Combatting Smoking in NYC

Data Warehousing and Analytics - CIS 9440 UTA

Belozertsev, Valeria

valerie.belozertsev@baruchmail.cuny.edu

Chen, Youming

youming.chen@baruchmail.cuny.edu

Ha, Melissa

melissa.ha@baruchmail.cuny.edu

Hakimian, Andrew

andrew.hakimian1@baruchmail.cuny.edu

Harris, Tevin

Tevin.Harris@baruchmail.cuny.edu

Table of Contents

Executive Summary

Smoking is classified as inhaling harmful carbons and chemicals that destroy the body overtime. It is no secret that smoking is also bad for people, who are around smokers inhaling toxic chemicals and for the environment, as a whole. New York City has been trying to combat smoking for a while, so much so that the complaints made to NYC 311 Service have been on the rise in recent years. The Smoke Free Air Act (SFAA) and the NYS Clean Indoor Air Act (CIAA) both prohibit smoking and the use of electronic cigarettes in most workplaces and public spaces. The CIAA also prohibits smoking and the use of e-cigarettes within 100 feet of entrances, exits or outdoor areas of public libraries and of public or private elementary or secondary schools. Lack of enforcement of the NYC SFAA and the NYS CIAA has caused an increase in illegal smoking in numerous locations that are supposedly protected.

Illegal smoking in New York City is a widespread issue affecting not only residential locations but also commercial locations such as offices, restaurants, and schools across all boroughs. With the data we gather and analyze, our goal is to determine which zip codes in each borough are most adversely affected by illegal smoking. To further refine our findings, we will also be analyzing when reports were submitted to 311 in order to determine if there are spikes in the number of reports for certain months of the year. Lastly, we will take into consideration the possibility of weather conditions playing a role in the total reports submitted for each borough during each month. Such should indicate whether there is a relationship between any spikes we find and the weather conditions that could have caused these spikes. Provided by the 311 smoking violations data in 2017, approximately 16% of the complaints filed were from Queens, 20% from the Bronx, 32% from Brooklyn, 28% from Manhattan, and 4% from Staten Island.

The Cloud Collective will be collaborating with the NYC 311 Service and the New York City Police Department to provide a structured platform that will be used to combat smoking in New York City. Throughout this project, we will focus on implementing two data sets into a data warehouse that will then be used to analyze and visualize data as the final step. Such a process will require the understanding of both data sets, as well as, the tools to carry out the task successfully. We hope our data findings can help law enforcements in each borough to better enforce smoking regulations throughout New York City, especially those boroughs with the highest number of complaints.

The below two lists consist of all of the available column names - or attributes - that were provided in each of the datasets we chose to work with. For better clarity, we profiled the data in each dataset and renamed the columns to better fit the data each column provided.

Attributes From 311 Dataset

- UNIQUE_KEY (Primary key)
- CREATED_DATE
- CREATED_TIME
- CLOSED_DATE
- AGENCY
- AGENCY_NAME
- COMPLAINT_TYPE
- DESCRIPTOR
- LOCATION_TYPE
- INCIDENT_ZIP
- INCIDENT_ADDRESS
- STREET_NAME
- CROSS_STREET_1
- CROSS_STREET_2
- INTERSECTION_STREET_1
- INTERSECTION_STREET_2
- ADDRESS_TYPE
- CITY
- LANDMARK
- FACILITY_TYPE
- STATUS
- DUE_DATE
- RESOLUTION_DESCRIPTION
- RESOLUTION_ACTION_UPDATED_DATE
- COMMUNITY_BOARD

- BBL
- BOROUGH
- X_COORDINATE
- Y_COORDINATE
- OPEN_DATA_CHANNEL_TYPE
- PARK_FACILITY_NAME
- PARK_BOROUGH
- VEHICLE_TYPE
- TAXI_COMPANY_BOROUGH
- TAXI_PICK_UP_LOCATION
- BRIDGE_HIGHWAY_NAME
- BRIDGE_HIGHWAY_DIRECTION
- ROAD_MAP
- BRIDGE_HIGHWAY_SEGMENT
- LATITUDE
- LONGITUDE
- LOCATION

Attributes From Weather Dataset

- WEATHER_ID (Primary key)
- ENTRY_DATE
- ENTRY_MONTH
- ENTRY_DAY
- ENTRY_YEAR
- PRCP
- SNOW
- TAVG
- TMAX
- TMIN
- WEATHER_DESC

Potential KPIs

- **Avg_temperature_per_quarter:** calculated by taking the average of average temperatures in every quarter of 2017
- **Avg_temperature_per_borough:** calculated by taking the average of average temperatures in every borough for the entire year of 2017
- **Complaints_per_borough:** calculated by taking the count of complaints in all of 2017 for each borough
- **Complaints_per_quarter:** calculated by taking the count of complaints in every quarter of 2017

Initial Dimensional Model

This was the first draft of our Dimensional Model. We had initially thought the 2 datasets we have would each need their own fact table and that every column in our dataset we intended to use for our analysis would be its own dimension table.

Figure 1

Revised Dimensional Model

For our second draft of the dimensional model, we created only one fact table in the middle that had all of our KPIs, and then grouped our attributes into 3 different categories - NYC Weather, Complaints, and Location. From those 3 categories, we created our dim tables and assigned the appropropriate attributes to each dim table.

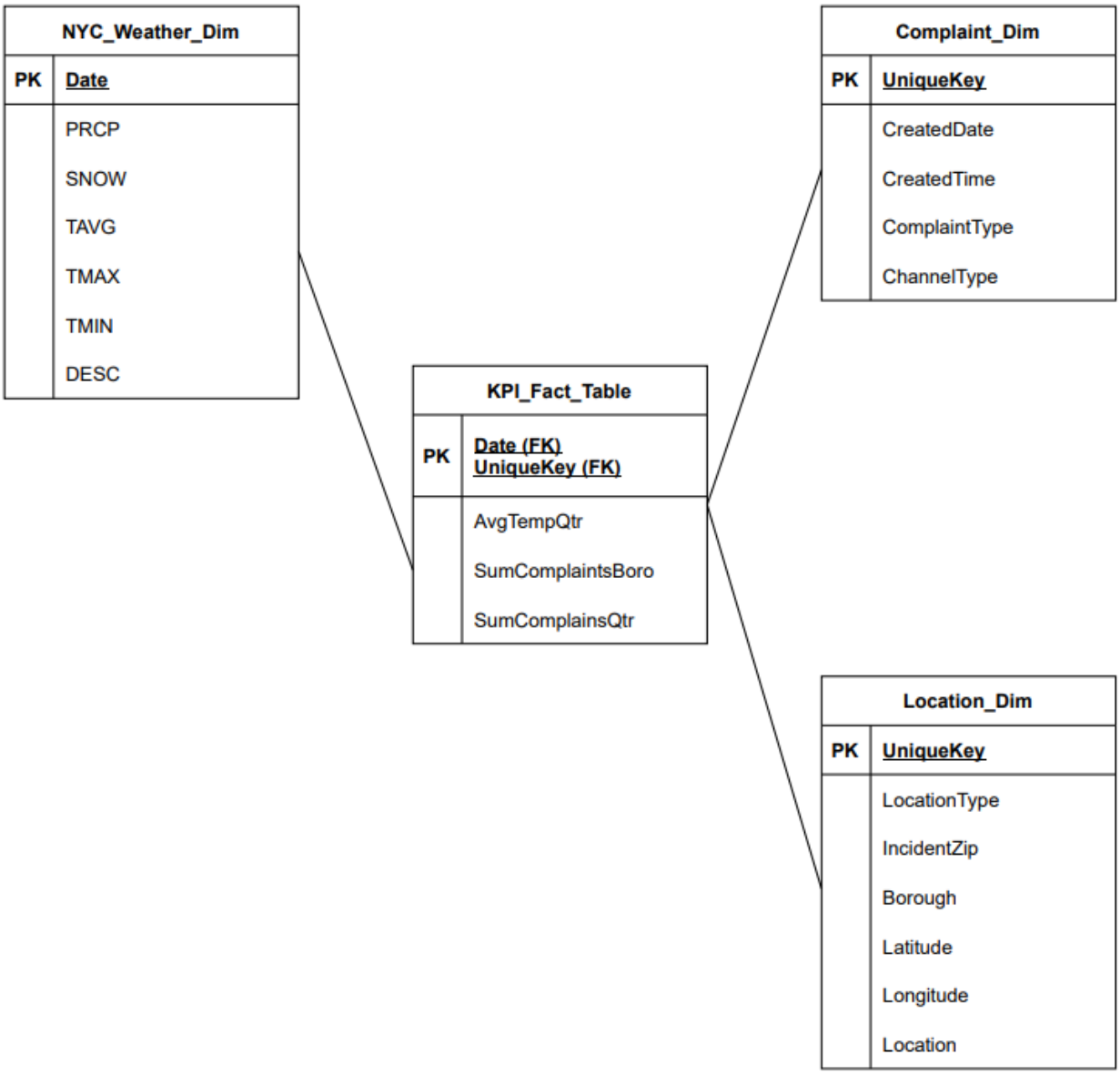


Figure 2

Finalized Dimensional Model

For our final draft, we moved all of our numerical values from the weather dataset into our facts table, and then also included a count of complaints in the facts table in order to calculate our KPIs. Now all of our numerical values that we intend to perform calculations on to obtain our KPIs are in the facts table, and the rest of the contextual descriptions can be found in their respectful dimensional tables if needed.

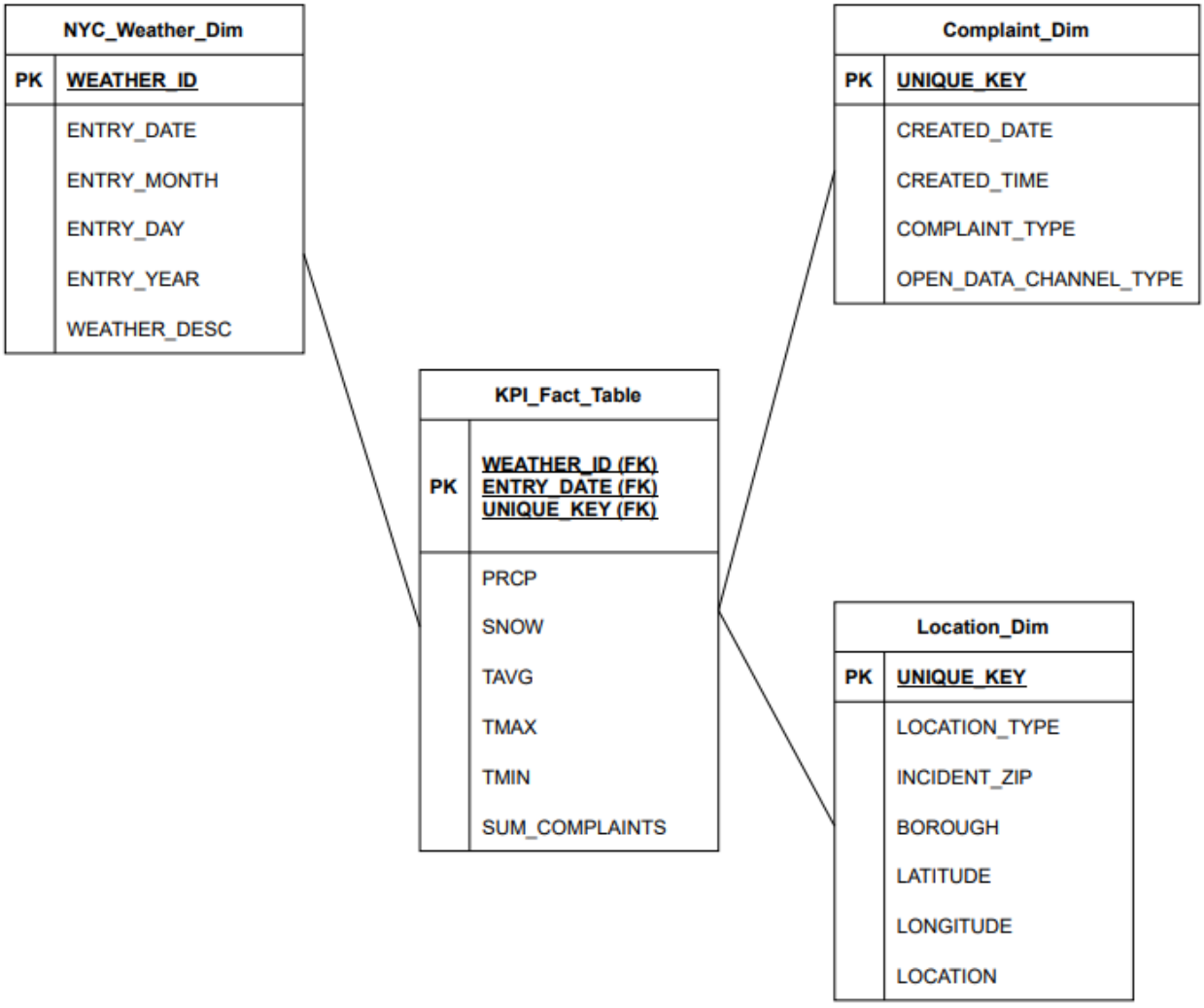
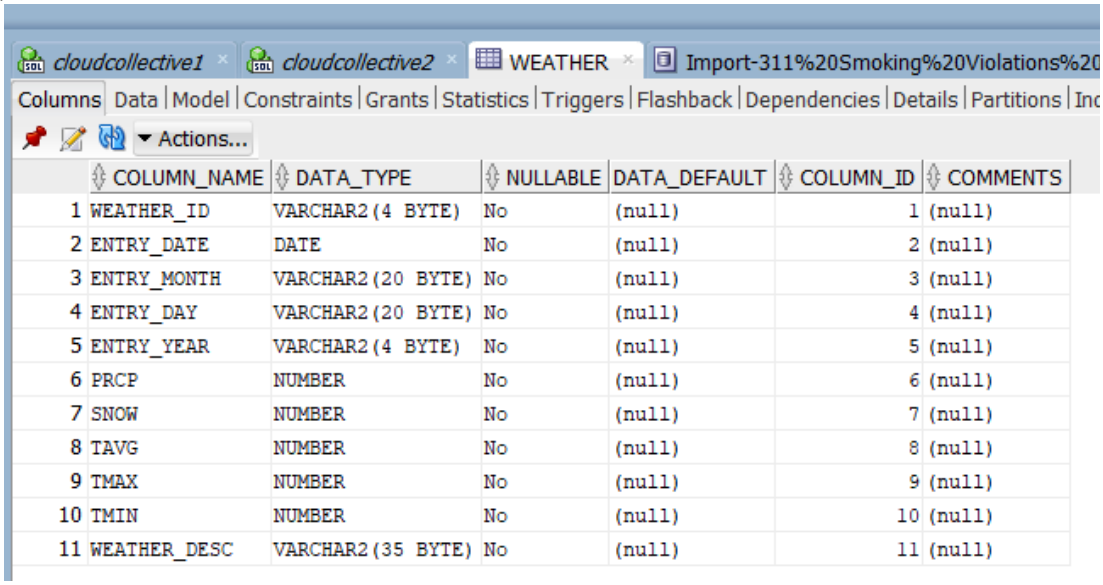


Figure 3

ETL Process

On our first attempt at our ETL process, we chose Oracle Cloud as our DBMS and set up our raw data tables using the CREATE TABLE SQL codes below. We then loaded our data for each table from the respective CSV file. We ran into a small error where a few of our rows from the CSV file weren't inserted but since the amount of unsuccessful rows was so small, the effect on the data warehouse as a whole was negligible. We kept a record of the error message we received and it can be found under the "Errors" section of this paper.

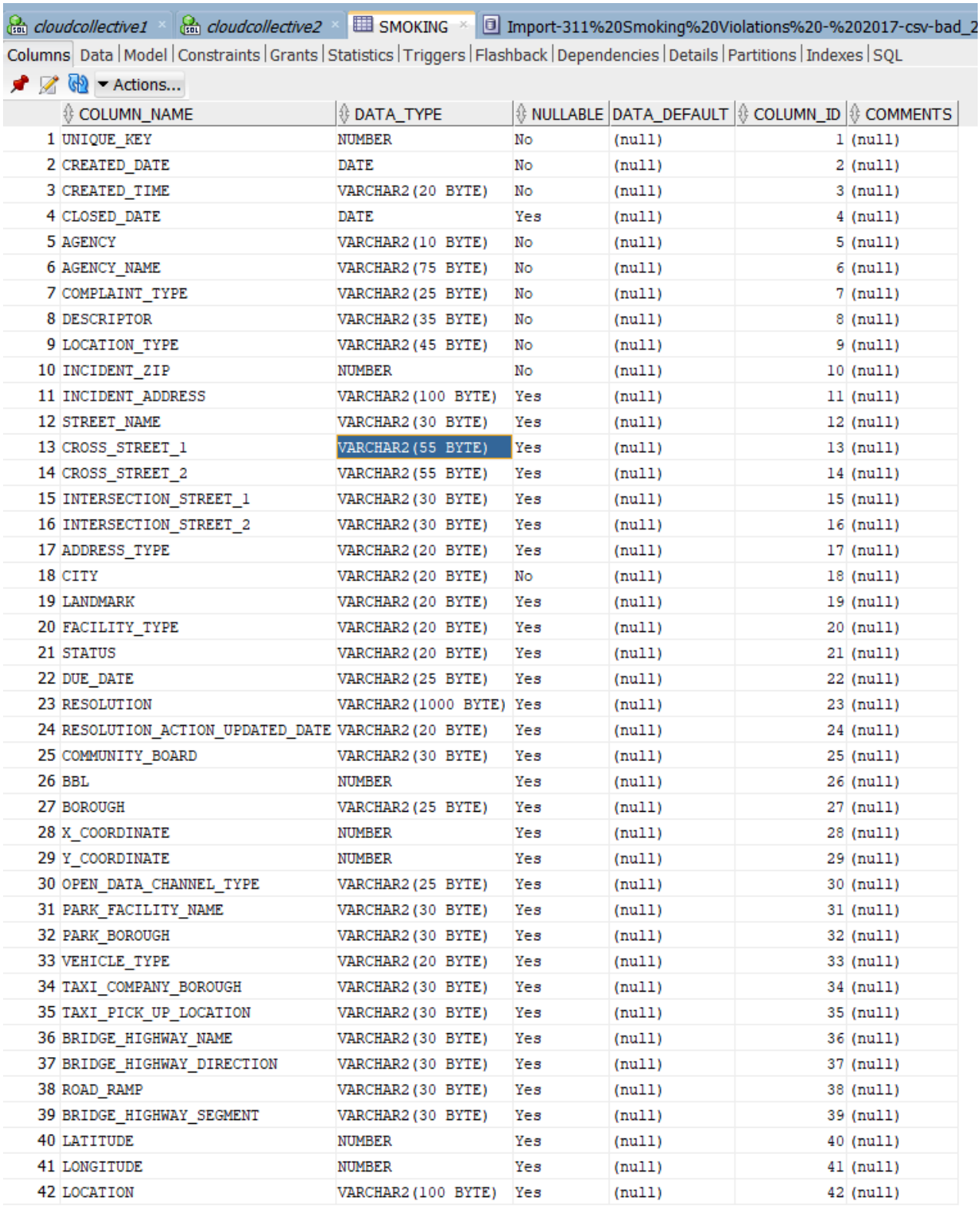
```
CREATE TABLE Weather
(
WEATHER_ID VARCHAR(4) NOT NULL PRIMARY KEY,
ENTRY_DATE DATE NOT NULL,
ENTRY_MONTH VARCHAR(20) NOT NULL,
ENTRY_DAY VARCHAR(20) NOT NULL,
ENTRY_YEAR VARCHAR(4) NOT NULL,
PRCP NUMBER NOT NULL,
SNOW NUMBER NOT NULL,
TAVG NUMBER NOT NULL,
TMAX NUMBER NOT NULL,
TMIN NUMBER NOT NULL,
WEATHER_DESC VARCHAR(20) NOT NULL
);
```



	COLUMN_NAME	DATA_TYPE	NULLABLE	DATA_DEFAULT	COLUMN_ID	COMMENTS
1	WEATHER_ID	VARCHAR2(4 BYTE)	No	(null)	1	(null)
2	ENTRY_DATE	DATE	No	(null)	2	(null)
3	ENTRY_MONTH	VARCHAR2(20 BYTE)	No	(null)	3	(null)
4	ENTRY_DAY	VARCHAR2(20 BYTE)	No	(null)	4	(null)
5	ENTRY_YEAR	VARCHAR2(4 BYTE)	No	(null)	5	(null)
6	PRCP	NUMBER	No	(null)	6	(null)
7	SNOW	NUMBER	No	(null)	7	(null)
8	TAVG	NUMBER	No	(null)	8	(null)
9	TMAX	NUMBER	No	(null)	9	(null)
10	TMIN	NUMBER	No	(null)	10	(null)
11	WEATHER_DESC	VARCHAR2(35 BYTE)	No	(null)	11	(null)

Figure 4

```
CREATE TABLE SMOKING
(
   UNIQUE_KEY NUMBER PRIMARY KEY,
   CREATED_DATE DATE NOT NULL,
   CREATED_TIME VARCHAR(30) NOT NULL,
   CLOSED_DATE DATE,
   AGENCY VARCHAR(10) NOT NULL,
   AGENCY_NAME VARCHAR(75) NOT NULL,
   COMPLAINT_TYPE VARCHAR(25) NOT NULL,
   DESCRIPTOR VARCHAR(35) NOT NULL,
   LOCATION_TYPE VARCHAR(45) NOT NULL,
   INCIDENT_ZIP NUMBER NOT NULL,
   INCIDENT_ADDRESS VARCHAR(100),
   STREET_NAME VARCHAR(30),
   CROSS_STREET_1 VARCHAR(55),
   CROSS_STREET_2 VARCHAR(55),
   INTERSECTION_STREET_1 VARCHAR(30),
   INTERSECTION_STREET_2 VARCHAR(30),
   ADDRESS_TYPE VARCHAR(20),
   CITY VARCHAR(20) NOT NULL,
   LANDMARK VARCHAR(20),
   FACILITY_TYPE VARCHAR(20),
   STATUS VARCHAR(20),
   DUE_DATE VARCHAR(25),
   RESOLUTION VARCHAR(1000),
   RESOLUTION_ACTION_UPDATED_DATE VARCHAR(25),
   COMMUNITY_BOARD VARCHAR(30),
   BBL NUMBER,
   BOROUGH VARCHAR(25),
   X_COORDINATE NUMBER,
   Y_COORDINATE NUMBER,
   OPEN_DATA_CHANNEL_TYPE VARCHAR(25),
   PARK_FACILITY_NAME VARCHAR(30),
   PARK_BOROUGH VARCHAR(30),
   VEHICLE_TYPE  VARCHAR(20),
   TAXI_COMPANY_BOROUGH VARCHAR(30),
   TAXI_PICK_UP_LOCATION VARCHAR(30),
   BRIDGE_HIGHWAY_NAME VARCHAR(30),
   BRIDGE_HIGHWAY_DIRECTION VARCHAR(30),
   ROAD_RAMP VARCHAR(30),
   BRIDGE_HIGHWAY_SEGMENT VARCHAR(30),
   LATITUDE NUMBER,
   LONGITUDE NUMBER,
   LOCATION VARCHAR(100)
);
```



cloudcollective1 | cloudcollective2 | SMOKING | Import-311%20Smoking%20Violations%20-%202017-csv-bad_2

Columns | Data | Model | Constraints | Grants | Statistics | Triggers | Flashback | Dependencies | Details | Partitions | Indexes | SQL

Actions...

	COLUMN_NAME	DATA_TYPE	NULLABLE	DATA_DEFAULT	COLUMN_ID	COMMENTS
1	UNIQUE_KEY	NUMBER	No	(null)	1	(null)
2	CREATED_DATE	DATE	No	(null)	2	(null)
3	CREATED_TIME	VARCHAR2(20 BYTE)	No	(null)	3	(null)
4	CLOSED_DATE	DATE	Yes	(null)	4	(null)
5	AGENCY	VARCHAR2(10 BYTE)	No	(null)	5	(null)
6	AGENCY_NAME	VARCHAR2(75 BYTE)	No	(null)	6	(null)
7	COMPLAINT_TYPE	VARCHAR2(25 BYTE)	No	(null)	7	(null)
8	DESCRIPTOR	VARCHAR2(35 BYTE)	No	(null)	8	(null)
9	LOCATION_TYPE	VARCHAR2(45 BYTE)	No	(null)	9	(null)
10	INCIDENT_ZIP	NUMBER	No	(null)	10	(null)
11	INCIDENT_ADDRESS	VARCHAR2(100 BYTE)	Yes	(null)	11	(null)
12	STREET_NAME	VARCHAR2(30 BYTE)	Yes	(null)	12	(null)
13	CROSS_STREET_1	VARCHAR2(55 BYTE)	Yes	(null)	13	(null)
14	CROSS_STREET_2	VARCHAR2(55 BYTE)	Yes	(null)	14	(null)
15	INTERSECTION_STREET_1	VARCHAR2(30 BYTE)	Yes	(null)	15	(null)
16	INTERSECTION_STREET_2	VARCHAR2(30 BYTE)	Yes	(null)	16	(null)
17	ADDRESS_TYPE	VARCHAR2(20 BYTE)	Yes	(null)	17	(null)
18	CITY	VARCHAR2(20 BYTE)	No	(null)	18	(null)
19	LANDMARK	VARCHAR2(20 BYTE)	Yes	(null)	19	(null)
20	FACILITY_TYPE	VARCHAR2(20 BYTE)	Yes	(null)	20	(null)
21	STATUS	VARCHAR2(20 BYTE)	Yes	(null)	21	(null)
22	DUE_DATE	VARCHAR2(25 BYTE)	Yes	(null)	22	(null)
23	RESOLUTION	VARCHAR2(1000 BYTE)	Yes	(null)	23	(null)
24	RESOLUTION_ACTION_UPDATED_DATE	VARCHAR2(20 BYTE)	Yes	(null)	24	(null)
25	COMMUNITY_BOARD	VARCHAR2(30 BYTE)	Yes	(null)	25	(null)
26	BBL	NUMBER	Yes	(null)	26	(null)
27	BOROUGH	VARCHAR2(25 BYTE)	Yes	(null)	27	(null)
28	X_COORDINATE	NUMBER	Yes	(null)	28	(null)
29	Y_COORDINATE	NUMBER	Yes	(null)	29	(null)
30	OPEN_DATA_CHANNEL_TYPE	VARCHAR2(25 BYTE)	Yes	(null)	30	(null)
31	PARK_FACILITY_NAME	VARCHAR2(30 BYTE)	Yes	(null)	31	(null)
32	PARK_BOROUGH	VARCHAR2(30 BYTE)	Yes	(null)	32	(null)
33	VEHICLE_TYPE	VARCHAR2(20 BYTE)	Yes	(null)	33	(null)
34	TAXI_COMPANY_BOROUGH	VARCHAR2(30 BYTE)	Yes	(null)	34	(null)
35	TAXI_PICK_UP_LOCATION	VARCHAR2(30 BYTE)	Yes	(null)	35	(null)
36	BRIDGE_HIGHWAY_NAME	VARCHAR2(30 BYTE)	Yes	(null)	36	(null)
37	BRIDGE_HIGHWAY_DIRECTION	VARCHAR2(30 BYTE)	Yes	(null)	37	(null)
38	ROAD_RAMP	VARCHAR2(30 BYTE)	Yes	(null)	38	(null)
39	BRIDGE_HIGHWAY_SEGMENT	VARCHAR2(30 BYTE)	Yes	(null)	39	(null)
40	LATITUDE	NUMBER	Yes	(null)	40	(null)
41	LONGITUDE	NUMBER	Yes	(null)	41	(null)
42	LOCATION	VARCHAR2(100 BYTE)	Yes	(null)	42	(null)

Figure 5

We then attempted to connect Pentaho to our Oracle Database but ran into numerous issues due to the instructions provided being outdated and the files available not matching up with the instructions. After numerous unsuccessful attempts and countless hours spent trying to integrate Pentaho with Oracle Cloud, we decided to switch over to Google BigQuery for our DBMS and dbt for our ETL tool.

After importing our NYC weather and 311 Complaints datasets onto BigQuery *(figure 6)* and ensuring all data types were accurate within the schema of each dataset, we began constructing our dimensional tables - *NYC_weather_dim (figure 7), complaint_dim (figure 8), and location_dim (figure 9)* - and fact table - *facts_table (figure 10)* - in dbt. In our SQL coding for each table, we made sure to include a config block at the top of the code that materialized each SQL code as a table and not a view.

Once our SQL tables were all set up, we began working on our KPI calculations. The first KPI we calculated was the total complaints by borough - we calculated this by counting the unique keys in our complaints dim table and then grouping by the boroughs *(figure 11)*. Next was the average annual temperature by borough where we found the mean of the daily average temperatures and grouped by borough. The results came out similar between all of them due to being around the same geographical location *(figure 12)*. We created separate SQL statements for each individual quarter to find the average of the temperatures for those months *(figures 13-16)*. For our remaining figures, we created separate SQL statements to calculate the count of total complaints for each individual quarter as well *(figures 17-20)*.

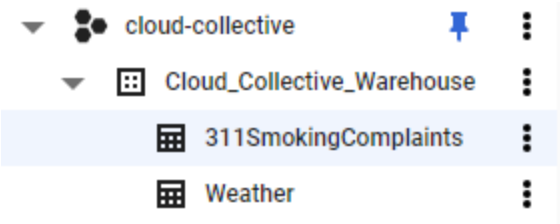


Figure 6

```
{{ config (
    materialized="table"
)}}

with weather as (

    select
        weather_id,
        entry_date,
        entry_month,
        entry_day,
        entry_year,
        weather_desc

    from cloud-collective.Cloud_Collective_Warehouse.Weather
    order by entry_date
)

select * from weather
```

Figure 7

complaint_dim.sql

```
{{ config (
    materialized="table"
)}}

with complaint as (

    select

        unique_key,
        Created_Date,
        Created_Time,
        Complaint_Type,
        OPEN_DATA_CHANNEL_TYPE as Channel_Type

    from cloud-collective.Cloud_Collective_Warehouse.311SmokingComplaints
)

select * from complaint
```

Figure 8

location_dim.sql

```
{{ config (
    materialized="table"
)}}

with location as (

    select

        unique_key,
        Location_Type,
        Incident_Zip,
        Borough,
        Latitude,
        Longitude,
        Location

    from cloud-collective.Cloud_Collective_Warehouse.311SmokingComplaints
)

select * from location
```

Figure 9

```
{{ config (
    materialized="table"
)}}

with kpifcts as (
    select
        unique_key,
        weather_id,
        entry_date,
        PRCP as precipitation,
        SNOW as snow,
        TAVG as avg_temp,
        TMAX as max_temp,
        TMIN as min_temp

    from cloud-collective.Cloud_Collective_Warehouse.Weather, cloud-collective.Cloud_Collective_Warehouse.311SmokingComplaints
    order by entry_date

)

select * from kpifcts
```

Figure 10

```sql
select borough, count(complaint_dim.unique_key) as Total_Complaints
from cloud-collective.dbt_cloudcollective.complaint_dim, cloud-collective.dbt_cloudcollective.location_dim
group by borough
```

borough	Total_Complaints
MANHATTAN	1606142
STATEN ISLAND	224002
QUEENS	891242
BROOKLYN	1825378
BRONX	1131925

Figure 11

```sql
select borough,
avg(avg_temp) as avg_annual_temp
from cloud-collective.dbt_cloudcollective.facts_table, cloud-collective.dbt_cloudcollective.location_dim
Group by location_dim.borough
```

borough	avg_annual_temp
MANHATTAN	55.00547945205485
STATEN ISLAND	55.00547945205481
QUEENS	55.00547945205481
BROOKLYN	55.00547945205486
BRONX	55.005479452054814

Figure 12

```
1  select
2  NYC_weather_dim.entry_month,
3  avg(facts_table.avg_temp) as average_temp
4  from cloud-collective.dbt_cloudcollective.facts_table, cloud-collective.dbt_cloudcollective.NYC_weather_dim
5  where facts_table.entry_date between '2017-01-01' and '2017-03-31'
6  group by NYC_weather_dim.entry_month
7  having entry_month = 'January' or entry_month = 'February' or entry_month = 'March'
```

preview data | compile sql | Results | Compiled SQL

2.5 sec —Returned 3 rows.

entry_month	average_temp
February	37.288888888888884
March	37.288888888888884
January	37.288888888888884

Figure 13

```
1  select
2  NYC_weather_dim.entry_month,
3  avg(facts_table.avg_temp) as average_temp
4  from cloud-collective.dbt_cloudcollective.facts_table, cloud-collective.dbt_cloudcollective.NYC_weather_dim
5  where facts_table.entry_date between '2017-04-01' and '2017-06-30'
6  group by NYC_weather_dim.entry_month
7  having entry_month = 'April' or entry_month = 'May' or entry_month = 'June'
```

preview data | compile sql | Results | Compiled SQL

3.1 sec —Returned 3 rows.

entry_month	average_temp
April	62.131868131868124
May	62.131868131868124
June	62.131868131868124

Figure 14

```sql
select
NYC_weather_dim.entry_month,
avg(facts_table.avg_temp) as average_temp
from cloud-collective.dbt_cloudcollective.facts_table, cloud-collective.dbt_cloudcollective.NYC_weather_dim
where facts_table.entry_date between '2017-07-01' and '2017-09-30'
group by NYC_weather_dim.entry_month
having entry_month = 'July' or entry_month = 'August' or entry_month = 'September'
```

preview data | compile sql | Results | Compiled SQL

3.3 sec —Returned 3 rows.

entry_month	average_temp
August	72.18478260869567
September	72.18478260869566
July	72.18478260869567

Figure 15

```sql
select
NYC_weather_dim.entry_month,
avg(facts_table.avg_temp) as average_temp
from cloud-collective.dbt_cloudcollective.facts_table, cloud-collective.dbt_cloudcollective.NYC_weather_dim
where facts_table.entry_date between '2017-10-01' and '2017-12-31'
group by NYC_weather_dim.entry_month
having entry_month = 'October' or entry_month = 'November' or entry_month = 'December'
```

preview data | compile sql | Results | Compiled SQL

3.4 sec —Returned 3 rows.

entry_month	average_temp
December	48.10869565217391
October	48.10869565217391
November	48.10869565217391

Figure 16

```sql
select NYC_weather_dim.entry_month, count(complaint_dim.complaint_type) as Total_Complaints
from cloud-collective.dbt_cloudcollective.complaint_dim
full join cloud-collective.dbt_cloudcollective.NYC_weather_dim
on complaint_dim.Created_Date = NYC_weather_dim.entry_date
where NYC_weather_dim.entry_date between '2017-01-01' and '2017-03-31'
group by NYC_weather_dim.entry_month
having entry_month = 'January' or entry_month = 'February' or entry_month = 'March'
```

preview data | compile sql | Results | Compil

3.1 sec —Returned 3 rows.

entry_month	Total_Complaints
March	171
February	143
January	172

Figure 17

```sql
select NYC_weather_dim.entry_month, count(complaint_dim.complaint_type) as Total_Complaints
from cloud-collective.dbt_cloudcollective.complaint_dim
full join cloud-collective.dbt_cloudcollective.NYC_weather_dim
on complaint_dim.Created_Date = NYC_weather_dim.entry_date
where NYC_weather_dim.entry_date between '2017-04-01' and '2017-06-30'
group by NYC_weather_dim.entry_month
having entry_month = 'April' or entry_month = 'May' or entry_month = 'June'
```

preview data | compile sql | Results | Compi

3.5 sec —Returned 3 rows.

entry_month	Total_Complaints
May	300
June	205
April	216

Figure 18

```sql
select NYC_weather_dim.entry_month, count(complaint_dim.complaint_type) as Total_Complaints
from cloud-collective.dbt_cloudcollective.complaint_dim
full join cloud-collective.dbt_cloudcollective.NYC_weather_dim
on complaint_dim.Created_Date = NYC_weather_dim.entry_date
where NYC_weather_dim.entry_date between '2017-07-01' and '2017-09-30'
group by NYC_weather_dim.entry_month
having entry_month = 'July' or entry_month = 'August' or entry_month = 'September'
```

preview data | compile sql | Results | Compi

4.1 sec —Returned 3 rows.

entry_month	Total_Complaints
July	140
August	166
September	231

Figure 19

```sql
select NYC_weather_dim.entry_month, count(complaint_dim.complaint_type) as Total_Complaints
from cloud-collective.dbt_cloudcollective.complaint_dim
full join cloud-collective.dbt_cloudcollective.NYC_weather_dim
on complaint_dim.Created_Date = NYC_weather_dim.entry_date
where NYC_weather_dim.entry_date between '2017-10-01' and '2017-12-31'
group by NYC_weather_dim.entry_month
having entry_month = 'October' or entry_month = 'November' or entry_month = 'December'
```

preview data | compile sql | Results | Compi

3.4 sec —Returned 3 rows.

entry_month	Total_Complaints
November	228
December	215
October	196

Figure 20

Dashboard Programming

We chose to integrate Tableau with our BigQuery database to create our visualizations after numerous failed attempts at integrating Pentaho with Oracle. What we were most interested in seeing from our raw data is which boroughs had the most complaints in 2017, and then how the weather may have had an impact on the number of complaints filed in 2017. *Figures 21 and 22* show the total number of complaints broken up by borough. From the heatmap in *figure 21*, we can see that Brooklyn and Manhattan have the highest complaint counts, and the statistics provided in *figure 22* confirm that. *Figure 23* is a bar line graph that shows the monthly average of the average temperatures for all the boroughs (depicted by the line graph) overlaying the total sum of monthly complaints for all the boroughs (depicted by the bar graphs). We wanted to further break down the statistics shown in *figure 23*, so we adjusted the line graph data to be a daily measure rather than a monthly measure, as depicted in *figure 24*. Next, we wanted to see if the amount of precipitation and snow had any solid correlation with the amount of complaints filed for all boroughs per month. *Figure 25* shows that precipitation had some inverse correlation, as the complaint counts went up during most months when precipitation was high and went down during the months when precipitation was low. *Figure 25* also shows us that snow did have a more direct correlation. We can see that in the months of January, February and March there was a lot of snow, and there were also less complaints filed. Lastly, we wanted to include a visualization that grouped smoking complaints by the weather description, since only having precipitation and snow measures was not enough for us to really understand if smoking complaints went down when the weather was not sunny and warm. *Figure 26* shows us that smoking complaints were highest when the sky was clear and sunny, but it also shows us that the complaints were still high during cloudy days as well.

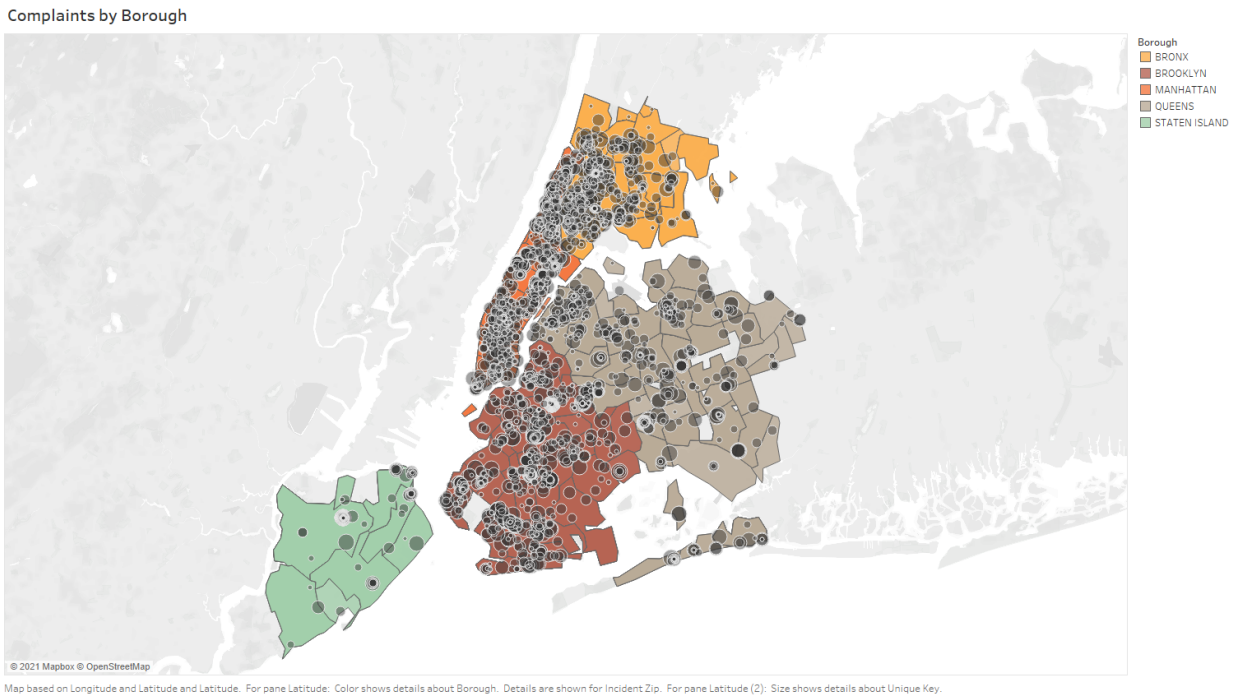

Complaints by Borough
Borough
BRONX
BROOKLYN
MANHATTAN
QUEENS
STATEN ISLAND
© 2021 Mapbox © OpenStreetMap
Map based on Longitude and Latitude and Latitude. For pane Latitude: Color shows details about Borough. Details are shown for Incident Zip. For pane Latitude (2): Size shows details about Unique Key.

Figure 21

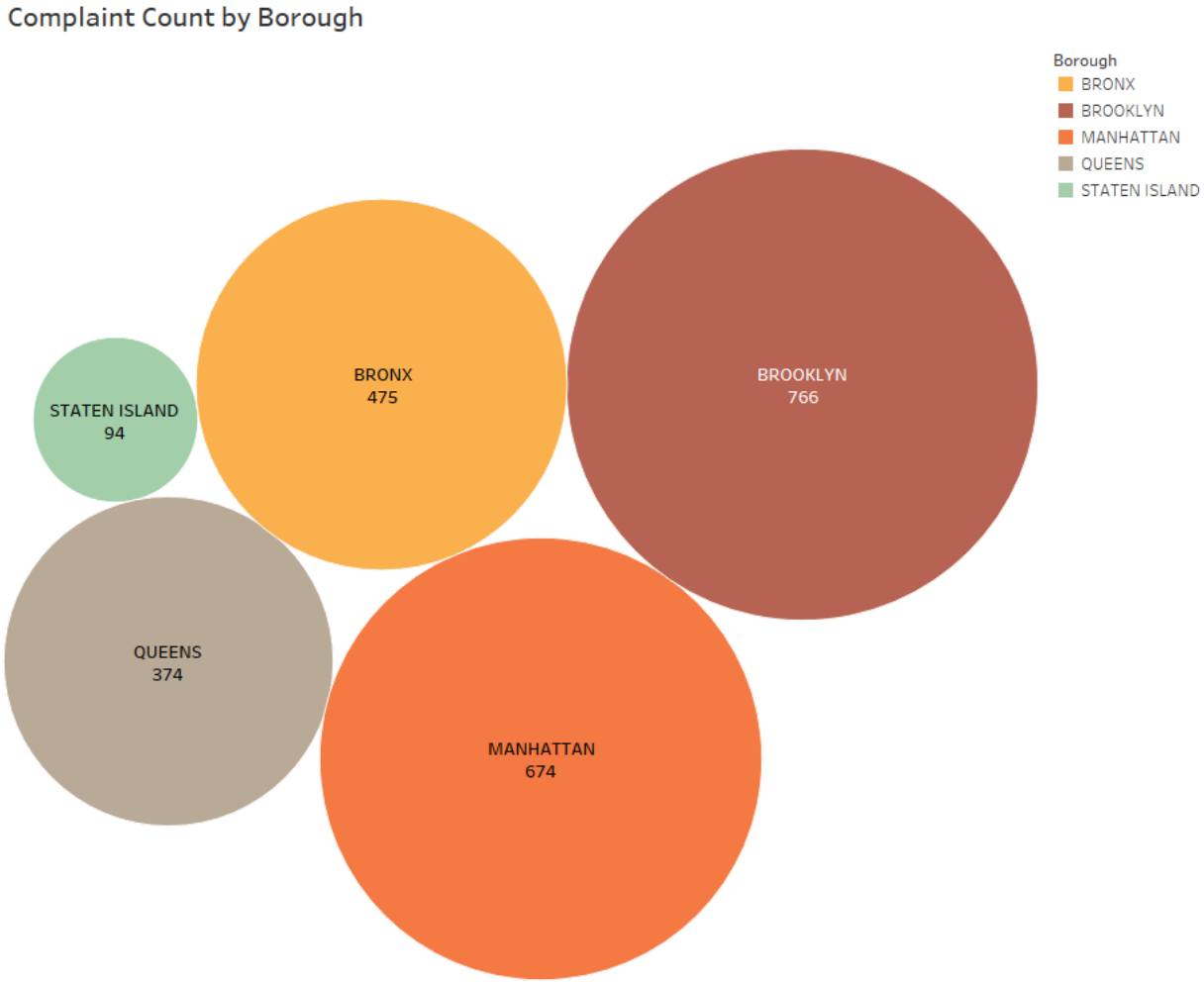

Complaint Count by Borough
Borough
BRONX
BROOKLYN
MANHATTAN
QUEENS
STATEN ISLAND
BRONX
475
BROOKLYN
766
STATEN ISLAND
94
QUEENS
374
MANHATTAN
674
Borough and count of Unique Key. Color shows details about Borough. Size shows count of Unique Key. The marks are labeled by Borough and count of Unique Key.

Figure 22

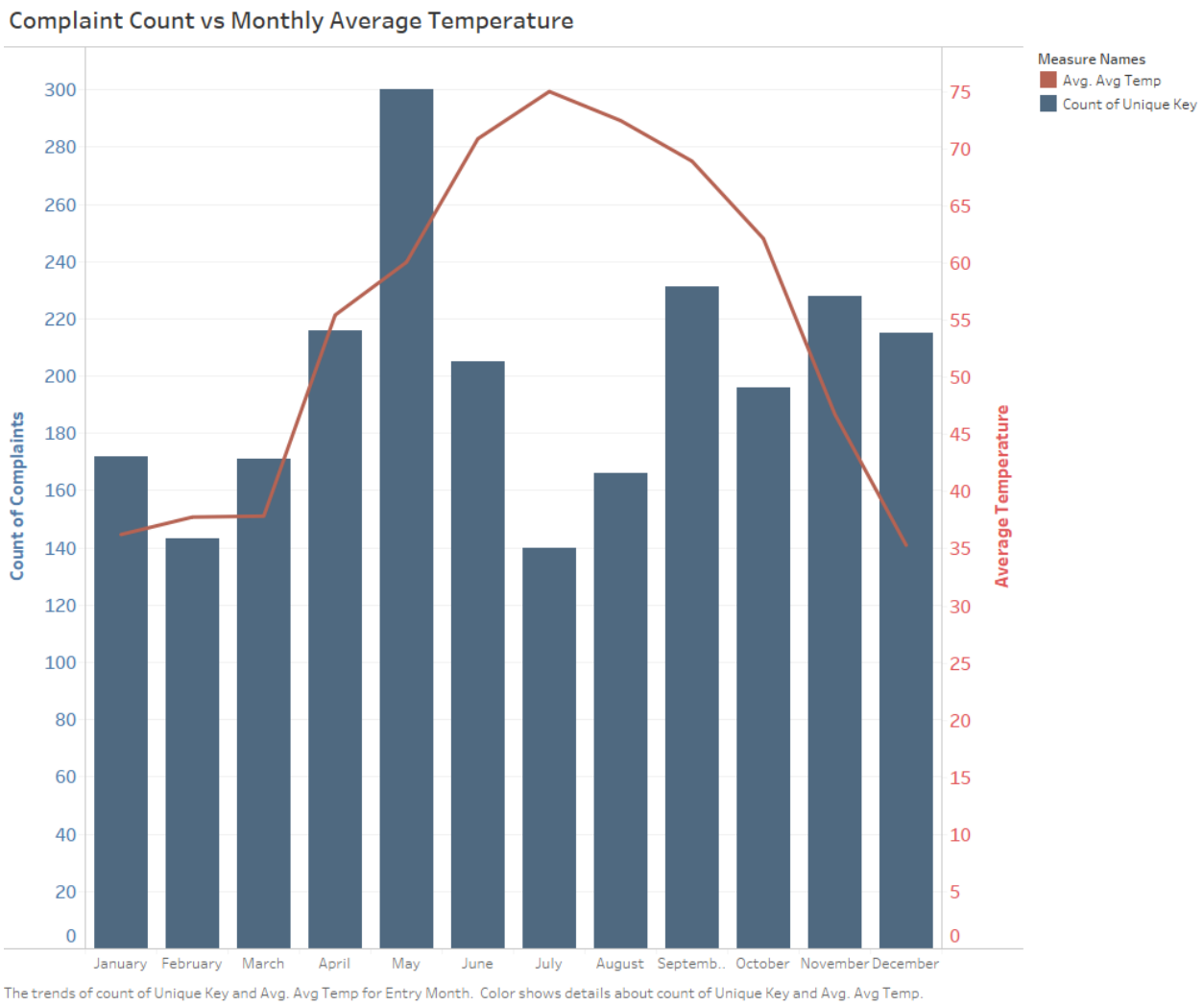


Figure 23

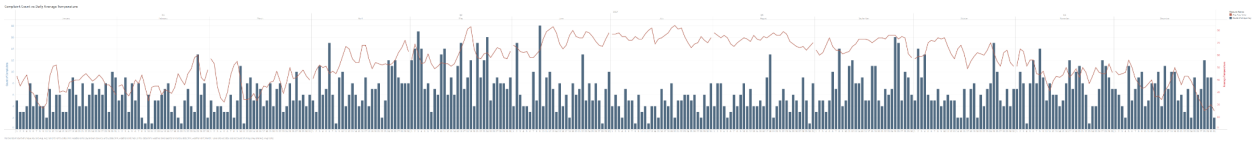


Figure 24

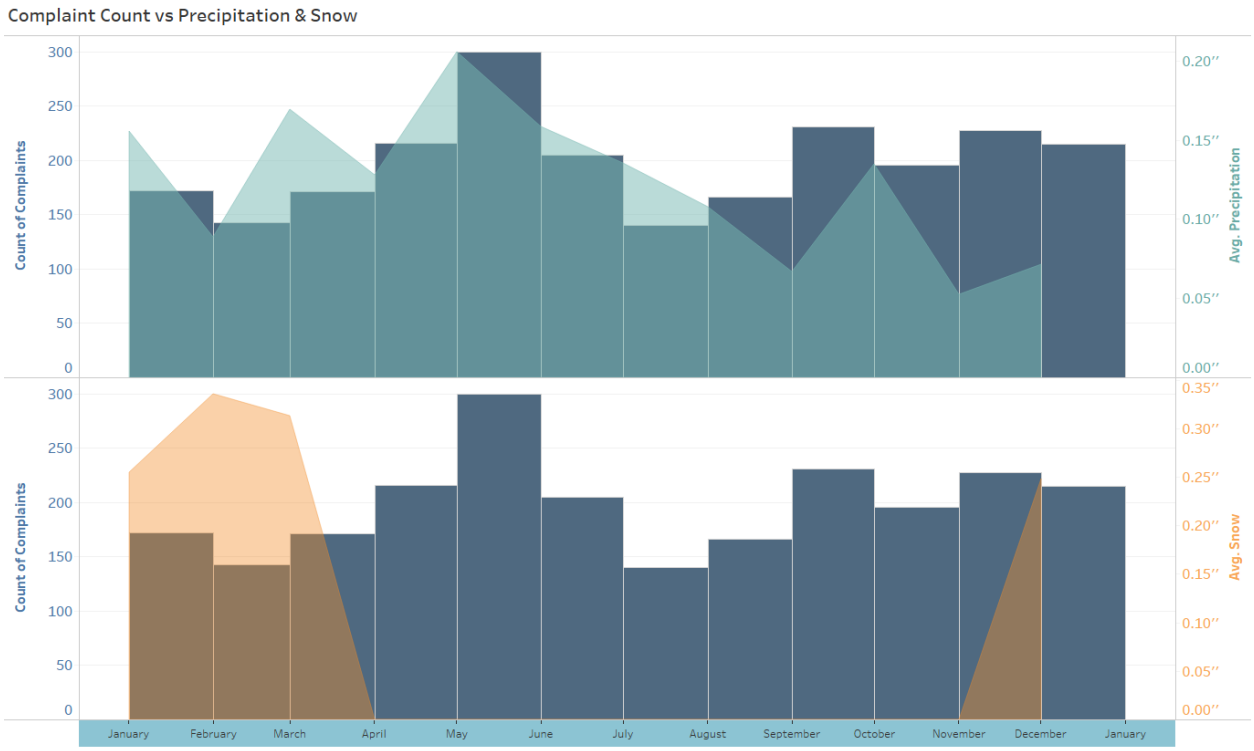

Complaint Count vs Precipitation & Snow
Count of Complaints
300
250
200
150
100
50
0
0.20''
0.15''
0.10''
0.05''
0.00''
Avg. Precipitation
Count of Complaints
300
250
200
150
100
50
0
0.35''
0.30''
0.25''
0.20''
0.15''
0.10''
0.05''
0.00''
Avg. Snow
January
February
March
April
May
June
July
August
September
October
November
December
January
The plots of count of Unique Key, Avg. Precipitation, count of Unique Key and Avg. Snow for entry date (NYC weather dim) Month. Color shows details about count of Unique Key, Avg. Precipitation, count of Unique Key and Avg. Snow.

Figure 25

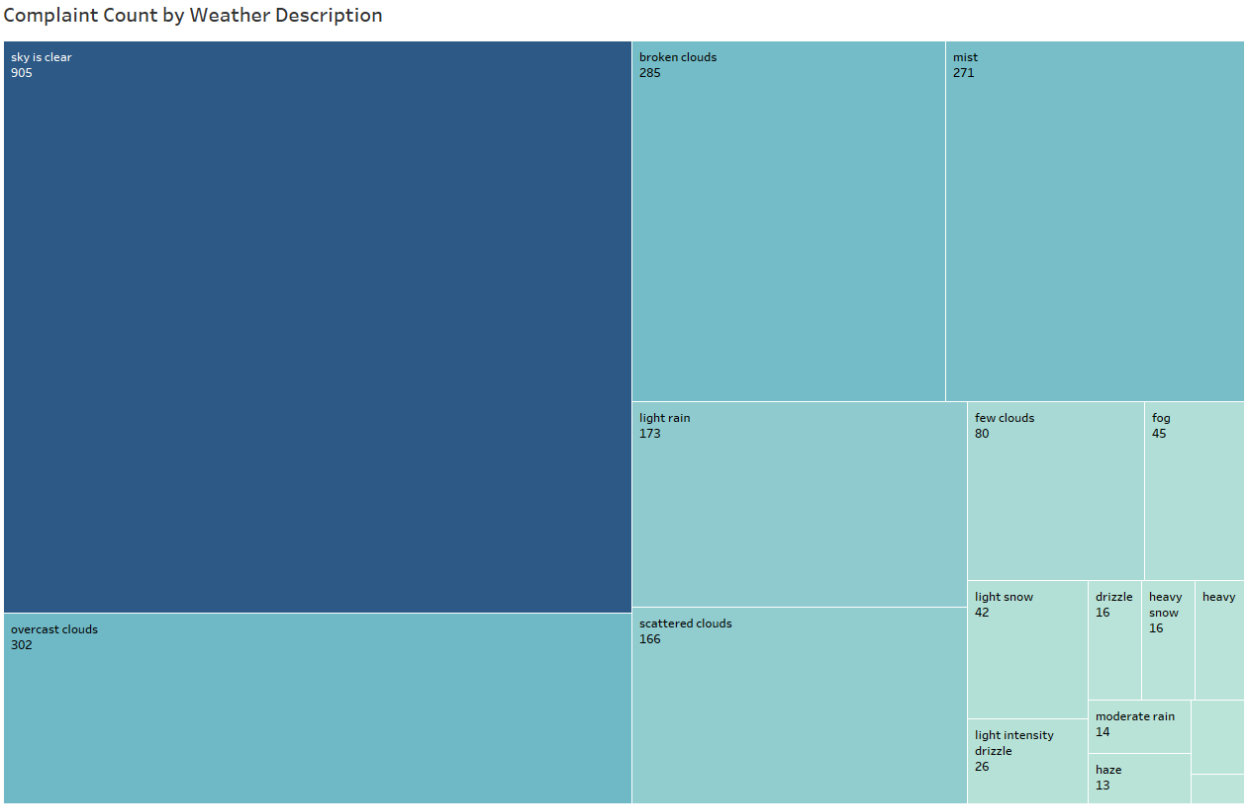

Complaint Count by Weather Description
sky is clear
905
broken clouds
285
mist
271
light rain
173
few clouds
80
fog
45
light snow
42
drizzle
16
heavy snow
16
heavy
overcast clouds
302
scattered clouds
166
moderate rain
14
light intensity drizzle
26
haze
13
Weather Desc and count of Unique Key. Color shows count of Unique Key. Size shows count of Unique Key. The marks are labeled by Weather Desc and count of Unique Key.

Figure 26

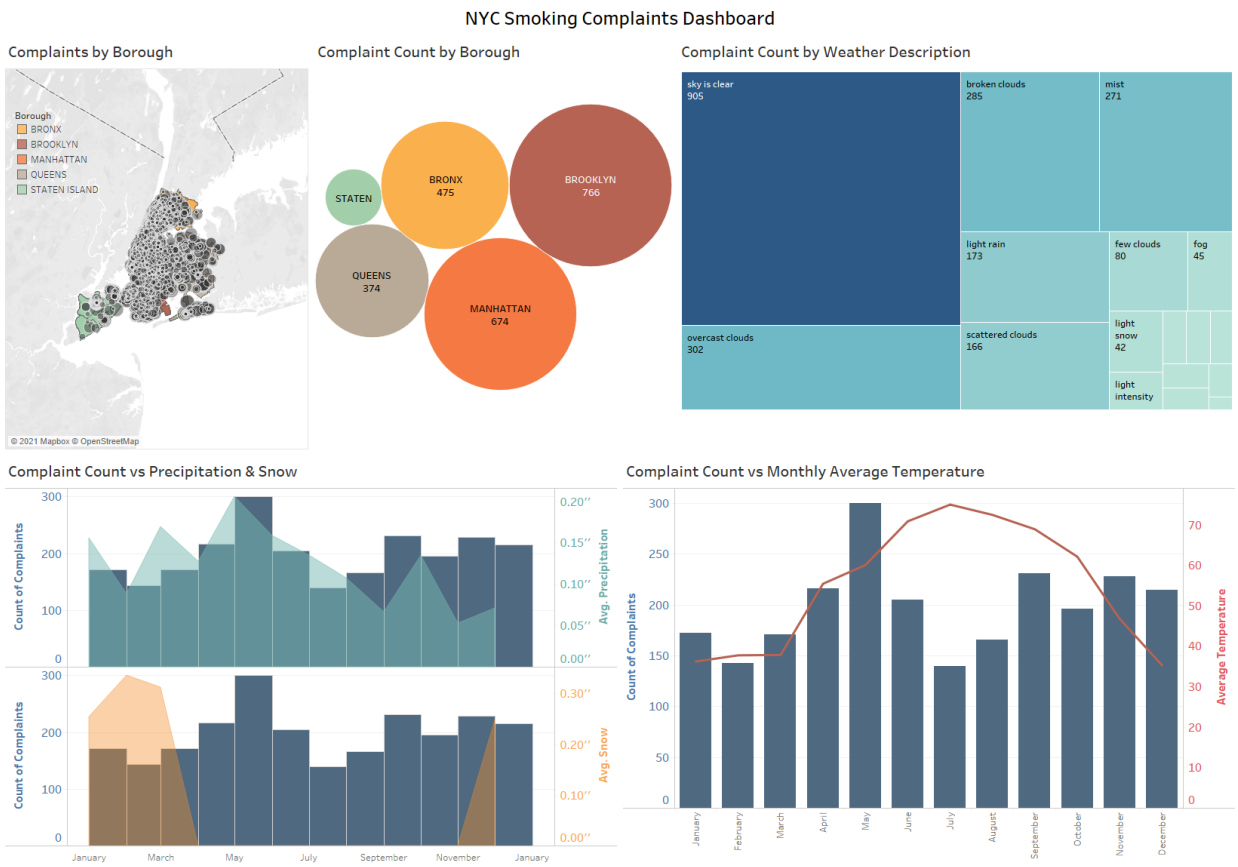


Figure 27

https://public.tableau.com/profile/andrew.hakimian#!/vizhome/CIS9440-CloudCollective-NYCSmokingComplaints/Dashboard1

Conclusion

After our analysis of the NYC 311 Smoking Complaints and Weather dataset, we found that our predictions were not fully correct. There were less complaints in the summer than there were in the winter, which we assume is due to the possibility that people may be happier in the warmer weather and choose to not focus on those around them. We also saw that smoking complaints did not decrease when precipitation was high - which is what we initially thought would be the case as most people prefer to not go outside and be in public areas when it is raining. Furthermore, despite Manhattan being more densely populated with approximately 73,000 people per square mile, we saw through our analysis and visualizations that it had less smoking complaints in 2017 than Brooklyn, which has a population density of approximately 36,732 people per square mile - half of Manhattan's population density. With the above findings in mind, we would suggest that the police force focus on increasing their efforts in the boroughs of Brooklyn and Manhattan, specifically during the months of May, September, November and December.

While working on this project, we learned a lot about the limitations and shortfalls of data availability, consistency, and reliability. We saw how dirty and incomplete data can be, which is most likely due to many firms and companies not having a predefined and set way of entering data to ensure consistency throughout the years, and we learned just how tedious and time consuming it can be to clean this data. Lastly, we learned that not every platform and product that is available for use would fit our needs, and that some may not even be compatible with each other. In our first attempt at the ETL process, we saw that our Oracle Database was not connecting to Pentaho due to outdated and mismatched files, so we opted for more modern and agile applications instead.

Sources

1. *311 Service Requests From 2010 to Present*. (2017). NYC Open Data. https://data.cityofnewyork.us/Social-Services/311-Service-Requests-from-2010-to-Present/erm2-nwe9/data
2. *Historical Hourly Weather Data 2012-2017*. (2017). Kaggle. https://www.kaggle.com/selfishgene/historical-hourly-weather-data?select=weather_description.csv
3. *Dimensional Modeling*. Draw.io https://app.diagrams.net/

Meeting Log

1. **Date & Time:** February 7, 2021 at 11:00 am - 12:15 pm

   Attendees:

   Valeria Belozertsev, Youming Chen, Melissa Ha, Andrew Hakimian, Tevin Harris

   Topic Discussed:

   At our first group meeting on February 7th, we met and introduced ourselves. In addition, we brainstormed our initial topic as a group as well as discussed what approach we will be taking in order to successfully complete the project.

2. **Date & Time:** February 14, 2021 at 11:00 am - 12:30 pm

   Attendees:

   Valeria Belozertsev, Youming Chen, Melissa Ha, Andrew Hakimian, Tevin Harris

   Topic Discussed:

   At our second group meeting on February 14th, we discussed what data sets would be the best to use as well as what KPIs we would use to monitor the data sets. Also, we discussed what BI tool we will be using in order to visualize the results from our Data warehouse.

3. **Date & Time:** February 22, 2021 at 9:00 pm - 10:30 pm

   Attendees:

   Valeria Belozertsev, Youming Chen, Melissa Ha, Andrew Hakimian, Tevin Harris

   Topic Discussed:

   At our third group meeting on February 22nd, we discussed our second dataset. We found two different datasets on weather that we wanted to combine, we performed a bit of data manipulation in excel in order to make sure the two datasets would be able to merge without any issues.

4. **Date & Time:** February 23, 2021 at 1:00 pm - 1:40 pm

   Attendees:

   Valeria Belozertsev, Youming Chen, Melissa Ha, Andrew Hakimian, Tevin Harris

   Topic Discussed:

   At our fourth group meeting on February 23rd, we reviewed our combined weather dataset to ensure all data is complete. We also chose the KPIs we will be focusing on for the NYC Weather dataset and added them to our project proposal. Lastly, we completed our first draft of the dimensional models for both the 311 Smoking Complaints dataset and our NYC Weather dataset, and added a screenshot to our group proposal.

5. **Date & Time:** February 25, 2021 at 9:00 pm - 9:30 pm

   Attendees:

   Valeria Belozertsev, Youming Chen, Melissa Ha, Andrew Hakimian, Tevin Harris

   Topic Discussed:

   At our fifth group meeting, we briefly met to finalize our datasets and dimensional modeling draft to be sent in as milestone #2. Sources, from where the data had been collected, have been included. The latest dimensional diagram includes both the weather and 311 service complaints with its primary keys.

6. **Date & Time:** April 13, 2021 at 9:00 pm - 11:30 pm

   Attendees:

   Valeria Belozertsev, Youming Chen, Melissa Ha, Andrew Hakimian, Tevin Harris

   Topic Discussed:

   At our sixth group meeting, we met to establish what data warehousing platform/program to use and created the setup on Oracle Cloud for the ETL programming process. We created our Oracle connections and through SQL, we created tables for our weather datasets and inserted the collected information.

7. **Date & Time:** April 16, 2021 at 5:30 pm - 8:00 pm

   Attendees:

   Valeria Belozertsev, Youming Chen, Melissa Ha, Andrew Hakimian, Tevin Harris

   Topic Discussed:

   At our seventh group meeting, we worked on our ETL programming by creating our weather and smoking tables. We uploaded our weather and 311 smoking datasets into our Oracle DBMS and started downloading pentaho.

8. **Date & Time:** April 18, 2021 at 3:15 pm - 6:00 pm

   Attendees:

   Valeria Belozertsev, Youming Chen, Melissa Ha, Andrew Hakimian, Tevin Harris

   Topic Discussed:

   At our eighth meeting, we continue to install and set up pentaho. Ran into a few hiccups due to the tutorials on Holowczak's website being outdated and the links no longer being available.

9. **Date & Time:** April 20, 2021 at 8:00 pm - 10:00 pm

   Attendees:

   Valeria Belozertsev, Youming Chen, Melissa Ha, Andrew Hakimian, Tevin Harris

   Topic Discussed:

   At the ninth meeting, we restarted the pentaho installation process because drivers and oracle are not connecting.

10. **Date & Time:** April 28, 2021 at 9:30 pm - 11:15 pm

    Attendees:

    Valeria Belozertsev, Youming Chen, Melissa Ha, Andrew Hakimian, Tevin Harris

    Topic Discussed:

    At our tenth meeting, we decided to restart our data warehouse and repositories through big query and dbt. We insert our data into Big Query and finally start to produce our dimensional tables through dbt cloud.

11. **Date & Time:** May 2, 2021 at 8:00 pm - 9:15 pm

    Attendees:

    Valeria Belozertsev, Youming Chen, Melissa Ha, Andrew Hakimian, Tevin Harris

    Topic Discussed:

    At our eleventh meeting, we revised our weather dim table to exclude all of the numerical values that should be part of our fact table, and then added in the month, day and year from the entry date to allow for our quarterly calculations to be done by month. We then coded in our facts table and coded our first KPI - the total complaints per borough. We attempted to work on the average temperatures by quarter KPI as well but ran into an issue where BigQuery wasn't displaying our new weather dim and facts tables.

12. **Date & Time:** May 4, 2021 at 9:30 pm - 11:35 pm

    Attendees:

    Valeria Belozertsev, Youming Chen, Melissa Ha, Andrew Hakimian, Tevin Harris

    Topic Discussed:

    At our twelfth meeting, we revised and edited our ETL statements to calculate our KPIs. We found our average temperatures per quarter, average temperature by borough and total complaints per month. There were many issues in our statements during the process but managed to figure out what we were stuck on. We were able to find total complaints per quarter as well, where we made separate statements for each of the quarters. Once we input screenshots of our figures, we provided explanations on its specific details and what we did.

13. **Date & Time:** May 17, 2021 at 6:30 pm - 8:20 pm

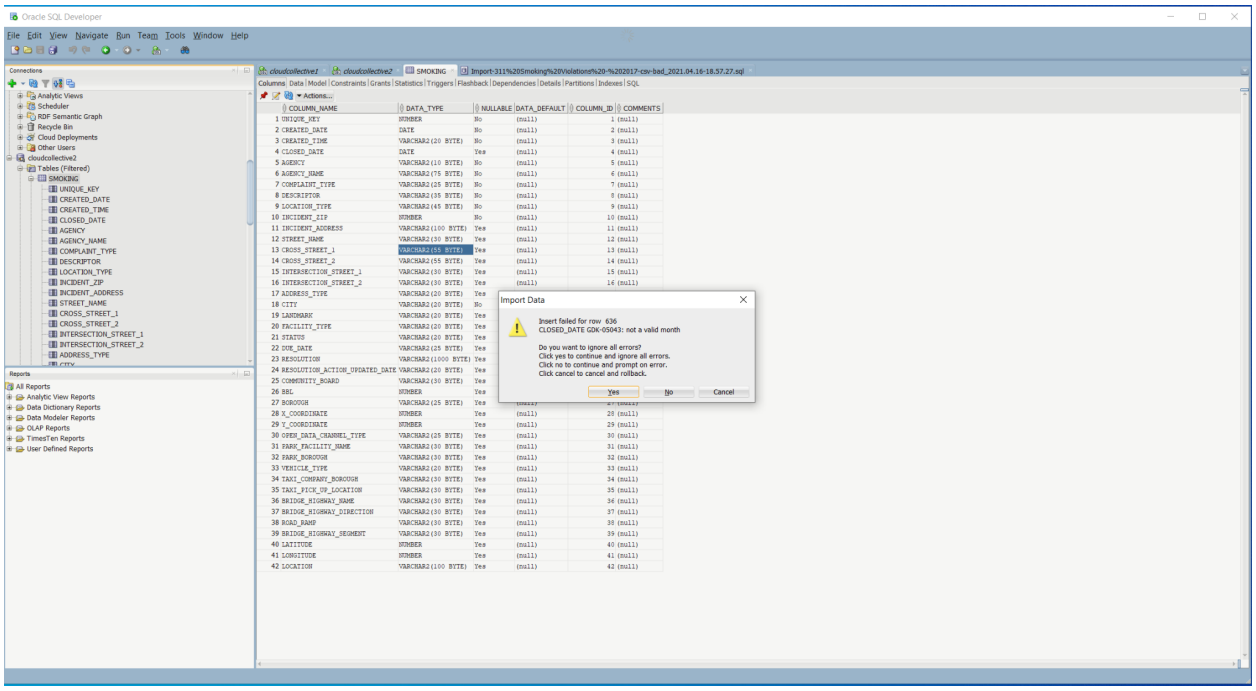
    Attendees:

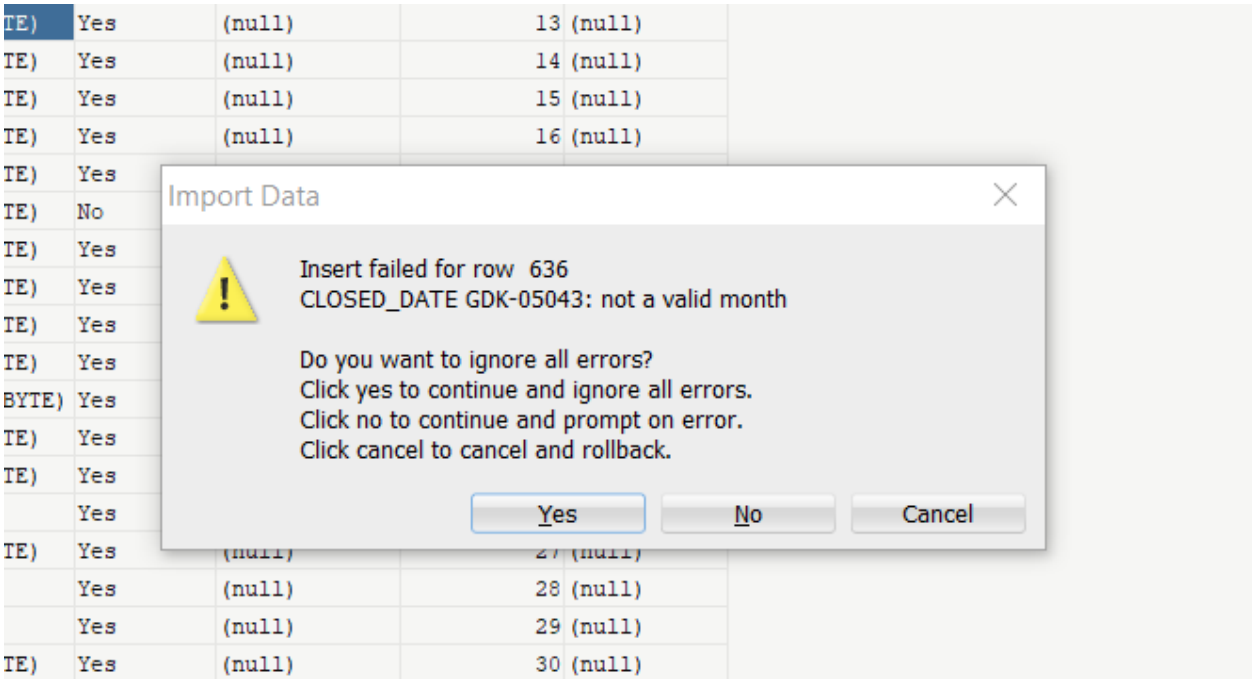
    Valeria Belozertsev, Youming Chen, Melissa Ha, Andrew Hakimian, Tevin Harris

    Topic Discussed:

    At our final meeting, we met to build the visuals from the data in Big Query into Tableau. We decided to create several graphs to illustrate the correlation of the weather and smoking complaints in New York City. After creating the visuals, we placed them in the dashboard, finalizing our dashboard programming and our project.

Errors




Import Data
Insert failed for row 636
CLOSED_DATE GDK-05043: not a valid month
Do you want to ignore all errors?
Click yes to continue and ignore all errors.
Click no to continue and prompt on error.
Click cancel to cancel and rollback.
Yes
No
Cancel

--Insert failed for row 636
--CLOSED_DATE GDK-05043: not a valid month
--Row 636
INSERT INTO SMOKING (UNIQUE_KEY, CREATED_DATE, CREATED_TIME, CLOSED_DATE, AGENCY, AGENCY_NAME, COMPLAINT_TYPE, DESCRIPTOR, LOCATION_TYPE, INCIDENT_ZIP, INCIDENT_ADDRESS, STREET_NAME, CROSS_STREET_1, CROSS_STREET_2, INTERSECTION_STREET_1, INTERSECTION_STREET_2, ADDRESS_TYPE, CITY, LANDMARK, FACILITY_TYPE, STATUS, DUE_DATE, RESOLUTION, RESOLUTION_ACTION_UPDATED_DATE, COMMUNITY_BOARD, BBL, BOROUGH, X_COORDINATE, Y_COORDINATE, OPEN_DATA_CHANNEL_TYPE, PARK_FACILITY_NAME, PARK_BOROUGH, VEHICLE_TYPE, TAXI_COMPANY_BOROUGH, TAXI_PICK_UP_LOCATION, BRIDGE_HIGHWAY_NAME, BRIDGE_HIGHWAY_DIRECTION, ROAD_RAMP, BRIDGE_HIGHWAY_SEGMENT, LATITUDE, LONGITUDE, LOCATION) VALUES (36016575,to_date('4/24/2017', 'MM/DD/RRRR'),'9:40:11 AM',to_date('1/1/1900 0:00', 'DD-MON-RR'),'DOHMH','Department of Health and Mental Hygiene','Smoking','Smoking Violation','Restaurant/Bar/Deli/Bakery',11416,'96-03 101 AVENUE','101 AVENUE','95 STREET','97 STREET',',','','ADDRESS','OZONE PARK',',','N/A','Assigned','5/31/2017 9:40','The Department of Health and Mental Hygiene has sent official written notification to the Owner/Landlord warning them of potential violations and instructing them to correct the situation. If the situation persists 21 days after your initial complaint, please make a new complaint.','4/24/2017 11:23','09 QUEENS',4090710116,'QUEENS',1027292,188739,'PHONE','Unspecified','QUEENS',',',',',',',',',',40.68461707,-73.84480682,'(40.68461707218578, -73.84480682059991)');
--Insert failed for row 882
--CLOSED_DATE GDK-05043: not a valid month
--Row 882
INSERT INTO SMOKING (UNIQUE_KEY, CREATED_DATE, CREATED_TIME, CLOSED_DATE, AGENCY, AGENCY_NAME, COMPLAINT_TYPE, DESCRIPTOR, LOCATION_TYPE, INCIDENT_ZIP, INCIDENT_ADDRESS, STREET_NAME, CROSS_STREET_1, CROSS_STREET_2, INTERSECTION_STREET_1, INTERSECTION_STREET_2, ADDRESS_TYPE, CITY, LANDMARK, FACILITY_TYPE, STATUS, DUE_DATE, RESOLUTION, RESOLUTION_ACTION_UPDATED_DATE, COMMUNITY_BOARD, BBL, BOROUGH, X_COORDINATE, Y_COORDINATE, OPEN_DATA_CHANNEL_TYPE, PARK_FACILITY_NAME, PARK_BOROUGH, VEHICLE_TYPE, TAXI_COMPANY_BOROUGH, TAXI_PICK_UP_LOCATION, BRIDGE_HIGHWAY_NAME, BRIDGE_HIGHWAY_DIRECTION, ROAD_RAMP,

BRIDGE_HIGHWAY_SEGMENT, LATITUDE, LONGITUDE, LOCATION) VALUES (36226417,to_date('5/19/2017', 'MM/DD/RRRR'),'10:52:46 AM',to_date('1/1/1900 0:00', 'DD-MON-RR'),'DOHMH','Department of Health and Mental Hygiene','Smoking','Smoking Violation','Residential Building',10460,'1574 BEACH AVENUE','BEACH AVENUE','GUERLAIN STREET','EAST TREMONT AVENUE',',','ADDRESS','BRONX',',','N/A','Assigned','6/25/2017 10:52','The Department of Health and Mental Hygiene has sent official written notification to the Owner/Landlord warning them of potential violations and instructing them to correct the situation. If the situation persists 21 days after your initial complaint, please make a new complaint.','5/22/2017 16:30','09 BRONX',2039240007,'BRONX',1020950,245177,'PHONE','Unspecified','BRONX',',',',',',',',',40.83955321,-73.86736552,'(40.83955321308456, -73.86736551870723)');

--Insert failed for row 883

--CLOSED_DATE GDK-05043: not a valid month

--Row 883

INSERT INTO SMOKING (UNIQUE_KEY, CREATED_DATE, CREATED_TIME, CLOSED_DATE, AGENCY, AGENCY_NAME, COMPLAINT_TYPE, DESCRIPTOR, LOCATION_TYPE, INCIDENT_ZIP, INCIDENT_ADDRESS, STREET_NAME, CROSS_STREET_1, CROSS_STREET_2, INTERSECTION_STREET_1, INTERSECTION_STREET_2, ADDRESS_TYPE, CITY, LANDMARK, FACILITY_TYPE, STATUS, DUE_DATE, RESOLUTION, RESOLUTION_ACTION_UPDATED_DATE, COMMUNITY_BOARD, BBL, BOROUGH, X_COORDINATE, Y_COORDINATE, OPEN_DATA_CHANNEL_TYPE, PARK_FACILITY_NAME, PARK_BOROUGH, VEHICLE_TYPE, TAXI_COMPANY_BOROUGH, TAXI_PICK_UP_LOCATION, BRIDGE_HIGHWAY_NAME, BRIDGE_HIGHWAY_DIRECTION, ROAD_RAMP, BRIDGE_HIGHWAY_SEGMENT, LATITUDE, LONGITUDE, LOCATION) VALUES (36226857,to_date('5/19/2017', 'MM/DD/RRRR'),'12:13:43 PM',to_date('1/1/1900 0:00', 'DD-MON-RR'),'DOHMH','Department of Health and Mental Hygiene','Smoking','Smoking Violation','Residential Building',10016,'154 EAST 29 STREET','EAST 29 STREET','LEXINGTON AVENUE','3 AVENUE',',','ADDRESS','NEW YORK',',','N/A','Assigned','6/25/2017 12:13','The Department of Health and Mental Hygiene has sent official written notification to the Owner/Landlord warning them of potential violations and instructing them to correct the situation. If the situation persists 21 days after your initial complaint, please make a new complaint.','5/22/2017 16:30','06 MANHATTAN',1008840048,'MANHATTAN',989410,209968,'ONLINE','Unspecified','MANHATTAN',',',',',',',',',40.74298876,-73.9813787,'(40.74298875626896, -73.98137870003814)');

ROW 636, 882, 883 NOT IMPORTED