NYC Road Preparation Warehouse

CIS 4400
Section PTRA

Nicole Arugay
manicole.arugay@baruchmail.cuny.edu
Stella Chung
stella.chung@baruchmail.cuny.edu
Richard Taveras
richard.taveras@baruchmail.cuny.edu
Rifat Khan
rifat.khan@baruchmail.cuny.edu

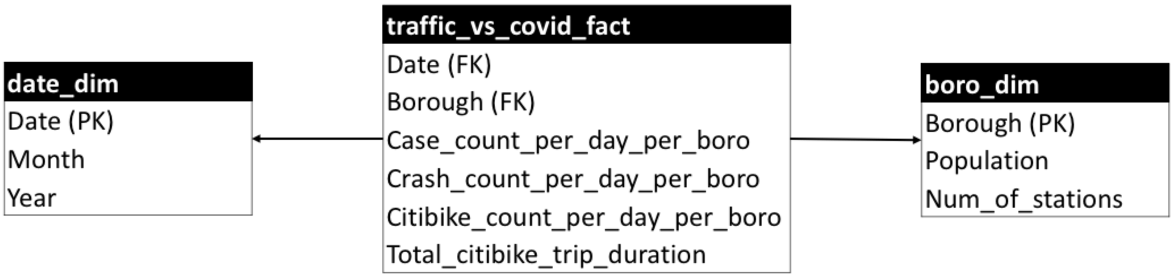
Description

The purpose of the warehouse is to answer questions and provide analysis that would be relevant in preparing motor vehicles on New York City roads as COVID-19 cases decrease over time. The warehouse includes COVID-19 case count data, Citi Bike ridership data, and motor vehicle collisions data. The correlation between these datasets could be useful in finding trends and providing analysis.

Problem Statement

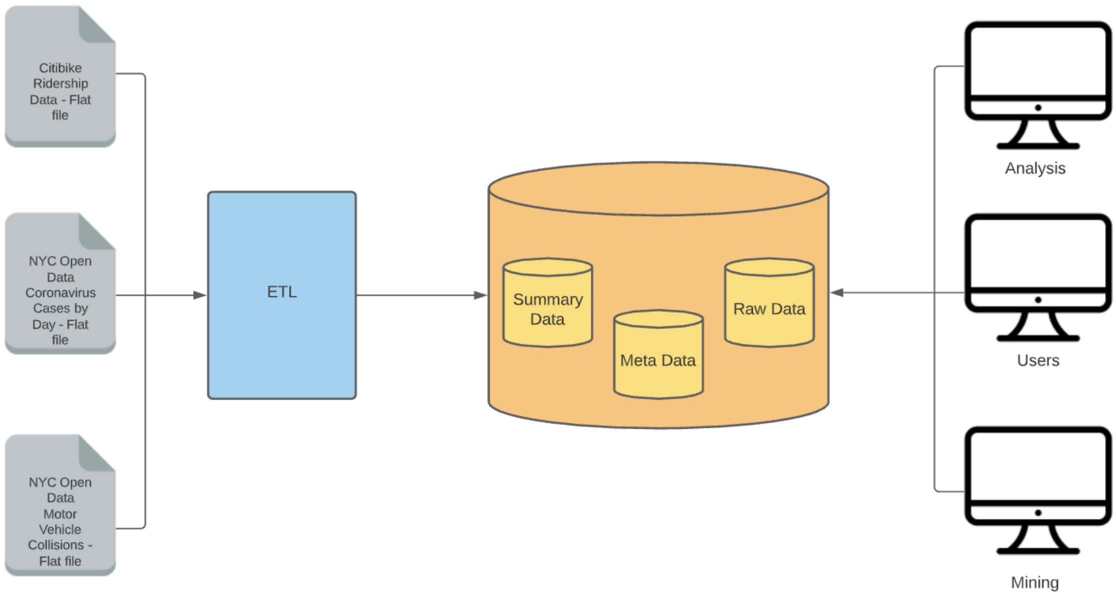
There is uncertainty in future traffic volume and motor vehicle collisions due to external factors related to the ongoing COVID-19 pandemic. The system will attempt to provide insight and answers to questions that will help bring a better understanding of COVID-19 case counts, Citi Bike ridership data, and motor vehicle collisions. Insights on the current relationship between motor vehicle collisions and the COVID-19 pandemic help predict the possible future relationship. Failure to understand and predict possible future scenarios can result in an unexpected increase in number of motor vehicle collisions.

Dimensional Model Diagram


date_dim
Date (PK)
Month
Year
traffic_vs_covid_fact
Date (FK)
Borough (FK)
Case_count_per_day_per_boro
Crash_count_per_day_per_boro
Citibike_count_per_day_per_boro
Total_citibike_trip_duration
boro_dim
Borough (PK)
Population
Num_of_stations

Grain: COVID case count, citibike ride count, motor vehicle crash count per day per borough

Architecture Diagram


Citibike Ridership Data - Flat file
NYC Open Data Coronavirus Cases by Day - Flat file
NYC Open Data Motor Vehicle Collisions - Flat file
ETL
Summary Data
Meta Data
Raw Data
Analysis
Users
Mining

Technical Decision Discussion

Data Source	Reason of Inclusion
Citi Bike Ridership Data	To make inferences of future Citi Bike ridership based on current ridership data and find trends and correlations given the other datasets
NYC Open Data Coronavirus Cases by Day	To make inferences on the current and future relationship between COVID-19 case counts and other vehicle data and find trends and correlations given the other datasets
NYC Open Data Motor Vehicle Collisions	To make inferences of the future number of motor vehicle collisions based on current motor vehicle collision data and find trends and correlations given the other datasets

Previously, our team wanted to use four datasets but decided to only keep the three remaining. We wanted to focus on using data that was available for the years significant COVID-19 case counts were present and the traffic volume data only had data for the years leading up to 2019. In hindsight, perhaps it would've been useful to keep the traffic volume dataset to extract analysis for when COVID-19 case counts reach zero, but since this data warehouse is COVID-19 focused, we ultimately thought the three datasets would suffice.

For the extraction and transformation process, our team originally wanted to use Colab so we could all contribute to this part of the ETL process in a seamless way. We wanted to build the schema and load into BigQuery using Colab. Although the sharing feature on Colab was useful, we found that it was easier for us to simply share our Jupyter Notebooks and combine code into one Python file; this was because most of us were more familiar with Jupyter Notebook. We also found it difficult to connect our Colab notebook to BigQuery which was the first reason as to why we switched to Jupyter Notebook.

For the loading process, our team originally wanted to use a localhost MySQL due to our difficulty connecting our Colab notebook to BigQuery. Ultimately, we wanted to be able to share the data warehouse easier among the group - due to this, we figured out that we can simply load our transformed datasets directly to BigQuery which led to our use of the software as the storage for our data warehouse.

Finally, Tableau was used to create data visualizations due to the team's familiarity with the software.

Ethical Issues

Possible concerns include the collection of data. More specifically, there are questions around whether Citi Bike riders and those counted in the COVID-19 case count dataset consented to having their data collected. It would also be surprising if there were any blatant consent decisions made for the collection of the motor vehicle collision data. In terms of the COVID-19 case count dataset, there are no unique id's for each row that was made public. There are however, unique identifiers present in the Citi Bike and motor vehicle collisions datasets that were made public - this could potentially be an issue if it is traceable to any individual person and if consent was not obvious. If the unique identifiers were included in the data warehouse, one example scenario could be that those involved in a motor vehicle collision accident are recognized and are then exposed to litigation issues involving the accident. The ability to possibly identify individuals through the Citi Bike ridership and motor vehicle collision dataset is a privacy concern; one way to eliminate this issue is to simply keep the data that is useful to our data warehouse and get rid of the unique identifiers.

Technical Documentation of ETL Processes

The purpose of the code is to extract and transform the datasets, COVID-19 case counts, motor vehicle collisions, and Citi Bike ridership, to ultimately load the transformed datasets manually into BigQuery which houses the data warehouse. The schema created in this code ultimately led to the creation of the dimension tables and fact table used in the data warehouse.

Link to full code with comments made shareable on Colab

- https://colab.research.google.com/drive/1xRiOdstCOUpqKdubicmWhyISqzIJ_rXM?usp=sharing

Public links used to download original datasets

- https://www.citibikenyc.com/system-data
  - Download each month from 02/2020-04/2021
- https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95
- https://github.com/nychealth/coronavirus-data/tree/master/archive
  - Download the csv titled 'boro.csv'

Python installation

Any IDE that supports Python should suffice. This code was written on Jupyter Notebook due to ease of use.

Running the scripts

You will need to install the geopy package if you haven't before. To do so, simply type the following in your IDE then run it:

```
Pip install geopy
```

Follow the instructions on your IDE to ensure the package will be used. In Jupyter Notebook, it may say to restart the kernel.

Geopy package

The geopy package was installed to be able to add a borough column to the Citi Bike ridership dataset. Adding a borough column to this dataset made for ease of comparison to the other tables in our schema. In order to get a borough column, we had to extract the borough from the addresses of the given coordinates of the various start stations. The following block of code may take around 10-20 minutes to complete due to the size of the file and slow iterating process.

```
locations=[]
for index, row in citibike_stations.iterrows():
      locations.append(reverse("{}, {}".format(row['latitude'],\
      row['longitude'])).raw['address'])
pd.DataFrame(locations[:10])
```

Downloading the transformed datasets

In Jupyter Notebook, running the code will automatically download the transformed datasets into csv format, which you can find in the folder in which the ipynb file with the code is placed in.

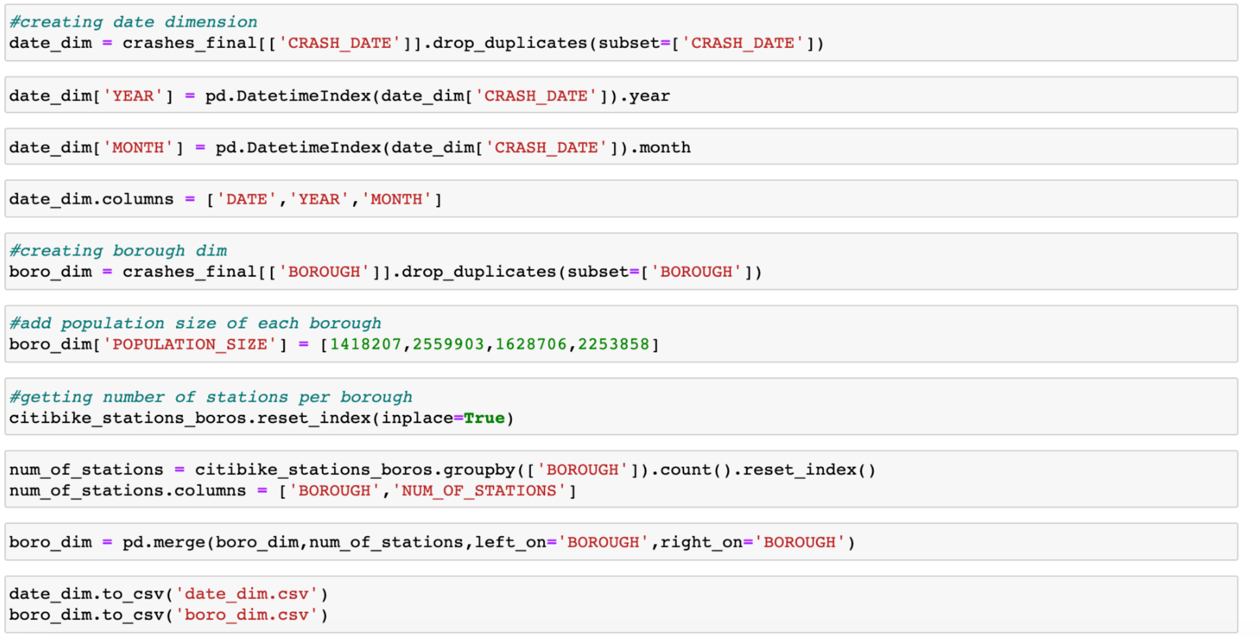
BigQuery installation

After creating an account on Google, go to https://cloud.google.com/bigquery and follow the instructions on the site. You may need to set up a billing process. After creating a project in the Google Cloud console, create a dataset. After creating a dataset, manually load the transformed datasets one by one by adding tables to that dataset. You are then ready to query the transformed datasets.

Getting in touch

For collaboration purposes, feel free to email any member of the team as seen in the cover letter of this document.

Final Schema



```python
#creating date dimension
date_dim = crashes_final[['CRASH_DATE']].drop_duplicates(subset=['CRASH_DATE'])
```

```python
date_dim['YEAR'] = pd.DatetimeIndex(date_dim['CRASH_DATE']).year
```

```python
date_dim['MONTH'] = pd.DatetimeIndex(date_dim['CRASH_DATE']).month
```

```python
date_dim.columns = ['DATE','YEAR','MONTH']
```

```python
#creating borough dim
boro_dim = crashes_final[['BOROUGH']].drop_duplicates(subset=['BOROUGH'])
```

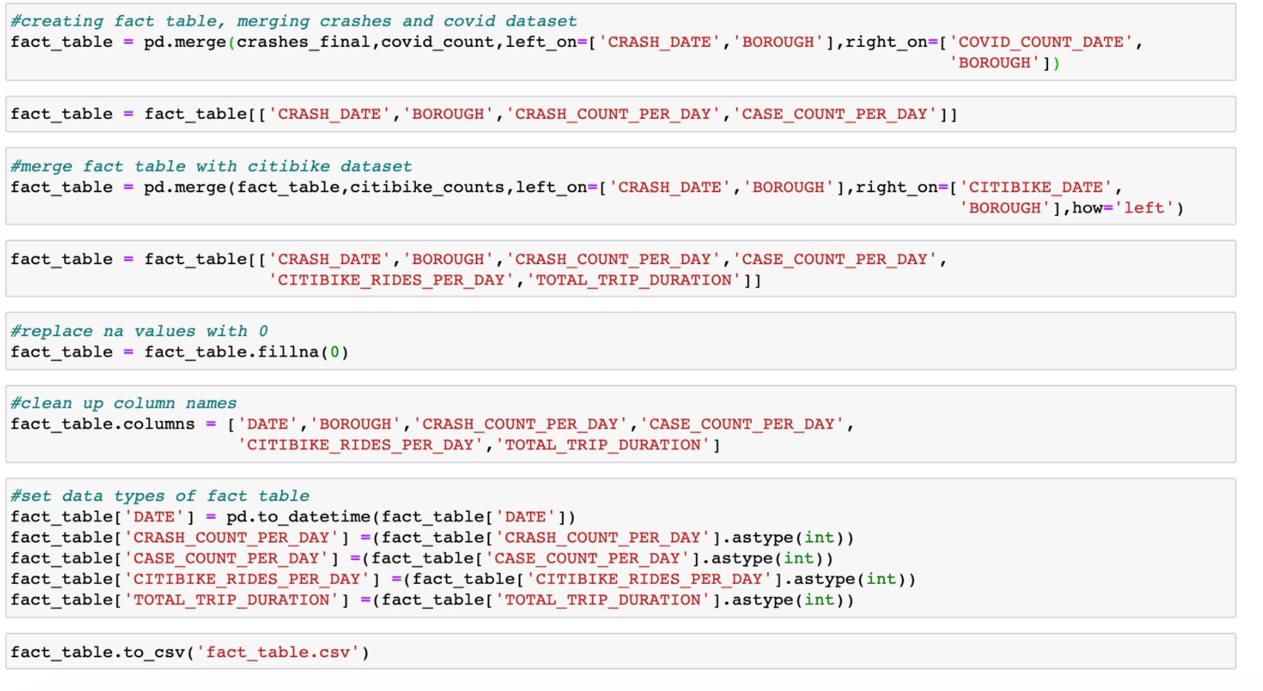
```python
#add population size of each borough
boro_dim['POPULATION_SIZE'] = [1418207,2559903,1628706,2253858]
```

```python
#getting number of stations per borough
citibike_stations_boros.reset_index(inplace=True)
```

```python
num_of_stations = citibike_stations_boros.groupby(['BOROUGH']).count().reset_index()
num_of_stations.columns = ['BOROUGH','NUM_OF_STATIONS']
```

```python
boro_dim = pd.merge(boro_dim,num_of_stations,left_on='BOROUGH',right_on='BOROUGH')
```

```python
date_dim.to_csv('date_dim.csv')
boro_dim.to_csv('boro_dim.csv')
```



```python
#creating fact table, merging crashes and covid dataset
fact_table = pd.merge(crashes_final,covid_count,left_on=['CRASH_DATE','BOROUGH'],right_on=['COVID_COUNT_DATE',
                                                                                           'BOROUGH'])
```

```python
fact_table = fact_table[['CRASH_DATE','BOROUGH','CRASH_COUNT_PER_DAY','CASE_COUNT_PER_DAY']]
```

```python
#merge fact table with citibike dataset
fact_table = pd.merge(fact_table,citibike_counts,left_on=['CRASH_DATE','BOROUGH'],right_on=['CITIBIKE_DATE',
                                                                                            'BOROUGH'],how='left')
```

```python
fact_table = fact_table[['CRASH_DATE','BOROUGH','CRASH_COUNT_PER_DAY','CASE_COUNT_PER_DAY',
                         'CITIBIKE_RIDES_PER_DAY','TOTAL_TRIP_DURATION']]
```

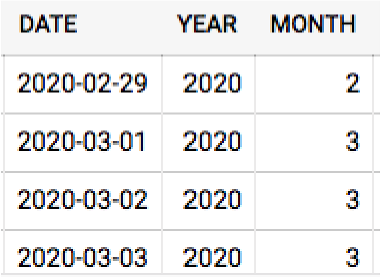
```python
#replace na values with 0
fact_table = fact_table.fillna(0)
```

```python
#clean up column names
fact_table.columns = ['DATE','BOROUGH','CRASH_COUNT_PER_DAY','CASE_COUNT_PER_DAY',
                      'CITIBIKE_RIDES_PER_DAY','TOTAL_TRIP_DURATION']
```

```python
#set data types of fact table
fact_table['DATE'] = pd.to_datetime(fact_table['DATE'])
fact_table['CRASH_COUNT_PER_DAY'] =(fact_table['CRASH_COUNT_PER_DAY'].astype(int))
fact_table['CASE_COUNT_PER_DAY'] =(fact_table['CASE_COUNT_PER_DAY'].astype(int))
fact_table['CITIBIKE_RIDES_PER_DAY'] =(fact_table['CITIBIKE_RIDES_PER_DAY'].astype(int))
fact_table['TOTAL_TRIP_DURATION'] =(fact_table['TOTAL_TRIP_DURATION'].astype(int))
```

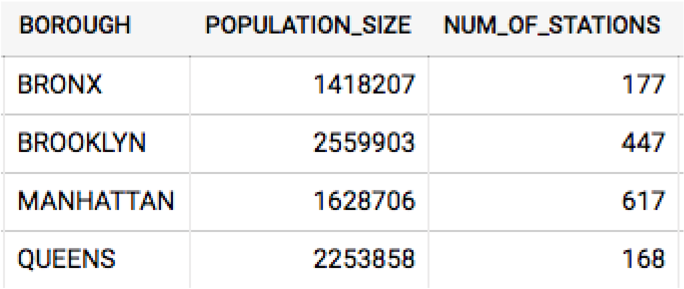
```python
fact_table.to_csv('fact_table.csv')
```

Date dim



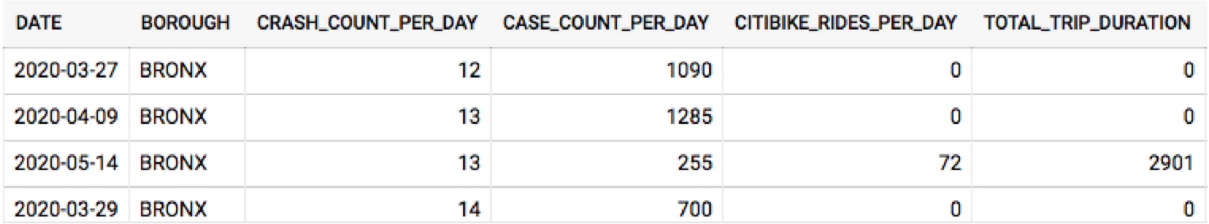
DATE	YEAR	MONTH
2020-02-29	2020	2
2020-03-01	2020	3
2020-03-02	2020	3
2020-03-03	2020	3

Borough dim



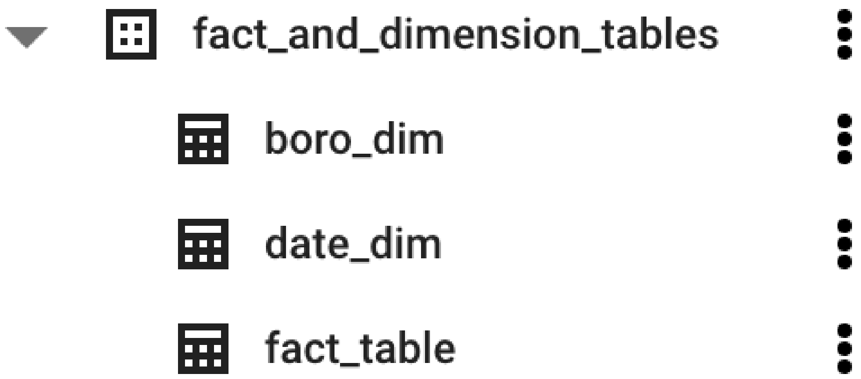
BOROUGH	POPULATION_SIZE	NUM_OF_STATIONS
BRONX	1418207	177
BROOKLYN	2559903	447
MANHATTAN	1628706	617
QUEENS	2253858	168

Fact table



DATE	BOROUGH	CRASH_COUNT_PER_DAY	CASE_COUNT_PER_DAY	CITIBIKE_RIDES_PER_DAY	TOTAL_TRIP_DURATION
2020-03-27	BRONX	12	1090	0	0
2020-04-09	BRONX	13	1285	0	0
2020-05-14	BRONX	13	255	72	2901
2020-03-29	BRONX	14	700	0	0

BigQuery



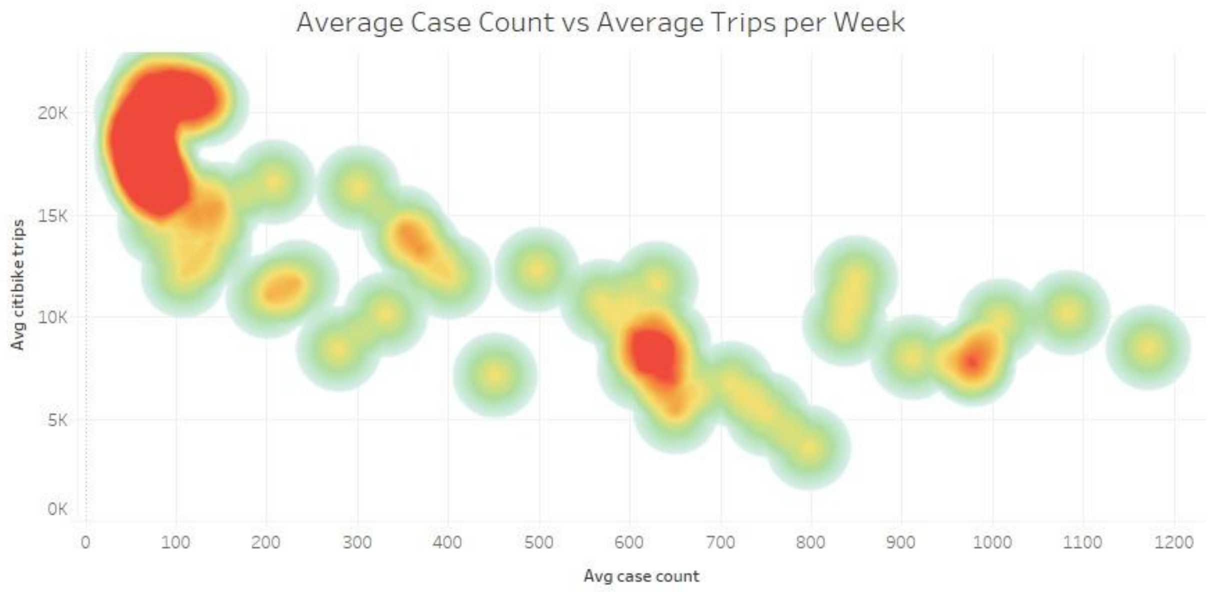
fact_and_dimension_tables

boro_dim

date_dim

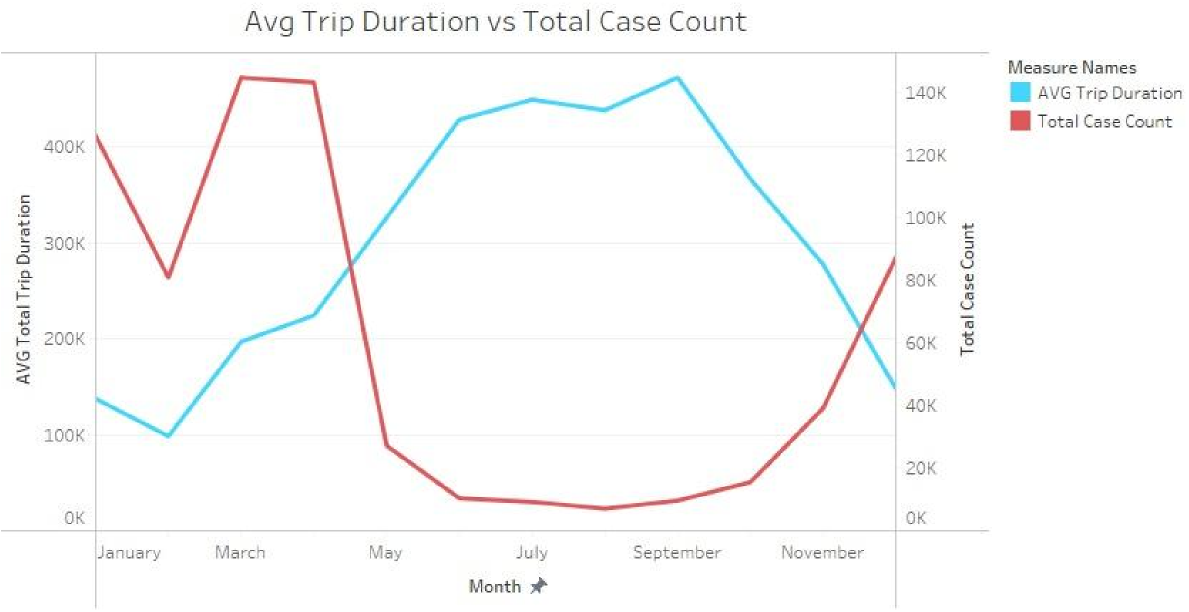
fact_table

Data Visualizations



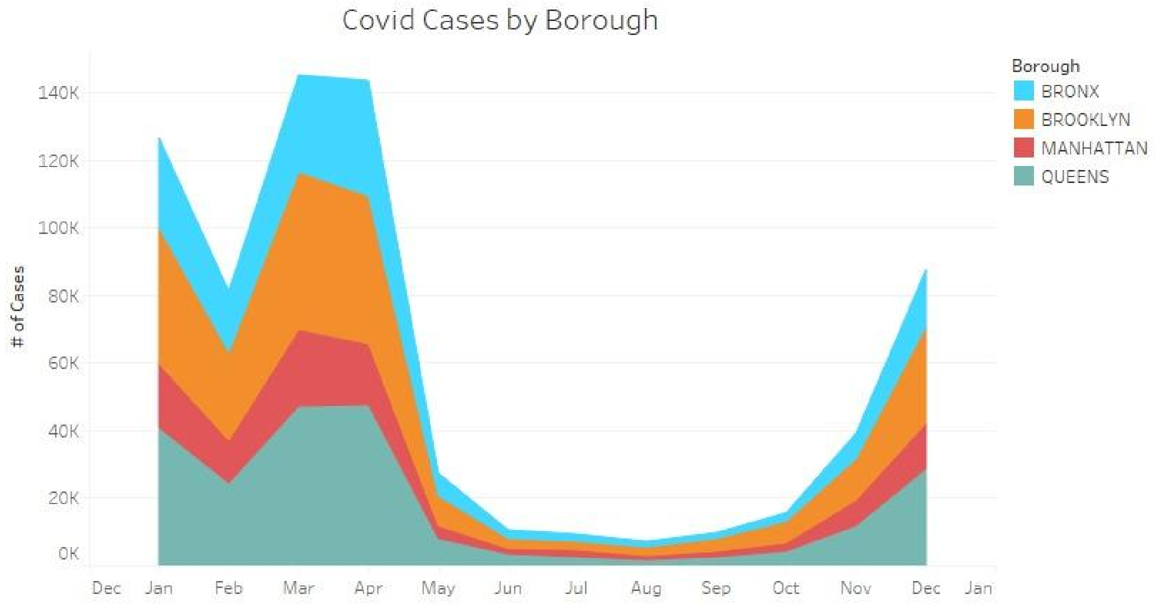
A heatmap correlation between average COVID-19 case counts and average Citi Bike trips per week across NYC. There seems to be a moderate negative correlation: the more cases of COVID-19, the less Citi Bike trips per week.

Steps:
-Created 2 calculated fields “Average Case Count” and “Average Trips”
-Placed “Average Case Count” onto the X axis
-Placed “Average Trips” onto the Y axis
-Placed “Date” dimension onto the Marks and adjusted the date to weeks
-Changed the mark type to “Density”
-Changed mark color to “Density Multi-Color” and adjusted the mark size to the maximum



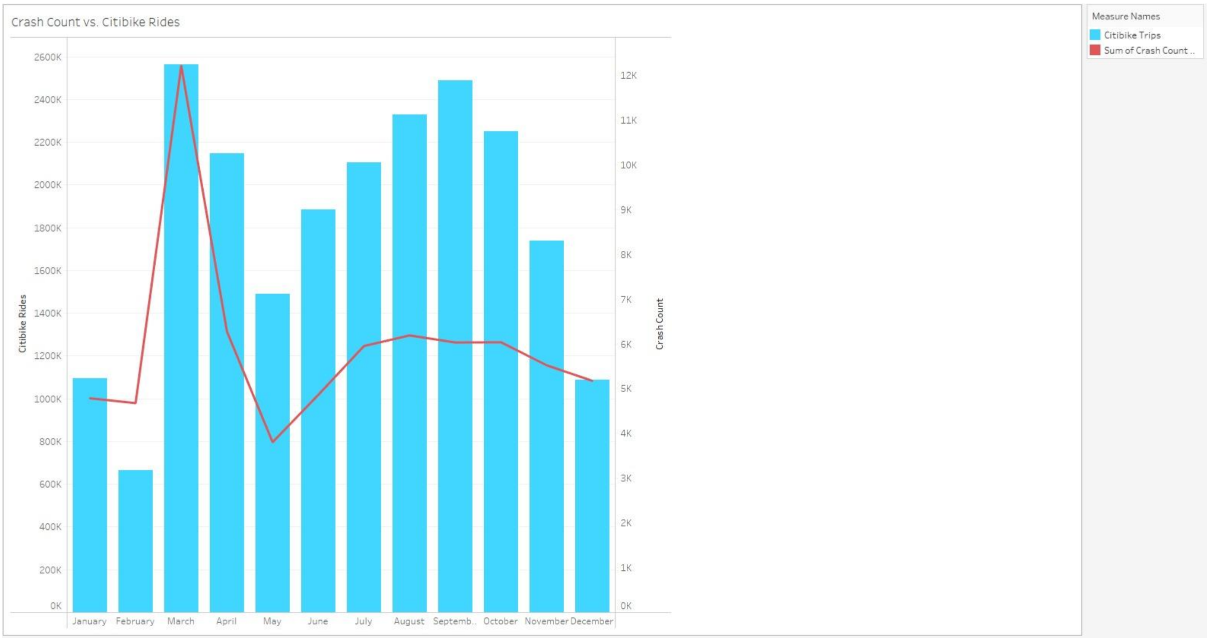
Dual line graph showing the relationship between average Citi Bike trip duration and total COVID-19 case counts. There seems to be an inverse relationship; trip duration increased as total case counts decreased.

```
Steps:
-Selected the "lines continuous" layout option
-Placed "date" dimension into columns
-Placed "Total Case Count" and "Average Trip Duration" into rows
-Right clicked X-axis and unmarked "Synchronize Axis" option
-Placed both measures into the marks
-Under marks selected "color" > "edit colors" and adjusted colors
```



A filled-in time series line graph showing COVID-19 cases within each borough. Manhattan seemed to have the least amount of COVID-19 cases among the boroughs due to its width on the graph. The case counts within each borough followed a similar pattern throughout the months which is to be expected. There was a peak of COVID-19 cases in March of 2020.

```
Steps:
-Placed the calculated measure "Sum of Case Count" into the columns
-Placed the "Date(Month)" dimensions into the rows
-Placed "Borough" dimension into the labels
-Changed the Mark type to "Area"
-Added filters for both "Date" and "Borough"
-Right clicked the filters and under "Apply to Worksheets" selected "All
using related data sources"
```



Bar and line graph of motor vehicle crashes and Citi Bike rides. There seems to be a moderate correlation between Citi Bike rides and motor vehicle collisions - the more Citi Bike trips the more motor vehicle collisions. There was an uptick of motor vehicle collisions during the months in which COVID-19 case counts were low.

Steps:
-Placed the calculated measures “Sum of Citibikes Trips” and “Sum of Crashes” into the rows
-Placed the date dimension into the columns
-Placed both measures into the marks
-Changed the mark type to “multiple”
-Changed the mark type for “Sum of Citibike Trips” to “bar”
-Changed the mark type for “Sum of Crashes” to “line”
-Edited the aliases for the measures from “Sum of Citibike Trips” and “Sum of Crashes” to “Citibike Trips” and “Crashes” respectively

Conclusion

Personal experience

My contributions included helping with the sourcing of the datasets, extracting of the datasets, helping with reverse geocoding on the Citi Bike ridership dataset, and helping to figure out which software to house our data warehouse in. Before using BigQuery, I was open to hosting the data warehouse in my localhost MySQL. The most difficult steps were figuring out where we were going to host the data warehouse and how to construct the dimensional tables and fact table given the datasets. During this project, I couldn't have imagined having to reverse geocode a dataset. When choosing the datasets and thinking about having to join them later on, I didn't realize that the latitude and longitude given would not suffice as keys that allow for joins. Thankfully, with the help of the professor, we realized we would have to find another way to create the tables from our dimensional model and thankfully we could do so using boroughs.

If I had to do this project over again, I would have perhaps spent more time figuring out how to be able to run a Python script on Google's Virtual Machine to be able to connect to the instance's MySQL so I could learn more about using the cloud.

Proposed benefits

What we hoped to accomplish was to find trends within our datasets to be able to give recommendations regarding steps to reduce overcrowding and motor vehicle collisions as COVID-19 cases decrease. Findings I was hoping we would find include boroughs where there were more motor vehicle collisions, if New Yorkers generally stopped going out as COVID-19 cases increased and so forth. Given more time, I would have wanted to do more SQL queries, perhaps utilizing analytical functions, to gain more insights about our datasets. I also would have wanted to add a traffic volume dataset and MTA ridership dataset if they had data from 2020 to present date.

Final comments

From our analysis, we definitely did see that New Yorkers have gone on less trips as COVID-19 cases increased. Motor vehicle collisions and COVID-19 case counts occurred the most in Brooklyn. If COVID-19 case counts decrease this summer, I believe there will be a spike in motor vehicle collisions and Citi Bike ridership. The NYC Department of Transportation should perhaps look into preventing collisions in Brooklyn, perhaps by implementing policies, analyzing certain locations where accidents happen the most, and possibly reducing the amount of Citi Bikes. If COVID-19 cases spike again, at least we can trust that New Yorkers will be more reluctant to go out.