Google Capstone Project

Cyclistic Bike-Share Analysis

1) Project Brief:
   - **Company activity:** Cyclistic is a bike-sharing company that has a fleet of 5,824 bicycles geotracked and locked into a network of 692 stations across Chicago. These bikes can be unlocked from one station and returned to any other station in the system anytime.
     The company offers three flexibility 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**.
   - **Problem:** Annual members are much more profitable than casual riders.
   - **Strategic goal:** Maximizing the number of annual memberships by designing marketing strategies aimed at converting casual riders into annual members.
   - **Key Stakeholders:** Cyclistic executives, Marketing Director (Lily Moreno) and marketing analytics team.

2) Ask Phase – Business Task:
   - **Business Task Statement:** Understand how annual members and casual riders use Cyclistic bikes differently so targeted marketing strategies can be recommended for conversion.
   - **Guiding questions:**
     - How do casual riders and annual members differ in their usage patterns?
     - What are the key differences in ride duration, time of day, and day of the week usage?
     - Are there any preferences in bike types used by each group?
     - What are the most popular start and end stations for each group?
     - Why might casual riders choose not to become annual members?
     - How can digital media be used to influence casual riders to become members?

3) Prepare Phase - Data Sources Description:
   - The data is first-party collected directly from Cyclistic's historical trip data, it is updated regularly, and the data source/license are available.
   - The data we will be working with is a public (open) data that is respecting privacy measures considering that we don't have access to riders' personally identifiable information. The company is making the data available under a license that allows us to access, reproduce, analyze, copy, modify and distribute the data.
   - **For our analysis we are using 12 csv files, each of them includes monthly bike rides starting from December 2023 to November 2024.**

4) Process Phase – Cleaning and Manipulation of Data:

We will be using **R** to clean and analyze the data as the datasets are too large and it will help us organize, modify and clean dataframes. The full processing and analysis are documented in an R Markdown file called **"***Google Capstone Project-Cyclistic Bike-Share Analysis (RStudio).html***".** Nevertheless, the main steps we followed are summarized below:

- Installing and loading required packages such as tidyverse, lubridate, janitor, etc.
- Importing csv files by using read_csv and changing their names to make them more understandable. These files contain data regarding the monthly trips from 12-2023 to 11-2024.
- Examining individual datasets using str() to display their structure and rowtotal() to identify the total rows of all tables combined.
- Merging all 12 datasets into one dataframe called "Full_Year_df"
- Examining the combined dataframe and checking for missing values, duplicates and invalid data types.

- Based on our data examination, we identified a significant number of missing values in station data columns. Before proceeding with data cleaning, we need to examine patterns by rideable_type (classic bike, electric bike, electric scooter) and member_casual (casual, member).
  - Key Observations:
    - **Station columns:** The missing values in "start_station_name", "start_station_id", "end_station_name", and "end_station_id" are primarily due to electric bikes (~93%) and scooters (~6%), which can be parked anywhere, not requiring docking stations.
    - **Coordinates:** The (7,340 rows, ~0.12%) missing "end_lat"/"end_lng" values for classic_bike trips suggest minor logging errors, possibly at specific stations or due to system issues.
    - **User Patterns:** Members have a higher share of missing station data compared to casual riders (61.7% vs. 38.3%), suggesting higher electric vehicle use among members.
- Cleaning the dataframe by following these steps:
  - Dropping station data columns as they are unreliable for station-based analysis.
  - Excluding rows where "end_lat" or "end_lng" are NA to ensure accurate geospatial mapping.
  - Removing duplicated rows based on ride_id to ensure each trip is unique and prevent skewed counts.
  - Cleaning column names using clean_names() function to ensure consistency and compatibility.
  - Key Observation:
    - The updated dataset contains 5,898,758 rows and 9 columns, reduced from 5,906,269 rows and 13 columns, by removing 4 station columns, filtering out 7,340 rows with missing coordinates, eliminating 171 duplicate ride_id rows, and standardizing column names.
- Manipulating the dataframe by following these steps:
  - Adding columns for: Date, Weekday, Day, Month, Year, Time, Hour and Ride Duration.
  - Rounding `start_lat`, `start_lng`, `end_lat`, and `end_lng` to 4 decimal places for consistent mapping.
  - Renaming the columns of “rideable_type” and “member_casual” to “bike_type” and “user_type” to improve clarity.
  - Saving the cleaned dataset.

5) Analyze Phase - Analysis Summary:

After processing and cleaning the dataset, we will keep using R to calculate key metrics and identify patterns differentiating casual riders and annual members. This supports our business task of enhancing membership conversion strategies. The main steps in our analysis are outlined below:

- Calculating the total rides and user type breakdown including rides count and percentage.
  - Key Observations:
    - **Total rides:** The dataset contains 5,897,917 rides.
    - **User type breakdown:** Members account for 3,739,306 rides (63.4%), while casual users account for 2,158,611 rides (36.6%), indicating a strong member base for targeted marketing.
- Calculating the overall average ride duration as well duration statistics by user type. These include average, median, minimum and maximum durations.
  - Key Observations:
    - **Overall average duration:** The average ride duration across all Cyclistic users is 15.48 minutes.
    - **By user type:**
      - **Average:** Casual riders average 21.09 minutes per trip, significantly longer than members who average 12.23 minutes. This represents an ~8.86-minute difference.

- **Median:** The typical (median) ride duration for casual users is 12.00 minutes, which is also considerably higher than members' median of 8.69 minutes, a gap of ~3.31 minutes.
- **Maximum:** Both user types exhibit extreme maximum ride durations (Casual: 1509.37 minutes; Member: 1499.93 minutes), indicating the presence of outliers that heavily influence the average.
- **Minimum:** Both user types have a minimum ride duration of 0.01 minutes (~0.6 seconds), which could represent very short trips or data recording anomalies.

Data Insights:

- The notable difference in average and median ride durations strongly suggests a divergence in primary usage patterns: casual riders likely favor longer, potentially leisure or exploratory trips, while members' shorter durations are more consistent with regular commuting or utilitarian use.
- The significant impact of maximum ride durations on the average highlights the median as a more robust indicator of typical ride behavior for both casual and member user types.
- The presence of extremely short rides (0.01 min) warrants consideration. While some may be valid, their frequency and context should be explored to understand their nature as either genuine quick trips or potential data artifacts.

- Investigating outliers and anomalies: To ensure our duration findings accurately reflect typical behavior, an analysis of ride duration outliers (rides > 120 minutes & rides < 1 minute) was conducted.

Key Observations:

- Outliers (>120 minutes):
  - **Outlier counts:** A total of 35,548 rides exceeds 120 minutes. Casual riders account for the vast majority of these extended trips, representing 80.26% (28,530 rides), compared to 19.74% (7,018 rides) from members.
  - **Outlier details:** Both casual (1509.37 minutes) and member (1499.93 minutes) riders exhibit extreme maximum ride durations with some rides exhibiting identical start/end coordinates (e.g.: 41.9466, -87.6946).
  - **Identical coordinates:** Among these high outliers, 8,706 rides (~0.15% of total rides) feature identical start and end coordinates (7,526 casual, 1,180 member).
- Outliers (<1 minute):
  - **Outlier counts:** A total of 131,944 rides were under 1 minute. The distribution of these very short trips is nearly even, with casual riders accounting for 49.35% (65,113 trips) and members for 50.65% (66,831 trips).
  - **Outlier details:** These low outliers consistently show ride durations just under 1 minute (0.99 minutes) and frequently feature identical or nearly identical start and end geographical coordinates (e.g.: 41.9300, -87.6400).
  - **Identical coordinates:** Among these low outliers, 90,275 rides (1.53% of total rides) exhibit identical start and end geographical coordinates. Casual riders account for 54.82% (49,486 trips) of these specific anomalies, while members account for 45.18% (40,789 trips).

Data Insights:

- Outliers (>120 minutes):
  - The overwhelming dominance of casual users among rides exceeding two hours strongly indicates their greater propensity for leisure-driven or exploratory trips. This contrasts sharply with members' usage, reinforcing a core difference in how each group engages with the bike share service. While a subset of these long rides, particularly those with identical start/end coordinates, might point to potential system anomalies or data artifacts (e.g., faulty docking during a very long trip), their impact

on overall average ride durations is minimal (0.6% of total rides), so no adjustment is applied. This means that the primary finding – that casual riders frequently take significantly longer trips – remains robust and actionable for conversion strategies.

- Outliers (<1 minute):
  - The nearly even distribution of these extremely short rides across both user types, coupled with their frequent identical start/end coordinates, strongly suggests they stem from common factors affecting all riders. These likely represent data recording errors, accidental checkouts, or immediate re-docking attempts due to issues, rather than genuine travel trips. This finding is more relevant for system diagnostics and data quality improvement than for differentiating member vs. casual usage patterns to drive conversion strategies.

- Given the limited impact of high and low outliers on differentiating usage, it makes sense to pivot to analyses that better highlight behavioral differences between casual and member users.

- Analyzing time-based trends in bike usage by calculating the ride count per user type by month, weekday and hour.
  - Key Observations:
    - **Month:** Both member and casual rides are peaking in September at 474,282 and 345,874 respectively. However, casual rides drop sharply to just 24,339 by January, while member rides decline more gradually to 120,149.
    - **Weekday:** Member rides peak on Wednesdays (617,738) and are lowest on Sundays (418,953), while casual rides peak on Saturdays (447,481) and hit their lowest on Tuesdays (231,410).
    - **Hour:** Both members and casuals reach their highest ride counts at 5 PM, with members having 395,814 rides and casuals 204,332 rides. Members show clear peaks during early morning and late afternoon hours, while casual rides are more evenly spread throughout the afternoon. Ride volumes drop sharply after 9 PM for both users.
  - Data Insights:
    - Although both groups ride more in warmer months, casual users show a steeper seasonal drop, indicating more weather-sensitive and recreational behavior. Members maintain steadier usage, likely reflecting routine or commuting habits.
    - Members show a strong weekday usage pattern, likely driven by work commutes. In contrast, casual riders are more active on weekends, pointing to leisure-driven behavior with minimal weekday use.
    - The strong morning and evening peaks suggest members mainly use the service for commuting purposes. In contrast, the more evenly distributed casual ride pattern indicates leisure or flexible usage rather than strict travel times. Low overnight activity highlights limited demand during late hours across both groups.

- Comparing bike type preferences across user types by calculating total rides and the average ride duration per bike type and user type.
  - Key Observation:
    - **Bike usage count:** Electric bikes are the most popular choice across both user types, with 1,097,153 rides by casual users and 1,894,242 rides by members, significantly outnumbering classic bikes (976,243 casual, 1,785,942 member) and electric scooters (85,215 casual, 59,122 member).
    - **Bike usage average:** Casual riders consistently have a longer average ride duration than annual members across all bike types. The largest duration gap between user types is observed for classic bikes (casual: 29.28 mins vs. member: 13.43 mins), while electric scooters show the shortest average duration for both groups (casual: 11.94 mins vs. member: 8.24 mins).

- Data Insight:
  - Annual members are the high-frequency users, driving significantly more total rides across all bike types compared to casuals, likely for consistent, goal-oriented trips (e.g., commuting or errands). While both user types strongly prefer electric bikes, casual riders consistently take significantly longer average trips across all bike types (especially classic bikes), suggesting their usage leans more towards leisure or exploratory purposes. These distinctions in frequency and duration of use, despite shared bike type preferences, are crucial for tailoring conversion strategies.

- Conducting a geospatial analysis to identify usage patterns:
  - Identifying activity hotspots (Start & End Points).
    - Key observation:
      - **Hotspot location differences & concentration:** Analysis of ride start and end coordinates reveals distinct hotspot clusters for casual versus annual members. Casual riders account for a higher number of total rides within the identified top 10 hotspots, with specific high-count clusters at coordinates such as (41.8923, -87.6120) and (41.8810, -87.6167). Annual members, while having higher overall total rides, show their highest-count clusters in distinct areas identified by coordinates like (41.8892, -87.6385) and (41.8834, -87.6412), indicating a more distributed pattern of activity compared to casual riders' concentration in specific top locations. This geographical separation is evident for both ride starts and ends.
    - Data Insight:
      - The striking concentration of casual rider activity in specific hotspots (e.g., coordinates near Navy Pier or the Lakefront Trail) suggests casual users frequently return to and prioritize these locations for leisure or tourism activities. This contrasts with members' more distributed usage across locations, including areas consistent with business and residential zones, indicative of commuting or regular utilitarian trips. This geographical divergence in usage patterns provides a valuable foundation for developing location-based targeted marketing strategies for conversion.
  - Identifying temporal trends within geospatial hotspots by defining peak weekdays and hours in top hotspots and determining the major hotspots during peak days and hours.
    - Key Observations:
      - **Peak day/hour in top 10 hotspots:** Casual riders exhibit their highest activity in the top 10 hotspots mostly on Saturdays between 14:00 and 16:00 (e.g.: 1,390 rides at 41.8923, -87.6120 at 15:00). Conversely, member riders show peak activity predominantly on Wednesdays and Tuesdays between 16:00 and 17:00 (e.g.: 880 rides at 41.8834, -87.6412 at 17:00), with an additional morning peak at 8:00 on Tuesday (377 rides at 41.9030, -87.6313).
      - **Top 10 hotspots by peak days:** When focusing on peak days (Saturdays for casuals and Wednesday for members), 6 of the top 10 hotspots are associated with casual riders with the highest ride count of 12,163 at 41.8923, -87.6120. The remaining 4 hotspots are associated with member riders, with the highest ride count of 4,886 at 41.8834, -87.6412.
      - **Top 10 Hotspots by Peak Hour:** Analysis of hotspots peaking at 17:00 reveals that 8 of the top 10 entries are associated with member riders, including 3,505 rides at 41.8834, -87.6412. Casual riders account for the remaining 2 hotspots with 4,548 rides at 41.8923, -87.6120 and 2,700 rides at 41.8810, -87.6167.

- Data Insight:
  - Casual riders are most active in specific recreational spots on weekends and during afternoons/evenings, suggesting they use the service for leisure or tourism. In contrast, annual members show peak usage primarily in business/residential areas on weekdays during typical commute times (Morning and 17:00), indicating they rely on bikes for utilitarian or commuting purposes. This clear difference in when and where each group rides provides a robust foundation for developing highly targeted marketing campaigns, allowing promotions to be tailored to specific geographic locations at the precise times casual riders are most active, thereby maximizing conversion potential.

- Exporting key analysis outputs as CSV files for Tableau visualizations.

6) Share Phase - Supporting Visualizations:

Based on our data analysis, we will use Tableau for our visualizations to create interactive maps, charts, and dashboards that highlight usage patterns and support membership conversion strategies.

- The final dashboard is published in Tableau website (Link: "Asmae Addab-Google Capstone Project_Cyclistic Bike-Share Analysis")

7) Project Recommendations:

- **Boost Member Conversion:** Target casual riders in weekend/leisure hotspots with promotions and trial memberships to encourage annual subscriptions.
- **Optimize Bike Allocation:** Prioritize electric bike availability at casual hotspots on weekends and member hotspots on weekdays.
- **Seasonal & Location-Based Campaigns:** Run promotions during spring/summer and near high-traffic recreational areas to maximize casual rider engagement.
- **Enhance System Reliability:** Strengthen docking station maintenance to prevent short and long rides and ensure reliability at high-traffic hotspots for both user types.
- **Tailor Engagement by User Behavior:** Send personalized offers to members for weekday commutes and to casual riders for weekend leisure trips, aligned with peak hours.
- **Plan for Growth & Maintenance:** Use hotspot and duration trends to guide station placement, bike type distribution, and maintenance scheduling.