CIS 4400 Group Project

Section:CMWA

Khilola Rustamova Khilola.Rustamova@baruchmail.cuny.edu
JASON MEI JASON.MEI3@baruchmail.cuny.edu
JACKY CHEN JACKY.CHEN5@baruchmail.cuny.edu
YIZUO ZHENG YIZUO.ZHENG@baruchmail.cuny.edu

Project Proposal

New York City has been very loud for decades due to the mass amount of people, its crime rates have also been also increasing throughout the years. In some regions of NYC crime rates can be much higher than others which can correspond with that region having more noise complaints in Residential, Street/Sidewalks, Commercial buildings, and Parks all around NYC.

The crime dataset can link the number of noise complaints per city/borough with the number of crimes per city/borough. The goal is to explore whether areas with higher crime rates experience more noise complaints or not. Areas where there's a lot of sex crimes/harassment may contribute to the noise complaints that occur in those areas. Another possibility is that areas where there's theft occurring in a Resident - Apt. House may result in a filing for noise complaint as Residential noise type for our primary dataset.

Choice of type of 311 complaints:

311 Noise Complaints | NYC Open Data

- 230227,Noise - Residential
- 73116,Noise - Street/Sidewalk
- 47420,Noise - Commercial
- 4676,Noise - Park

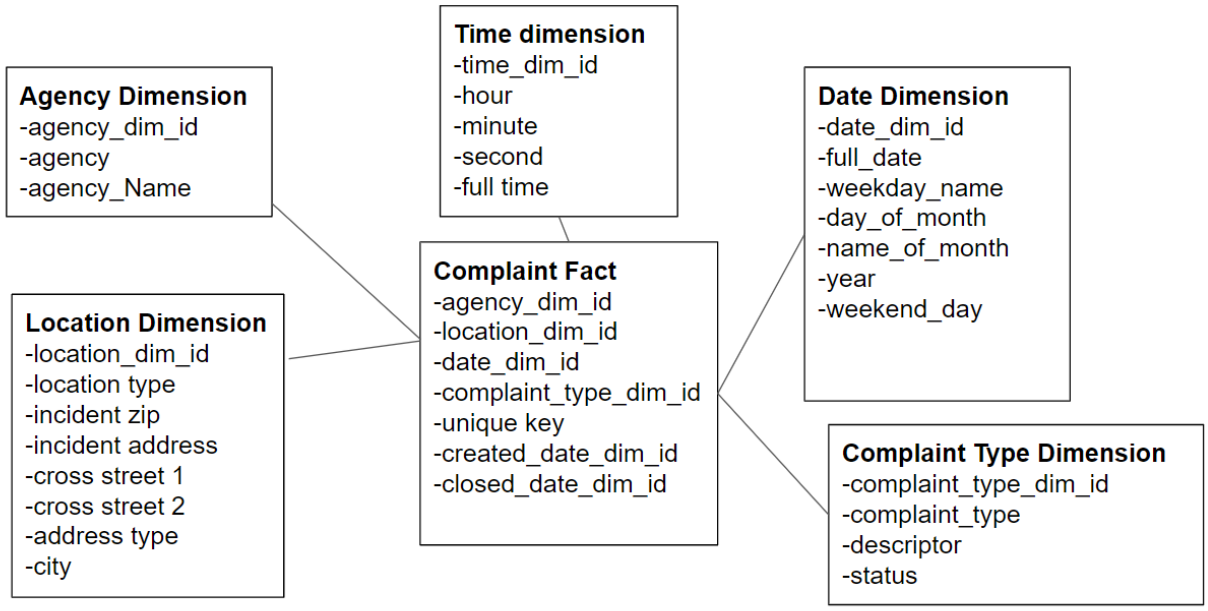
Second Dataset:
https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Current-Year-To-Date-/5uac-w243

KPI:

- Number of noise complaints per borough
- Number of noise complaints per city
- Number of noise complaints per location type
- Number of crimes per borough
- Complain increase in boroughs with increasing crime rates
- The most type of noise complaints in each borough
- Number of complaints per descriptor

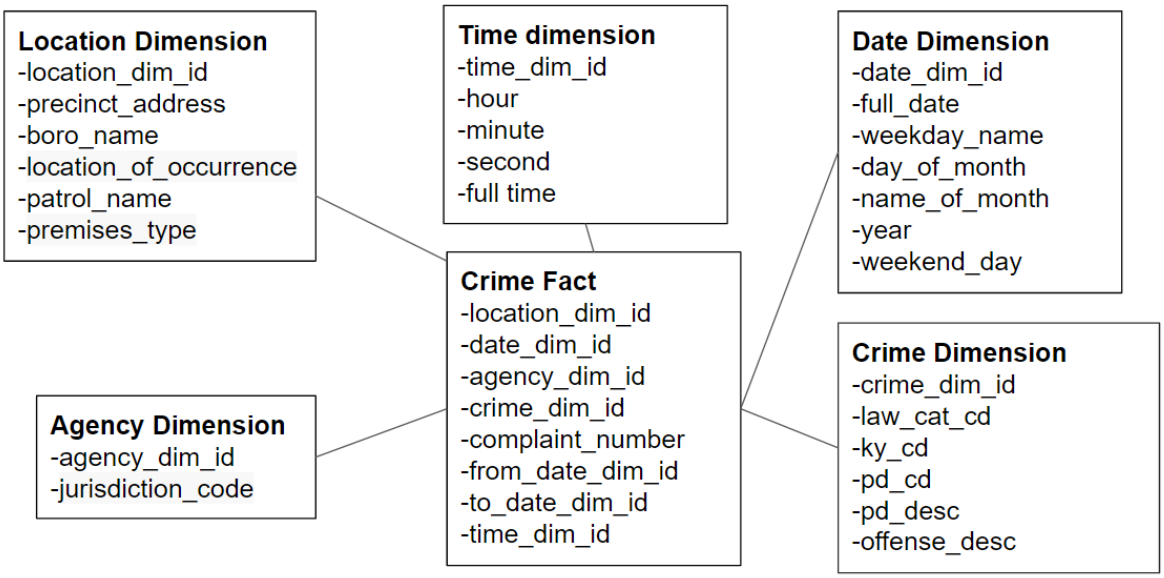
Dimension model drafts

1st data set



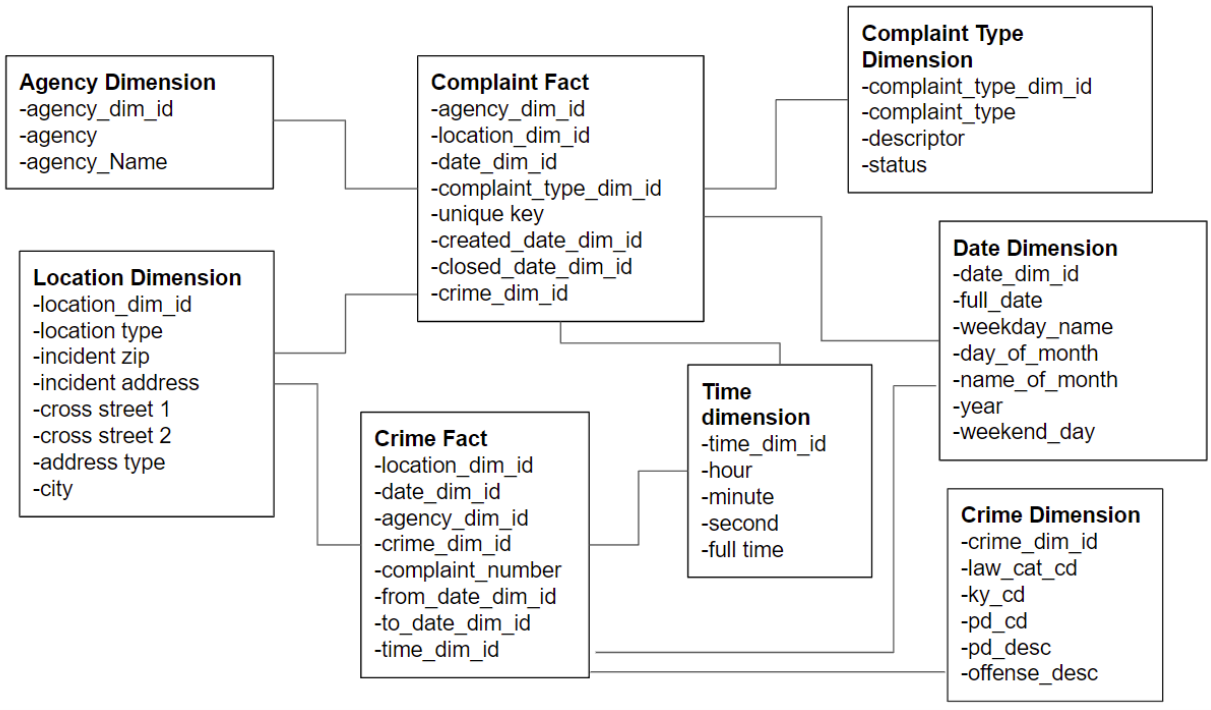
We used transaction grain for data set 1.

2nd data set



We used transaction grain for data set 2nd.

Integrated data warehouse model



Data Profiling for NYPD Data/Second Data Set

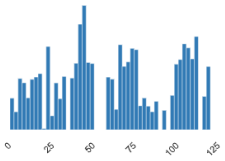
ADDR_PCT_CD

Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION

Distinct	77
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	63.66867628

Minimum	1
Maximum	123
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	3.4 MiB



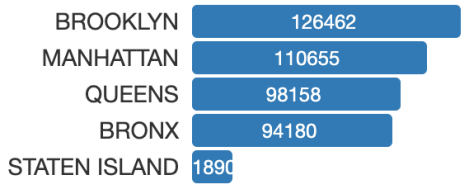
Toggle details

BORO_NM

Categorical

HIGH CORRELATION
HIGH CORRELATION

Distinct	5
Distinct (%)	< 0.1%
Missing	1151
Missing (%)	0.3%
Memory size	3.4 MiB



Toggle details

CMPLNT_FR_...
Categorical

HIGH CARDINALITY

Distinct	1979
Distinct (%)	0.4%
Missing	0
Missing (%)	0.0%
Memory size	3.4 MiB

01/01/2021	1593
10/15/2021	1537
10/01/2021	1530
12/01/2021	1489
10/20/2021	1484
Other values (1974)	441873

Toggle details

CMPLNT_FR_...
Categorical

HIGH CARDINALITY

Distinct	1440
Distinct (%)	0.3%
Missing	0
Missing (%)	0.0%
Memory size	3.4 MiB

12:00:00	11914
15:00:00	9016
18:00:00	8379
17:00:00	8283
16:00:00	7961
Other values (1435)	403953

Toggle details

Data Profiling for NY 311 Data Set

Overview

Overview | Alerts 88 | Reproduction

Dataset statistics

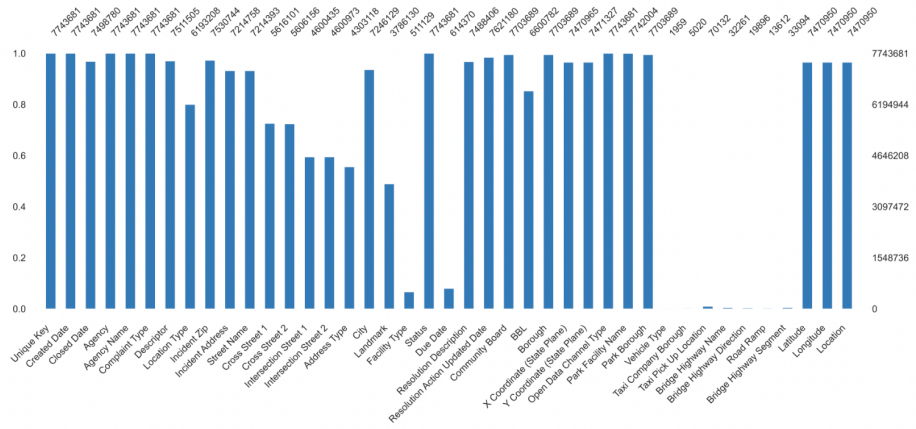
Number of variables	41
Number of observations	7743681
Missing cells	93142670
Missing cells (%)	29.3%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	2.4 GiB
Average record size in memory	328.0 B

Variable types

Numeric	6
Categorical	34
Unsupported	1

Missing values

Count | Matrix | Heatmap | Dendrogram



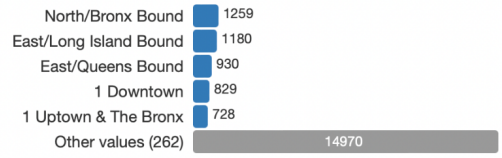
A simple visualization of nullity by column.

Bridge Highway Direction

Categorical

HIGH CARDINALITY
MISSING

Distinct	267
Distinct (%)	1.3%
Missing	7723785
Missing (%)	99.7%
Memory size	59.1 MiB



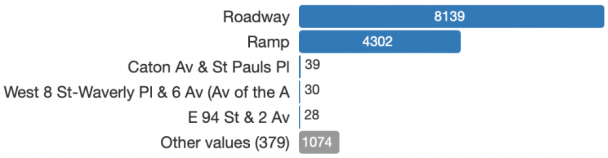
Toggle details

Road Ramp

Categorical

HIGH CARDINALITY
MISSING

Distinct	384
Distinct (%)	2.8%
Missing	7730069
Missing (%)	99.8%
Memory size	59.1 MiB



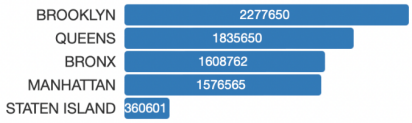
Toggle details

Borough

Categorical

HIGH CORRELATION

Distinct	6
Distinct (%)	< 0.1%
Missing	39992
Missing (%)	0.5%
Memory size	59.1 MiB



Toggle details

ETL Tools and Target DBMS

For this step of the project, we will be using **Python** and **dbt** for performing ETL .For our target DBMS for this project, we will be using **Google BigQuery**

311 Noise Complaint Models

311 agency dimension model

```
SELECT
    row_number() OVER () AS agency_dim_id,
    agency, agency_name
FROM
   ( SELECT DISTINCT agency, agency_name
     FROM `bigquery-public-data.new_york_311.311_service_requests`
     WHERE complaint_type IN ('Noise')
    )
```

311 location dimension model

```
SELECT
    row_number() OVER () AS location_dim_id,
    location_type, incident_zip,
    incident_address,
    street_name,
    cross_street_1,
    cross_street_2,
    address_type,
    city
FROM `bigquery-public-data.new_york_311.311_service_requests`
WHERE complaint_type IN ('Noise')
```

Date Dimension Model

```
SELECT
    ROW_NUMBER() OVER() as date_dim_id,
    d AS full_date,
    FORMAT_DATE('%A', d) AS weekday_name,
    EXTRACT(DAY FROM d) AS day_of_month,
    FORMAT_DATE('%B', d) as name_of_month,
    EXTRACT(YEAR FROM d) AS year,
(CASE WHEN FORMAT_DATE('%A', d) IN ('Sunday', 'Saturday')
```

```
THEN 0 ELSE 1 END) AS weekend_day,
FROM (
SELECT * FROM
UNNEST (GENERATE_DATE_ARRAY('2018-12-01', '2022-05-01', INTERVAL 1 DAY) ) AS d
)
```

Time Dimension Model

```
SELECT
  ROW_NUMBER() OVER() as time_dim_id,
  t AS full_time,
  EXTRACT(HOUR FROM t) as hour,
  EXTRACT(MINUTE FROM t) AS minute,
  EXTRACT(SECOND FROM t) AS second,
FROM (
  SELECT
    *
  FROM
    UNNEST(GENERATE_TIMESTAMP_ARRAY('2018-12-01 00:00:00', '2022-05-01
00:00:00', INTERVAL 10 MINUTE)) AS t )
```

Complaint Type Dimension Model

```
SELECT
    row_number() OVER () AS complaint_type_dim_id,
    complaint_type,
    descriptor,
    status
FROM
   ( SELECT DISTINCT complaint_type, status, descriptor
     FROM `bigquery-public-data.new_york_311.311_service_requests`
     WHERE complaint_type IN ('Noise')
    )
```

NYPD Dimensional Models

NYPD location dimension model

```
SELECT
    row_number() OVER () AS location_dim_id,
    ADDR_PCT_CD as precinct_address,
    BORO_NM as boro_name,
    latitude as latitude,
    Longitude as longitude,
FROM `handy-bonbon-142723.nypd_complaints.nypd_complaints_2020`
```

NYPD crime dimension model

```
SELECT
    row_number() OVER () AS crime_dim_id,
    law_cat_cd as level_of_offense,
    ky_cd as offense_code,
    pd_cd as internal_classification_code,
    pd_desc as descriptor,
    ofns_desc as offense_description
FROM `handy-bonbon-142723.nypd_complaints.nypd_complaints_2020`
```

Time Dimension Model

```
SELECT
  ROW_NUMBER() OVER() as time_dim_id,
  t AS full_time,
  EXTRACT(HOUR FROM t) as hour,
  EXTRACT(MINUTE FROM t) AS minute,
  EXTRACT(SECOND FROM t) AS second,
FROM (
  SELECT
    *
  FROM
    UNNEST(GENERATE_TIMESTAMP_ARRAY('2018-12-01 00:00:00', '2022-05-01
00:00:00', INTERVAL 10 MINUTE)) AS t )
```

Date Dimension Model

```
SELECT
    ROW_NUMBER() OVER() as date_dim_id,
    d AS full_date,
    FORMAT_DATE('%A', d) AS weekday_name,
    EXTRACT(DAY FROM d) AS day_of_month,
    FORMAT_DATE('%B', d) as name_of_month,
    EXTRACT(YEAR FROM d) AS year,
(CASE WHEN FORMAT_DATE('%A', d) IN ('Sunday', 'Saturday')
THEN 0 ELSE 1 END) AS weekend_day,
FROM (
SELECT * FROM
UNNEST (GENERATE_DATE_ARRAY('2018-12-01', '2022-05-01', INTERVAL 1 DAY) ) AS d
```

Agency Dimension Model

```
SELECT
    row_number() OVER () AS agency_dim_id,
    jurisdiction_code
FROM
   ( SELECT DISTINCT jurisdiction_code
     FROM `handy-bonbon-142723.nypd_complaints.nypd_complaints_2018`
    UNION ALL
    SELECT DISTINCT jurisdiction_code
    FROM `handy-bonbon-142723.nypd_complaints.nypd_complaints_2019`
    UNION ALL
    SELECT DISTINCT jurisdiction_code
    FROM `handy-bonbon-142723.nypd_complaints.nypd_complaints_2020`
    UNION ALL
    SELECT DISTINCT jurisdiction_code
    FROM `handy-bonbon-142723.nypd_complaints.nypd_complaints_2021`
    )
```

Integrated Dimensional Models

Agency Dimension Model

Agency_Dimension.sql

save

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

with integrated AS (
SELECT *
FROM {{ref('nypd_agency')}}
LEFT JOIN {{ref('agency')}} USING (agency_dim_id)
)

SELECT agency_dim_id, agency, agency_name
FROM integrated
```

Preview | Compile | Query Results | Compiled SQL | Lineage

4.9 sec —Returned 78 rows. Download CSV

agency_dim_id	agency	agency_name
1	DEP	Department of Environmental Protection
2	NULL	NULL
3	NULL	NULL
4	NULL	NULL
5	NULL	NULL
6	NULL	NULL

Enter

ready

Location Dimensional Model

Location_Dimension.sql

save

```sql
{{config(
    materialized = "table"
)}}
WITH integrated AS(
    SELECT *
    FROM {{ref('location')}}
    LEFT JOIN {{ref('nypd_location')}} ON (location.location_dim_id = nypd_location.location_dim_id)
)

SELECT DISTINCT *, ROW_NUMBER() OVER (ORDER BY (SELECT NULL)) AS location_dim_id
FROM integrated
```

Preview | Compile | Query Results | Compiled SQL | Lineage

14.5 sec —Results limited to 500 rows.

Download CSV

location_dim_id	location_type	incident_zip	incident_address	street_name	cross_street_1	cross_street_2	address
381449	NULL	11104	43-12 46 STREET	46 STREET	43 AVE	QUEENS BLVD	ADDRESS
423058	NULL	11375	107-40 QUEENS BO…	QUEENS BOUL…	70 RD	71 AVE	ADDRESS
162672	NULL	11235	3000 BRIGHTON 12…	BRIGHTON 12…	OCEAN VIEW AVE	BRIGHTON BEACH…	ADDRESS
173653	NULL	11221	87 PALMETTO STRE…	PALMETTO ST…	BUSHWICK AVE	EVERGREEN AVE	ADDRESS

Enter

ready

Complaint Type Dimension

Complaint_Type.sql

save

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

SELECT *
FROM {{ref('complaint_type')}}
```

Preview | Compile | Query Results | Compiled SQL | Lineage

4.5 sec —Returned 64 rows. Download CSV

complaint_type_dim_id	complaint_type	descriptor	status
1	Noise	Noise, Other Animals …	Closed
2	Noise	Noise: Private Cartin…	Closed
3	Noise	Noise: Air Condition/…	Closed
4	Noise	Noise: Air Condition/…	Closed

Enter ready

Crime Dimension

Crime_Dimension.sql

save

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

SELECT *
FROM {{ref('nypd_crime')}}
```

Preview | Compile | Query Results | Compiled SQL | Lineage

4.2 sec —Results limited to 500 rows. Download CSV

crime_dim_id	level_of_offense	offense_code	internal_classification_code	descriptor	offens
1	FELONY	101	NULL	NULL	MURDER
2	MISDEMEANOR	351	256	MISCHIEF, …	CRIMIN
3	MISDEMEANOR	351	256	MISCHIEF, …	CRIMIN
4	MISDEMEANOR	351	256	MISCHIEF, …	CRIMIN

Enter ready

Fact Tables

Complaint Fact:

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

WITH complaint_table AS (
    SELECT * FROM {{ref('Complaint_Type')}}
),
agency_table AS (
    SELECT * FROM {{ref('Agency_Dimension')}}
),
location_table AS (
    SELECT * FROM {{ref('Location_Dimension')}}
),
date_table AS (
    SELECT * FROM {{ref('date')}}
),
crime_table AS (
    SELECT * FROM {{ref('Crime_Dimension')}}
),
created_date_table AS(
    SELECT ROW_NUMBER() OVER() AS created_date_dim_id,
    FROM `bigquery-public-data.new_york_311.311_service_requests`
    WHERE complaint_type IN ('Noise')
),
closed_date_table AS (
   SELECT ROW_NUMBER() OVER() AS closed_date_dim_id, unique_key
    FROM `bigquery-public-data.new_york_311.311_service_requests`
    WHERE complaint_type IN ('Noise')
),
total_table AS (
    SELECT * FROM complaint_table, agency_table,
    location_table, date_table, crime_table, created_date_table,
closed_date_table
)
```

```
SELECT agency_dim_id,
location_dim_id,
date_dim_id,
complaint_type_dim_id,
created_date_dim_id,
closed_date_dim_id,
crime_dim_id,
unique_key
FROM total_table
```

10.1 sec —Results limited to 500 rows. Download CSV

agency_dim_id	location_dim_id	date_dim_id	complaint_type_dim_id	created_date_dim_id	closed_date_dim_id	crime_dim_id	unique_key
47	117744	291	57	76	16	5543	39548833
47	117744	291	57	76	85	5543	41235380
47	117744	291	57	76	108	5543	39453500
47	117744	291	57	76	139	5543	39915674
47	117744	291	57	76	177	5543	39508735
47	117744	291	57	76	198	5543	39604705
47	117744	291	57	76	219	5543	18531888

Crime Fact:

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

WITH location_table AS(
    SELECT * FROM {{ref('Location_Dimension')}}
),
date_table AS (
    SELECT *
    FROM {{ref('date')}}
),
agency_table AS (
    SELECT *
    FROM {{ref('Agency_Dimension')}}
),
crime_table AS (
    SELECT *
```

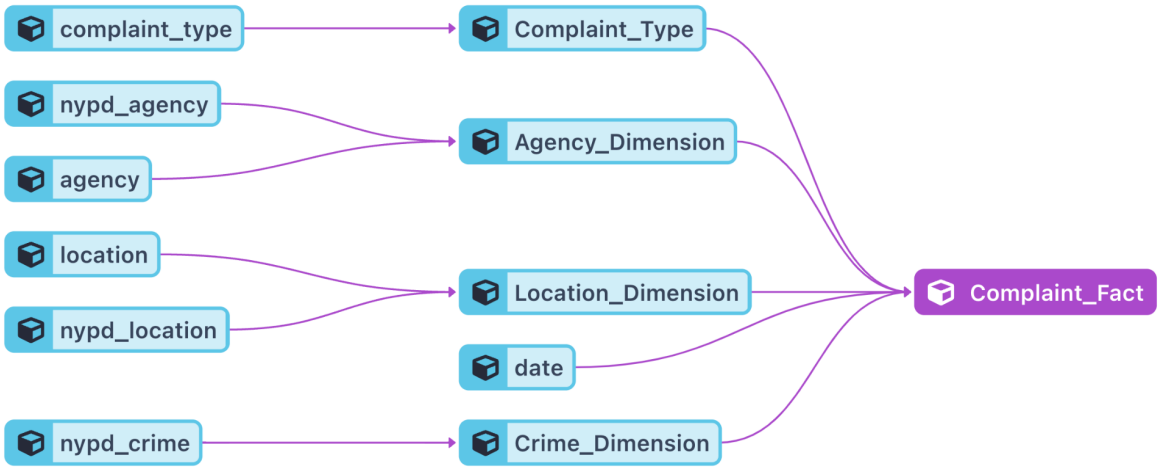
```
    FROM {{ref('Crime_Dimension')}}
),
from_date_table AS(
    SELECT ROW_NUMBER() OVER() as from_date_dim_id, CMPLNT_TO_TM, cmplnt_num
    FROM `handy-bonbon-142723.nypd_complaints.nypd_complaints_2019`
    WHERE CMPLNT_FR_DT IS NOT NULL
),
to_date_table AS(
    SELECT ROW_NUMBER() OVER() as to_date_dim_id, CMPLNT_TO_TM
    FROM `handy-bonbon-142723.nypd_complaints.nypd_complaints_2019`
    WHERE CMPLNT_TO_TM IS NOT NULL
),
time_table AS (
    SELECT ROW_NUMBER() OVER() as time_dim_id
    FROM {{ref('time')}}
),
total_table AS(
    SELECT * FROM location_table,
    date_table,
    agency_table,
    crime_table,
    from_date_table,
    to_date_table,
    time_table
)
SELECT location_dim_id,
date_dim_id,
agency_dim_id,
crime_dim_id,
cmplnt_num AS complaint_number,
from_date_dim_id,
to_date_dim_id,
time_dim_id,
FROM total_table
```

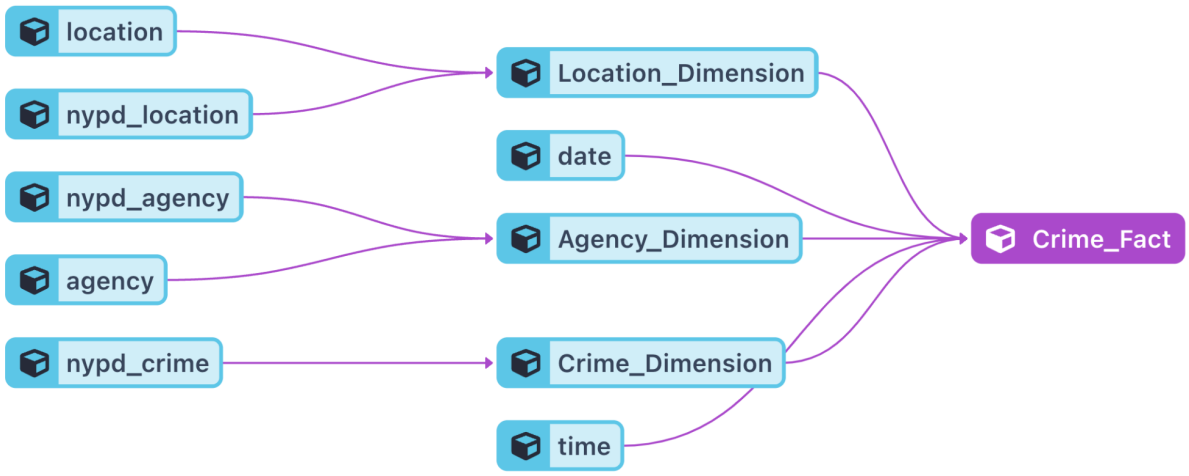
10.1 sec —Results limited to 500 rows.

Download CSV

agency_dim_id	location_dim_id	date_dim_id	complaint_type_dim_id	created_date_dim_id	closed_date_dim_id	crime_dim_id	unique_key
47	117744	291	57	76	16	5543	39548833
47	117744	291	57	76	85	5543	41235380
47	117744	291	57	76	108	5543	39453500
47	117744	291	57	76	139	5543	39915674
47	117744	291	57	76	177	5543	39508735
47	117744	291	57	76	198	5543	39604705
47	117744	291	57	76	219	5543	18531888

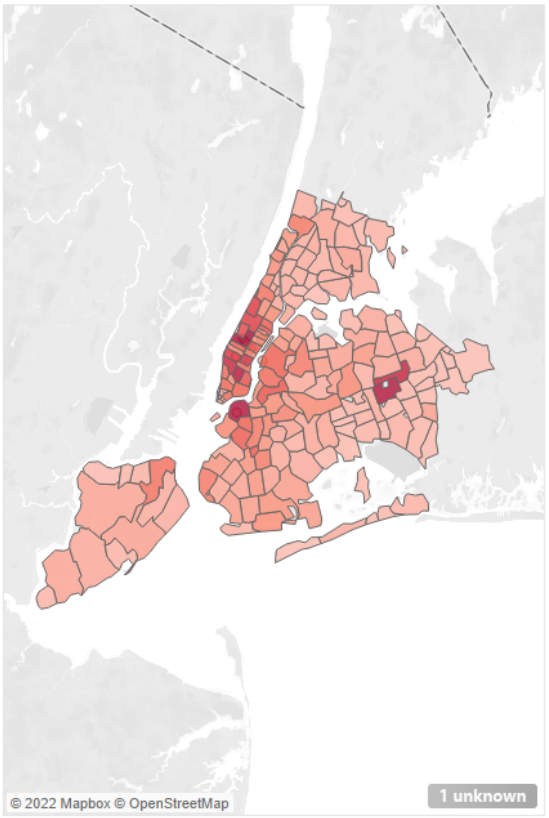
Lineages




location
nypd_location
nypd_agency
agency
nypd_crime
Location_Dimension
date
Agency_Dimension
Crime_Dimension
time
Crime_Fact

Dashboard

Number of Noise complaints per zip(density)



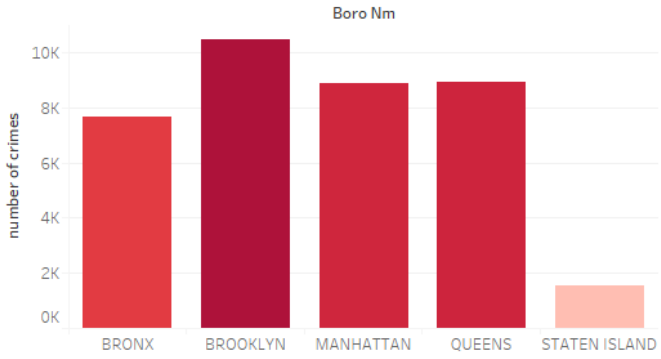
count of noisie complaints

1 799

Number of complaints per descriptor

Descriptor	
Noise, Barking Dog (NR5)	3,238
Noise, Ice Cream Truck (NR4)	1,511
Noise, Other Animals (NR6)	92
Noise: air condition/ventilation equipment (NV1)	2,349
Noise: Alarms (NR3)	3,001
Noise: Boat(Engine,Music,Etc) (NR10)	234
Noise: Construction Before/After Hours (NM1)	14,579
Noise: Construction Equipment (NC1)	4,016
Noise: Jack Hammering (NC2)	1,675
Noise: lawn care equipment (NCL)	260
Noise: Loud Music/Daytime (Mark Date And Time) (NN1)	3
Noise: Loud Music/Nighttime(Mark Date And Time) (NP1)	15
Noise: Manufacturing Noise (NK1)	68
Noise: Other Noise Sources (Use Comments) (NZZ)	6
Noise: Private Carting Noise (NQ1)	616
Noise: Vehicle (NR2)	1

Number of crimes per borough

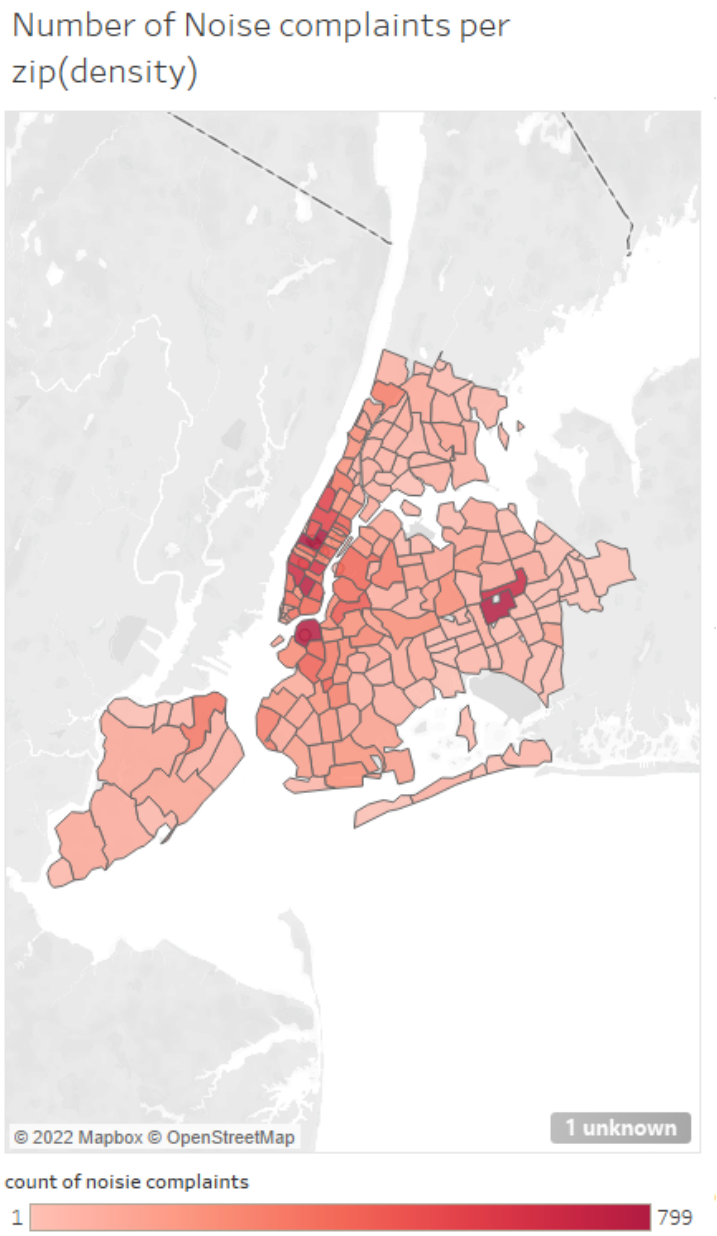


Count of crimes

1,535 10,453

The dashboard above was created in Tableau to visualize ouor kpi's. We chose to use gradient colors instead of static colors to make our visualizations stand out more.

KPI 1



This visualization is a map that shows the density of number of noise complaints by zipcodes. The visualization was created by using the longitude and latitude to get the map and then used the measure values to count of the zipcodes in the dataset.

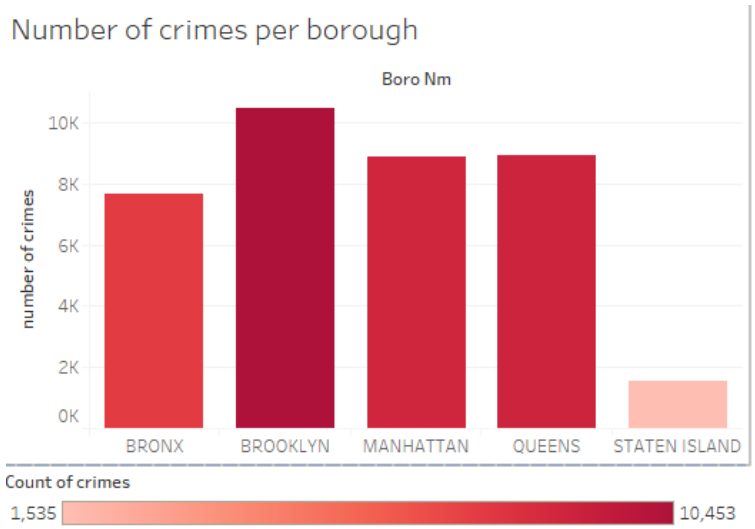
KPI 2

Number of complaints per descriptor

Descriptor	
Noise, Barking Dog (NR5)	3,238
Noise, Ice Cream Truck (NR4)	1,511
Noise, Other Animals (NR6)	92
Noise: air condition/ventilation equipment (NV1)	2,349
Noise: Alarms (NR3)	3,001
Noise: Boat(Engine,Music,Etc) (NR10)	234
Noise: Construction Before/After Hours (NM1)	14,579
Noise: Construction Equipment (NC1)	4,016
Noise: Jack Hammering (NC2)	1,675
Noise: lawn care equipment (NCL)	260
Noise: Loud Music/Daytime (Mark Date And Time) (NN1)	3
Noise: Loud Music/Nighttime(Mark Date And Time) (NP1)	15
Noise: Manufacturing Noise (NK1)	68
Noise: Other Noise Sources (Use Comments) (NZZ)	6
Noise: Private Carting Noise (NQ1)	616
Noise: Vehicle (NR2)	1

This visualization is a table that shows the density of number of noise complaints per descriptor. This table was created to show which type of noise complaint was most frequent. After creating the visualization it is construction before/after hours.

KPI 3



This visualization is a table that shows the density of number of crimes per borough. This vizualiation was created to see which Borough had the most crimes according to the data. Brooklyn has the most reported crimes to the NYPD dataset.

Conclusion

Descriptions of the tools:

In terms of tool selection, Python mainly uses APIs to scrape data and clean it by reformatting, removing missing values, and data analysis. We chose dbt to perform the extract, transform, and load (ETL) part, which helps us organize and analyze the data in our warehouse. Then, we chose Google BigQuery as our target database. By the end, we chose to complete the visualization part in tableau.

Tools used for this project:

- Python: Python mainly uses APIs to scrape data and clean it by reformatting, removing missing values, and data analysis.
- Excel: Excel is mainly used for simple analysis of data so that we can choose the data we need.
- dbt Cloud: dbt Cloud mainly runs ETL programming and completes the extract, transform, and load (ETL) part.
- Google BigQuery: Google BigQuery is used as our data warehouse, storing all our data.
- Tableau: Tableau mainly used for KPI calculations and visualization.

Group's Experience

This group project was very challenging for us as it was the first time we were on the same project from the beginning to the end of the semester.
The most difficult part of this project is that we need to maintain the consistency between the two data. In order to maintain the consistency of the two data, we first need to ensure the consistency of the Dimension model, and then clean and transform the two data. We spend a lot of time on this part.
In addition, we also spent some time on data collation. First, the data is too big, and we need to delete a lot of data that we don't need. Second, some missing values or too many invalid cells will affect our operation.

Benefits can be realized by the new system

As our Project Proposal says, our system tries to correlate the number of noise complaints per city/borough with the number of crimes per city/borough. New York City is a city with a huge population, and as the population continues to rise, so does the crime rate. We noticed areas with high crime rates that corresponded to more noise complaints from surrounding homes,

streets/sidewalks, commercial buildings and parks. On the new system, we can analyze the noise complaint data in various areas of New York City to identify the areas with severe noise and strengthen the security management of the area. We believe that the new system will reduce the crime rate in New York City.

Final comments and conclusions

In the group project, what we feel most strongly is teamwork. We keep in touch on WhatsApp and have regular zoom meetings, which makes us work more like a team and will be great for us in the future. We all use our skills to make this group project work better. This group project made us understand the importance of teamwork and provided us with a good project experience. Most importantly, this group project allowed us to review and reinforce all the skills we learned in college.

Reference Link

https://holowczak.com/

https://cloud.google.com/architecture/performing-etl-from-relational-database-into-bigquery

Analytics, Mjr. “Column-Level Data Profiling for Google BigQuery Datasets Using DBT.” Rittman Analytics, Rittman Analytics, 3 June 2020, https://rittmananalytics.com/blog/2020/6/4/column-level-data-profiling-for-google-bigquery-datasets-using-dbt.

https://holowczak.com/drawing-entity-relationship-diagrams-with-uml-notation-using-lucidchart/

https://holowczak.com/data-warehousing-projects/

http://holowczak.com/getting-started-with-nyc-opendata-and-the-socrata-api/

https://docs.getdbt.com/tutorial/getting-started

Meeting Log:
Date & time of meeting: 6 PM
End of meeting: 7:30PM
Attendees: Khilola Rustamova, Jason Mei, Jacky Chen, Yizuo Zheng
*Can't get a hold of Francesco
Main topic discussed or work done: Finalized/proposed ideas for the project proposal

Meeting Log 3/2/22:
Date & time of meeting: 7 PM
End of meeting: 9:30PM
Attendees: Khilola Rustamova, Jason Mei, Jacky Chen, Yizuo Zheng
Main topic discussed or work done: Worked on draft for dimensional modeling

Meeting Log 3/15/22:
Date & time of meeting: 7 PM
End of meeting: 9:30PM
Attendees: Khilola Rustamova, Jason Mei, Jacky Chen, Yizuo Zheng
Main topic discussed or work done: Finalized the dimensional modeling

Meeting Log 4/12/22:
Date & time of meeting: 7 PM
End of meeting: 9:30PM
Attendees: Khilola Rustamova, Jason Mei, Jacky Chen, Yizuo Zheng
Main topic discussed or work done: Worked on draft for dimensional modeling

Meeting Log 5/10/22:
Date & time of meeting: 7 PM
End of meeting: 9:30PM
Attendees: Khilola Rustamova, Jason Mei, Jacky Chen, Yizuo Zheng
Main topic discussed or work done: ETL Tools and Target DBMS Selected

Meeting Log 5/18/22:
Date & time of meeting: 7 PM
End of meeting: 9:30PM
Attendees: Khilola Rustamova, Jason Mei, Jacky Chen, Yizuo Zheng
Main topic discussed or work done: Working on the final process