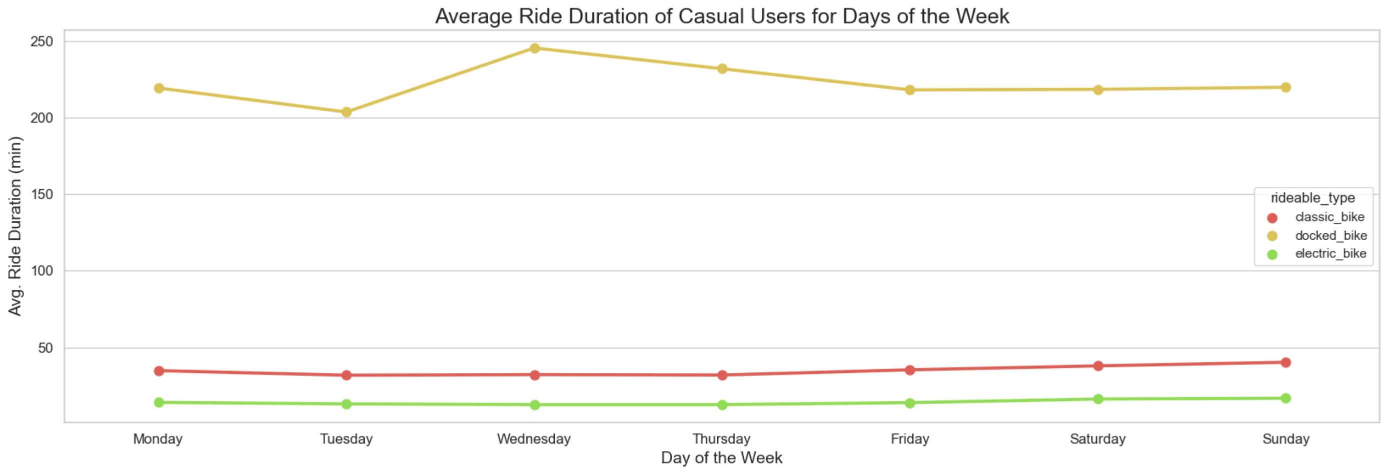


Findings:

- 64.3% of total users of Cyclistic are Annual Members while the rest are Casual members.
- Most popular bike category is Classic Bikes (49.8%) closely followed by Electric Bikes (49.6%). Docked Bikes are being used in very low numbers (0.6%).
- While annual members prefer classic bikes over electric bikes, casual riders prefer electric bikes over classic bikes.
- While Casual members are choosing all three types of bikes for their rides, annual members only prefer Classic and Electric Bikes.

Average Ride Duration:

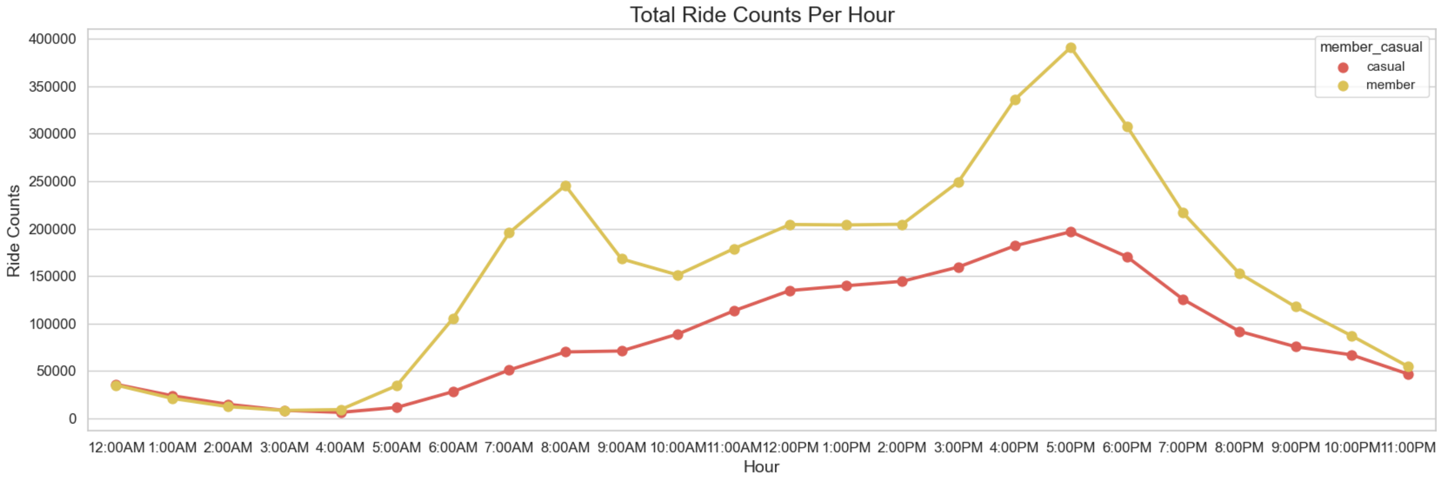


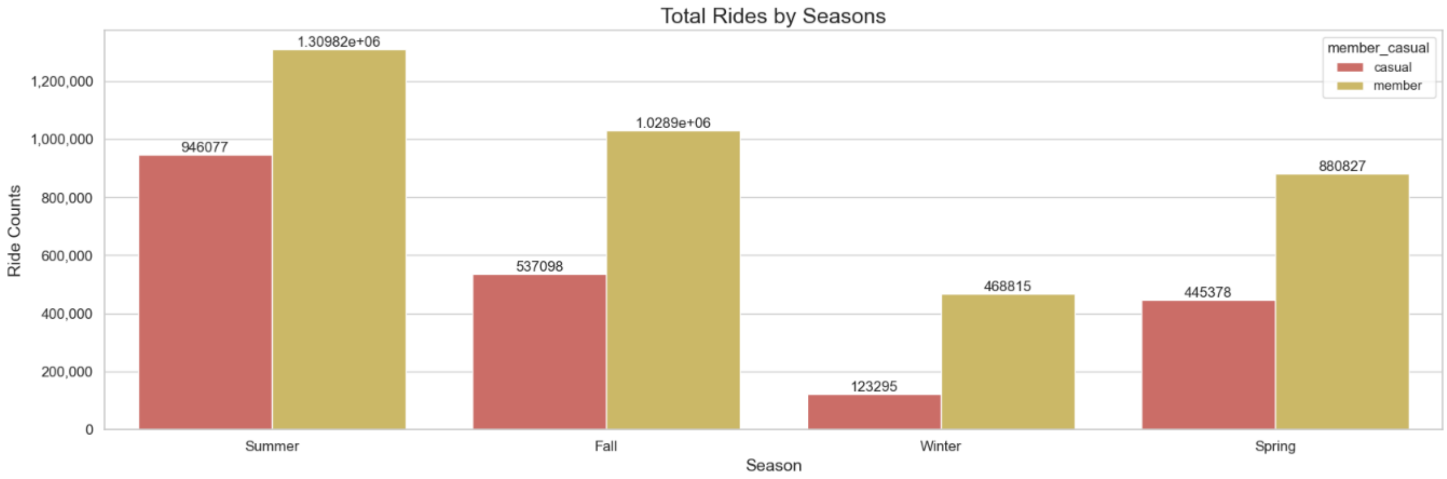


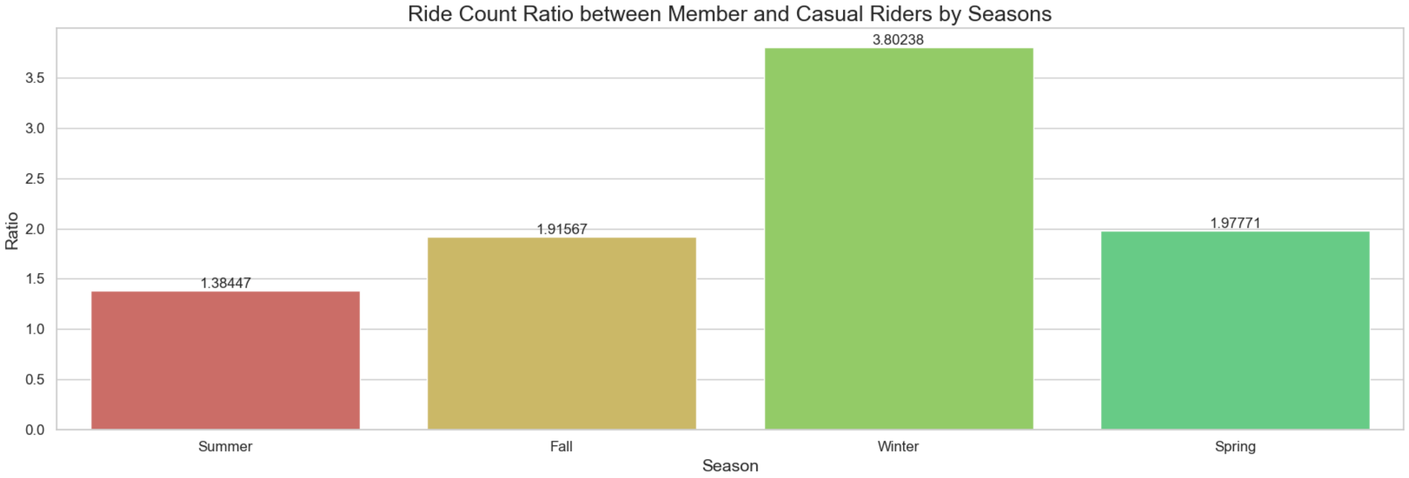
Findings:

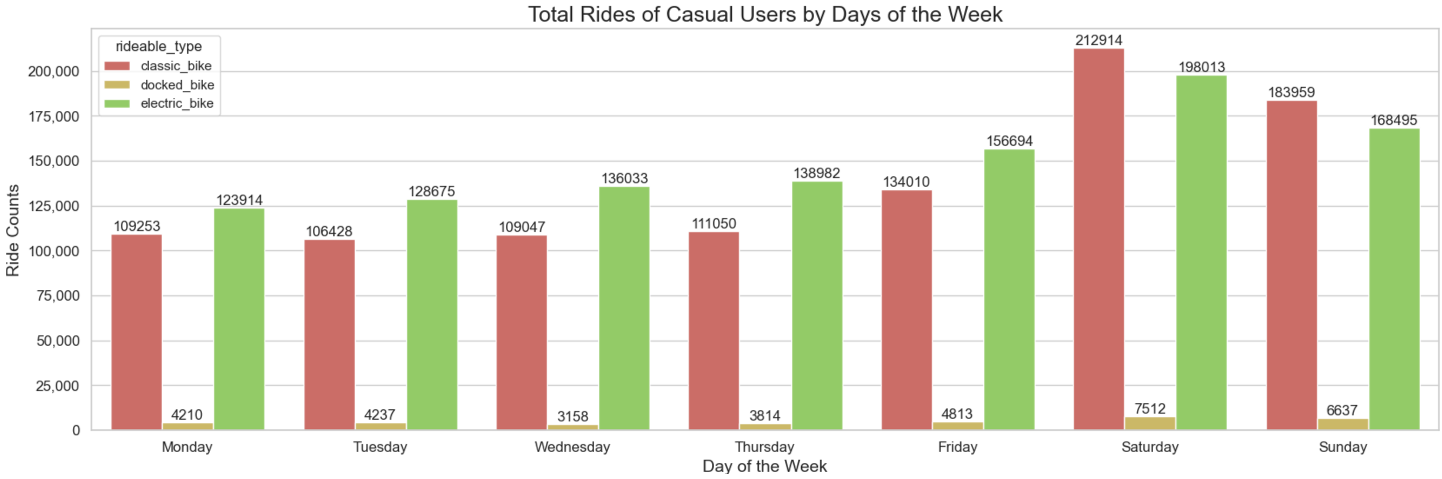
- Average ride duration of annual members remains almost same throughout the weekdays and increases during the weekend.
- Casual members are riding approx. 2x times the annual members.
- There is a significant rise in ride duration of casual members during the weekends.
- Casual riders are riding Classic bikes 2x times the electric bikes and average ride durations for both bikes remain almost same from Tuesday to Thursday and start increasing from Friday to Monday.
- Docked bike rented on Wednesday has significantly high average ride duration.

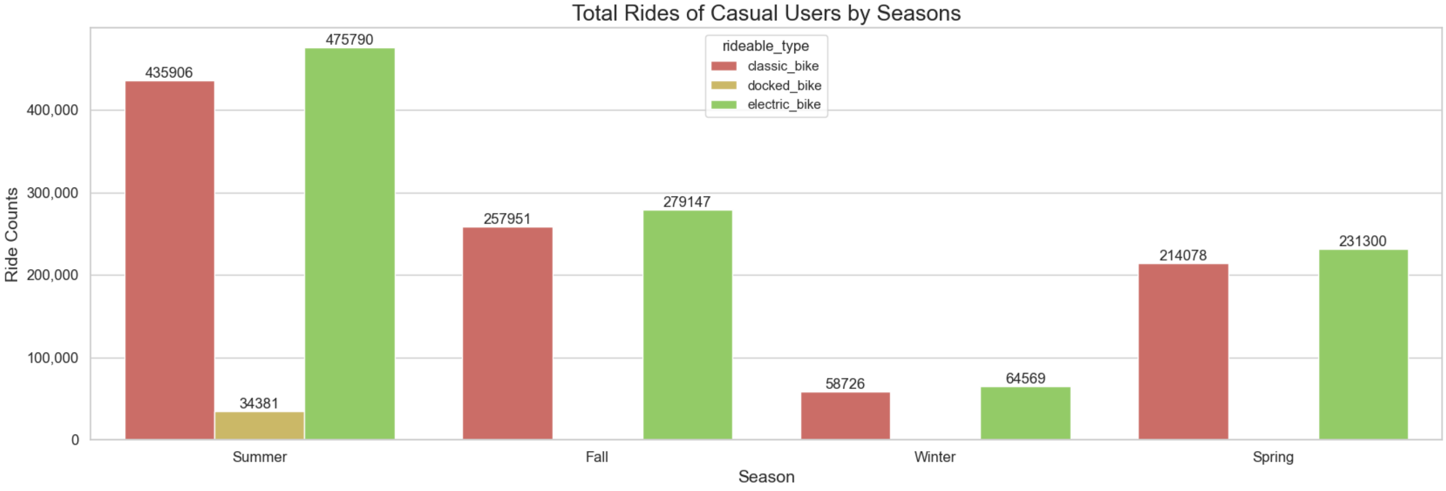
Total Number of Rides:




Total Rides by Seasons
member_casual
casual
member
Ride Counts
0
200,000
400,000
600,000
800,000
1,000,000
1,200,000
946077
1.30982e+06
537098
1.0289e+06
123295
468815
445378
880827
Summer
Fall
Winter
Spring
Season


Ride Count Ratio between Member and Casual Riders by Seasons
Ratio
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
1.38447
1.91567
3.80238
1.97771
Summer
Fall
Winter
Spring
Season

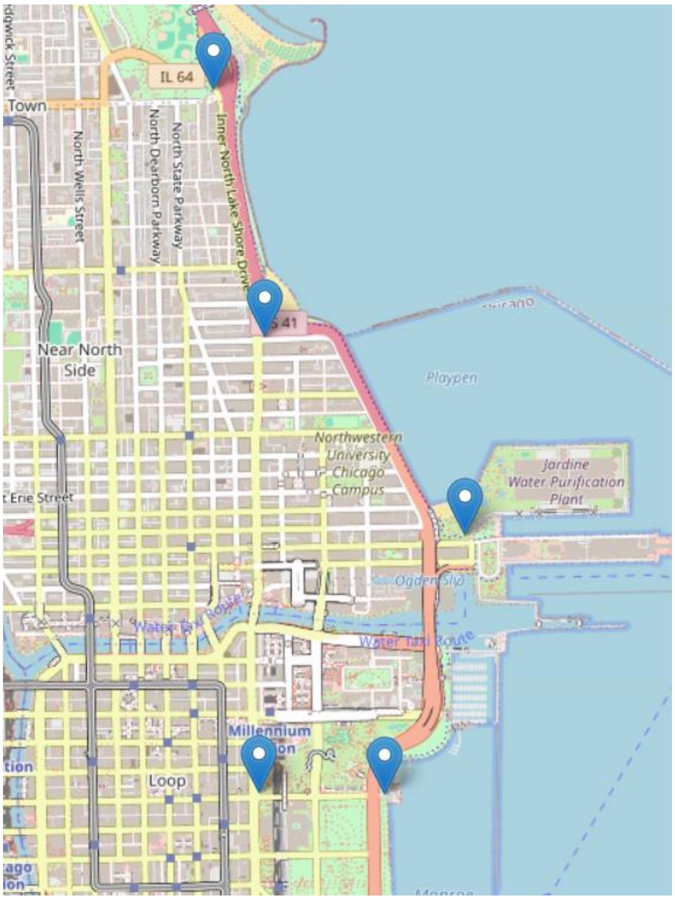

Total Rides of Casual Users by Days of the Week
rideable_type
classic_bike
docked_bike
electric_bike
Ride Counts
0
25,000
50,000
75,000
100,000
125,000
150,000
175,000
200,000
109253
4210
123914
106428
4237
128675
109047
3158
136033
111050
3814
138982
134010
4813
156694
212914
7512
198013
183959
6637
168495
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Day of the Week


Total Rides of Casual Users by Seasons
rideable_type
classic_bike
docked_bike
electric_bike
Ride Counts
0
100,000
200,000
300,000
400,000
435906
34381
475790
257951
279147
58726
64569
214078
231300
Summer
Fall
Winter
Spring
Season

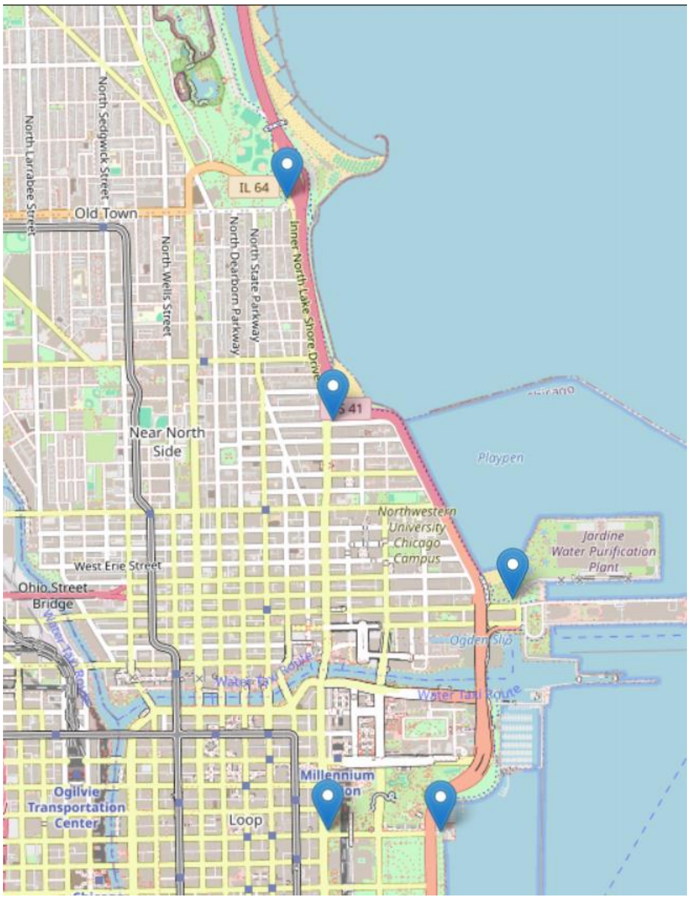
- **Findings:**
  - Peak hours for annual members to ride is between 6AM-8AM (peak hours for office) and 4PM-6PM. There is a slight increase in rides during 11PM-1PM.
  - Number of rides of casual members starts increasing from 6AM till 5PM and then starts declining.
  - Member riders take more rides than casual riders throughout the year.
  - Highest number of rides recorded during summer season (Jun-Aug) for both member and casual riders followed by Fall (Sep-Nov) and Spring (Mar-May). Lowest number of rides recorded in winter season (Dec-Feb).
  - Difference between member and casual riders' rides is highest in winter season (approx. 3.8x times more than casual riders).
  - Both types of riders prefer to ride more in warmer seasons. Although, there is significant drop in number of rides for both the users, Annual members ride significantly more than casual riders even in winter season. Main reason behind this may be that the Annual members use Cyclistic bikes for day-to-day essential purposes while casual riders use the bikes for leisure activities.
  - Casual riders are using Cyclistic bikes more on weekends (i.e. Saturday and Sunday).
  - Although casual riders prefer electric bikes over classic bike in general, riders are using classic bike more than electric bikes on weekends.
  - Docked bikes are being used only in summer season by casual riders.

- **Popular Ride Location of Cyclistic Riders:**

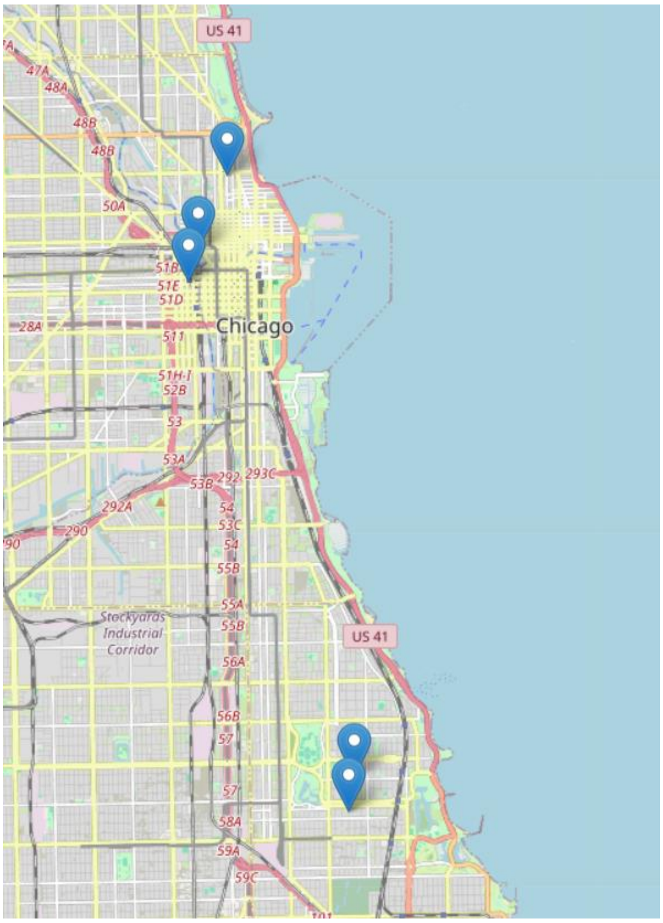
Casual Riders Trip Start Location



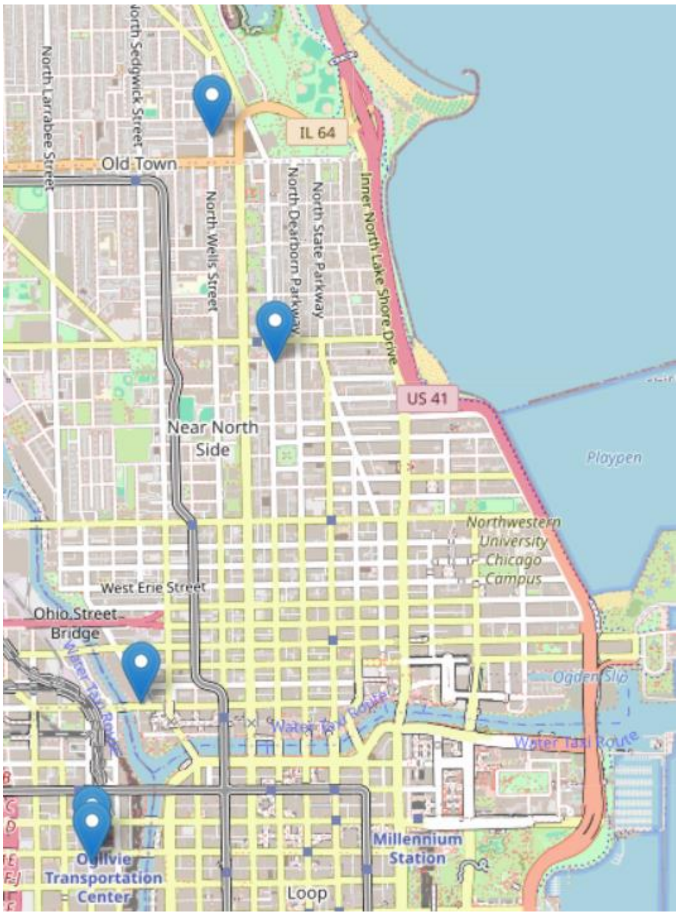
Casual Riders Trip End Location



Member Riders Trip Start Location



Member Riders Trip End Location



- **Findings:**
  - Top 5 popular locations for casual riders to start and end their trips are near the parks, gardens, beach, harbour points, museums.
  - Top 5 popular locations for member riders to start and end their trips are near the offices, universities, research centres, restaurants.
  - These findings further support my hypothesis that casual riders tend to use the Cyclistic bikes for leisure activities while annual members use the bikes for day-to-day work related activities.

<u>Act Stage:</u>

Guiding questions:

- What is my conclusion based on my analysis?
- I have concluded that Casual Riders use Cyclistic for longer duration than the annual members and for leisure activities while Member Riders use it for work related activities.

- How could my team and business apply my insights?
- My team could use these insights to create targeted marketing strategies to convert casual riders into member riders.

- What next steps would I or my stakeholders take based on my findings?
- ➢ The next step should be to apply these findings into marketing strategies.

- Is there additional data I could use to expand on my findings?
- ➢ I could use the data if casual riders live in the Cyclistic service area or if they have purchased multiple single passes, this would be beneficial to create marketing strategies:
  - Decisions on expanding service areas or improving infrastructure in high-demand regions.
  - Creating targeted marketing campaigns.
    For example, promotional offers can be tailored to specific neighbourhoods to encourage casual riders to become annual members.

Key tasks:

- Create my portfolio.
- ➢ I have created my portfolio. This can be viewed here.

- Add my case study.
- ➢ I have added my case study to my portfolio. This can be viewed here.

Deliverables:

- Top three recommendations based on the analysis
- ➢ As casual users ride Cyclistic bikes for longer duration than the annual members throughout the year, special awareness campaign should be run to spread awareness the benefits (such as cost) of annual membership program amongst casual riders.
- ➢ As casual riders preferably ride in the vicinity of parks, gardens, beach, harbour points, museums etc., special promotional offers can be given to annual members for these locations. Casual riders take trips in large number during the weekends. Weekend special promotional offers can be given to annual members. So that casual riders can be compelled to use annual membership program.
- ➢ Docked bikes are only being used by casual riders and that too only in summer. Special promotions can be given for docked bikes to annual members.