Stakeholders
Lily Moreno- Director of Marketing- Reporting boss.
Cyclistic executive team- Go / No Go

ASK

Problem statement: How do annual members and Casual riders use cyclistic differently?
Business task statement: Based on the past 12 months, what measurable usage pattern differentiates casual riders from annual riders?
This step could potentially help in designing a marketing strategy to convert more casual riders to annual memberships.
Time duration: 12 months from Aug 2021 to July 2022
Definition
Casual:
Member:

PREPARE

- Data source: Public data set made available by Motivate International Inc. under license. The original owners of the data is the City of Chicago. Data privacy laws do not allow us to relate each ride to a particular person however, a unique ride id has been assigned to each ride. Being public data, the data is unbiased, accessible and credible.
- Import data organization: Monthly data files name tagged in .xls (MS Excel) format going back to one year from current month.
- Data contents:
    - Data column headers of all files are structured in a similar format.
    - Column headers are
        - Unique identified: ride_id
        - Bike type: Cycle / Electric / Docked
        - Type of customer: Casual or Member
        - Time: Start datetime and End datetime
        - Station id: starting and ending of ride
        - Geographic co-ordinates of Location:
- How will the data help answer the business task: This data is appropriate to answer the business task since it describes the rider activity for 12 months.

PROCESS

Documentation of cleaning and manipulation of data

Tools chosen for

- Data cleaning and processing: SQL SERVER is used as a tool, because I was able to load data on the first try.
- Documentation with SQL codes: Obsidian was chosen because I am used to the formatting and it based on markdown.

Data integrity check: Data stored in our secure servers and backup made on regular frequency. All data cleaning work is done on temporary files so that actual data is not tampered with.

Steps taken to process data

1. Individual monthly excel files were unioned to make one table (5,901,463 rows).
2. Data cleaning in each column involves checks for
    1. Unique / distinct values
    2. NULL values count.
    3. Any peculiarities in the column
    4. Check for spell errors and use of trim.
3. Check for relations between columns that seem to be impossible.
4. Clean the data with assumptions clearly labeled.

Observations based on columns.

Column id	Data type	Unique values	Data range within limits	NULL	Comments
Ride_id (alphanumeric)	varchar(50)	59,01,463	Yes	N.A.	N.A.
Rideable_type (category)	varchar(50)	3	Yes	N.A	N.A
started_at (datetime)	datetime2	Not required	Aug 2021 to July 2022	N.A.	N.A.
ended_at (datetime)	datetime2	Not required	Aug 2021 to July 2022	N.A	N.A.
start_station_name (alphanumeric)	varchar(100)	1,382	Yes	8,60,786	
end_station_name (alphanumeric)	varchar(100)	1,397	Yes	9,19,896	
start_station_id (aplhanumeric)	varchar(50)	1,227	Yes	8,60,784	.0' suffix, alpha numeric with different char lengths
end_station_id (alphanumeric)	varchar(50)	1,237	Yes	9,19,896	.0' suffix, alpha numeric with different char lengths
member_casual	varchar(50)	2	Yes	N.A.	member / casual

1. Ride_id is a unique field and can be used as a primary key.
2. Rideable_type has 3 categories namely Classic / Electric / Docked
3. Member_Casual column
4. 4. Latitude and Longitude are long floats with 13 decimal places and 12 decimal places respectively
5. Start station ids and End station ids
    - Start station ids are alpha numeric codes with varying character lengths. (Will be checked for analysis later)
    - Start_station_id and end_station_id have fields with '.0' suffix. If this suffix is removed, then it matches an already present station id, so it needs to be replaced.
    - Observations
        1. The unique station ids are 1243 in nos.
        2. The numeric coded station_id (without alphabets) represent 605 stations which match the company profile "stating around 600 stations" with length of station_id ranging between 3 and 4.
        3. The alphanumeric coded stations are mainly consisting of
            1. KA and TA prefix followed by numbers: The rides associated with these stations were checked but no valid reason to disqualify these rows.
            2. Station names with keywords 'char', 'charging','repair','DIVVY','warehouse': The rides associated with these keywords seem to be service stations but both members and casuals use these stations as drop and pick up, hence there is no strong reason to disqualify these rows.

Observations based on relation between columns

1. Start time and End time relate to ride duration. Difference between the two values are negative in some rows. In some cases ride duration are greater than 16 hours.
2. Curious case of station id and their attributes:
   The peculiarities are as mentioned below
    1. Station ids have more than one station name .
    2. Station ids have more than one latitude and longitude combination. In one instance, one station id had 16 latitude - longitude combinations.
       In order to rationalize these oddities, a master table had to be created for each station using the station_id as a primary key with 3 other unique attributes- station name, latitude and longitude.

Assumptions made for cleaning process

1. Average latitude and average longitude is treated as a unique location point each station
2. Station ids with multiple station names have been coalesced into one station name.
3. 16 hours (960 minutes) of ride duration from ride start is only considered valid for the purpose of analysis.

Actions taken

1. Data modified
    1. Station ids with .0 suffix were fixed. (40,597 rows)

    2. Station ids with valid station_id but no name were filled in using a master table.
    3. Master table data of station id, station name and average latitude and longitudes were brought into the source data to rationalize the data.
    4. Rows with valid station id and missing station name were filled.
2. Data excluded
    1. Time difference between start time and end time
        1. Ride duration was negative.(75 rows)
        2. Ride duration was greater than 960 minutes (1564 rows)
    2. Rides which start from 'warehouse' and have no ending station (628 rows)
    3. Row items where the start station name and start station id were NULL had to be dropped despite having a latitude and longitude value for start station as no correlation was possible. (860,784 rows)
    4. Row items where the end station name and end station id were NULL had to be dropped despite having latitude and longitude for end station as no correlation was possible. (410,820 rows)
3. Revised data stored into a "Cleaned data" file.