NYC Air Quality
Data Warehousing Project

CIS 4400 Group 3 Members

John Cruz john.cruz@baruchmail.cuny.edu

Sergey Khegay sergey.khegay@baruchmail.cuny.edu

Mohammed Ali mohammed.ali6@baruchmail.cuny.edu

Reann Wilson REANN.WILSON@baruchmail.cuny.edu

Eli Mogorichev ELI.MOGORICHEV@baruchmail.cuny.edu

CIS 4400-CMWA

Introduction

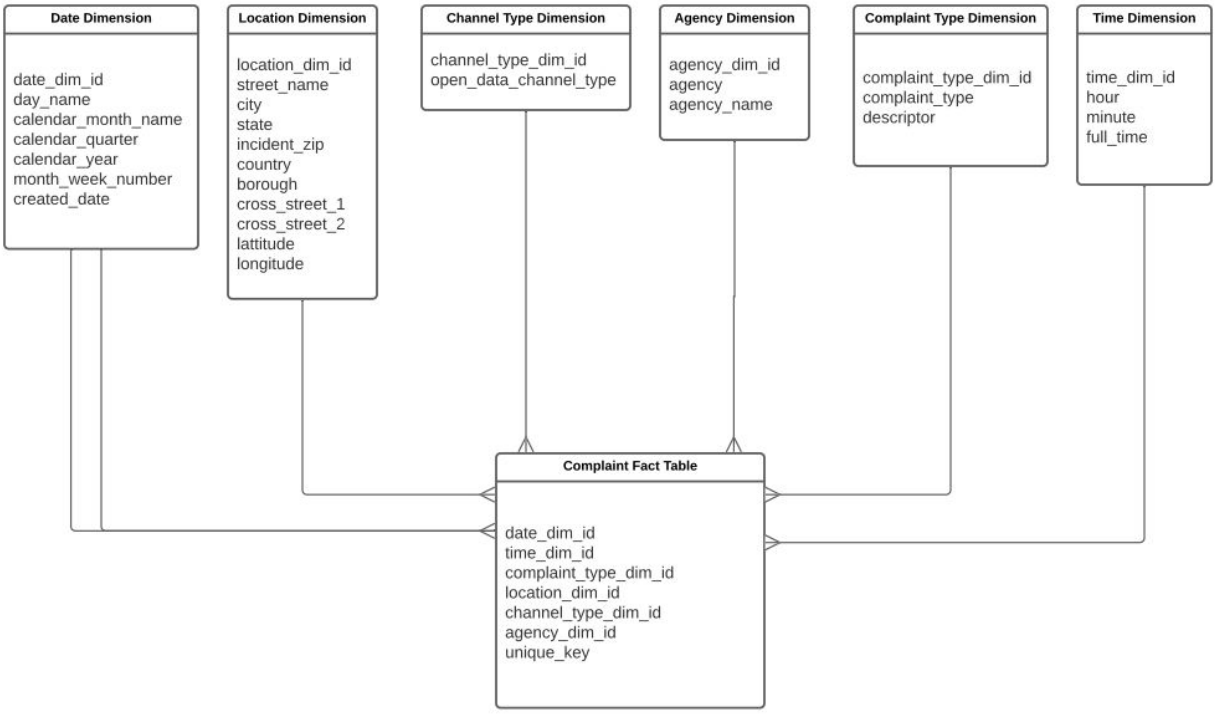
With over eight million Americans living in New York City and being a city that never sleeps, you can guarantee that the city is dealing with issues and complaints regarding air quality. Complaints ranging from businesses, vehicles, constructions and home appliances dominate the category of complaints regarding air quality. For our group project we decided to create a data warehouse that stores air quality complaint data[1]. Our data is sourced from the New York City 311 database that we extracted by using python with the socrata and pandas library. We obtained over 127,000 records of dating with over 40 columns with a date range from 2010 to 2020.

From our data warehouse we want to analyze and draw conclusions of how air quality complaints affect the city air quality. Within our data warehouse we want to research if there are any trends with the air quality complaints from changing seasons, holidays, time of day and location. We want to know what category of complaint is most often reported and how it correlates to time and location. We want to visualize where these complaints are being reported and what type of neighborhood/area plays a role in the reporting of complaints. We are hoping to find which neighborhoods have the best air quality and finding the most polluted areas of New York City.

KPI'S For Air Quality

1. What time of day are the complaints being recorded ?
    a. Morning, afternoon, evening or night
2. Where are the most complaints being recorded ?
    a. Which borough has the most and least recorded complaints?
    b. Which neighborhood has the highest and lowest recorded complaints?
    c. What type of area? Suburban, Commercial, private area etc..
3. Does season's have an effect on air quality complaints ?
    a. Fall, Summer, Winter and Spring
    b. Do holiday seasons have an effect on air quality complaints?

Dimensional Model



What type of grain will we use?

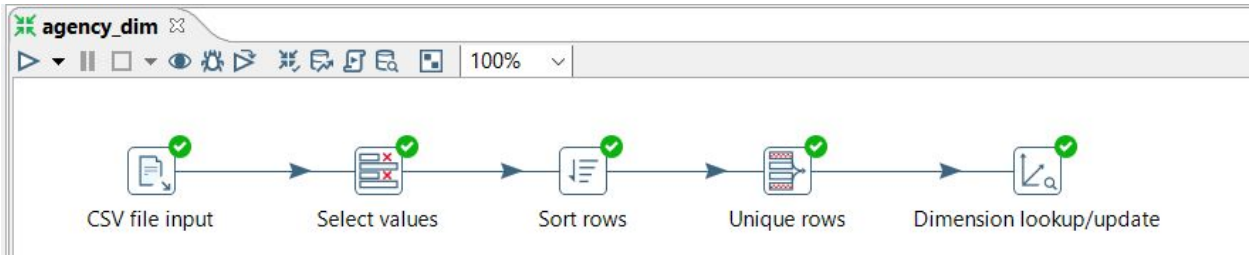
For our data warehouse project we will be using the transaction grain in our fact table. A transaction grain creates a fact record for each business transaction. We chose transaction grain because we want to analyze the date and time of when the air quality complaint was made so we can find any correlation between the air complaints and date/time. From this we can further analyze if certain air quality complaints are seasonal, have patterns in a certain time of day and many more.

Python Code for Data Extraction

```
!pip install sodapy
import pandas as pd
from sodapy import Socrata
data_url='data.cityofnewyork.us'
data_set='erm2-nwe9'    # The data set at the API endpoint (311 data in this case)
app_token='....'  # The app token hidden for privacy
client = Socrata(data_url,app_token)      # Create the client to point to the API endpoint
query_condition = "complaint_type = 'Air Quality' AND created_date between '2010-01-01' and
'2019-12-31' OR complaint_type = 'Indoor Air Quality' AND created_date between '2010-01-01'
and '2019-12-31'"
results = client.get(data_set, where = query_condition, limit=400000)
df = pd.DataFrame.from_records(results)
df.shape
df['complaint_type'].value_counts()
df.to_csv("my_311_data.csv")
```

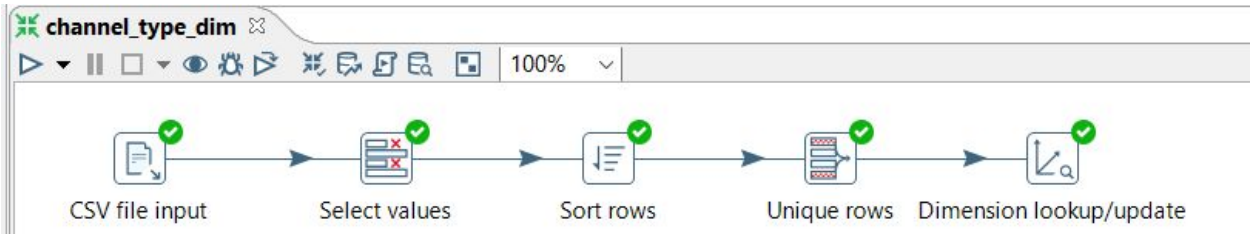
ETL

agency_dim



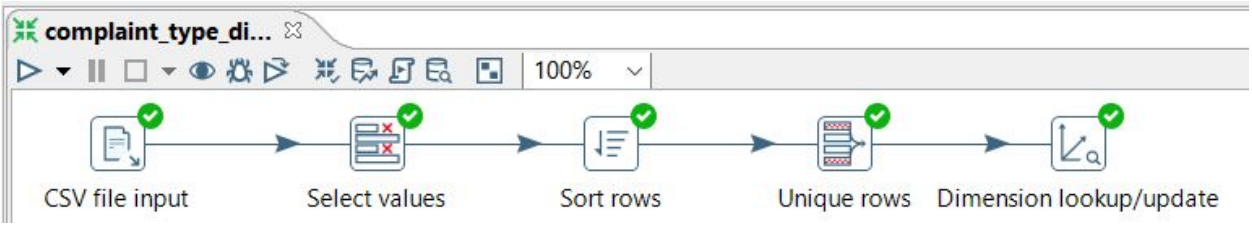
- *Creating the agency dimension*

channel_type_dim



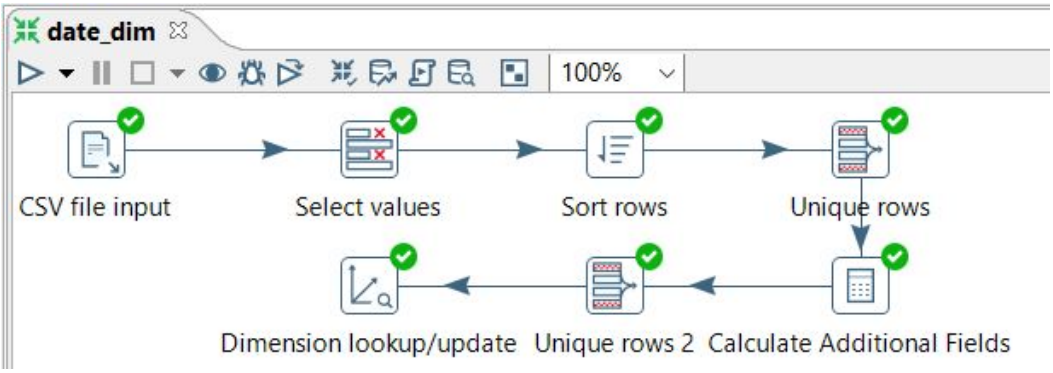
- *Creating the channel type dimension*

complaint_type_dim



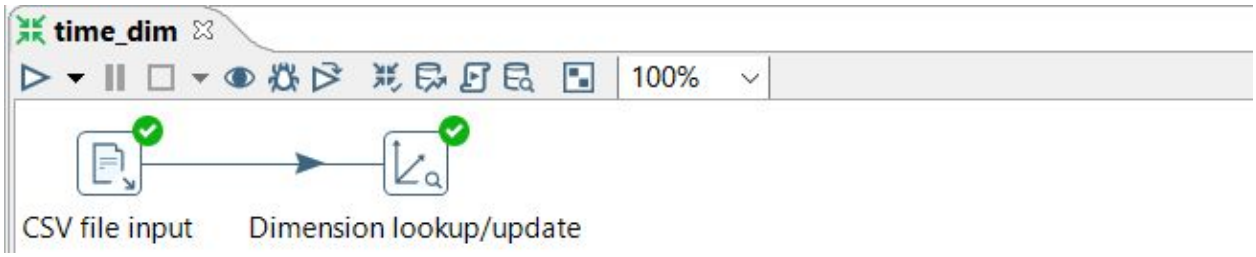
- *Creating the complaint type dimension*

date_dim



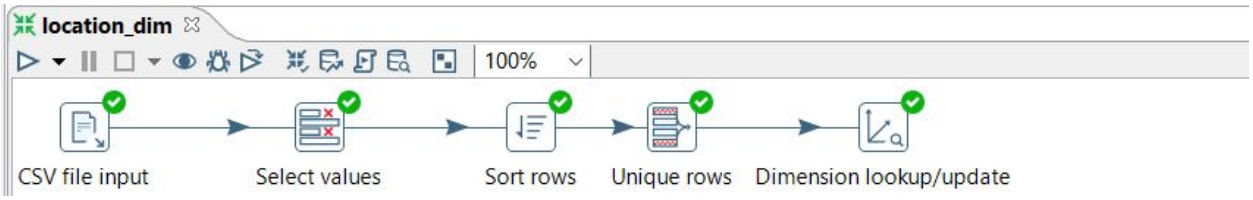
- *Establishing the date dimension*

time_dim



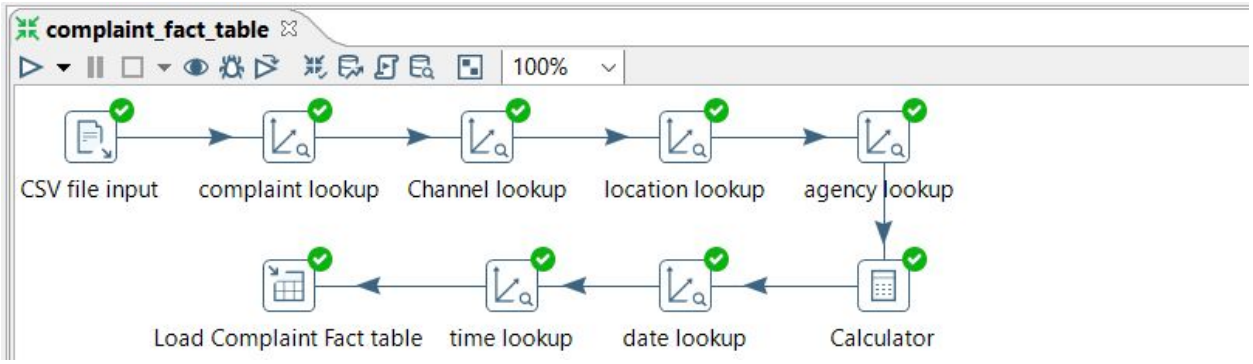
- *Establishing the time dimension*

location_dim



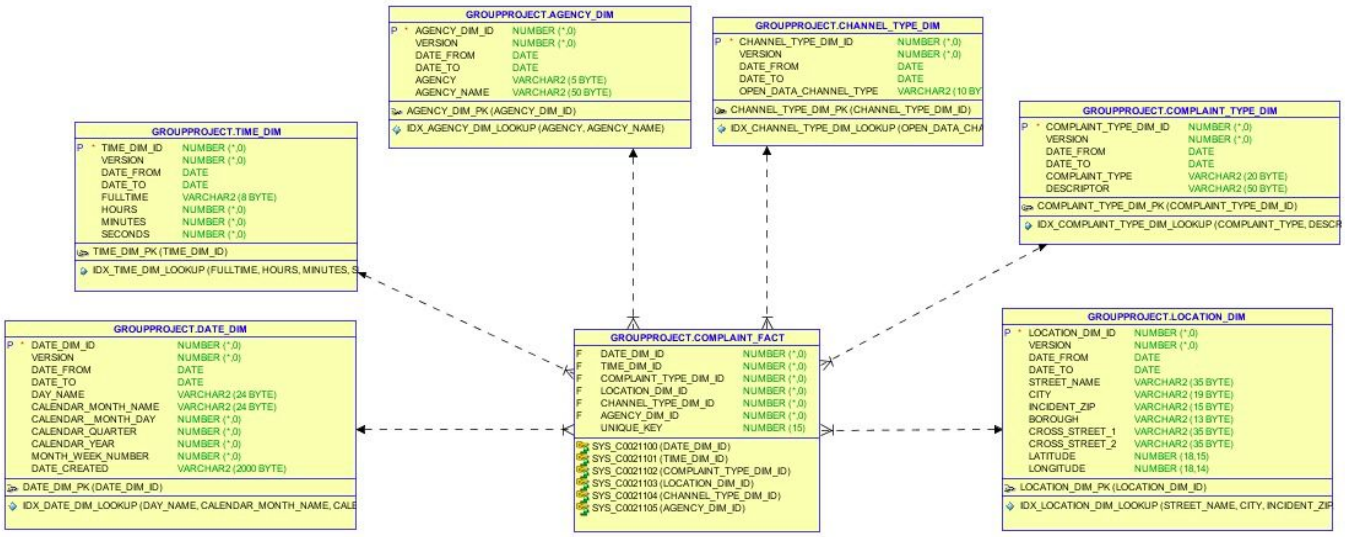
- *Location dimension being created*

complaint_fact_table



- *Gathering all dimensions and making the complaint fact table.*

Physical Model



- *This is our final physical model that we reversed engineered from Oracle SQL developer. This model was used to produce our dashboard and other visuals.*

SQL CODE

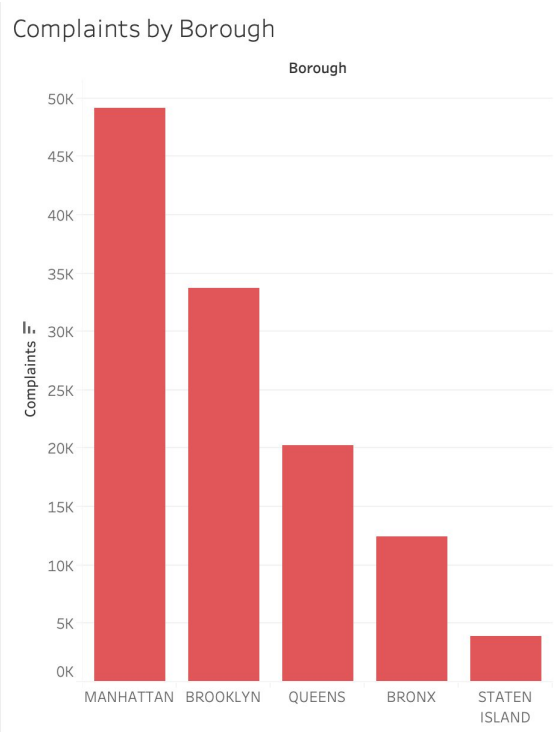
```
ALTER TABLE complaint_fact
ADD FOREIGN KEY (date_dim_id) REFERENCES date_dim(date_dim_id);
ALTER TABLE complaint_fact
ADD FOREIGN KEY (time_dim_id) REFERENCES time_dim(time_dim_id);
ALTER TABLE complaint_fact
ADD FOREIGN KEY (complaint_type_dim_id) REFERENCES
complaint_type_dim(complaint_type_dim_id);
ALTER TABLE complaint_fact
ADD FOREIGN KEY (location_dim_id) REFERENCES
location_dim(location_dim_id);
ALTER TABLE complaint_fact
ADD FOREIGN KEY (channel_type_dim_id) REFERENCES
channel_type_dim(channel_type_dim_id);
ALTER TABLE complaint_fact
ADD FOREIGN KEY (agency_dim_id) REFERENCES
agency_dim(agency_dim_id);
```

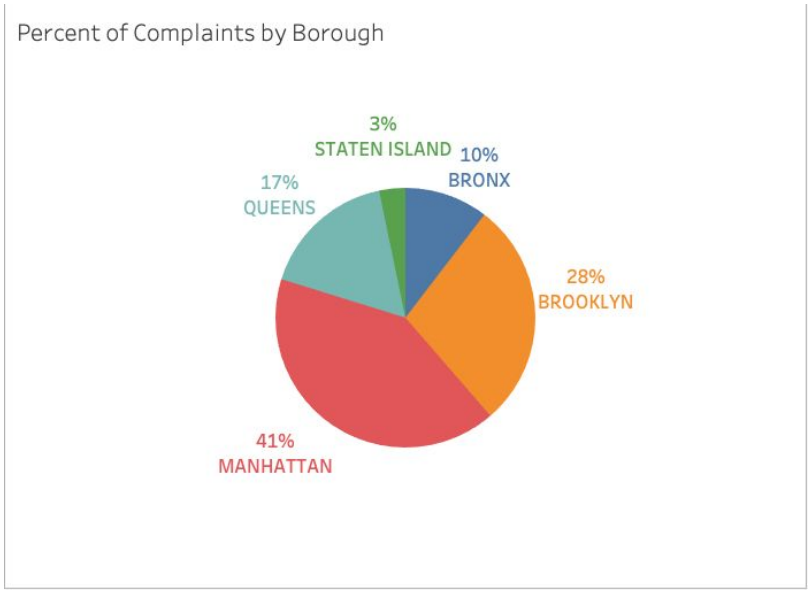
SQL View Code

```
CREATE OR REPLACE VIEW NYC_311 AS
    SELECT
        ad.agency_dim_id, ad.agency, ad.agency_name,
        chd.channel_type_dim_id, chd.open_data_channel_type,
        cf.unique_key,
        ct.complaint_type_dim_id, ct.complaint_type, ct.descriptor,
        dd.date_dim_id, dd.calendar_quarter, dd.day_name,
            dd.calendar_month_name, dd.calendar__month_day, dd.calendar_year,
            dd.date_created, dd.month_week_number,
        ld.location_dim_id, ld.street_name, ld.city, ld.incident_zip, ld.borough,
            ld.cross_street_1, ld.cross_street_2, ld.latitude, ld.longitude,
        td.time_dim_id, td.fulltime, td.hours, td.minutes, td.seconds
    FROM complaint_fact cf
        INNER JOIN agency_dim ad
            ON cf.agency_dim_id = ad.agency_dim_id
        INNER JOIN channel_type_dim chd
            ON chd.channel_type_dim_id = cf.channel_type_dim_id
        INNER JOIN complaint_type_dim ct
            ON ct.complaint_type_dim_id = cf.complaint_type_dim_id
        INNER JOIN date_dim dd
            ON dd.date_dim_id = cf.date_dim_id
        INNER JOIN location_dim ld
            ON ld.location_dim_id = cf.location_dim_id
        INNER JOIN time_dim td
            ON td.time_dim_id = cf.time_dim_id;
```

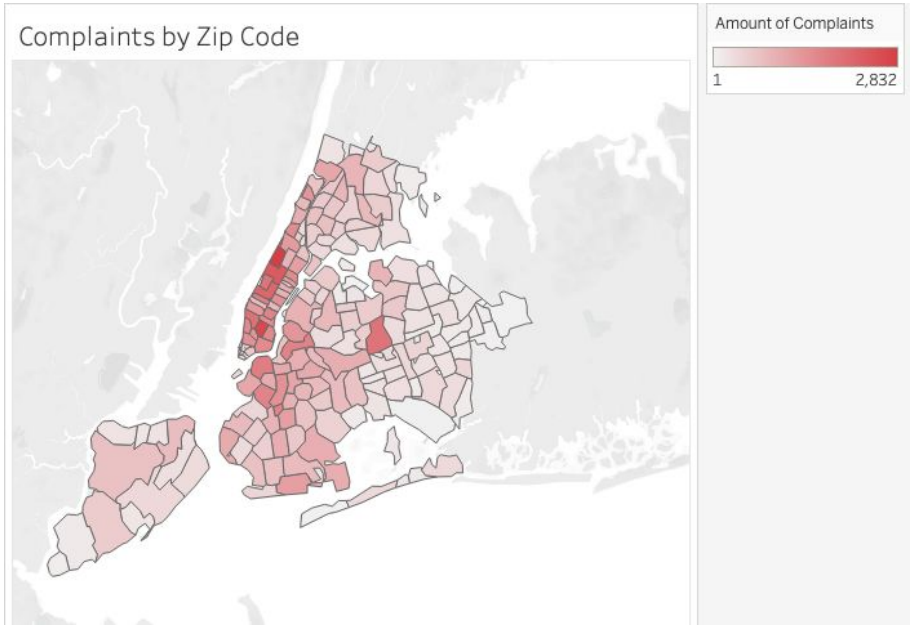
	AGENCY_DIM_ID	AGENCY	AGENCY_NAME	CHANNEL_TYPE_DIM_ID	OPEN_DAT...	UNIQUE_KEY	COMPL...	COMPLAINT_TYPE	DESCRIPTOR	DATE_DIM_ID	CALENDAR_QUARTER	DA
1	2	DOHMH	Department of Heal...	3	PHONE	17614200	22	Indoor Air Quality	(null)	190	3	Frid
2	2	DOHMH	Department of Heal...	3	PHONE	40185800	22	Indoor Air Quality	(null)	3168	3	Tues
3	1	DEP	Department of Envi...	1	ONLINE	41872500	1	Air Quality	Air: Dust, Commercial (AE2)	3349	1	Mond
4	1	DEP	Department of Envi...	1	ONLINE	42797600	1	Air Quality	Air: Dust, Commercial (AE2)	3434	2	Tues
5	1	DEP	Department of Envi...	3	PHONE	45002700	1	Air Quality	Air: Dust, Commercial (AE2)	3613	4	Satu
6	1	DEP	Department of Envi...	4	UNKNOWN	15654797	1	Air Quality	Air: Dust, Commercial (AE2)	4	1	Mond
7	1	DEP	Department of Envi...	4	UNKNOWN	15654816	1	Air Quality	Air: Dust, Commercial (AE2)	4	1	Mond
8	1	DEP	Department of Envi...	4	UNKNOWN	15655266	1	Air Quality	Air: Dust, Commercial (AE2)	4	1	Mond
9	1	DEP	Department of Envi...	4	UNKNOWN	15683459	1	Air Quality	Air: Dust, Commercial (AE2)	7	1	Thur
10	1	DEP	Department of Envi...	4	UNKNOWN	15683490	1	Air Quality	Air: Dust, Commercial (AE2)	7	1	Thur
11	1	DEP	Department of Envi...	4	UNKNOWN	15683492	1	Air Quality	Air: Dust, Commercial (AE2)	7	1	Thur
12	2	DOHMH	Department of Heal...	3	PHONE	15696755	22	Indoor Air Quality	(null)	10	1	Sund
13	2	DOHMH	Department of Heal...	3	PHONE	15730021	22	Indoor Air Quality	(null)	14	1	Thur
14	1	DEP	Department of Envi...	4	UNKNOWN	15734423	1	Air Quality	Air: Dust, Commercial (AE2)	14	1	Thur
15	1	DEP	Department of Envi...	4	UNKNOWN	15743256	1	Air Quality	Air: Dust, Commercial (AE2)	15	1	Frid
16	1	DEP	Department of Envi...	4	UNKNOWN	15744206	1	Air Quality	Air: Dust, Commercial (AE2)	15	1	Frid
17	1	DEP	Department of Envi...	4	UNKNOWN	15744213	1	Air Quality	Air: Dust, Commercial (AE2)	15	1	Frid
18	1	DEP	Department of Envi...	4	UNKNOWN	15744320	1	Air Quality	Air: Dust, Commercial (AE2)	15	1	Frid
19	1	DEP	Department of Envi...	4	UNKNOWN	15744863	1	Air Quality	Air: Dust, Commercial (AE2)	15	1	Frid
20	1	DEP	Department of Envi...	4	UNKNOWN	15744865	1	Air Quality	Air: Dust, Commercial (AE2)	15	1	Frid
21	2	DOHMH	Department of Heal...	3	PHONE	15746087	22	Indoor Air Quality	(null)	16	1	Satu
22	2	DOHMH	Department of Heal...	3	PHONE	15746304	22	Indoor Air Quality	(null)	16	1	Satu
23	1	DEP	Department of Envi...	4	UNKNOWN	15756732	1	Air Quality	Air: Dust, Commercial (AE2)	18	1	Mond
24	1	DEP	Department of Envi...	4	UNKNOWN	15756744	1	Air Quality	Air: Dust, Commercial (AE2)	18	1	Mond
25	2	DOHMH	Department of Heal...	3	PHONE	15757678	22	Indoor Air Quality	(null)	19	1	Tues
26	2	DOHMH	Department of Heal...	1	ONLINE	15758182	22	Indoor Air Quality	(null)	19	1	Tues
27	2	DOHMH	Department of Heal...	3	PHONE	15758469	22	Indoor Air Quality	(null)	19	1	Tues
28	2	DOHMH	Department of Heal...	1	ONLINE	15758535	22	Indoor Air Quality	(null)	19	1	Tues
29	2	DOHMH	Department of Heal...	3	PHONE	15758635	22	Indoor Air Quality	(null)	19	1	Tues
30	2	DOHMH	Department of Heal...	3	PHONE	15759520	22	Indoor Air Quality	(null)	19	1	Tues
31	2	DOHMH	Department of Heal...	3	PHONE	15759571	22	Indoor Air Quality	(null)	19	1	Tues
32	1	DEP	Department of Envi...	4	UNKNOWN	15806306	1	Air Quality	Air: Dust, Commercial (AE2)	25	1	Mond
33	2	DOHMH	Department of Heal...	3	PHONE	15808282	22	Indoor Air Quality	(null)	26	1	Tues
34	2	DOHMH	Department of Heal...	3	PHONE	15808477	22	Indoor Air Quality	(null)	26	1	Tues
35	1	DEP	Department of Envi...	4	UNKNOWN	15823134	1	Air Quality	Air: Dust, Commercial (AE2)	27	1	Wedn
36	2	DOHMH	Department of Heal...	3	PHONE	15826909	22	Indoor Air Quality	(null)	28	1	Thur
37	1	DEP	Department of Envi...	4	UNKNOWN	15901570	1	Air Quality	Air: Dust, Commercial (AE2)	31	1	Sund
38	2	DOHMH	Department of Heal...	3	PHONE	15902417	22	Indoor Air Quality	(null)	32	1	Mond
39	2	DOHMH	Department of Heal...	3	PHONE	15902682	22	Indoor Air Quality	(null)	32	1	Mond
40	2	DOHMH	Department of Heal...	3	PHONE	15902798	22	Indoor Air Quality	(null)	32	1	Mond
41	1	DEP	Department of Envi...	4	UNKNOWN	15910953	1	Air Quality	Air: Dust, Commercial (AE2)	32	1	Mond

Visualizations

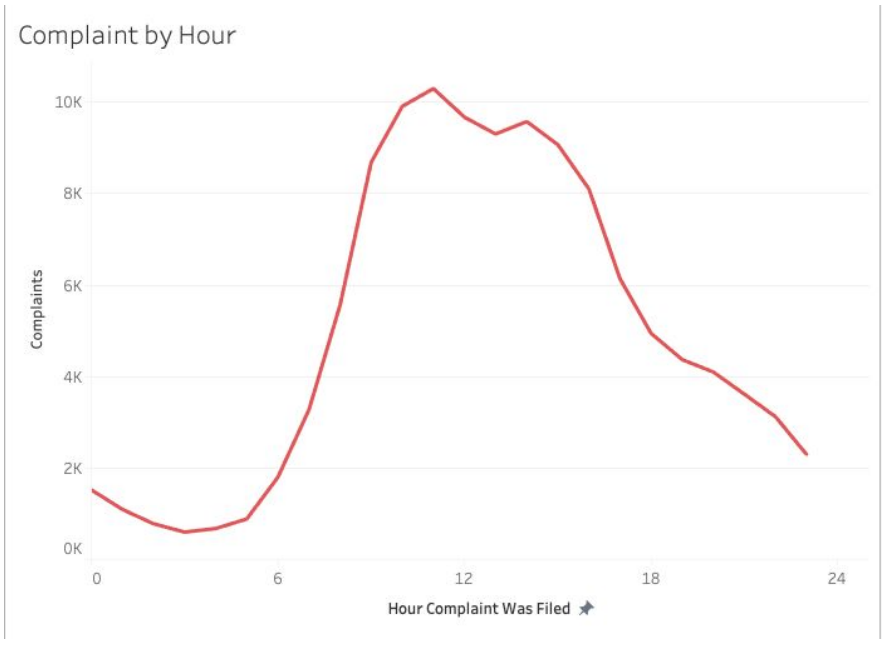




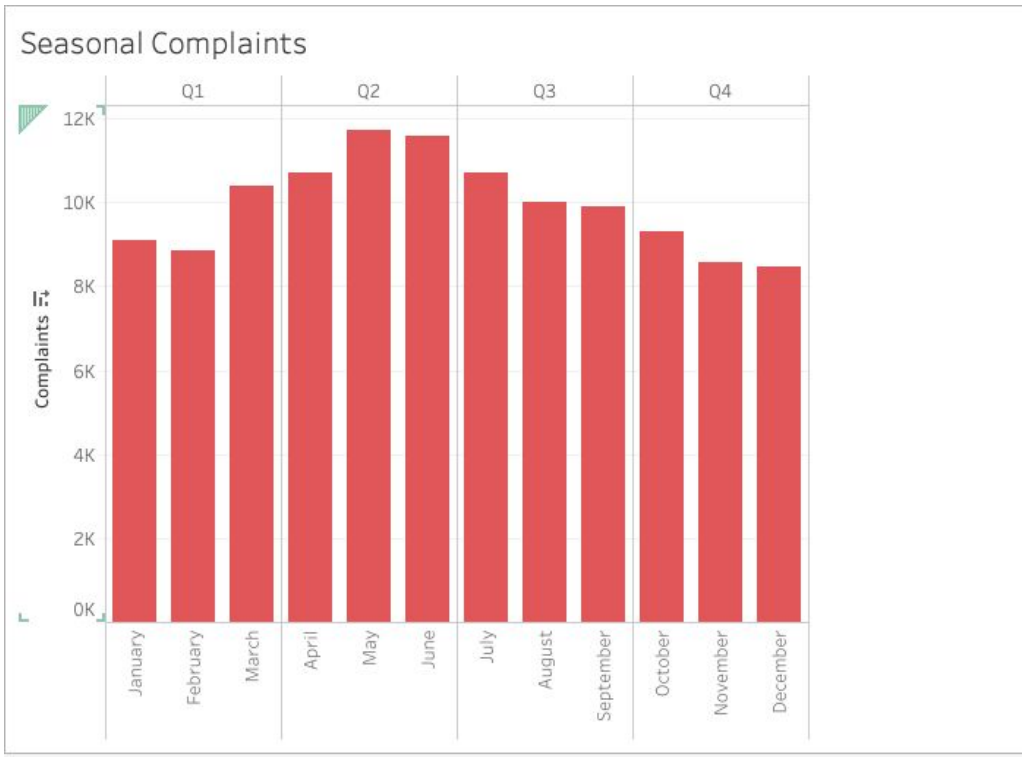
The visualizations above show how many total complaints were recorded by each Borough. We can see that Manhattan has a total of roughly around 49,000 complaints or 41% of the total complaints.



Above we can see total complaints broken down by each NYC Zip Code.



The line plot above shows the total complaints by the hour. We can see that most complaints were being filed 9am - 3pm, with a peak at 11am.



From this visualization, we can see that the amount of complaints do shift as seasons change. The season with the most complaints seems to be Spring, most likely because projects that were shutdown for the Winter weather started back up again.

What are the technological tools we utilized ?

One of the first tools we used in our project was python. We needed python to help us easily extract data from the 311 website. If we were to do the data retrieval manually from the website it would be very inefficient and ineffective. Utilizing the pandas and socrata library to extract our data and import into pandas dataframe in which exported to a csv file. Excel helped us have a full view of our raw data. It allowed us to make our dimensional model by analyzing which columns to use and take note of the data types required for each column. Next we used Lucidchart to make our dimensional model, which provided us the appropriate symbols and notations, while also cataloging every version of the model, as we updated it to our specific needs. Oracle has been the database management system of choice for this project and this course. This was the obvious database management system to use because each individual already has experience using Oracle to create ETL transformation. What went hand in hand with Oracle was the Oracle SQL developer which allowed us to query our results, create our views and develop our final schema. Pentaho Data integration is a business intelligence software that provides users to create and manage their own data warehouse. This is where the heavy work for the project was completed, the ETL development was created in Pentaho Data Integration. After creating our data warehouse in Oracle we exported the data into a csv file and used Tableau to develop our dashboard and visualizations.

For communication and tracking our files we utilized a wide toolset to aid in our collaboration.

1. Zoom: We were able to use this to meet together via video conferencing, screen sharing and talking out the topic we all could agree on and wanted to find answers about.
2. WhatsApp: This gave every individual the autonomy of working on their portion of the project while being able to share and communicate updates through chat such as providing PDFs or screenshots of the progression. It also allowed for our various time constraints to be flexible without having numerous rigid meetup times.
3. Google Drive: The bulk of our data storage was conveniently available to us by utilizing cloud storage. This provided a centralized location everyone could access files. We were able to store PDI files, the schemas and modeling revisions, and the visualizations.

Conclusion

Communication was the key to completing our data warehouse, utilizing various applications to reach out to each other. Depending on the individuals in each group, each team has a different method of communication. Some groups want to speak via voice or video calls while others rely on instant messaging or emails. Here is a list and brief description of what our group utilized.

1. ***WhatsApp:***
    a. This was our primary tool of communicating with one another. Since there is a mix of members who are using android and apple phones within our group it made sense to use whatsapp because it connected all of us. It allowed us to send files of our SQL & python code, images of our dimensional and final schema model, website links and other various documents.
2. ***Zoom****:*
    a. Zoom was a platform we used to have group meetings with one another, this was primarily used in the early stages of the project when we were discussing the type of complaint we wanted to research on.
3. ***Google Sheets:***
    a. Google sheets allowed us to document our group and individual progress pertaining to the completion of the project.
4. ***Jupyterlab:***
    a. Juypter lab notebook is a web based python IDE that we used to write our data extraction code.
5. ***Oracle*****:**
    a. Oracle cloud service gave us the ability to store our data and connect the data warehouse from pentaho data integration.
6. ***Oracle SQL Developer:***
    a. Gave us the function to query our data and create our final schema.
7. ***Pentaho Data Integration***
    a. Business intelligence software that provides users to create and manage our data warehouse.

Developing the dimensional model and the iterations it went through, we found that as simple it may sound, without prior experience developing dimensional models it can easily lead to some disorganization and errors later on into the ETL process. Embellishing the dimensions for categories you may have not first thought of, or how you would want to normalize the model, but realize the pitfalls of doing so. We did not expect how each stage of this project could be further developed into wanting a more complex end result of KPIs, but then realizing the sheer complexity and project time constraints pulling in the opposite direction. We would have wanted to develop the KPIs and its potential visualizations much more in the beginning, to better have an understanding of what data we really needed.

One of the steps in the ETL process that were the most difficult was creating a date and time dimension. Whereas the extract and transform portion showed the outcome that we were looking for, but during the load part of the ETL there were constant problems with the formatting. Nevertheless, after further evaluation and help from professor Holowczak and his website the problem was solved[2]. Which left us with the easiest portion of creating other dimensions, as well as getting data using OpenData API. During the entire process, we learned how to use sodapy library specifically Socrata, that allowed the use of SQL like query. In addition to different ways to run transformations using Pentaho DI. At last, if we were to do it over again, one of the things that we would have done differently would have been finding a better and faster way to load data into the database.

Creating the visualizations was a very interesting part of the project. At this stage, we got to see our data come to life and as a result, we were able to find many interesting things about our data, for example, we found out that most complaints were filed during Spring. This might be partly attributed to construction sites starting back up after the Winter. We didn't find the overall process of visualizing difficult and so if we had to do it again, we wouldn't do anything different.

Our data warehouse is a valuable tool for the state and city government to utilize because it can help aid the air pollution issues within New York City. The data warehouse is constructed in a specific way that the user can view the date, time, location of the complaint and the type of complaint it is. What helped us the most completing this project was doing second homework assignments for this course. It gave us the hands on experience of creating a data warehouse by using pentaho data integration. Executing each procedure of the ETL in the homework gave us the foundational knowledge and confidence to complete the ETL process for our air quality data warehouse.

Reference

1. New York, C. (2011, October 18). *311 Service Requests from 2010 to Present* (United States of America, 311). Retrieved September 8, 2020, from https://data.cityofnewyork.us/Social-Services/311-Service-Requests-from-2010-to-Present/erm2-nwe9
2. Holowczak, R. (2020, June 25). Data Warehouse Dimensional Modeling. Retrieved September 8, 2020, from http://holowczak.com/data-warehouse-dimensional-modeling/

Milestones

Date	Start Time	End Time	Attendees	Connected Via	Notes
8/31/2020	4:00 PM	4:30 PM	John, Eli, Sergey, Reann, Mohammed	WhatsApp	Discussed potential categories of interest. Will reconvene on problems we want to analyze
9/8/2020	1:25 PM	2:45 PM	Mohammed, John, Sergey, Reann	WhatsApp	Reviewing all categories and looking for overlaps between waste
9/8/2020	6:00 PM	8:00 PM	John, Mohammed, Sergey	WhatsApp	Finalized waste categories and prepared proposal submission
9/28/2020	7:00 PM	8:30 PM	John, Eli, Sergey, Reann, Mohammed	WhatsApp, Zoom	Discussed potential KPIs and scheduling talk with prof.
9/29/2020	12:00 PM	12:15 PM	John, Eli, Sergey, Reann, Mohammed	Zoom	Spoke to Prof.H about our changes in project proposal and KPI's
9/29/2020	1:00 PM	8:00 PM	Reann,John, Eli, Sergey, Mohammed	Google sheets, whatsapp	Reann Developed Dimensional Model and group finalized the dimensional model
10/20/2020	8:00 PM	10:00 PM	John	Whatsapp	John redesigned dimesional model. Team decided to run transactional grain
10/21/2020	10:00 AM	10:30 AM	John, Sergey, Mohammed	Whatsapp	Reviewed and finalized dimensional model for submission
11/2/2020	7:00 PM	10:00 PM	Sergey	Whatsapp	Completed channel, complaint, location dims
11/3/2020	9:00 PM	10:00 PM	Sergey	Whatsapp	Completed date dim
11/4/2020	11:00 PM	11:30 PM	Sergey, John, Mohammed	Whatsapp	Troubleshooting facts dim
11/9/2020	9:00 AM	9:30 AM	Sergey, Mohammed, John, Eli	Whatsapp	Reevaluted date and time dimensions required in model

11/17/2020	1:00 PM	6:00 PM	Eli	Whatsapp	Started developing data visuals for the project
12/7/2020	5:00 PM	9:00 PM	Eli	Whatsapp	Completed the data visuals
12/14/2020	5:00 PM	11:00 PM	Mohammed, John, Sergey, Eli	Whatsapp, Google docs	Started writing the final report
12/14/2020	8:00 PM	9:00 PM	Sergey, John	Whatsapp	Added constraints to the final schema.
12/14/2020	11:00 PM	11:45PM	John	WhatsApp	Created VIEW within Oracle DB as NYC_311 and added to doc
12/16/2020	11:30 AM	12:30 PM	Mohammed, John, Sergey, Eli	WhatsApp	Completed the Final report