Cyclistic Journey Analytics: A Data-Driven Study of Ridership

Data Analytics Project

Prepared by: Kushal Jeet Kumawat

Course: Google Business Intelligence

Year: 2025

Table of Contents

Project Overview

Project Background

In this fictitious workplace scenario, the imaginary company Cyclistic has partnered with the city of New York to provide shared bikes. Currently, there are bike stations located throughout Manhattan and neighboring boroughs. Customers are able to rent bikes for easy travel among stations at these locations.

Cyclistic's Customer Growth Team is creating a business plan for next year. The team wants to understand how their customers are using their bikes; their top priority is identifying customer demand at different station locations. The dataset includes millions of rides, so the team wants a dashboard that summarizes key insights. Business plans that are driven by customer insights are more successful than plans driven by just internal staff observations. The executive view must include key data points that are summarized and aggregated in order for the leadership team to get a clear vision of how customers are using Cyclistic.

Aim / Objective

The aim of this project is to analyze Cyclistic’s existing bike usage data to understand customer behavior and station performance across New York City. The findings will support operational efficiency, marketing strategies, and customer experience improvements. These insights will support data-driven decision making and can further assist in future planning for new station development and overall business expansion.

Objectives:

- Examine trip patterns across different boroughs and time periods.
- Compare ride behavior between subscriber and non-subscribers.
- Analyse how external factors (like weather and time) affect demand.
- Provide summarized dashboards that help leadership understand usage trends.

Key Questions:

- How do different users utilize Cyclistic bikes?
- What factors (season, weather, location) influence ride demand?
- What do usage trends in summer 2015 indicate about customer behavior?
- Which stations experience the highest and lowest usage levels?
- How can insights guide Cyclistic’s marketing strategies?

Dataset Information

Source and Timeframe

Four datasets were used in this project. Three are publicly available on Google BigQuery and one additional zipcode spreadsheet was provided by the team.

Primary Dataset: NYC Citi Bike Trips
Secondary Dataset: Census Bureau US Boundaries
Secondary Dataset: GSOD - NOAA
Additional Dataset: Zip Code Spreadsheet
[File details : Type Google Sheets, Size 2 KB, Storage used 2 KB Owned by Grow with Google Certificates]

Time period: January to December 2015.

Key Columns / Data Description

Cyclistic has captured data points in their primary dataset for every trip taken by their customers, including:

- Trip start time and location (station number, and its latitude/longitude).
- Trip end time and location (station number, and its latitude/longitude).
- The type of customer (either a one-time customer, or a subscriber).
- The rented bike's identification number.

The dataset includes latitude and longitude of stations but does not identify more geographic aggregation details such as zip code, neighborhood, and borough. To address this, a separate zip code spreadsheet was used and joined with other datasets to obtain complete and accurate location information.

Data Limitations

The weather data provided does not include what time precipitation occurred; it's possible that on some days, it precipitated during off-peak hours. However, for the purpose of this dashboard, we have assumed any amount of precipitation that occurred on the day of the trip could have an impact.

Methodology

1. **Data Preparation**

Overview

The data preparation phase focused on extracting, cleaning, and combining multiple public datasets to build a unified analytical table for Cyclistic's 2015 operations. The final dataset integrates trip, weather, and geographical data, enabling multi-dimensional analysis across time, location, and user segments.

Data Sources and Ingestion

The project utilizes one primary dataset, two secondary datasets, and one additional zipcode spreadsheet, all of which support comprehensive geographic and trip-level analysis.

- New York Citi Bike Trips - Core dataset containing trip-level information such as start and stop times, station IDs and usertypes.
- NOAA GSOD Weather Data - Daily temperature, wind speed, and precipitation values recorded at the Central Park weather station (wban = 94728).
- U.S. ZIP Code Boundaries (Census Bureau) - Geographical shapefiles used to associate trip start and end coordinates with zipcode areas.
- Additional ZIP Code Spreadsheet - Supplementary file containing ZIP, Borough, and Neighborhood details. This file was joined with the Census Bureau ZIP Boundaries dataset to enrich it with missing borough and neighborhood information.

All datasets were accessed directly via Google BigQuery using SQL queries. Data was filtered to include only records from the 2015 calendar year to maintain a consistent analysis window.

Data Cleaning

Data cleaning ensured only valid, relevant, and interpretable records were retained. Weather data was constrained to valid daily observations matching trip start dates.

Standardization

All datetime fields were standardized to date-level granularity. Trip dates were standardized by adding five years using the DATE_ADD function to make the dataset appear more recent while preserving the original time patterns for analysis. Temperature, precipitation, and wind speed fields were normalized to daily averages. Zip codes were converted to consistent string formats, and borough and neighborhood names were aligned with standardized geographical references.

Joins and Merging

Spatial joins were performed between the trip coordinates and Zip code polygons using the ST_WITHIN function to identify the start and end zip codes for each ride. Date based joins linked trip data to NOAA weather records using

the trip start date. Additionally, neighborhood and borough names were merged from the zip reference table to enrich geographical context.

2. Data Transformation and Analysis

Feature Engineering

Several new analytical fields were derived to support advanced insights:

- Trip Duration (Minutes) - Calculated using TIMESTAMP_DIFF between trip start and stop times.
- Is Rainy Day - Binary flag (1=rainy, 0=non-rainy) based on daily precipitation.
- Congestion Ratio - Computed as MAX(hourly starts) / AVG(hourly starts) to measure peak hour load at each station.
- Trips per Bike - Proxy for utilization efficiency, calculated as total trips divided by the number of unique bikes at each station.
- Temperature, Wind Speed, Precipitation - Merged weather attributes for daily environment context.

Aggregations and KPI Definitions

The final dataset was aggregated by usertype, zipcode, neighborhood, borough, and day, allowing multi-level analysis.

Key performing indicators (KPIs) were defined as:

- Trip Count - Total number of rides per geographic and user segment.
- Average Trip Duration - Mean trip length in minutes.
- Average Hourly Starts / Maximum Hourly Starts - Used to compute congestion ratios.
- Trips per Bike - Evaluates operational efficiency and station utilization.

3. Tools and Environment

Tools Used:

- Google BigQuery - For SQL based data extraction, cleaning, joining, and feature creation.
- Tableau Public - For visualization, dashboard building, and storytelling.
- Excel & Google Sheets - For quick data validation and CSV exports.

Analytical Environment

The analysis was conducted entirely in Google BigQuery Sandbox for computation and Tableau Public (web version) for visualization. The final dataset (final_cyclistic_2015.csv) was approximately 2.3 million records, exported as CSV to Google Drive for use in Tableau.

Visualization and Dashboard

Dashboard Overview

The project consists of four interactive dashboards developed in Tableau Public each designed to address key business questions from the Cyclistic bike-share analysis. Together, they provide a complete view of user behavior, operational efficiency, and demand trends across locations and seasons.

Interactive filters and parameters enhance the analytical experience:

Metric and Month Selectors allow switching between trip count and duration and exploring trends across summer months. Usertype Filters (Subscriber vs Customer) enable behavioral segmentation. Weather Filters (Non Rainy vs Rainy) reveal external factors influencing ridership patterns. Dynamic tooltips, highlights, and map interactivity is used to support deeper exploration of specific neighborhoods and congestion zones.

Season and Weather Trends

Purpose: To analyze how trip patterns change across different seasons and weather conditions

Key features:

- Seasonal segmentation of ride counts and average trip duration.
- Comparative line and bar charts highlighting fluctuations throughout the year.
- Visualization of weather impact (non-rainy days vs rainy days) on total rides.

Focus: Understanding demand variation influenced by season and precipitation to support marketing and resource planning.

Operations and Efficiency

Purpose: To assess system performance in terms of congestion and bike utilization across locations.

Key features:

- Maps and bubble charts illustrating congestion ratios and utilization indices by zip code.
- Identification of high-pressure areas and stations with maximum usage.
- Segmentation of top performing and under-utilized zones.

Focus: Providing insights for operational optimization, resource allocation, and station management.

Top Trip Locations

Purpose: To identify the most popular starting and ending points for rides and understand location level usage.

Key features:

- Combined analysis of trip count and average trip duration.
- Data categorized by neighborhood, ranking zip codes from most to least popular based on trip duration.
- Filtering by user type for detailed behavioral comparison.

Focus: Highlighting key origin-destination clusters to inform expansion and service improvement decisions.

Summer Ride Insights

Purpose: To examine system performance during the high demand summer period (July-September)

Key Features:

- Interactive map with month and metric selection (trip count or duration).
- Comparative analyses of location level activity across summer months.
- Summary table of rides and durations by borough, neighbourhood, and zip code.

Focus: Capturing peak-season performance patterns and evaluating the effect of summer demand on system capacity.

Insights and Findings

Season and Weather Trends

- Trip volume was the lowest during winter, rose slightly in spring, and peaked through summer and fall, following a clear seasonal pattern.
- The ideal riding temperature ranged between 50 F - 70 F, showing optimal conditions for both short and long rides. Riders preferred non-rainy days, with trip activity declining during rainfall.
- During summer, ride frequency on rainy days nearly matched non-rainy days, whereas in fall, rainy days saw a noticeable dip in ridership.

Operations and Efficiency

Congestion Analysis:

- A ratio above 8 indicates high congestion, below 5 indicates low congestion, and values between 5 and 8 represent normal congestion levels.
- The Sunset Park (11220) neighborhood had the lowest ratio (1.00), while Lower Manhattan (10278) recorded the highest (18.90).
- Congested neighborhoods included Chelsea & Clinton, Gramercy Park & Murray Hill, and Lower Manhattan, whereas Bushwick & Williamsburg, and Sunset Park showed lower congestion.
- Among the top 15 most congested areas, Chelsea & Clinton and Gramercy Park & Murray Hill dominated, each with multiple zipcodes ranking high.

Bike Utilization:

- Average trips per bike were below 2 for over 18 zipcodes, indicating underuse in several areas.
- The highest utilization was observed in Chelsea & Clinton (10199) with 10.69 trips/bike, the lowest in Sunset Park (11220) with 1.00 trip/bike.
- High performing neighborhoods included Chelsea & Clinton, Gramercy Park & Murray Hill, and Lower Manhattan, highlighting concentrated demand zones.

Top Trip Locations

Top Starting Points:

- Chelsea & Clinton (10011, 10019, 10001) and Greenwich Village & Soho (10014, 10013, 10012) led in both trip count and duration.
- The highest metrics were recorded in 10011 (Duration: 4,048,668; Trips: 826,457) and 10014 (Duration: 4,173,838; Trips: 916,469).

Top Destination Points:

- Lower East Side (10002, 10003) and Chelsea & Clinton (10011, 10019, 10001) were the most popular destinations.
- The top performers included 10002 (Duration: 4,117,433; Trips: 521,232) and 10011 (Duration: 2,900,150; Trips: 840,139).

These findings reflect the dominance of downtown and central neighborhoods as both trip origins and destinations.

Summer Ride Insights

Monthly Ride Activity (Map Analysis):

- Across July, August, and September, Lower East Side (10003) and Chelsea & Clinton (10011) consistently led in both trip count and duration.
- September recorded the highest ridership and total duration, followed by August, with July being relatively lower.
- The trend reflects peak travel behavior in late summer, driven by favourable weather and tourism.

Summer Ride Performance (Summary Table):

This table consolidates trip count and average trip duration for all locations across the three months, segmented by usertype and weather conditions, providing a comprehensive view of summer performance patterns.

Business Recommendations and Data-Driven Actions

Based on the analysis, Cyclistic can enhance operational efficiency and user engagement through several data-driven insights. High demand zones such as Chelsea & Clinton, Lower Manhattan, and the Lower East Side should be prioritized for bike availability and station capacity expansion, while low utilization areas like Sunset Park could benefit from promotional campaigns or dynamic pricing to boost usage.

Seasonal demand peaks in summer and fall suggest allocating maintenance and fleet resources during these periods. Incorporating weather forecasts into demand prediction models can optimize bike distribution on rainy versus clear days. Finally, focusing retention campaigns on high duration riders and subscribers in central neighborhoods can strengthen loyalty and improve overall trip volume.

Conclusion

This project involved a comprehensive analysis of Cyclistic's 2015 bike-share data, integrating trip, weather, and geographical information to uncover patterns in ridership behavior and system performance. The process included end-to-end data preparation in Google BigQuery, feature engineering to derive meaningful metrics, and the creation of interactive dashboards in Tableau Public to visualize insights across seasons, operations, and locations.

The key outcomes include a cleaned and structured analytical dataset, well-defined KPIs such as trip count, average trip duration, congestion ratio, and trips per bike, and four interactive dashboards that present seasonal trends, operational efficiency, top trip locations, and summer performance.

This analysis provides data-driven visibility into Cyclistic's dynamics, supporting more informed decisions in areas like bike allocation, station expansion, and marketing strategy. It demonstrates how analytical workflows can translate raw operational data into actionable business insights.

Limitations: The analysis was restricted to a single year (2015), and weather data lacked hourly precision, limiting insight into intraday weather effects. Additionally, the absence of pricing and demographic information constrained deeper behavioral segmentation, and the lack of total trip minutes data limited detailed duration-based analysis.

Future Scope: The study can be extended by incorporating data from additional years to analyze long term growth patterns, or by applying predictive modeling to forecast demand and optimize resource distribution.

Project Links

Tableau Dashboard

GitHub Repository