Google Business Intelligence Professional Certificate Case Study

Ismoil Ismoilov

2024-09-30

Case Study: Cyclistic

A company's success hinges on its ability to quickly identify potential issues and seize opportunities. Business intelligence (BI) plays a pivotal role in transforming vast amounts of data into actionable insights. By automating processes and information channels, BI empowers decision-makers with timely and relevant information.

Let's explore a real-world example from Google Coursera Data Analytics. Using a public dataset from Google BigQuery, we'll walk through the steps involved in analyzing a business, from the initial meeting to the final delivery of a dashboard that addresses specific business needs

Background Cyclistic, a bike-sharing service, has partnered with New York City to provide bikes for easy travel. To understand customer usage patterns and inform their business plan, Cyclistic's Customer Growth Team wants to analyze millions of ride data points.

The data includes trip start/end times and locations, bike IDs, and customer types (one-time or subscriber). The team seeks a dashboard that summarizes key insights, such as customer demand at different stations. This data-driven approach will help Cyclistic develop a more successful business plan.

Project goal: Attract more riders

- Gain insights into customer preferences, identify the key elements of a successful product, and evaluate how expanding new stations can address demand in various geographic regions.
- Analyze the usage patterns of the current bike fleet.
- Leverage customer usage data to guide decisions on expanding station networks.
- Customer growth efforts require understanding the behavior of different user groups (subscribers vs. casual users). A comprehensive study of a large and diverse user base will ensure a well-rounded representation across different locations and activity levels.
- Consider that inclement weather may impact bike usage, and this should be clearly displayed in the dashboard for accurate insights.

Project Organization After evaluating the scenario and project background provided by stakeholders, we develop our strategy, establish a timeline, and record any questions related to the project. Since BI projects are complex, staying organized from the start to the finish is crucial for success. A good way to manage this is by creating detailed BI documentation that captures the overall project requirements, ensures organization, and drives impactful results within the organization.

Key documents include the Stakeholder Requirements Document, Project Requirements Document, and Strategy Document. Each document builds upon the last, forming three phases of our project planning process.

Stakeholder Requirements Document This document is essential as it allows us to pinpoint the stakeholders' needs and confirm that our project addresses those needs. By gathering these requirements early in the process, we can prevent expensive modifications later on. I have provided a stakeholder requirements document **here** for your reference.

Project Requirements Document After establishing the stakeholder requirements, we can start to define the project requirements necessary to fulfill those needs. By identifying these requirements at an early stage, we can ensure our project's success and that it addresses the stakeholders' needs. You can find a project requirements document **here.**

Strategy Document Lastly, we will develop a Strategy Document for our project. This document serves as a collaborative space to align with stakeholders regarding project deliverables. We will collaborate to define details about dashboard functionality, as well as the related metrics and charts. You can find this document **here**.

Data Preparation The next phase of the project involves preparing the data. You may be familiar with the ETL concept. In the context of BI, ETL stands for Extract, Transform, Load. So, what does it entail? ETL is a data warehousing process that includes extracting data from various sources, transforming it to meet operational requirements, and loading it into a target system, such as a data warehouse. This step is crucial for integrating data from different sources and ensuring it is usable for our dashboard.

For this project, we will utilize three public datasets that can be found in the public data section of the Explorer pane in our Google Cloud Console:

- **NYC Citi Bike Trips**
- **Census Bureau US Boundaries**
- **GSOD from the National Oceanic and Atmospheric Administration**
- Additionally, we need to upload the zip code data from this **spreadsheet**.

Next, we will create a target table by combining the datasets. While we could use Google's DataFlow for job scheduling and fetching into our target table, I will explain how to retrieve a dataset using Google BigQuery based on the established requirements. This process involves querying the data using specific parameters and then loading the results into our target table for the next phase of the project.

1. Execute the following SQL query to generate a summary table for the entire year, and then save the results as a BigQuery table:

```
##
## SELECT
##   TRI.usertype,
##   ZIPSTART.zip_code AS zip_code_start,
##   ZIPSTARTNAME.borough AS borough_start,
##   ZIPSTARTNAME.neighborhood AS neighborhood_start,
##   ZIPEND.zip_code AS zip_code_end,
##   ZIPENDNAME.borough AS 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,
##   WEA.wdsp AS day_mean_wind_speed,
##   WEA.prcp AS day_total_precipitation,
```

```
##   ROUND(TRI.tripduration / 60, -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', FORMAT('%04d%02d%02d', WEA.year, WEA.mo, WEA.da)) = DATE(TRI.starttime)
## INNER JOIN
##   `your_project.your_dataset.zipcodes` AS ZIPSTARTNAME
## ON ZIPSTART.zip_code = CAST(ZIPSTARTNAME.zip AS STRING)
## INNER JOIN
##   `your_project.your_dataset.zipcodes` AS ZIPENDNAME
## ON ZIPEND.zip_code = CAST(ZIPENDNAME.zip AS STRING)
## WHERE
##   WEA.wban = '94728'
##   AND EXTRACT(YEAR FROM DATE(TRI.starttime)) BETWEEN 2014 AND 2015
## GROUP BY
##   TRI.usertype,
##   ZIPSTART.zip_code,
##   ZIPSTARTNAME.borough,
##   ZIPSTARTNAME.neighborhood,
##   ZIPEND.zip_code,
##   ZIPENDNAME.borough,
##   ZIPENDNAME.neighborhood,
##   start_day,
##   stop_day,
##   WEA.temp,
##   WEA.wdsp,
##   WEA.prcp,
##   trip_minutes;
```

2. Furthermore, we need to run a query that retrieves data specifically from the summer season (July, August, and September only), and then save the results as a BigQuery table.

```
##
## 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 AS borough_start,
##   ZIPSTARTNAME.neighborhood AS neighborhood_start,
```

```
##   ZIPEND.zip_code AS zip_code_end,
##   ZIPENDNAME.borough AS borough_end,
##   ZIPENDNAME.neighborhood AS neighborhood_end
## 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
##   `legalbi.sandbox.zipcodes` AS ZIPSTARTNAME
## ON ZIPSTART.zip_code = CAST(ZIPSTARTNAME.zip AS STRING)
## INNER JOIN
##   `legalbi.sandbox.zipcodes` AS ZIPENDNAME
## ON ZIPEND.zip_code = CAST(ZIPENDNAME.zip AS STRING)
## WHERE
##   WEA.wban = '94728'
##   AND DATE(TRI.starttime) BETWEEN DATE('2015-07-01') AND DATE('2015-09-30');
```

As we prepare to develop the system that will supply data for our reports and dashboard, it's important to consider the various types of users, their needs, their inquiries, and how they currently utilize the data. This approach ensures that we create a solution that effectively serves everyone involved.

Data Preparation We have now reached the final stage of the BI project. After securing the required data resources and obtaining stakeholder approval to utilize them, it's advisable to create a mockup design first. This approach allows stakeholders to visualize the layout intended for the final project, reducing the likelihood of major changes once we start developing the dashboard with visualization software. Here's an example of a mockup I created, which has received approval from stakeholders (Figures 1-3):

Figure 1: Summer Trends

It's crucial for us to seek feedback on the mockup design from stakeholders. We should also explain our choices regarding the charts and color schemes, as these elements are part of the project planning documentation.

Figure 2: YoY Trends

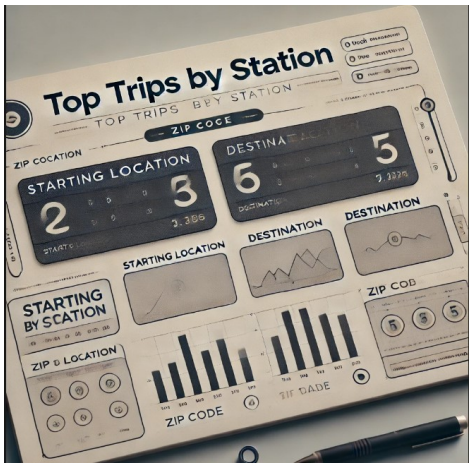


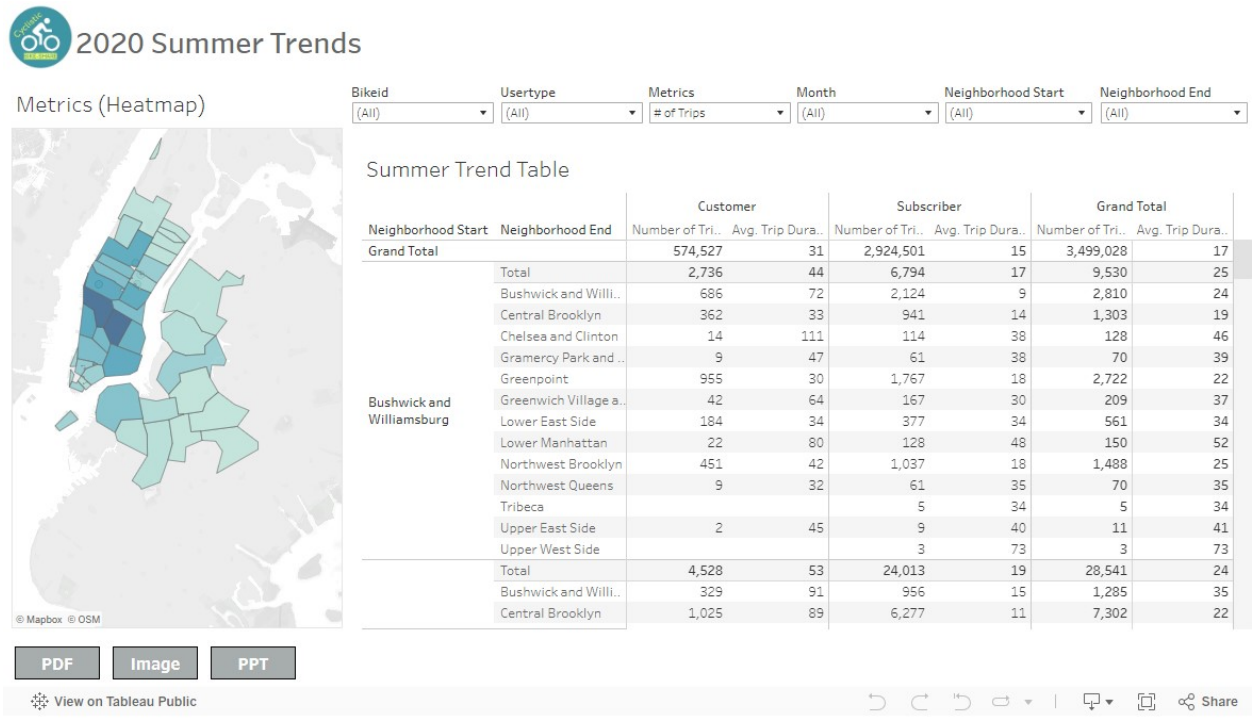
Figure 3: Top Trips by Station

Another key aspect is optimizing the loading speed of our dashboard. Since we have a substantial amount of data from previous retrievals, there are several strategies we can implement to enhance the dashboard's efficiency. This includes pre-aggregating and joining data, similar to what we did with the Summer Summary table.

Final output Now we can link our two target tables and visualizations using Tableau for the final dashboards.

1. *Summer Trends*

The first dashboard (Figure 4) features a map with interactive metric selection, including the number of trips, average trip duration, average temperature, average wind speed, and average precipitation for each borough. Additionally, there's a table that compares the number of trips and average trip duration for customers and subscribers in each neighborhood. Several filters are available to narrow down specific bike IDs, user types, metrics, months, starting neighborhoods, and ending neighborhoods. The map and table interact with each other, allowing for detailed insights. Currently, the data is limited to July, August, and September 2020, as specified by the stakeholders' requirements.



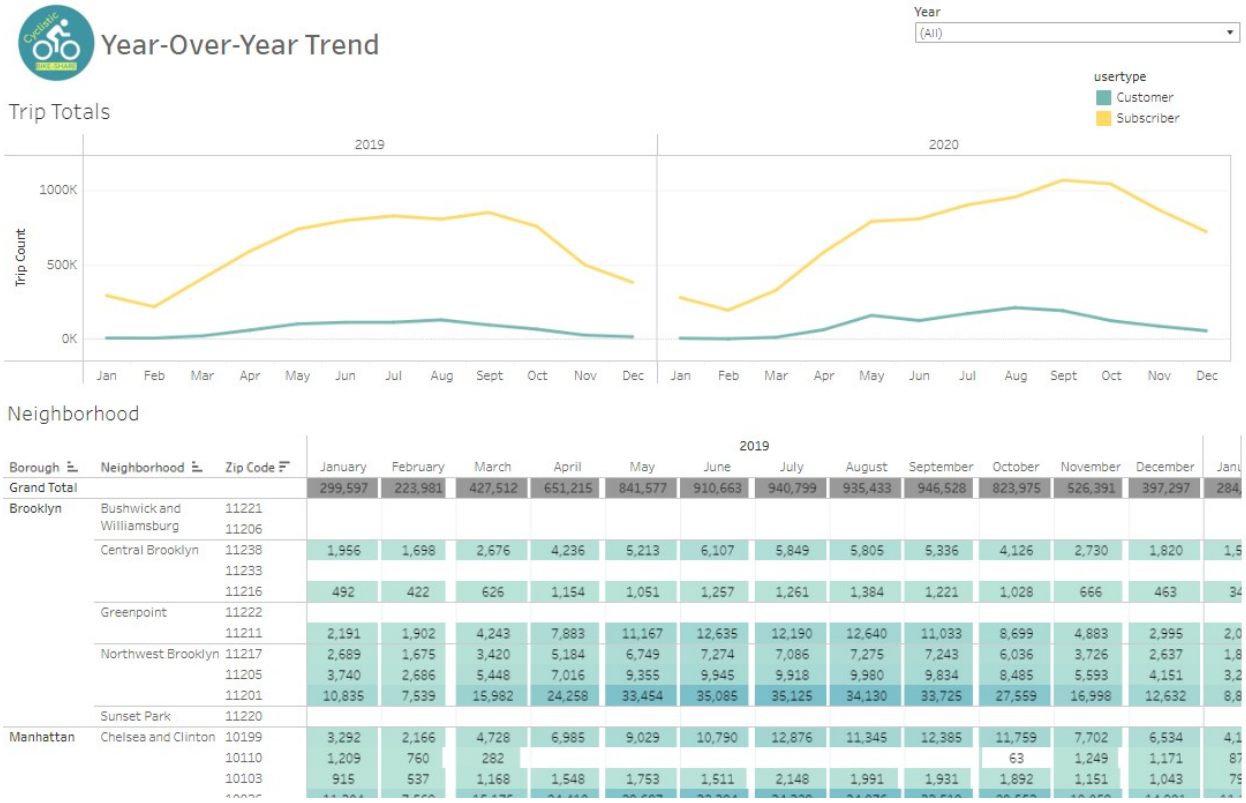
Neighborhood Start	Neighborhood End	Customer		Subscriber		Grand Total	
		Number of Tri..	Avg. Trip Dura..	Number of Tri..	Avg. Trip Dura..	Number of Tri..	Avg. Trip Dura..
Grand Total		574,527	31	2,924,501	15	3,499,028	17
Bushwick and Williamsburg	Total	2,736	44	6,794	17	9,530	25
	Bushwick and Willi..	686	72	2,124	9	2,810	24
	Central Brooklyn	362	33	941	14	1,303	19
	Chelsea and Clinton	14	111	114	38	128	46
	Gramercy Park and ..	9	47	61	38	70	39
	Greenpoint	955	30	1,767	18	2,722	22
	Greenwich Village a..	42	64	167	30	209	37
	Lower East Side	184	34	377	34	561	34
	Lower Manhattan	22	80	128	48	150	52
	Northwest Brooklyn	451	42	1,037	18	1,488	25
	Northwest Queens	9	32	61	35	70	35
	Tribeca			5	34	5	34
	Upper East Side	2	45	9	40	11	41
	Upper West Side			3	73	3	73
	Total	4,528	53	24,013	19	28,541	24
	Bushwick and Willi..	329	91	956	15	1,285	35
	Central Brooklyn	1,025	89	6,277	11	7,302	22

Figure 4: Cyclistic 2020 Summer Trends

The chart reveals that, despite varying weather conditions such as wind speed, precipitation, and high temperatures during the summer, Chelsea and Clinton consistently record the highest number of trips. However, Central Brooklyn stands out with the longest average trip duration.

2. *Year-over-Year Trends*

The second dashboard (Figure 5) emphasizes seasonality and trends throughout the year, featuring the Trip Totals chart categorized by user types and the Trip Counts by Starting Neighborhood table. There's also a minor interaction between the total trips chart and the selected user types, along with the seasonality table.



Borough	Neighborhood	Zip Code	2019 January	February	March	April	May	June	July	August	September	October	November	December	Janu
Grand Total			299,597	223,981	427,512	651,215	841,577	910,663	940,799	935,433	946,528	823,975	526,391	397,297	284
Brooklyn	Bushwick and Williamsburg	11221													
		11206													
	Central Brooklyn	11238	1,956	1,698	2,676	4,236	5,213	6,107	5,849	5,805	5,336	4,126	2,730	1,820	1,5
		11233													
		11216	492	422	626	1,154	1,051	1,257	1,261	1,384	1,221	1,028	666	463	34
	Greenpoint	11222													
		11211	2,191	1,902	4,243	7,883	11,167	12,635	12,190	12,640	11,033	8,699	4,883	2,995	2,0
	Northwest Brooklyn	11217	2,689	1,675	3,420	5,184	6,749	7,274	7,086	7,275	7,243	6,036	3,726	2,637	1,8
		11205	3,740	2,686	5,448	7,016	9,355	9,945	9,918	9,980	9,834	8,485	5,593	4,151	3,2
		11201	10,835	7,539	15,982	24,258	33,454	35,085	35,125	34,130	33,725	27,559	16,998	12,632	8,8
	Sunset Park	11220													
Manhattan	Chelsea and Clinton	10199	3,292	2,166	4,728	6,985	9,029	10,790	12,876	11,345	12,385	11,759	7,702	6,534	4,1
		10110	1,209	760	282							63	1,249	1,171	87
		10103	915	537	1,168	1,548	1,753	1,511	2,148	1,991	1,931	1,892	1,151	1,043	79

Figure 5: Cyclistic YoY Trends

Users with a subscription type constitute a significantly larger portion compared to regular customers. Throughout the year, the seasonal trend shows an increase during the summer months, particularly from July to October.

3. *Top Trips by Station*

The final dashboard (Figure 6) presents a comparison of the total number of trip minutes by starting and ending neighborhoods for both customers and subscribers. It consists of two horizontal stacked bar graphs, organized from highest to lowest total minutes for the combined data of customers and subscribers.

The charts indicate that the Lower East Side and Chelsea and Clinton neighborhoods have the highest total trip minutes for both starting and ending stations.

At the conclusion of the next BI project, we will present our findings to the stakeholders, but this is not the end of the project. We are expected to continuously iterate on the overall performance based on the progress we've made. Our goal is to optimize data loading performance to ensure that the dashboard we've created offers a seamless experience for users. Additionally, it's important to document how to use the dashboard for users.

Finally, please refer to the link provided below to access all relevant information and details about the project, which will give you a comprehensive understanding of its progress and outcomes.

Cyclistic Summer Trends

Cyclistic YoY Trend

Cyclistic Top Trips by Station

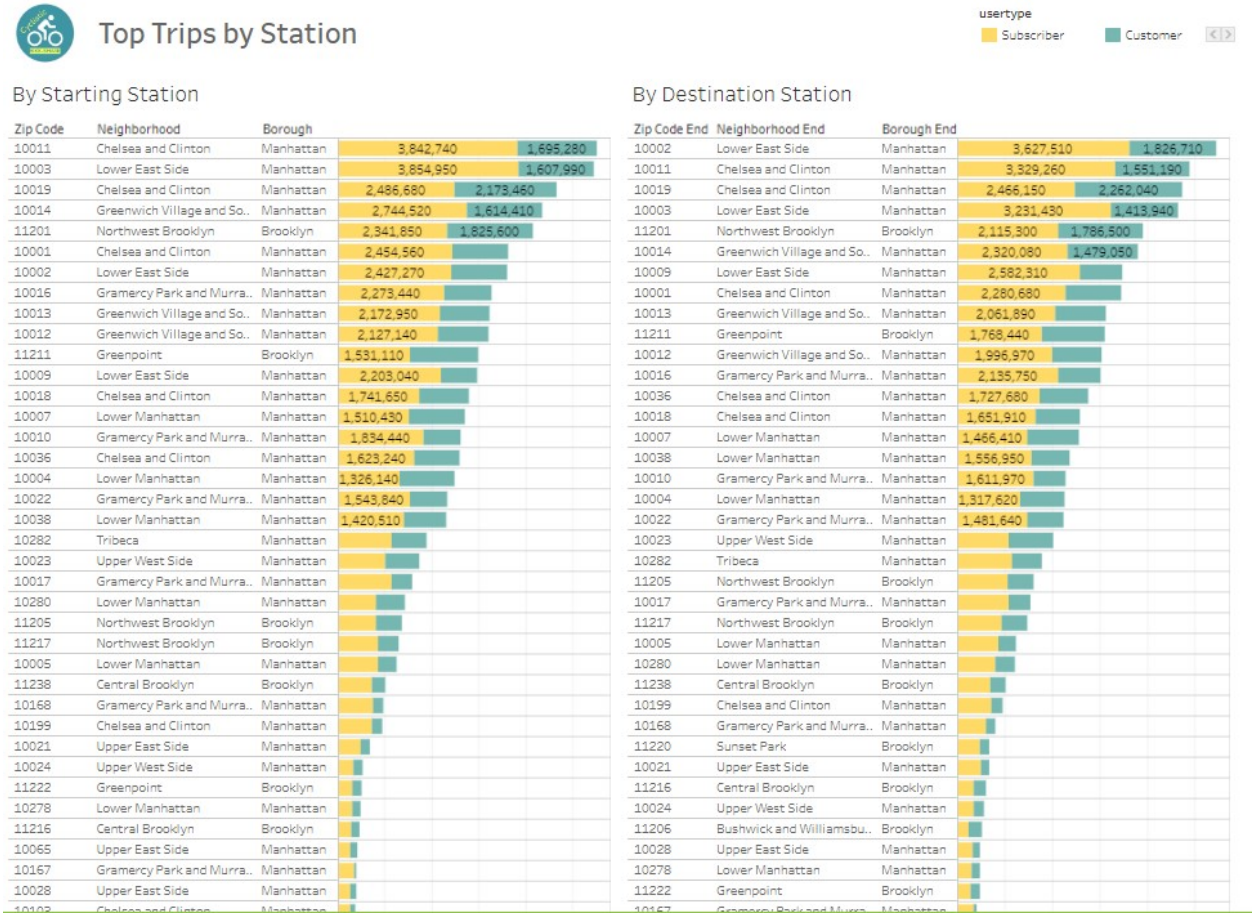


Figure 6: Cyclistic Top Trips by Station