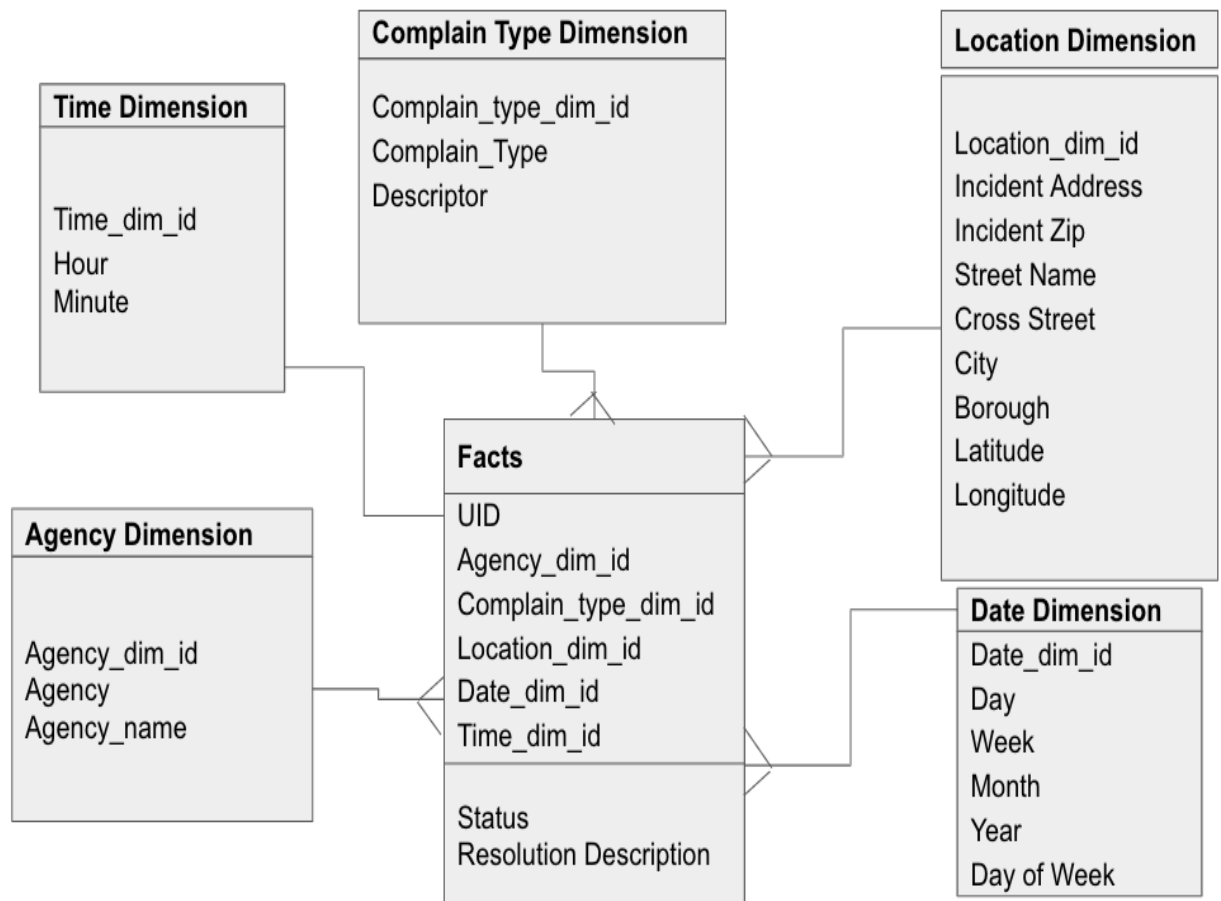**KPIs:**

Our first step is to list Key Performance Indicators (KPI's) that our group will use to address the issue or problem. Our KPIs are the following:

- ❏ Illegal Parking Complaint per hour per borough
- ❏ Illegal Parking Complaint per hour per borough from 2017 to now
- ❏ Illegal Parking Complaint Type from 2017 to now
- ❏ Illegal Parking Complaint Map from 2017 to now

Time frame is from 2017 to now. We create a Dimension Model to demonstrate the information including each entity and the relationships between dimension tables and facts table. Please see the Picture for Dimension Model.

Dimension Model:

Complain Type Dimension

Complain_type_dim_id
Complain_Type
Descriptor

Location Dimension

Location_dim_id
Incident Address
Incident Zip
Street Name
Cross Street
City
Borough
Latitude
Longitude

Time Dimension

Time_dim_id
Hour
Minute

Facts

UID
Agency_dim_id
Complain_type_dim_id
Location_dim_id
Date_dim_id
Time_dim_id

Status
Resolution Description

Date Dimension

Date_dim_id
Day
Week
Month
Year
Day of Week

Agency Dimension

Agency_dim_id
Agency
Agency_name

Dimension and fact tables:
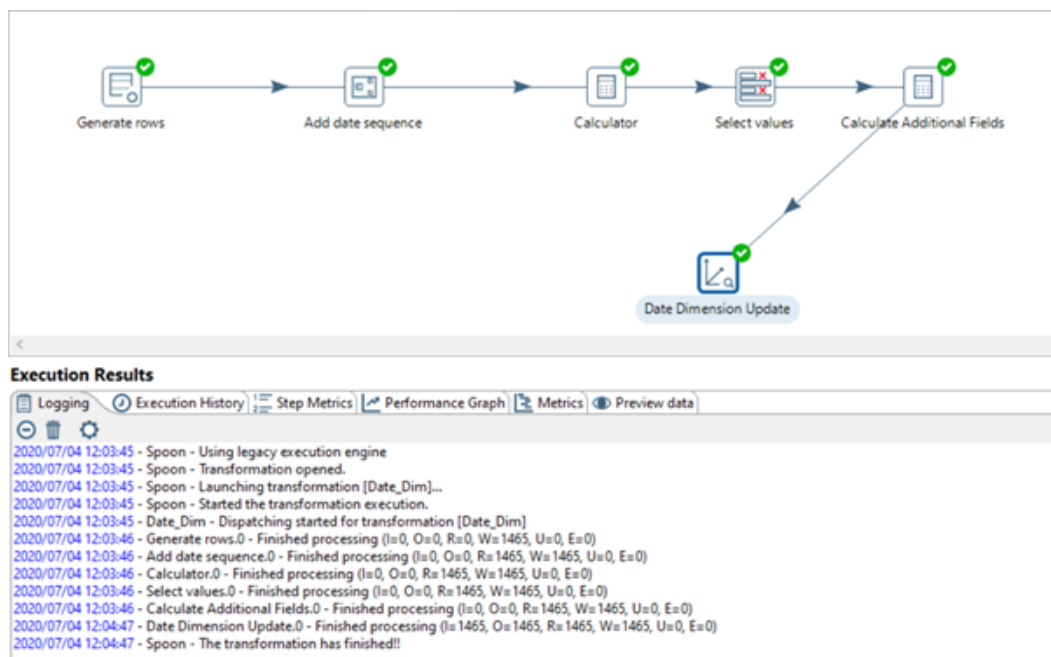
Below are each tables transformations and configurations:

Time Dimension:

we create a spreadsheet from 0:00 to 23:59 and then input it to Pentaho to create a Time

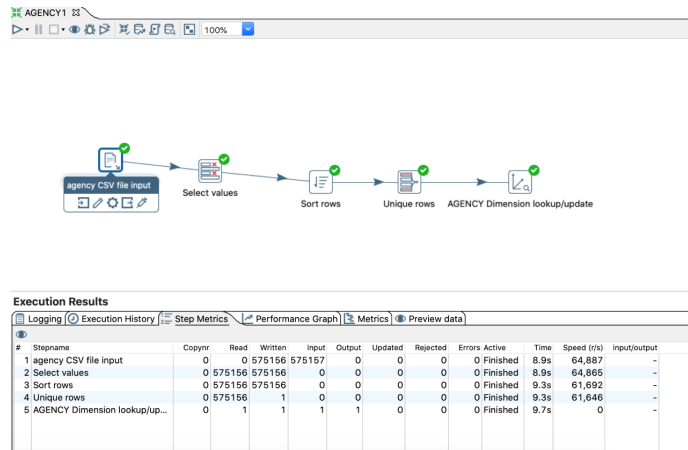Dimension.

CSV file input   Time lookup/update
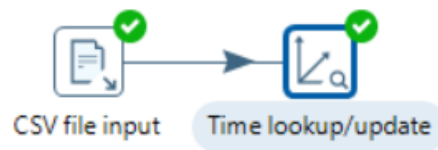
Date Dimension:

we are using Generate Rows as the first step to create the date dimension. we generate the 1465 rows from 1/1/2017 as our data is from 2017. Then we increase the date by day through Add Sequence and Calculator modules. After this, we pare down the field by Select Values and embellish the date with additional calculated date attributes.
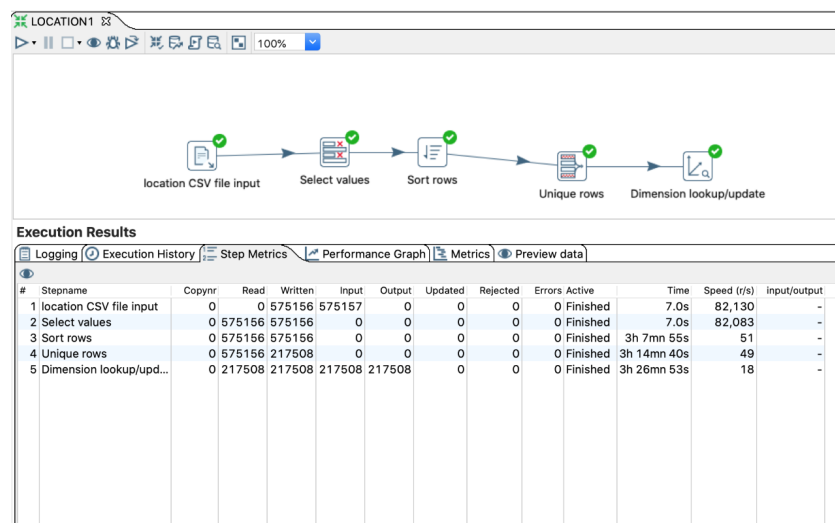

Generate rows   Add date sequence   Calculator   Select values   Calculate Additional Fields

Date Dimension Update

Execution Results

Logging   Execution History   Step Metrics   Performance Graph   Metrics   Preview data

2020/07/04 12:03:45 - Spoon - Using legacy execution engine
2020/07/04 12:03:45 - Spoon - Transformation opened.
2020/07/04 12:03:45 - Spoon - Launching transformation [Date_Dim]...
2020/07/04 12:03:45 - Spoon - Started the transformation execution.
2020/07/04 12:03:45 - Date_Dim - Dispatching started for transformation [Date_Dim]
2020/07/04 12:03:46 - Generate rows.0 - Finished processing (I=0, O=0, R=0, W=1465, U=0, E=0)
2020/07/04 12:03:46 - Add date sequence.0 - Finished processing (I=0, O=0, R=1465, W=1465, U=0, E=0)
2020/07/04 12:03:46 - Calculator.0 - Finished processing (I=0, O=0, R=1465, W=1465, U=0, E=0)
2020/07/04 12:03:46 - Select values.0 - Finished processing (I=0, O=0, R=1465, W=1465, U=0, E=0)
2020/07/04 12:03:46 - Calculate Additional Fields.0 - Finished processing (I=0, O=0, R=1465, W=1465, U=0, E=0)
2020/07/04 12:04:47 - Date Dimension Update.0 - Finished processing (I=1465, O=1465, R=1465, W=1465, U=0, E=0)
2020/07/04 12:04:47 - Spoon - The transformation has finished!!

Agency Dimension:

We selected the columns of agency and agency name, then sorted all rows and found the unique rows; finally, we created the agency dimension for three years.
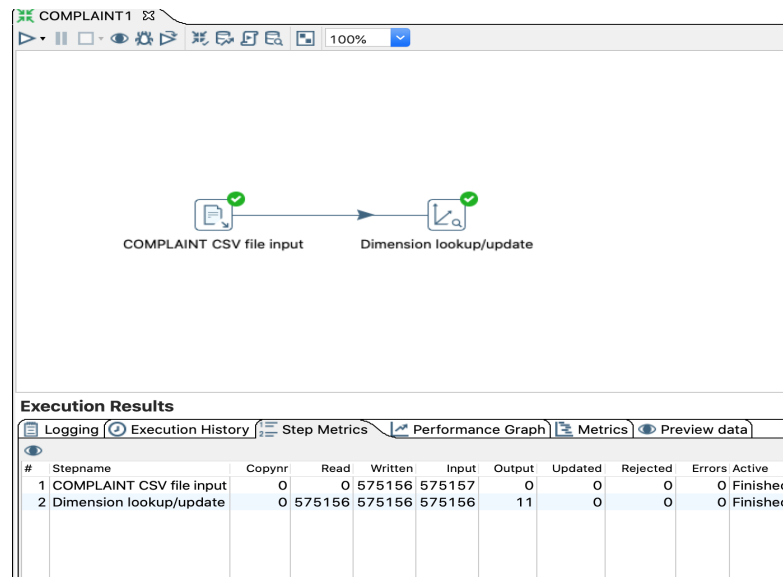
Execution Results

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 CSV file input	0	0	575156	575157	0	0	0	0	Finished	8.9s	64,887	-
2	Select values	0	575156	575156	0	0	0	0	0	Finished	8.9s	64,865	-
3	Sort rows	0	575156	575156	0	0	0	0	0	Finished	9.3s	61,692	-
4	Unique rows	0	575156	1	0	0	0	0	0	Finished	9.3s	61,646	-
5	AGENCY Dimension lookup/up...	0	1	1	1	1	0	0	0	Finished	9.7s	0	-

Location Dimension:

Similar with agency dimension, we selected the columns that we need like borough, city, latitude and so on; we created location dimension after we sort rows and finding the unique rows for three years.

Execution Results

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

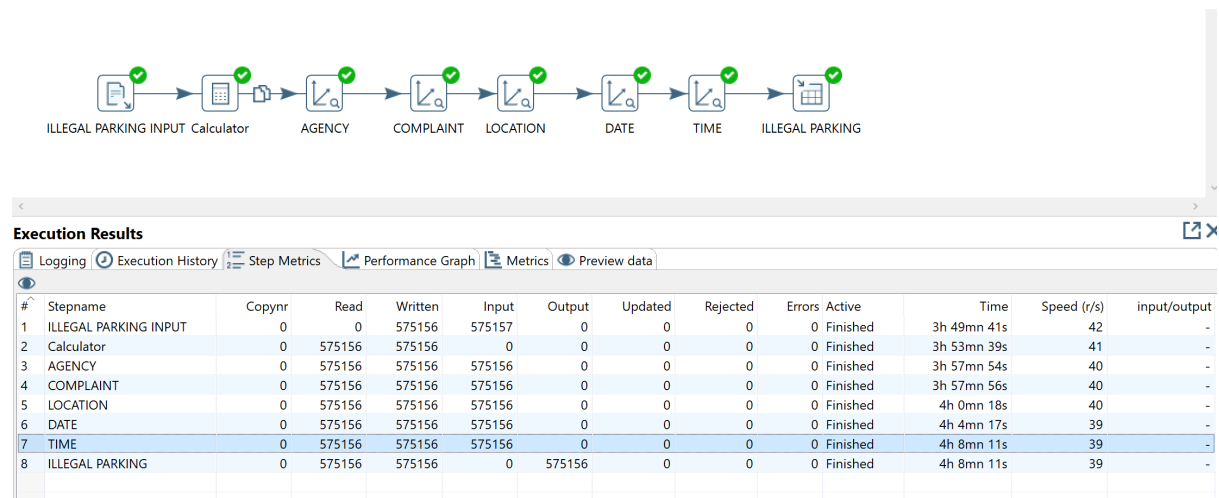| #	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	input/output
1	location CSV file input	0	0	575156	575157	0	0	0	0	Finished	7.0s	82,130	-
2	Select values	0	575156	575156	0	0	0	0	0	Finished	7.0s	82,083	-
3	Sort rows	0	575156	575156	0	0	0	0	0	Finished	3h 7mn 55s	51	-
4	Unique rows	0	575156	217508	0	0	0	0	0	Finished	3h 14mn 40s	49	-
5	Dimension lookup/upd...	0	217508	217508	217508	217508	0	0	0	Finished	3h 26mn 53s	18	-

Complaint Dimension:

We imported our datasets into complaint CSV file input; then we created our complaint dimension in a direct way that we skipped the steps of sorted rows and find the unique rows.

#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active
1	COMPLAINT CSV file input	0	0	575156	575157	0	0	0	0	Finished
2	Dimension lookup/update	0	575156	575156	575156	11	0	0	0	Finished

The final fact table with all the dimension tables:

#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	input/output
1	ILLEGAL PARKING INPUT	0	0	575156	575157	0	0	0	0	Finished	3h 49mn 41s	42	-
2	Calculator	0	575156	575156	0	0	0	0	0	Finished	3h 53mn 39s	41	-
3	AGENCY	0	575156	575156	575156	0	0	0	0	Finished	3h 57mn 54s	40	-
4	COMPLAINT	0	575156	575156	575156	0	0	0	0	Finished	3h 57mn 56s	40	-
5	LOCATION	0	575156	575156	575156	0	0	0	0	Finished	4h 0mn 18s	40	-
6	DATE	0	575156	575156	575156	0	0	0	0	Finished	4h 4mn 17s	39	-
7	TIME	0	575156	575156	575156	0	0	0	0	Finished	4h 8mn 11s	39	-
8	ILLEGAL PARKING	0	575156	575156	0	575156	0	0	0	Finished	4h 8mn 11s	39	-

Other configuration of dimension tables:

Table output — □ ✕

Field	Value
Step name	ILLEGAL PARKING
Connection	test4 ▾ Edit... New... Wizard...
Target schema	ADMIN Browse...
Target table	ILLEGAL PARKING FACT Browse...
Commit size	1000
Truncate table	☐
Ignore insert errors	☐
Specify database fields	☑

Main options | Database fields

Partition data over tables	☐
Partitioning field	
Partition data per month	◉
Partition data per day	○
Use batch update for inserts	☑
Is the name of the table defined in a field?	☐
Field that contains name of table:	
Store the tablename field	☑
Return auto-generated key	☐
Name of auto-generated key field	

Help   OK   Cancel   SQL

Dimension lookup/update — □ ✕

Field	Value
Step name	TIME
Update the dimension?	☐
Connection	TIME1 ▾ Edit... New... Wizard...
Target schema	ADMIN Browse...
Target table	TIME5_DIM Browse...
Commit size	100
Enable the cache?	☐
Pre-load the cache?	☐
Cache size in rows (0 = cache all)	

Keys | Fields

Key fields (to look up row in dimension):

#	Dimension field	Field in stream
1	HOURS	HOUR
2	MINUTES	MINUTE

Technical key field TIME_DIM_ID ▾   New name

Creation of technical key
◉ Use table maximum + 1
○ Use sequence
○ Use auto increment field

Version field version ▾
Stream Datefield

Dimension lookup/update — □ ✕

Field	Value
Step name	AGENCY
Update the dimension?	☐
Connection	AGENCY1 ▾ Edit... New... Wizard...
Target schema	ADMIN Browse...
Target table	AGENCY_DIMENSION Browse...
Commit size	100
Enable the cache?	☐
Pre-load the cache?	☐
Cache size in rows (0 = cache all)	

Keys | Fields

Key fields (to look up row in dimension):

#	Dimension field	Field in stream
1	AGENCY	agency
2	AGENCY_NAME	agency name

Technical key field AGENCY_DIM_ID ▾   New name

Creation of technical key
◉ Use table maximum + 1
○ Use sequence
○ Use auto increment field

Version field version ▾
Stream Datefield

Dimension lookup/update

Step name	COMPLAINT
Update the dimension?	☐
Connection	COMPLAINT1
Target schema	ADMIN
Target table	COMPLAINT_DIMENSION
Commit size	100
Enable the cache?	☐
Pre-load the cache?	☐
Cache size in rows (0 = cache all)	

Edit... New... Wizard...
Browse...
Browse...

Keys | Fields

Key fields (to look up row in dimension):

#	Dimension field	Field in stream
1	COMPLAINT_TYPE	complaint_type
2	DESCRIPTOR	descriptor

Technical key field	COMPLAINT_DIM_ID	New name	

Creation of technical key
- ◉ Use table maximum + 1
- ○ Use sequence
- ○ Use auto increment field

Version field	version
Stream Datefield	

⊙ Help

Calculator

Step name

Calculator

🔲

☑ Throw an error on non existing files

Fields:

#	New field	Calculation	Field A	Field B	Field C	Value type	Length	Precision	Remove	Conversion mask	Decimal symbol	Grouping symbol	Cu
1	DATE	Remove time from a date A	created_date			Date	10		N	MM/dd/yy			
2	HOUR	Hour of Day of Date A	created_date			String			N				
3	MINUTE	Minute of Hour of Date A	created_date			String			N				

Step name	ILLEGAL PARKING INPUT
Filename	C:\Users\Alice\Downloads\project data.csv
Delimiter	,
Enclosure	"
NIO buffer size	50000
Lazy conversion?	☐
Header row present?	☑
Add filename to result	☐
The row number field name (optional)	
Running in parallel?	☐
New line possible in fields?	☐
File encoding	

Browse...
Insert TAB

| #	Name	Type	Format	Length	Precision	Currency	Decimal	Group	Trim type
1	unique key	Integer	#	15	0	$	.	,	none
2	created_date	Date	MM/dd/yy HH:mm	16		$	.	,	none
3	closed date	String		14		$	.	,	none
4	agency	String		4		$	.	,	none
5	agency name	String		31		$	.	,	none
6	complaint_type	String		15		$	.	,	none
7	descriptor	String		30		$	.	,	none
8	location type	String		15		$	.	,	none
9	incident zip	Integer	#	15	0	$	.	,	none
10	incident address	String		39		$	.	,	none
11	street name	String		32		$	.	,	none
12	cross street 1	String		32		$	.	,	none
13	cross street 2	String		32		$	.	,	none
14	intersection street 1	String		32		$	.	,	none
15	intersection street 2	String		32		$	.	,	none
16	address type	String		12		$	.	,	none
17	city	String		19		$	.	,	none
18	landmark	String		31		$	.	,	none

Below is the query to create the view:

```
CREATE VIEW TOGETHER AS
SELECT *
FROM ILLEGAL_PARKING_FACT
NATURAL JOIN AGENCY_DIMENSION
NATURAL JOIN COMPLAINT_DIMENSION
NATURAL JOIN DATE_DIM
NATURAL JOIN LOCATION_DIMENSION
NATURAL JOIN TIME5_DIM;
```

**Final schema is one primary fact table with 5 foreign keys from 5 different dimension tables.**
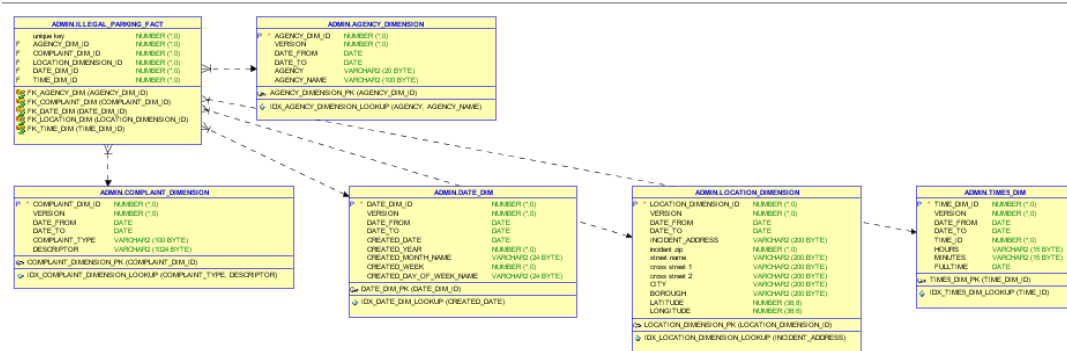
**According to the KPIs, we visualized four KPIs by Tableau.**

A. Illegal parking complaint per hour per borough: We can easily conclude that Brooklyn has the most complaints and 8am-10am and 8pm-10pm are the two peaks of illegal parking.
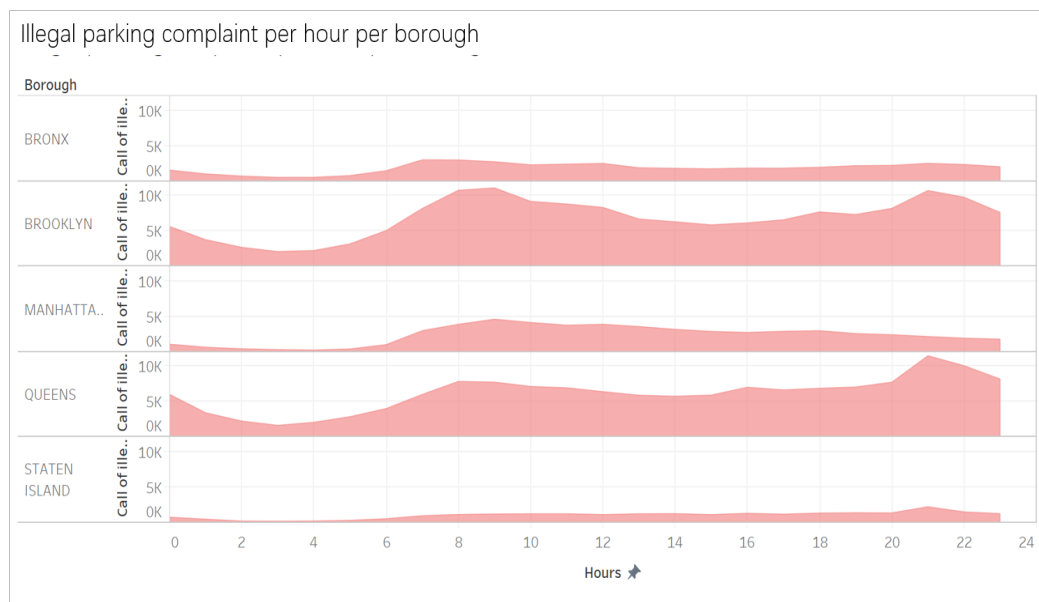
B. Illegal parking complaint per day per borough from 2017 to now: It's apparent that Manhattan and Queen have a growing illegal parking since last year.

Illegal Parking Complaint per Day per Borough
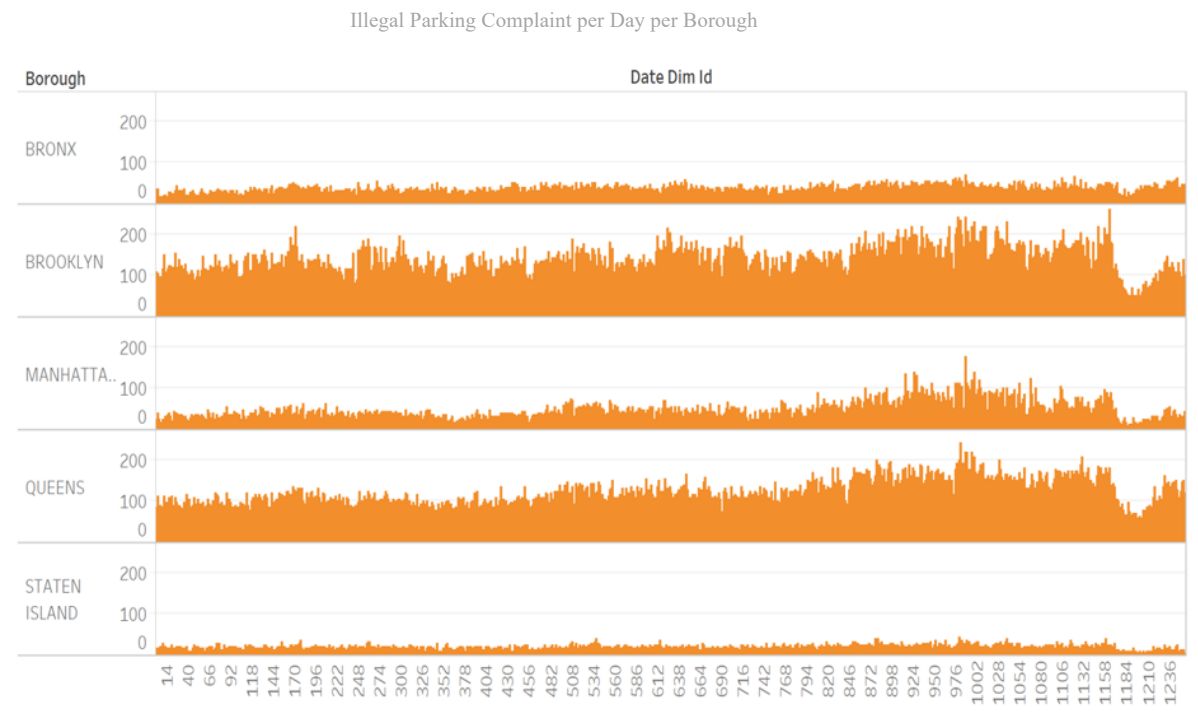
C. Illegal Parking Complaint Map
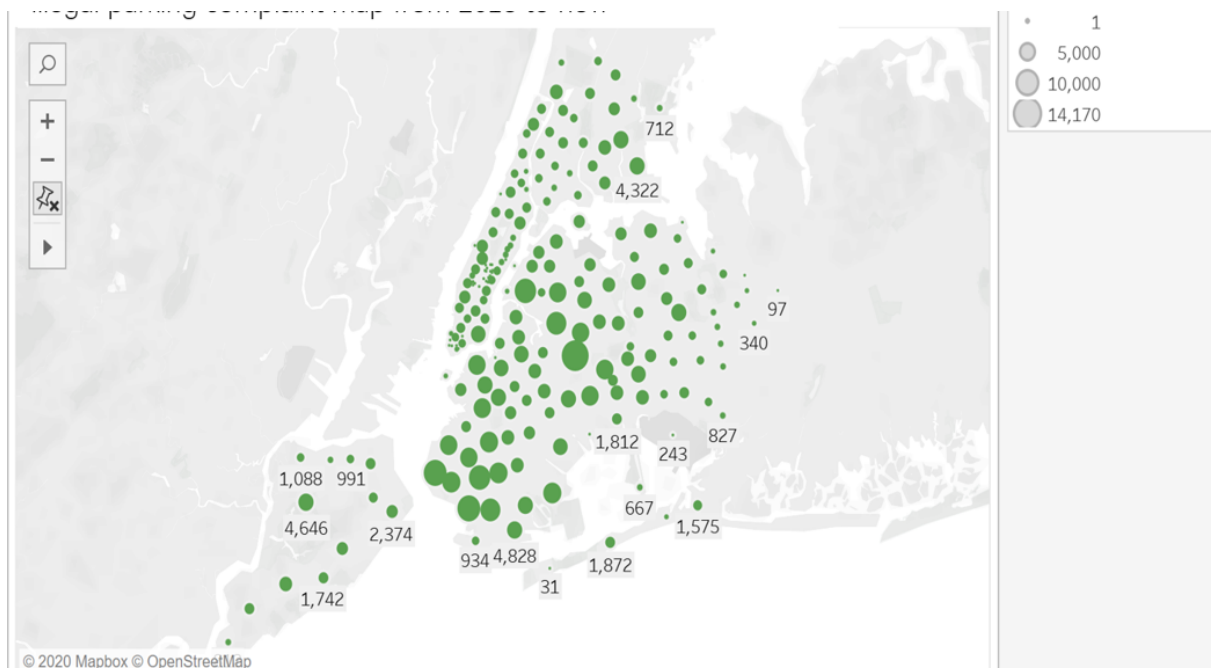
D. Illegal parking complaint type from 2017 to now: Blocking hydrant is the top complaint type of illegal parking.

Illegal parking complaint type
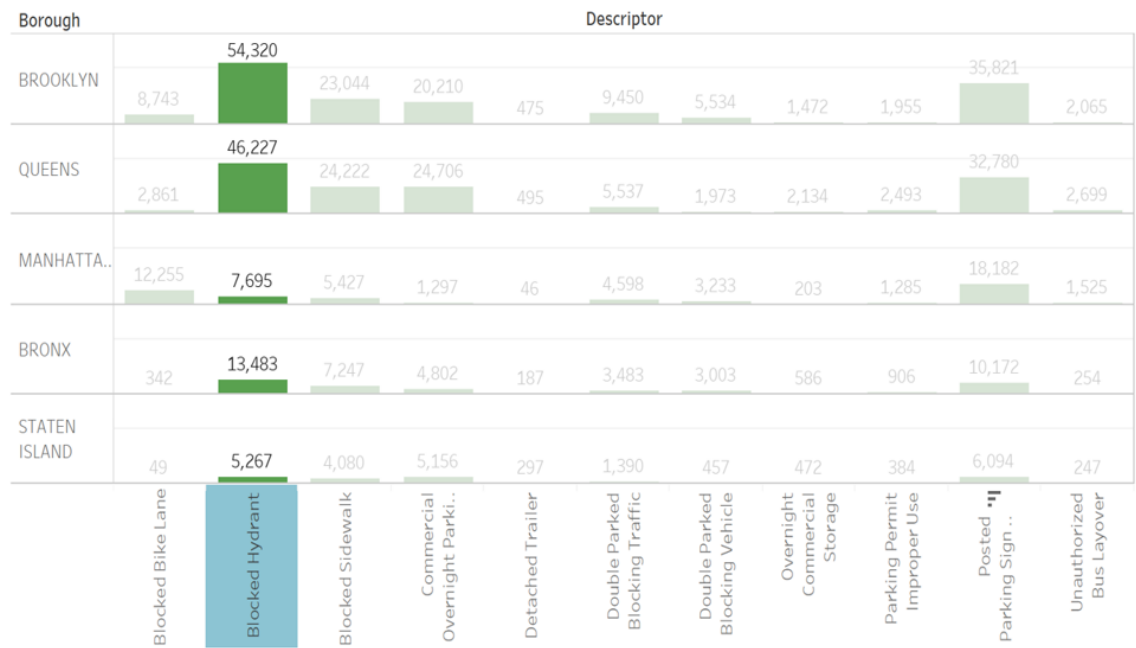
Other questions:

1. Descriptions of the tools (databases, analytics, ETL, programming languages, etc.) used to complete the project.

   We are using python to download the 311 database, conducting the ETL with Pentaho and finally visualizing the result/KPIs by Tableau.

2. A narrative conclusion section that describes:
   a) the software and database tools the group used to coordinate and manage the project as well as carry out the programming tasks

   Google drive is the main document share tool used to coordinate our group project while iMessage is another tool we are using to discuss our problem and process. Besides, we are also using Zoom to meet and discuss our project.

   b) the group's experience with the project (which steps were the most difficult? Which were the easiest? what did you learn that you did not imagine you would have? if you had to do it all over again, what would you have done differently?)

   The fact table is the hardest part to complete this project, because we have to make sure all the dimension tables can be combined together without any error and the dimension tables are created by different people. In order to keep the transaction level to leave us more flexibility, we decided to split the date column to date and time. However, it had some errors when we tried to link the date dimension to the fact table and therefore, it took more time than we originally expected to design the fact parts to fix the issue of conversion of the date.

   c) if the proposed benefits can be realized by the new system.

   Yes. We can suggest NYPD an optimized allocation of police resources to deal with illegal parking.

d) any final comments and conclusions.

This is a group project and needs a lot of communication and coordination. Also, this project provides us a chance to apply some real data to what we learned from class, which is fun!

3. A References list that provides the web sites and other sources for data, techniques, methods, software, etc. used to complete the project.