Traffic Signal Conditions

Group 5

Allan Yu: allan.yu1@baruchmail.cuny.edu

Jia Lau: jia.lau1@baruchmail.cuny.edu

Edward Grinberg: edward.grinberg@baruchmail.cuny.edu

Hua Cai: hua.cai@baruchmail.cuny.edu

Anil Poonai: anil.poonai@baruchmail.cuny.edu

CIS 4400-CMWA

LEC (36156)

Introduction/Description:

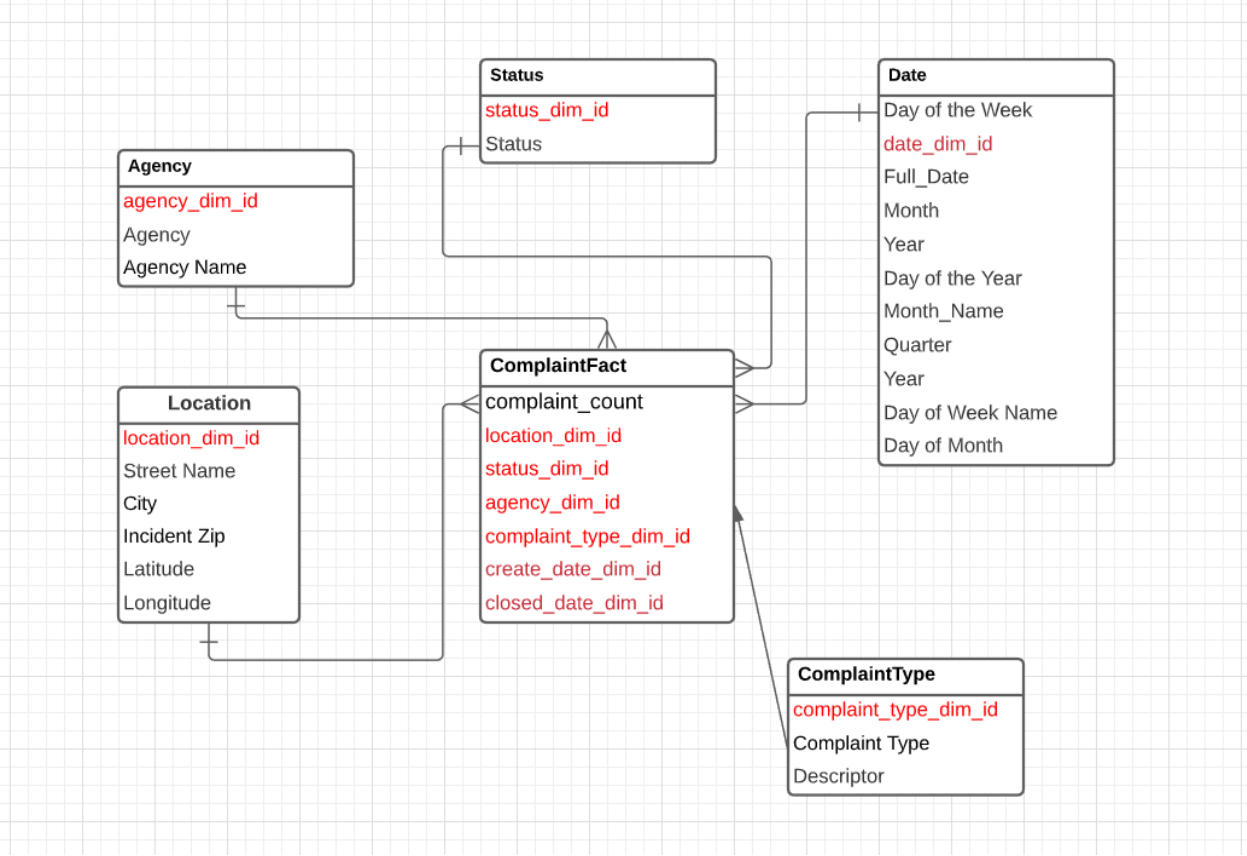
Traffic congestion is a problem in every metropolitan area, especially in one of the most densely packed destinations in the United States, New York City. It is an inescapable condition of modern societies and New York City is prone to traffic congestion regardless of borough or the time of the day. New Yorkers can regularly end up in rush hour traffic due to the sheer number of individuals heading home from work. As a result, the condition of traffic signals is crucial to reducing and controlling traffic by increasing efficiency and the effectiveness of traffic flow. Traffic signals are one of the main communication pathways drivers have with one another on the road. Our group has recognized this importance and we strongly believe traffic signal conditions should be a top priority for the city considering the danger, and the large number of complaints that were filed in this category.

To better understand the scope of the problem, we retrieved the traffic signal condition data from NYC Open Data on September 9, 2020. In addition, we used 311 data that further helped describe our dimensions, fact table and our keys. Utilizing the complaints from November 2017 to November 2020, there were over 7.2 million rows, showcasing that it is a frequent issue. Every commuter in the city has been frustrated at one point due to a faulty traffic light. It is so common because the economy and school system requires many people to commute to work, school or even run errands typically around the same time. This poses a critical issue with traffic signals during peak-hour congestion because it can directly cause many commuters to be late to their destination. Solving this issue will immensely help drivers at intersections, reduce accidents with pedestrians and other drivers, arrive at their destinations in a timely manner and create safety for all. Therefore, the importance of traffic signal functionality should be a top priority.

Dataset Link: https://data.cityofnewyork.us/Social-Services/Traffic-Signal-Condition/9ggn-w232/data

List of Confirmed KPIs:

1. The count of traffic signal complaints per borough (city is borough).
2. Amount of time it takes to fix traffic signal problems.
3. How many cases do each agency have open at any given time?
4. What month has the most complaints?

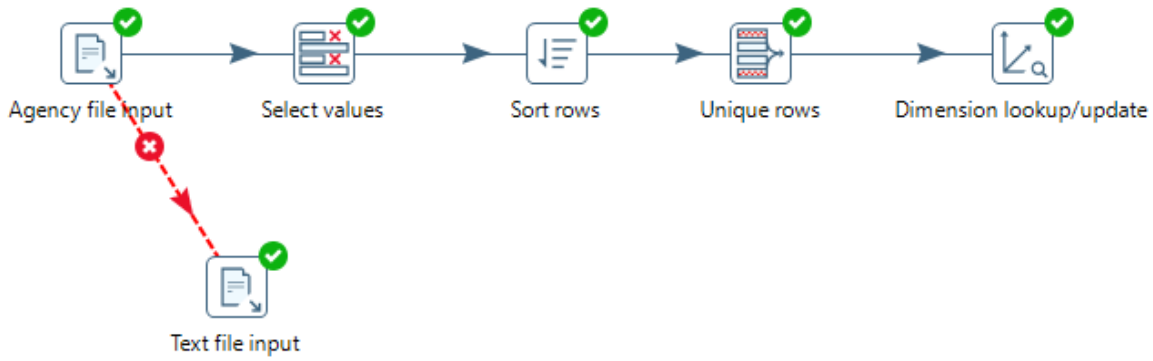


The dimensional model diagram above shows all the relative information about each incident, which we will use to answer our KPIs.

ETL Programming

We used Oracle Cloud as our cloud host service and Pentaho as our ETL tool. We implemented a periodic snapshot grain(date) and created the grain dimension table by sequencing all dates from April 2017 to May 2022. The other dimension tables were created by getting the data from the key fields in the 311 data. The fact table was made by matching the rows in the dataset to the rows in the dimension tables and adding the table keys together in the fact table. There is a step where we send less than 1,000 observations to a dummy transformation; this was needed as there were rows with just null values across all columns and they did not fit the data types or format, so all of the errors were sent there. We have foreign key constraints between the dimension keys and the columns in the fact table as well as a composite primary key involving the unique key, location, agency, date, and complaint type as they don't change.

Agency Table:

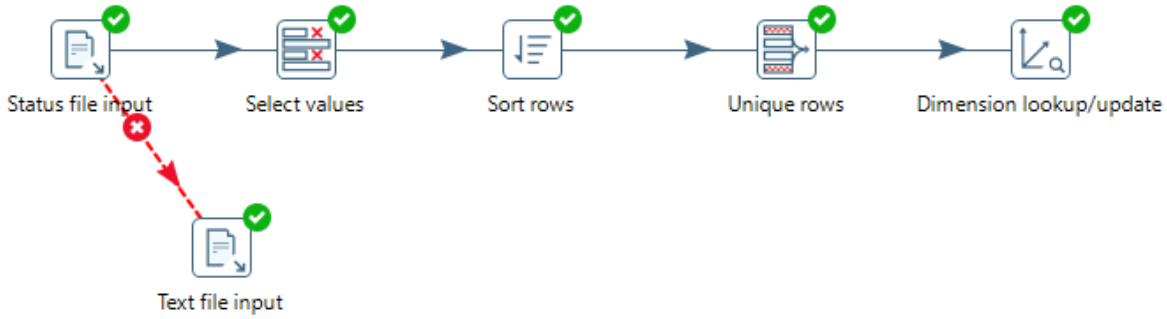


Execution Results

Logging | Execution History | Step Metrics | Performance Graph | Metrics | Preview data

#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	input/output
1	Agency file input	0	0	7257759	7258639	0	0	879	0	Finished	1mn 12s	100,324	-
2	Select values	0	7257759	7257759	0	0	0	0	0	Finished	1mn 12s	100,303	-
3	Text file input	0	0	0	0	0	0	0	0	Finished	0.0s	-	-
4	Sort rows	0	7257759	7257759	0	0	0	0	0	Finished	1mn 19s	91,440	-
5	Unique rows	0	7257759	999	0	0	0	0	0	Finished	1mn 19s	91,435	-
6	Dimension lookup/update	0	999	999	999	999	0	0	0	Finished	1mn 37s	10	-

Status Table (The table is shown below as an example for what the other tables look like):



Execution Results

Logging | Execution History | Step Metrics | Performance Graph | Metrics | Preview data

#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	input/output
1	Status file input	0	0	7257759	7258639	0	0	879	0	Finished	1mn 5s	111,515	-
2	Select values	0	7257759	7257759	0	0	0	0	0	Finished	1mn 5s	111,491	-
3	Sort rows	0	7257759	7257759	0	0	0	0	0	Finished	1mn 10s	103,406	-
4	Unique rows	0	7257759	9	0	0	0	0	0	Finished	1mn 10s	103,400	-
5	Dimension lookup/update	0	9	9	9	9	0	0	0	Finished	1mn 10s	0	-
6	Text file input	0	0	0	0	0	0	0	0	Finished	0.0s	0	-

	STATUS_DIM_ID	VERSION	DATE_FROM	DATE_TO	STATUS
1	0	1	(null)	(null)	(null)
2	1	1	01-JAN-00 12.00.00.000000000 AM	31-DEC-99 11.59.59.999000000 PM	Assigned
3	2	1	01-JAN-00 12.00.00.000000000 AM	31-DEC-99 11.59.59.999000000 PM	Closed
4	3	1	01-JAN-00 12.00.00.000000000 AM	31-DEC-99 11.59.59.999000000 PM	Closed - Testing
5	4	1	01-JAN-00 12.00.00.000000000 AM	31-DEC-99 11.59.59.999000000 PM	In Progress
6	5	1	01-JAN-00 12.00.00.000000000 AM	31-DEC-99 11.59.59.999000000 PM	Open
7	6	1	01-JAN-00 12.00.00.000000000 AM	31-DEC-99 11.59.59.999000000 PM	Pending
8	7	1	01-JAN-00 12.00.00.000000000 AM	31-DEC-99 11.59.59.999000000 PM	Started
9	8	1	01-JAN-00 12.00.00.000000000 AM	31-DEC-99 11.59.59.999000000 PM	Unassigned
10	9	1	01-JAN-00 12.00.00.000000000 AM	31-DEC-99 11.59.59.999000000 PM	Unspecified

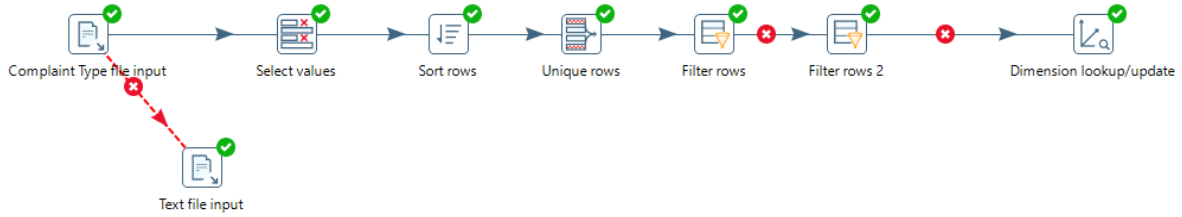
Date Table:

Execution Results

Logging | Execution History | Step Metrics | Performance Graph | Metrics | Preview data

#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	input/output
1	Generate rows	0	0	2000	0	0	0	0	0	Finished	0.0s	1,000,000	-
2	Add sequence	0	2000	2000	0	0	0	0	0	Finished	0.0s	250,000	-
3	Calculate Dates	0	2000	2000	0	0	0	0	0	Finished	0.0s	90,909	-
4	Select values	0	2000	2000	0	0	0	0	0	Finished	0.0s	71,428	-
5	Calculator	0	2000	2000	0	0	0	0	0	Finished	0.1s	26,316	-
6	Dimension lookup/update	0	2000	2000	2000	2000	0	0	0	Finished	49.1s	41	-

Complaint Type Table (We added all of the complaint types and filtered out which ones are related to traffic):

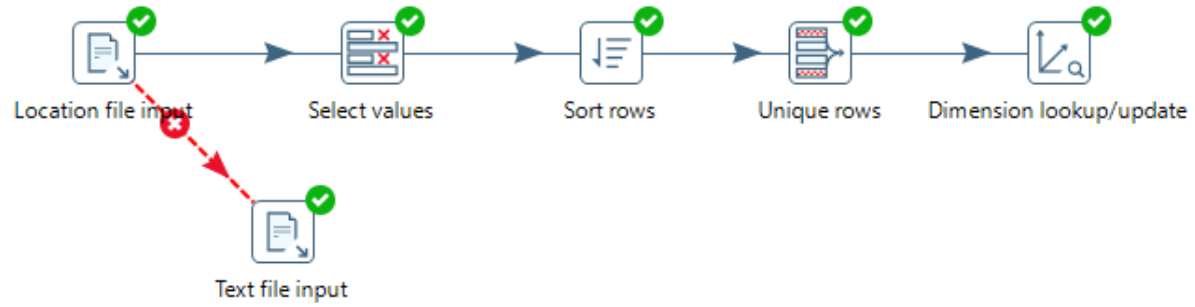


Execution Results

Logging | Execution History | Step Metrics | Performance Graph | Metrics | Preview data

#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	input/output
1	Complaint Type file input	0	0	7257759	7258639	0	0	879	0	Finished	1mn 15s	96,385	-
2	Text file input	0	0	0	0	0	0	0	0	Finished	0.0s	0	-
3	Select values	0	7257759	7257759	0	0	0	0	0	Finished	1mn 15s	96,365	-
4	Sort rows	0	7257759	7257759	0	0	0	0	0	Finished	1mn 22s	88,429	-
5	Unique rows	0	7257759	1349	0	0	0	0	0	Finished	1mn 22s	88,423	-
6	Filter rows	0	1349	1333	0	0	0	0	0	Finished	1mn 22s	16	-
7	Filter rows 2	0	1333	1318	0	0	0	0	0	Finished	1mn 22s	16	-
8	Dimension lookup/update	0	1318	1318	1318	1318	0	0	0	Finished	1mn 47s	12	-

Location Table:

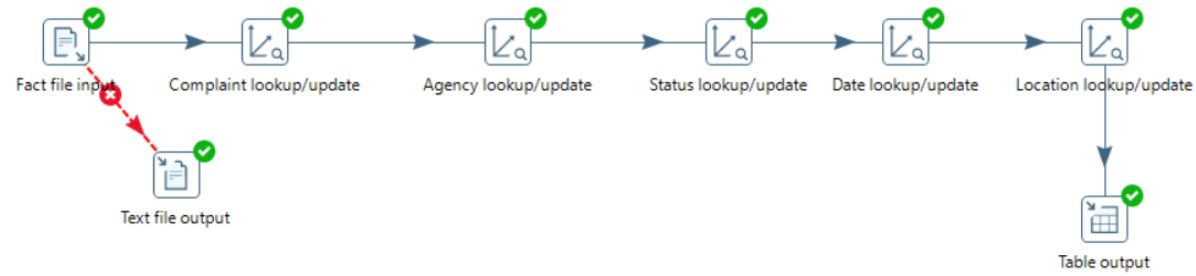


Execution Results

Logging | Execution History | Step Metrics | Performance Graph | Metrics | Preview data

#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	input/output
1	Location file input	0	0	7257759	7258639	0	0	879	0	Finished	1mn 24s	85,462	-
2	Select values	0	7257759	7257759	0	0	0	0	0	Finished	1mn 24s	85,446	-
3	Sort rows	0	7257759	7257759	0	0	0	0	0	Finished	8h 27mn 29s	238	-
4	Unique rows	0	7257759	1278757	0	0	0	0	0	Finished	8h 28mn 5s	238	-
5	Dimension lookup/update	0	1278757	1278757	1278757	1278757	0	0	0	Finished	8h 31mn 51s	42	-
6	Text file input	0	0	0	0	0	0	0	0	Finished	0.0s	0	-

Fact Table:



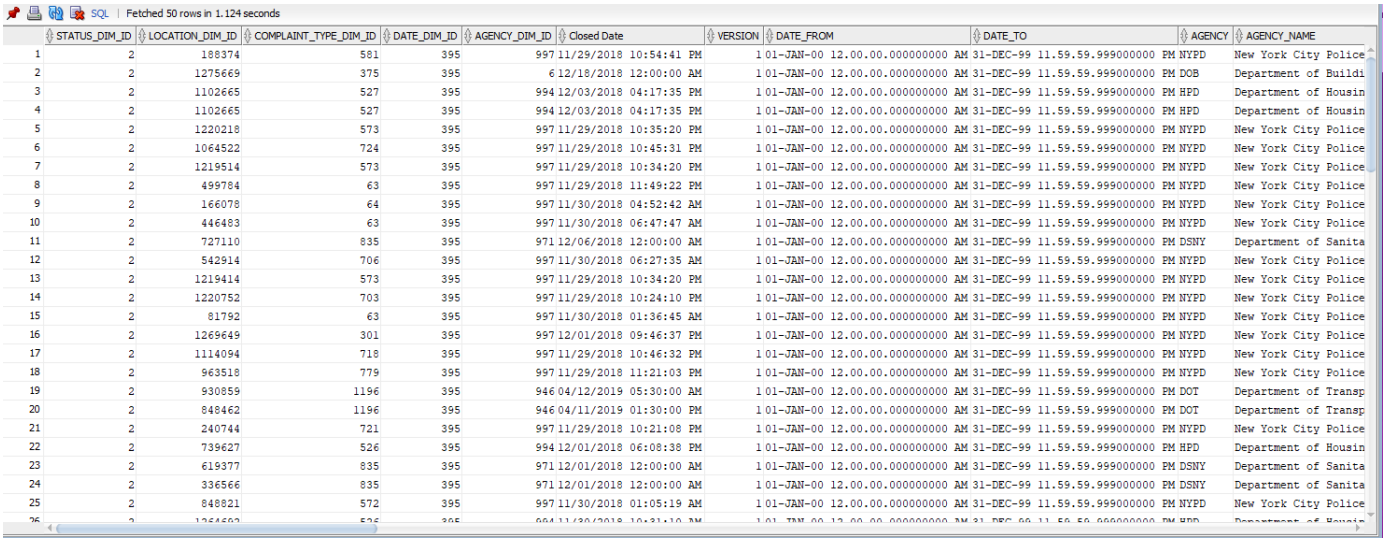
Execution Results

Logging | Execution History | Step Metrics | Performance Graph | Metrics | Preview data

#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	input/output
1	Fact file input	0	0	7258638	7258639	0	0	0	0	Finished	24h 42mn 18s	82	-
2	Complaint lookup/update	0	7258638	7258638	7258638	0	0	0	0	Finished	24h 46mn 4s	81	-
3	Agency lookup/update	0	7258638	7258638	7258638	0	0	0	0	Finished	24h 48mn 27s	81	-
4	Status lookup/update	0	7258638	7258638	7258638	0	0	0	0	Finished	24h 50mn 26s	81	-
5	Date lookup/update	0	7258638	7258638	7258638	0	0	0	0	Finished	24h 51mn 35s	81	-
6	Location lookup/update	0	7258638	7258638	7258638	0	0	0	0	Finished	24h 51mn 36s	81	-
7	Table output	0	7258638	7258638	0	7258638	0	0	0	Finished	24h 51mn 36s	81	-
8	Text file output	0	0	0	0	0	0	0	0	Finished	24h 42mn 18s	0	-

Below is the query we used to get all of the data and a sample of the data:

```
SELECT *

FROM TRAFFIC_FACT

JOIN AGENCY_DIM USING (AGENCY_DIM_ID)

JOIN DATE_DIM USING (DATE_DIM_ID)

JOIN COMPLAINT_TYPE_DIM USING (COMPLAINT_TYPE_DIM_ID)

JOIN LOCATION_DIM USING (LOCATION_DIM_ID)

JOIN STATUS_DIM USING (STATUS_DIM_ID);
```



SQL | Fetched 50 rows in 1.124 seconds

	STATUS_DIM_ID	LOCATION_DIM_ID	COMPLAINT_TYPE_DIM_ID	DATE_DIM_ID	AGENCY_DIM_ID	Closed Date	VERSION	DATE_FROM	DATE_TO	AGENCY	AGENCY_NAME
1	2	188374	581	395	997	11/29/2018 10:54:41 PM	1	01-JAN-00 12.00.00.000000000 AM	31-DEC-99 11.59.59.999000000 PM	NYPD	New York City Police
2	2	1275669	375	395	6	12/18/2018 12:00:00 AM	1	01-JAN-00 12.00.00.000000000 AM	31-DEC-99 11.59.59.999000000 PM	DOB	Department of Buildi
3	2	1102665	527	395	994	12/03/2018 04:17:35 PM	1	01-JAN-00 12.00.00.000000000 AM	31-DEC-99 11.59.59.999000000 PM	HPD	Department of Housin
4	2	1102665	527	395	994	12/03/2018 04:17:35 PM	1	01-JAN-00 12.00.00.000000000 AM	31-DEC-99 11.59.59.999000000 PM	HPD	Department of Housin
5	2	1220218	573	395	997	11/29/2018 10:35:20 PM	1	01-JAN-00 12.00.00.000000000 AM	31-DEC-99 11.59.59.999000000 PM	NYPD	New York City Police
6	2	1064522	724	395	997	11/29/2018 10:45:31 PM	1	01-JAN-00 12.00.00.000000000 AM	31-DEC-99 11.59.59.999000000 PM	NYPD	New York City Police
7	2	1219514	573	395	997	11/29/2018 10:34:20 PM	1	01-JAN-00 12.00.00.000000000 AM	31-DEC-99 11.59.59.999000000 PM	NYPD	New York City Police
8	2	499784	63	395	997	11/29/2018 11:49:22 PM	1	01-JAN-00 12.00.00.000000000 AM	31-DEC-99 11.59.59.999000000 PM	NYPD	New York City Police
9	2	166078	64	395	997	11/30/2018 04:52:42 AM	1	01-JAN-00 12.00.00.000000000 AM	31-DEC-99 11.59.59.999000000 PM	NYPD	New York City Police
10	2	446483	63	395	997	11/30/2018 06:47:47 AM	1	01-JAN-00 12.00.00.000000000 AM	31-DEC-99 11.59.59.999000000 PM	NYPD	New York City Police
11	2	727110	835	395	971	12/06/2018 12:00:00 AM	1	01-JAN-00 12.00.00.000000000 AM	31-DEC-99 11.59.59.999000000 PM	DSNY	Department of Sanita
12	2	542914	706	395	997	11/30/2018 06:27:35 AM	1	01-JAN-00 12.00.00.000000000 AM	31-DEC-99 11.59.59.999000000 PM	NYPD	New York City Police
13	2	1219414	573	395	997	11/29/2018 10:34:20 PM	1	01-JAN-00 12.00.00.000000000 AM	31-DEC-99 11.59.59.999000000 PM	NYPD	New York City Police
14	2	1220752	703	395	997	11/29/2018 10:24:10 PM	1	01-JAN-00 12.00.00.000000000 AM	31-DEC-99 11.59.59.999000000 PM	NYPD	New York City Police
15	2	81792	63	395	997	11/30/2018 01:36:45 AM	1	01-JAN-00 12.00.00.000000000 AM	31-DEC-99 11.59.59.999000000 PM	NYPD	New York City Police
16	2	1269649	301	395	997	12/01/2018 09:46:37 PM	1	01-JAN-00 12.00.00.000000000 AM	31-DEC-99 11.59.59.999000000 PM	NYPD	New York City Police
17	2	1114094	718	395	997	11/29/2018 10:46:32 PM	1	01-JAN-00 12.00.00.000000000 AM	31-DEC-99 11.59.59.999000000 PM	NYPD	New York City Police
18	2	963518	779	395	997	11/29/2018 11:21:03 PM	1	01-JAN-00 12.00.00.000000000 AM	31-DEC-99 11.59.59.999000000 PM	NYPD	New York City Police
19	2	930859	1196	395	946	04/12/2019 05:30:00 AM	1	01-JAN-00 12.00.00.000000000 AM	31-DEC-99 11.59.59.999000000 PM	DOT	Department of Transp
20	2	848462	1196	395	946	04/11/2019 01:30:00 PM	1	01-JAN-00 12.00.00.000000000 AM	31-DEC-99 11.59.59.999000000 PM	DOT	Department of Transp
21	2	240744	721	395	997	11/29/2018 10:21:08 PM	1	01-JAN-00 12.00.00.000000000 AM	31-DEC-99 11.59.59.999000000 PM	NYPD	New York City Police
22	2	739627	526	395	994	12/01/2018 06:08:38 PM	1	01-JAN-00 12.00.00.000000000 AM	31-DEC-99 11.59.59.999000000 PM	HPD	Department of Housin
23	2	619377	835	395	971	12/01/2018 12:00:00 AM	1	01-JAN-00 12.00.00.000000000 AM	31-DEC-99 11.59.59.999000000 PM	DSNY	Department of Sanita
24	2	336566	835	395	971	12/01/2018 12:00:00 AM	1	01-JAN-00 12.00.00.000000000 AM	31-DEC-99 11.59.59.999000000 PM	DSNY	Department of Sanita
25	2	848821	572	395	997	11/30/2018 01:05:19 AM	1	01-JAN-00 12.00.00.000000000 AM	31-DEC-99 11.59.59.999000000 PM	NYPD	New York City Police

We created a table and csv file of all of the data in the dimensional model.

We also got the number of times each date showed up with the following query:

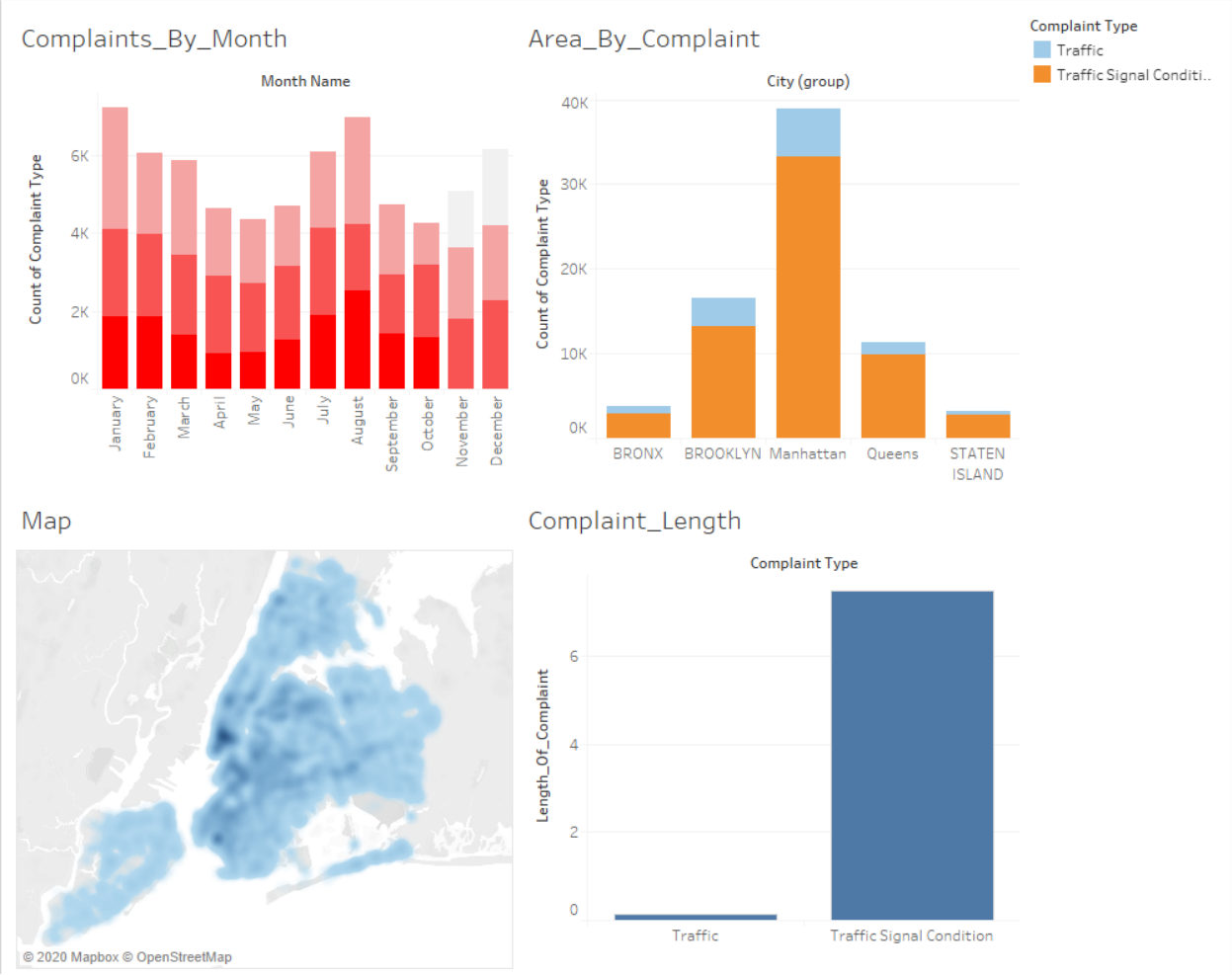
```
SELECT traffic_date,count(*)

FROM TRAFFIC_FACT

JOIN AGENCY_DIM USING (AGENCY_DIM_ID)

JOIN DATE_DIM USING (DATE_DIM_ID)

JOIN COMPLAINT_TYPE_DIM USING (COMPLAINT_TYPE_DIM_ID)

JOIN LOCATION_DIM USING (LOCATION_DIM_ID)
```

```
JOIN STATUS_DIM USING (STATUS_DIM_ID)
Group by traffic_date;
```

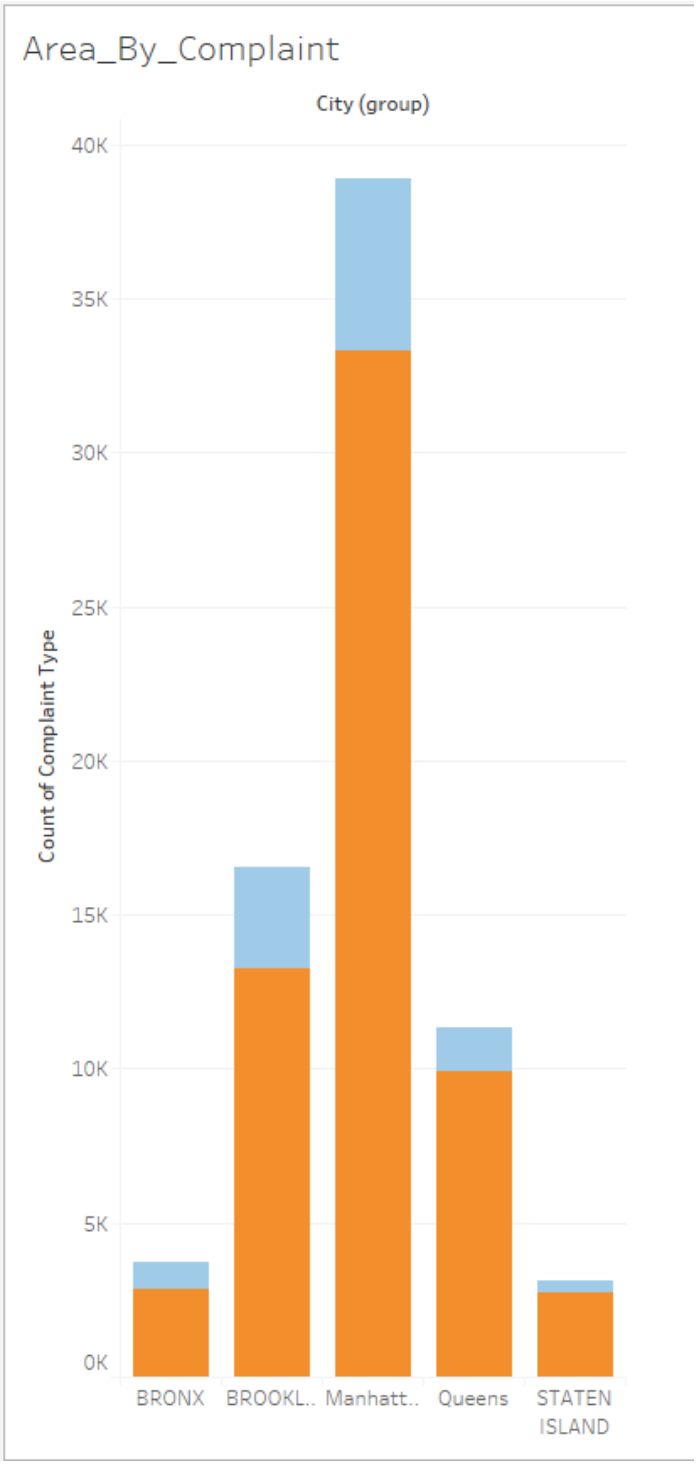
We did this as an alternative to using the group by function in Pentaho since it led to the same thing as it did not add another column or row about the number of each date.

Dashboard:

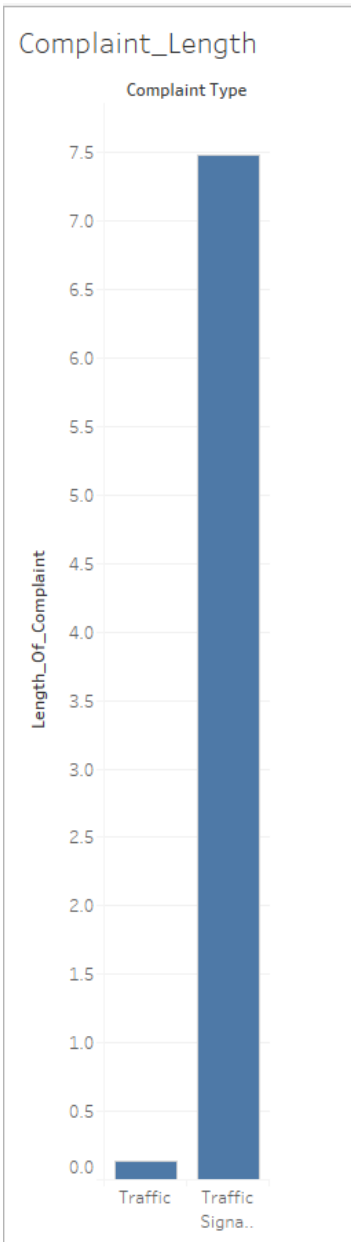
We used Tableau to create the dashboard that has the answers to our KPI's. Below is the dashboard with all four KPI's shown. We filtered everything by the complaint "Traffic Signal Condition" for our analysis, but we also filtered by "Traffic" just to have a comparison.



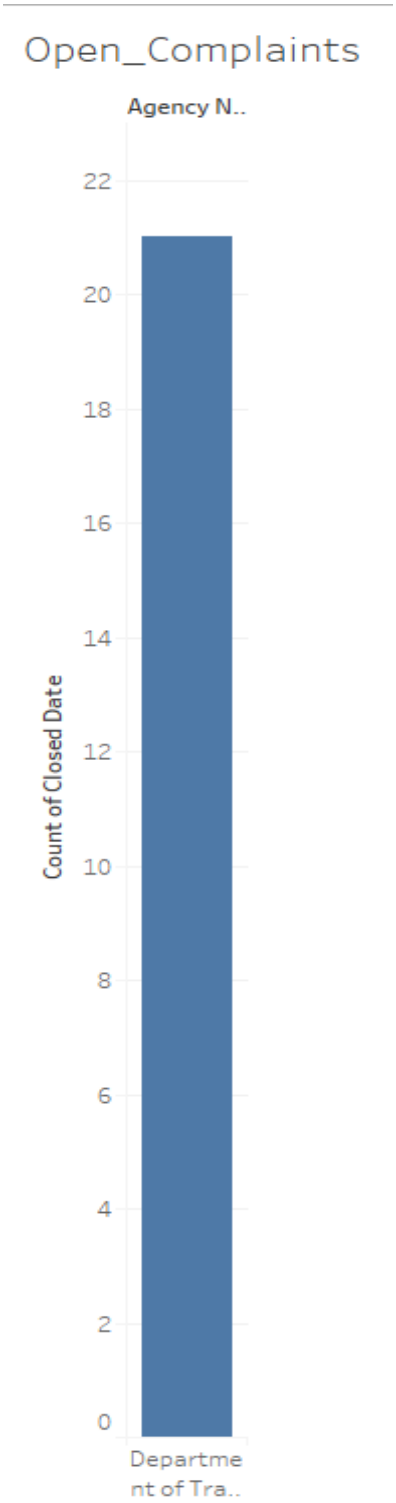
The graph below is the number of each complaint type by location. Manhattan has the most amount of traffic problems by far. Orange is the traffic signal problems and we also wanted to show the comparison to just traffic problems (Blue). We are only showing the data for traffic in our report, for analysis we are only using the traffic signal reports.



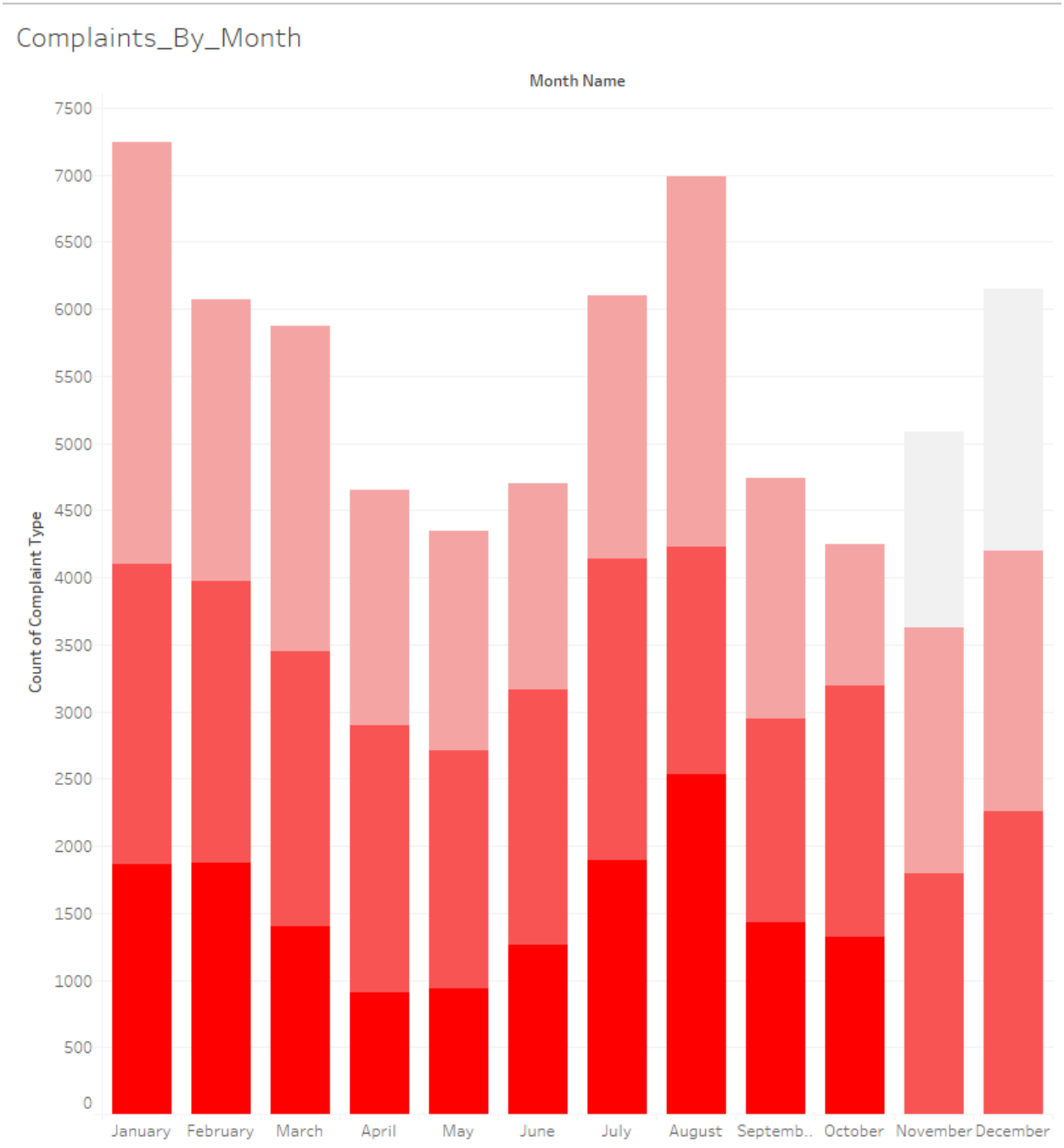
The next image shows the average time a complaint is open measured in days. Traffic signals get fixed in 7.5 days while traffic gets fixed in less than a day, which makes sense. Again, we showed the data for traffic problems as well, but our analysis only uses the traffic signal reports.



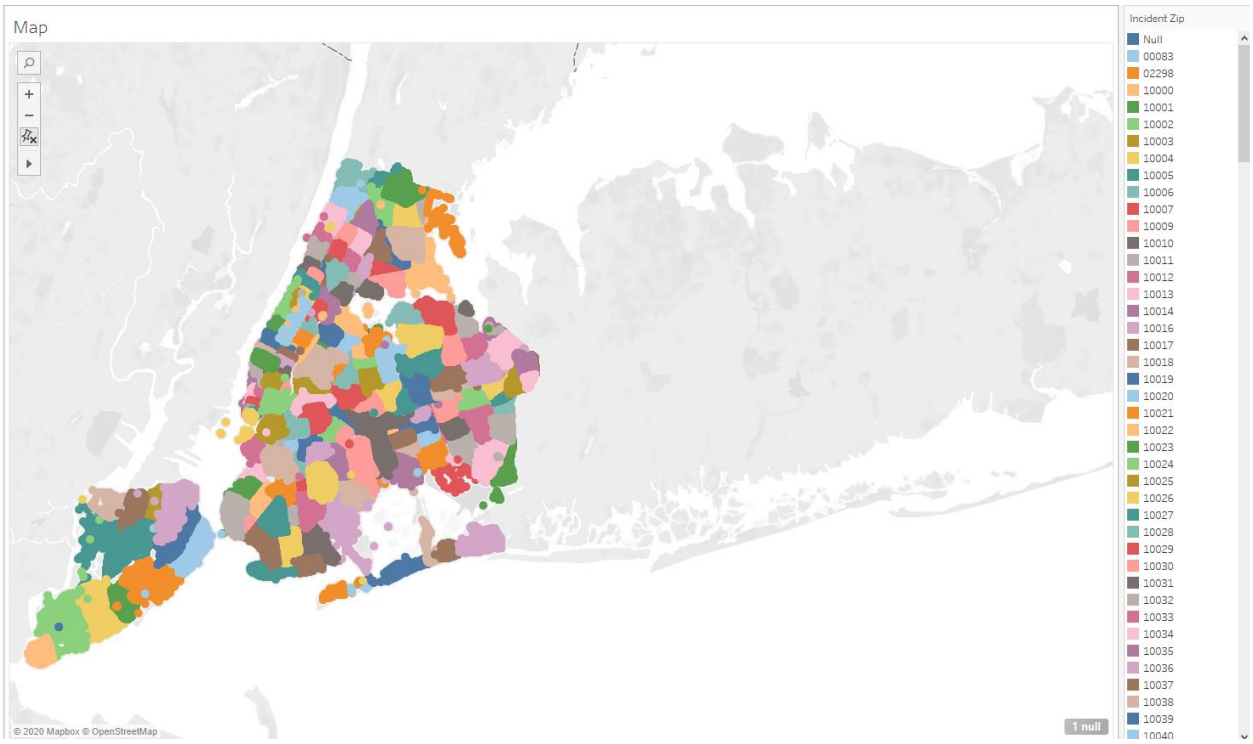
The next graph is the total amount of open complaints at any given time. This only uses traffic signal reports. 22 cases will usually be open at one time regarding traffic signal conditions.



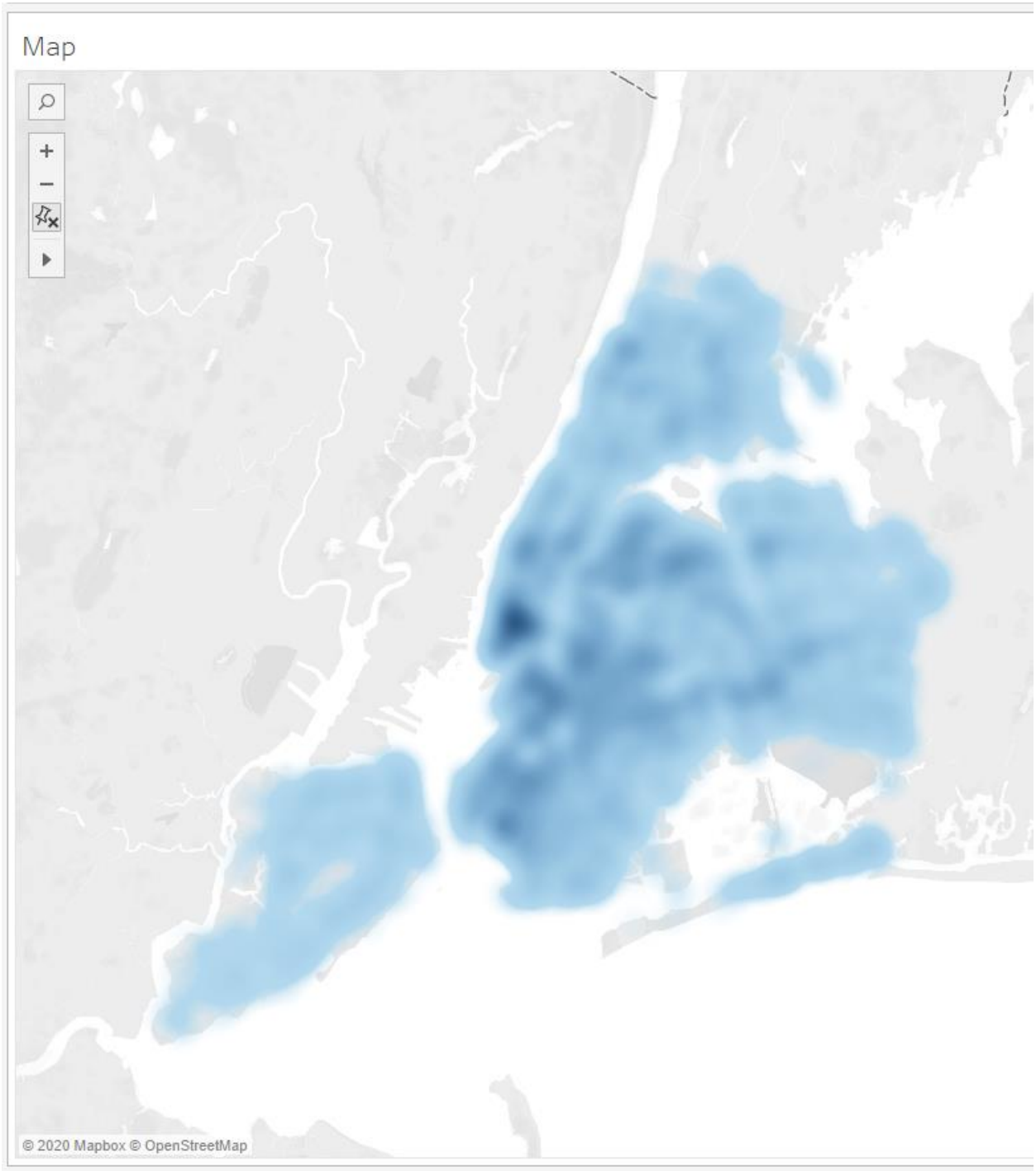
Below is the number of complaints by month graphed out. The brightest red is 2020, second brightest is 2019, third is 2018 and least is 2017 (in order from bottom to top). January and August are the most popular months for traffic signal problems.



Below is a map of all of the coordinates that have had a case reported along with a color grouping based on the zip code.



Below is the density of them. The darker spots are places where traffic signal problems happen the most.



Technology:

To complete this project, our group used Oracle Cloud's transnational database to host the data, Pentaho to complete the ETL and Tableau to visualize the data and create the dashboard.

Conclusion:

Over the course of the last 3 months, we had hands-on experience on what it is like to develop and use a data warehouse. To help us complete the project, we used a variety of tools like:

- Oracle: Used to host our traffic signals condition data
- Pentaho: Used to create the data dimensions and the final schema
- Tableau: Used to create a dashboard which helped visualize our KPIs
- WhatsApp: Used to communicate with one another on a weekly basis
- Google Drive: Used to share large files with one another and to have a location to all work on this document together

While working on the project, we came across easy and tough hurdles to get by. The most difficult step was the ETL, as it requires high attention to detail when setting our constraints on the tables and the fields. The easiest part was the dashboard as we planned out how to have our data in a way that would make it easy to visualize; we also had more experience with this part as we had experience with this part from homework one. During the second homework assignment, we had the practice ETL given to us and as a result, we did not have the full experience leading up to it. One thing that we learned that we would not have imagined learning is how powerful data and visualization can be. While we did practice with this in one of our homework assignments, we saw a ton of opportunities to showcase a story outside of our KPIs. We learned

that data in an excel file is important, but it does not tell a story. Using visualization can draw an image to a person's head and it creates a stronger argument. If we had to do this again, we would have broken down the dataset into different datasets of 100,000 rows each to help us find any errors quicker.

We believe the new system’s benefits can be used as it does pack the data to be as concise as possible. In addition, it makes it easy to navigate through not only Tableau, but Python and R worked well with it when we give it a test try. Overall, this project was a huge learning experience in terms of learning and skills, and we can see it being applicable to many industries.

We did not have any references for this project because the majority of what we did was to use what we learned in our homework assignments to this project.

Meeting Log:

9/1/2020, All Group Members: Meeting with all members through email and WhatsApp to introduce ourselves and talk about our initial thoughts of the project assignment.

9/8/2020, All Group Members: Group chat utilizing WhatsApp to review each of our suggestions on topics that we were interested in and come to a consensus. Begin bouncing ideas and start drafting the description of the project.

9/9/2020, All Group Members: Group chat utilizing WhatsApp and Google Docs to review, edit and finalize our description of the project and create an initial list of KPI's.

9/20, All Group Members: Meeting through WhatsApp to discuss and review our KPI's and identify our data set.

9/24, All Group Members: Started to create dimensional model

9/27, All Group Members: Final discussion through WhatsApp to review our dimensional model and KPI's and made any further finalized changes.

10/20, All Group Members: Discussion through WhatsApp to review our dimensional model based on the feedback from the professor.

10/21, All Group Members: Discussion through WhatsApp to finalize our dimensional model and begin discussion of the ETL steps.

10/23, All Group Members: Updated dimensional model.

11/01, All Group Members: Split up work for ETL portions

11/5, All Group Members: Finished ETL

11/26, All Group Members: Started and finished Dashboard using Tableau, we also tested out how well the data worked in Python and R.