Project Requirements Document: Cyclistic

BI Analyst: Rolex Brice Yimtsop

Client/Sponsor: Jamal Harris, Director, Customer Data

Purpose:

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.

Key dependencies:

This project will require a dataset of customer data, so the Director of Customer Data will need to approve the request. Approval should also be given by the teams that own specific product data including bike trip duration and bike identification numbers to validate that the data is being interpreted correctly. The primary contacts are Adhira Patel, Megan Pirato, Rick Andersson, and Tessa Blackwell.

Stakeholder requirements:

In order to continuously improve and effectively market products, the dashboard must help Cyclistic decision-makers understand how their customers are using the bikes and the demand at different locations, including factors that might influence that demand at different times.

- A table or map visualization exploring starting and ending station locations, aggregated by location.
- A visualization showing which destination (ending) locations are popular based on the total trip minutes.
- A visualization that focuses on trends from the summer of 2015.
- A visualization showing the percent growth in the number of trips year over year.
- Gather insights about congestion at stations.
- Gather insights about the number of trips across all starting and ending locations.

- Gather insights about peak usage by time of day, season, and the impact of weather.

Success criteria:

Specific: BI insights must clearly identify the specific characteristics of a successful product. They must demonstrate how customers are currently using bikes and what impacts demand at station locations. **Measurable:** Each trip should be evaluated using starting and ending location, duration, variables such as time of day, season, and weather. For example, do customers use Cyclistic less when it rains? Or does bikeshare demand stay consistent? Does this vary by location and user types (subscribers vs. non-subscribers)? **Action-oriented:** These outcomes must prove or disprove the theory that location, time, season, and weather impact user demand. Then, the Cyclistic team will use this knowledge to refine future product development. **Relevant:** All metrics must support the primary question: How can we build a better Cyclistic experience? **Time-bound:** Analyze data that spans at least one year to see how seasonality affects usage. Exploring data that spans multiple months will capture peaks and valleys in usage.

User journeys:

The main purpose of Cyclistic is to provide customers with a better bike-share experience. A deeper-dive into trip trends will help decision-makers explore how customers are currently using Cyclistic bikes and how that experience can be improved.

Assumptions:

The dataset includes latitude and longitude of stations but does not identify more geographic aggregation details like zip code, neighborhood name, or borough. The team will provide a separate database with this data.

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, you should assume any amount of precipitation that occurred on the day of the trip could have an impact.

Starting bike trips at a location will be impossible if there are no bikes available at a station, so we might need to consider other factors for demand.

Compliance and privacy:

The data must not include any personal data such as name, email address, phone number, or physical address. The user provides this data as part of their device activation but is not

necessary for this project. It is paramount that the users be anonymized to avoid any bias.

Accessibility:

The dashboards should offer text alternatives including large print and text-to-speech.

Roll-out plan:

The stakeholders have requested a completed BI tool in six weeks:

- Week 1: Dataset assigned. Initial design for fields and BikeIDs validated to fit the requirements.
- Weeks 2-3: SQL & ETL development
- Weeks 3-4: Finalize SQL. Dashboard design. 1st draft review with peers.
- Weeks 5-6: Dashboard development and testing

Stakeholder Requirements Document: Cyclistic

BI Professional: Rolex Brice Yimtsop

Client/Sponsor: Jamal Harris, Director, Customer Data

Business problem:

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 Primary question: How can we apply customer usage insights to inform new station growth?

Stakeholders:

- Sara Romero, VP, Marketing
- Ernest Cox, VP, Product Development
- Jamal Harris, Director, Customer Data
- Nina Locklear, Director, Procurement

Stakeholder usage details:

To effectively develop new station locations, the team wants to understand how customers use the current line of bikes. They will use this BI tool in order to gain insights related to data generated by the bikes when being used by customers. Then, this information will be used to understand what customers want, what makes a successful product, and how new stations might alleviate demand in different geographical areas.

Primary requirements:

- A table or map visualization exploring starting and ending station locations, aggregated by location.
- A visualization showing which destination (ending) locations are popular based on the total trip minutes.
- A visualization that focuses on trends from the summer of 2015.
- A visualization showing the percent growth in the number of trips year over year.
- Gather insights about congestion at stations.
- Gather insights about the number of trips across all starting and ending locations.
- Gather insights about peak usage by time of day, season, and the impact of weather.

Strategy Document: Cyclistic

Sign-off matrix:

Name	Team / Role	Date
Rolex Brice Yimtsop	BI Analyst	2025/01/18

Proposer: Jamal Harris, Director, Customer Data

Status: Draft > Under review > Implemented | Not implemented

Primary dataset: NYC Citi Bike Trips

Secondary dataset: Census Bureau US Boundaries

User Profiles

Sara Romero, VP, Marketing

Ernest Cox, VP, Product Development

Jamal Harris, Director, Customer Data

Nina Locklear, Director, Procurement

Adhira Patel, API Strategist

Megan Pirato, Data Warehousing Specialist

Rick Andersson, Manager, Data Governance

Tessa Blackwell, Data Analyst

Brianne Sand, Director, IT

Shareefah Hakimi, Project Manager

Dashboard Functionality

Dashboard Feature	Your Request
Reference dashboard	Build a new dashboard to display the starting and ending locations, aggregated by location. This should show the number of trips at starting locations.
Access	Access will be provided as read-only to the user profiles listed in this document.
Scope	Fields include: station, zip code, neighborhood, and/or borough, year, month, trip count, weather
Date filters and granularity	Data filters can be applied for the following: Date, Month, Year Granularity:

	Any chart with user detail metrics should have the ability to click on that metric to view specific information.

Metrics and Charts

Please create a table like the example below for each chart that you'd like to include in the dashboard. If you'd like to break the dashboard under different headers, feel free to list those here as well.

Chart 1

Chart Feature	Your Request
Chart title	Trip Totals
Chart type	Line
Dimension(s)	Date
Metric(s)	Trip count

Chart 2

Chart Feature	Your Request
Chart title	Trip Counts by Starting Neighborhood
Chart type	Table
Dimension(s)	Neighborhood, month
Metric(s)	Trip count

Chart 3

Chart Feature	Your Request
Chart title	Total Trip Minutes by Destination
Chart type	Bar
Dimension(s)	Zip code end, borough end, neighborhood end, user type
Metric(s)	Trip minutes

Chart 4

Chart Feature	Your Request
Chart title	Average Time to Arrive

Chart type	Table
Dimension(s)	Zip code end, borough end, neighborhood end, start day, grand total
Metric(s)	Trip minutes

Chart 5

Chart Feature	Your Request
Chart title	Seasonal trends
Chart type	Map
Dimension(s)	Neighborhood start, neighborhood end, number of rides, average trip duration, weather
Metric(s)	Trip minutes, weather, number of rides

PIPELINE(ETL) TO CREATE TAEGET TABLES FOR CYCLISTIC

SQL query to create a summary table for the entire year:

```sql
SELECT
TRI.usertype,
ZIPSTART.zip_code AS zip_code_start,
ZIPSTARTNAME.borough borough_start,
ZIPSTARTNAME.neighborhood AS neighborhood_start,
 ZIPEND.zip_code AS zip_code_end,
 ZIPENDNAME.borough borough_end,
ZIPENDNAME.neighborhood AS neighborhood_end,
 DATE_ADD(DATE(TRI.starttime), INTERVAL 5 YEAR) AS start_day,
 DATE_ADD(DATE(TRI.stoptime), INTERVAL 5 YEAR) AS stop_day,
 WEA.temp AS day_mean_temperature, -- Mean temp
WEA.wdsp AS day_mean_wind_speed, -- Mean wind speed
 WEA.prcp day_total_precipitation, -- Total precipitation
-- Group trips into 10 minute intervals to reduces the number of rows
 ROUND(CAST(TRI.tripduration / 60 AS INT64), -1) AS trip_minutes,
 COUNT(TRI.bikeid) AS trip_count
FROM
`bigquery-public-data.new_york_citibike.citibike_trips` AS TRI
INNER JOIN
 `bigquery-public-data.geo_us_boundaries.zip_codes` ZIPSTART
ON ST_WITHIN(
ST_GEOGPOINT(TRI.start_station_longitude, TRI.start_station_latitude),
ZIPSTART.zip_code_geom)
INNER JOIN
 `bigquery-public-data.geo_us_boundaries.zip_codes` ZIPEND
ON ST_WITHIN(
ST_GEOGPOINT(TRI.end_station_longitude, TRI.end_station_latitude),
ZIPEND.zip_code_geom)
INNER JOIN
`bigquery-public-data.noaa_gsod.gsod20*` AS WEA
ON PARSE_DATE("%Y%m%d", CONCAT(WEA.year, WEA.mo, WEA.da)) = DATE(TRI.starttime)
INNER JOIN
 -- Note! Add your zip code table name, enclosed in backticks: `example_table`
`(insert your table name) zipcodes` AS ZIPSTARTNAME
ON ZIPSTART.zip_code = CAST(ZIPSTARTNAME.zip AS STRING)
INNER JOIN
 -- Note! Add your zipcode table name, enclosed in backticks: `example_table`
`(insert your table name) zipcodes` AS ZIPENDNAME
ON ZIPEND.zip_code = CAST(ZIPENDNAME.zip AS STRING)
WHERE
-- This takes the weather data from one weather station
WEA.wban = '94728' -- NEW YORK CENTRAL PARK
-- Use data from 2014 and 2015
AND EXTRACT(YEAR FROM DATE(TRI.starttime)) BETWEEN 2014 AND 2015
GROUP BY
1,
```

```
2,
 3,
 4,
 5,
 6,
 7,
 8,
 9,
 10,
 11,
 12,
 13
```

SQL query that captured data from just the summer season:

```
SELECT
TRI.usertype,
TRI.start_station_longitude,
TRI.start_station_latitude,
TRI.end_station_longitude,
TRI.end_station_latitude,
ZIPSTART.zip_code AS zip_code_start,
ZIPSTARTNAME.borough borough_start,
ZIPSTARTNAME.neighborhood AS neighborhood_start,
ZIPEND.zip_code AS zip_code_end,
 ZIPENDNAME.borough borough_end,
 ZIPENDNAME.neighborhood AS neighborhood_end,
--Since we're using trips from 2014 and 2015, we will add 5 years to make it look recent
 DATE_ADD(DATE(TRI.starttime), INTERVAL 5 YEAR) AS start_day,
DATE_ADD(DATE(TRI.stoptime), INTERVAL 5 YEAR) AS stop_day,
 WEA.temp AS day_mean_temperature, -- Mean temp
WEA.wdsp AS day_mean_wind_speed, -- Mean wind speed
 WEA.prcp day_total_precipitation, -- Total precipitation
 -- We will group trips into 10 minute intervals, which also reduces the number of rows
ROUND(CAST(TRI.tripduration / 60 AS INT64), -1) AS trip_minutes,
TRI.bikeid
FROM
`bigquery-public-data.new_york_citibike.citibike_trips` AS TRI
INNER JOIN
`bigquery-public-data.geo_us_boundaries.zip_codes` ZIPSTART
ON ST_WITHIN(
ST_GEOGPOINT(TRI.start_station_longitude, TRI.start_station_latitude),
ZIPSTART.zip_code_geom)
INNER JOIN
`bigquery-public-data.geo_us_boundaries.zip_codes` ZIPEND
ON ST_WITHIN(
ST_GEOGPOINT(TRI.end_station_longitude, TRI.end_station_latitude),
ZIPEND.zip_code_geom)
INNER JOIN
```

```
-- https://pantheon.corp.google.com/bigquery?p=bigquery-public-data&d=noaa_gsod
`bigquery-public-data.noaa_gsod.gsod20*` AS WEA
ON PARSE_DATE("%Y%m%d", CONCAT(WEA.year, WEA.mo, WEA.da)) = DATE(TRI.starttime)
INNER JOIN
-- Note! Add your zipcode table name, enclosed in backticks: `example_table`
`legalbi.sandbox.zipcodes` AS ZIPSTARTNAME
ON ZIPSTART.zip_code = CAST(ZIPSTARTNAME.zip AS STRING)
INNER JOIN
-- Note! Add your zipcode table name below, enclosed in backticks: `example_table`
 `legalbi.sandbox.zipcodes` AS ZIPENDNAME
  ON ZIPEND.zip_code = CAST(ZIPENDNAME.zip AS STRING)
WHERE
-- Take the weather from one weather station
 WEA.wban = '94728' -- NEW YORK CENTRAL PARK
-- Use data for three summer months
AND DATE(TRI.starttime) BETWEEN DATE('2015-07-01') AND DATE('2015-09-30')
```

DASHBOARD

Link to view dashboards click here

Summer trends

The first tab of the dashboard is a map of seasonal trends of bike trips in each of the New York boroughs. The largest map shows each of the boroughs. The table compares the number of trips and average trip duration for customers and subscribers in each neighborhood. Three smaller maps focus on July, August, and September: the three months with the highest bike traffic.

This map features several filters to focus on specific bike IDs, user types, metrics, months, starting neighborhoods, and ending neighborhoods. Using any of these filters or clicking on a borough in one of the maps updates the table and maps to focus on your selection in greater detail.

Seasonality

The second tab of the dashboard focuses on seasonality, or trends throughout the year, with the Trip Totals chart and the Trip Counts by Starting Neighborhood table.

Trip Totals chart

The Trip Totals chart visualizes the total number of bike trips taken throughout 2019 and 2020, with a distinction between customers and subscribers. This chart shows that subscribers make up a significantly larger portion of Cyclistic's users than regular customers. It also shows that there are far more users in warmer months (May–October) than there are in colder months. This makes sense considering that people are less likely to ride bicycles in colder weather.

This chart was made by putting the Start Day (aggregated by month) in the columns field, the sum of Trip Counts in the rows filed, and UserType as color assignment.

Trip Counts by Starting Neighborhood table

The Trip Counts by Starting Neighborhood table lists the total number of bike trips started in each neighborhood in each month of 2019 and 2020. It is organized by zip code, borough, and neighborhood. It also uses a color gradient to emphasize the highest and lowest counts of monthly trips. The greater the number of trips, the darker the value is in the table. It also uses light text on the darker values to ensure that the table is readable and accessible.

Because the starting location is more indicative of where users look for a bike, it is more important to emphasize starting location when determining where to advertise. The most active stations are in the Lower East Side and the Chelsea and Clinton neighborhoods. The most active months are from May to October.

This table was created by putting the Start Day dimension (aggregated by Year and Month) in the Columns field, then the Borough Start and Neighborhood Start dimensions in the Rows field. Then, the color and labels can be set by putting the sum of the Trip Count measure into the Color and Label fields.

Top Trips

The third and final tab of the dashboard is a comparison of the total number of trip minutes by starting neighborhood and ending neighborhood for both customers and subscribers. The two charts are horizontal stacked bar graphs that are ordered from highest to lowest number of minutes (between customers and subscribers combined).

These charts lend insight into which locations users are most willing to travel long distances to. The charts show that the Lower East Side and Chelsea and Clinton neighborhoods have the highest total trip minutes for both start and end stations.

To make the starting neighborhood chart, you can put the sum of Trip Minutes in the columns field, and then the Zip Code Start, Neighborhood Start, and Borough Start dimensions in the rows field. Then, set UserType as the color assignments. To make the ending neighborhood chart, complete the same steps but use the Zip Code End, Neighborhood End, and Borough End dimensions.