Parking Violations and Complaints in NYC with Correlation to Sidewalk Cafes

Prakhyat Rana *prakhyat.rana@baruchmail.cuny.edu*

Melissa Ramirez Nunez *melissa.ramireznunez5@baruchmail.cuny.edu*

William Zhu- *william.zhu1@baruchmail.cuny.edu*

CIS 4400 [FTA 34788] Data Warehousing for Analytics

Dr. Isaac Vaghefi

11/21/2021

Introduction (Issues/Problems)

Living in the city, parking has always been an issue for all New Yorkers, for drivers and non-drivers. There are problems with accidents, parking tickets, and parking violations. With the introduction of coronavirus in early 2020, there has been an increase in these issues because of space allocated for outdoor dining. There have also been debates on how NYC should be handling the lack of parking spots and how they can better allocate the use of space to solve these issues[1].

To find a solution to the reason for these main issues we are using the "311 Service Requests from 2010 to Present" data set as the primary data set to focus on the issue of illegal parking. With the second and third data set, "Open Parking and Camera Violations" and "Local Law 8 of 2020 – Complaints of Illegal Parking of Vehicles Operated on Behalf of the City", to filter more specifically the types of violations, the fine amount, and the specific locations of violation. The core of our problem will be to find a correlation between the introduction of sidewalk cafes and illegal parking. The data set "Sidewalk Cafe Licenses and Applications" would give us an insight to the amount of sidewalk cafes opening up in the city, the amount of space they are occupying, as well as, the allowed space they are allowed to occupy by regulations of the DCA (Department of Consumers r Affairs). All data sets are from NYC Open Data.

Data Sources

https://data.cityofnewyork.us/City-Government/Open-Parking-and-Camera-Violations/nc67-uf89

https://data.cityofnewyork.us/Social-Services/311-Service-Requests-from-2010-to-Present/erm2-nwe9

https://data.cityofnewyork.us/Business/Sidewalk-Caf-Licenses-and-Applications/qcdj-rwhu

https://data.cityofnewyork.us/City-Government/Local-Law-8-of-2020-Complaints-of-Illegal-Parking-/cwy2-px8b

[1] Rosa, Amanda. "Yes, Parking in New York Has Gotten Worse." *The New York Times*, The New York Times, 6 Jan. 2021, https://www.nytimes.com/2021/01/06/nyregion/nyc-parking.html.

KPIs (Key Performance Indicators)

- Rate of incoming parking violations for the last 5 years (2016 - 2021).
- Count of parking violations by type for each borough
- Average fine amount by year
- Average amount of violations by time period (AM/PM)
- Popularity of different violation per borough
- Percent increase of square feet occupation of sidewalk cafes by borough for the last 5 years (2016-2021)
- Average of square feet occupied by sidewalk cafes.
- Correlation of sidewalk cafes and parking violations/complaints by zip code and borough

Dimensional Modeling Draft

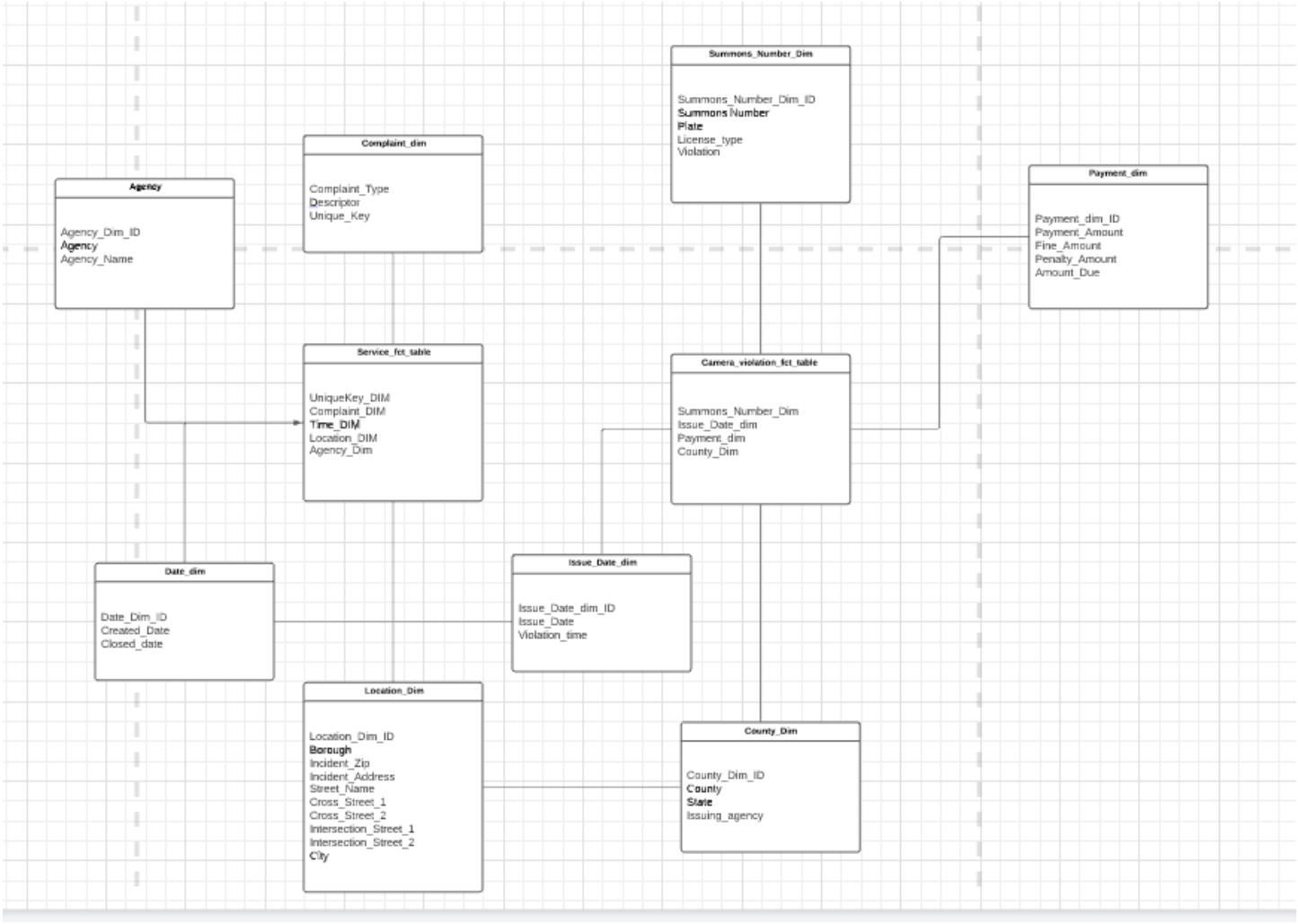


Figure 1. Parking Violations and Sidewalk Cafes Data Dimensional Model (First Draft)

Using the "311 service request from 2010 to present" and "open parking and camera violations" dataset, we created a dimensional model with the idea of using date and location dimensions to connect the two datasets. We decided to select a monthly grain to achieve a better understanding of parking violations on a month to month basis.

Finalized Dimensional Model

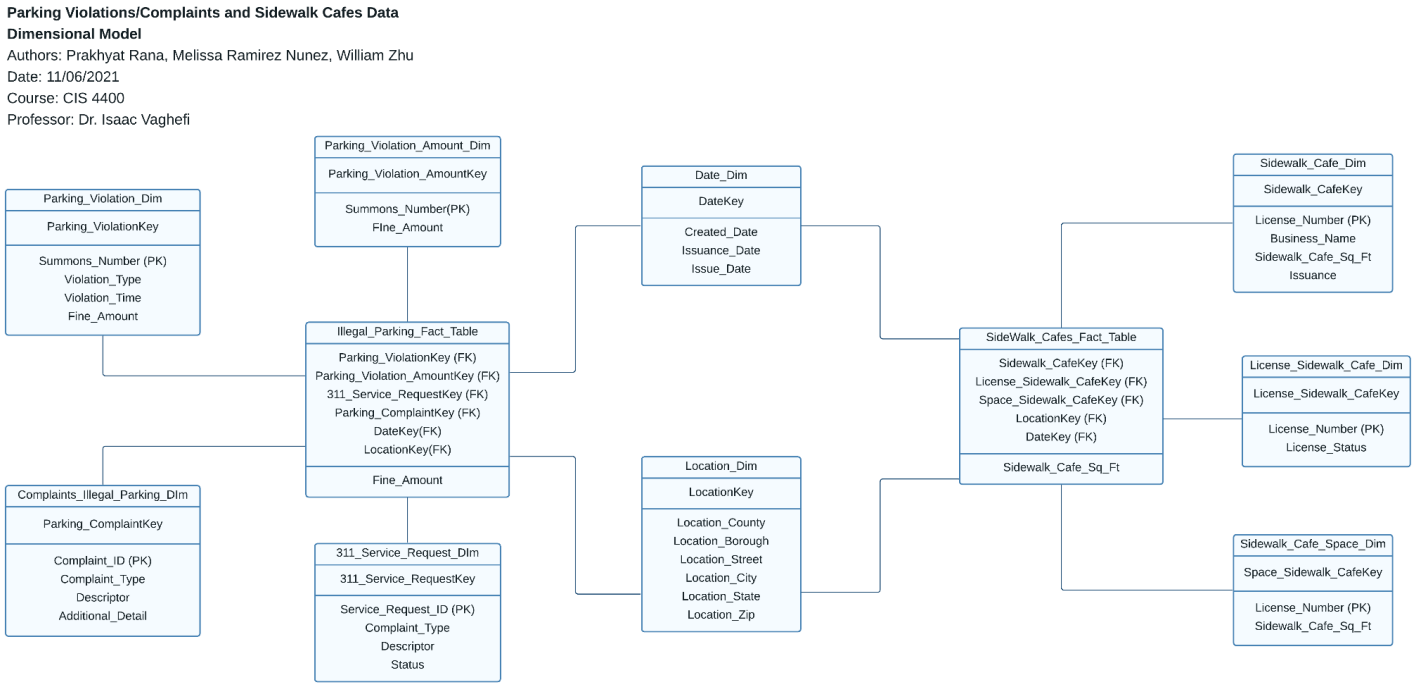


Figure 2. Parking Violations/Complaints and Sidewalk Cafes Data Dimensional Model (Final Model)

Link to Dimensional Model via LucidChart

From our previous draft, we decided to include 2 more data sets "Sidewalk Cafe Licenses and Applications" and "Local Law 8 of 2020 – Complaints of Illegal Parking of Vehicles Operated on Behalf of the City". We modified our KPI's to include average square feet, percent increase of square feet occupation, and correlation of sidewalk cafes to parking violations and parking complaints. To connect the two fact tables, we used the date and location dimensions. We decided to select a monthly grain to achieve a better understanding of parking violations and parking complaints on a month to month basis.

ETL Transfer Plan Using Pentaho Software

We plan on using Pentaho Software for our ETL process. Using the CSV file input feature, we will transfer the data into Pentaho and filter out null entries using SQL in the filter rows function. We will place null entries into a text file and no instances of null entries into a Microsoft Excel output. We will then continue to clean our data by making sure there are no repeated entries by using unique rows. After that, we will plan on editing some of our metadata using select values and finally using table output to open a connection to Google BigQuery.

ETL Process

To transform our public datasets and use it in the project, we performed the following:

- Used Input Functions in Pentaho to load "Sidewalk Cafe Licenses and Applications", "Local Law 8 of 2020 – Complaints of Illegal Parking of Vehicles Operated on Behalf of the City", "311 Service Requests from 2010 to Present", and "Open Parking and Camera Violations" CSV files.
- Reviewed the files, made adjustments to several columns and removed unnecessary ones.
- Renamed the important columns using the "Select Values" and "Renaming Fields" tool.
- Identified field names and field types by filtering out rows for each of them.
- Filtered all rows to remove null values.
- Used all the borough locations as a mainstream field to match with our Excel file using the "Fuzzy Match" tool for data set "Local Law 8 of 2020 – Complaints of Illegal Parking of Vehicles Operated on Behalf of the City".
- Filtered out null entries for "311 Service Requests from 2010 to Present" and "Open Parking and Camera Violations" data sets, however the size of these files were extremely large and took us several minutes to run.

ETL Transformations

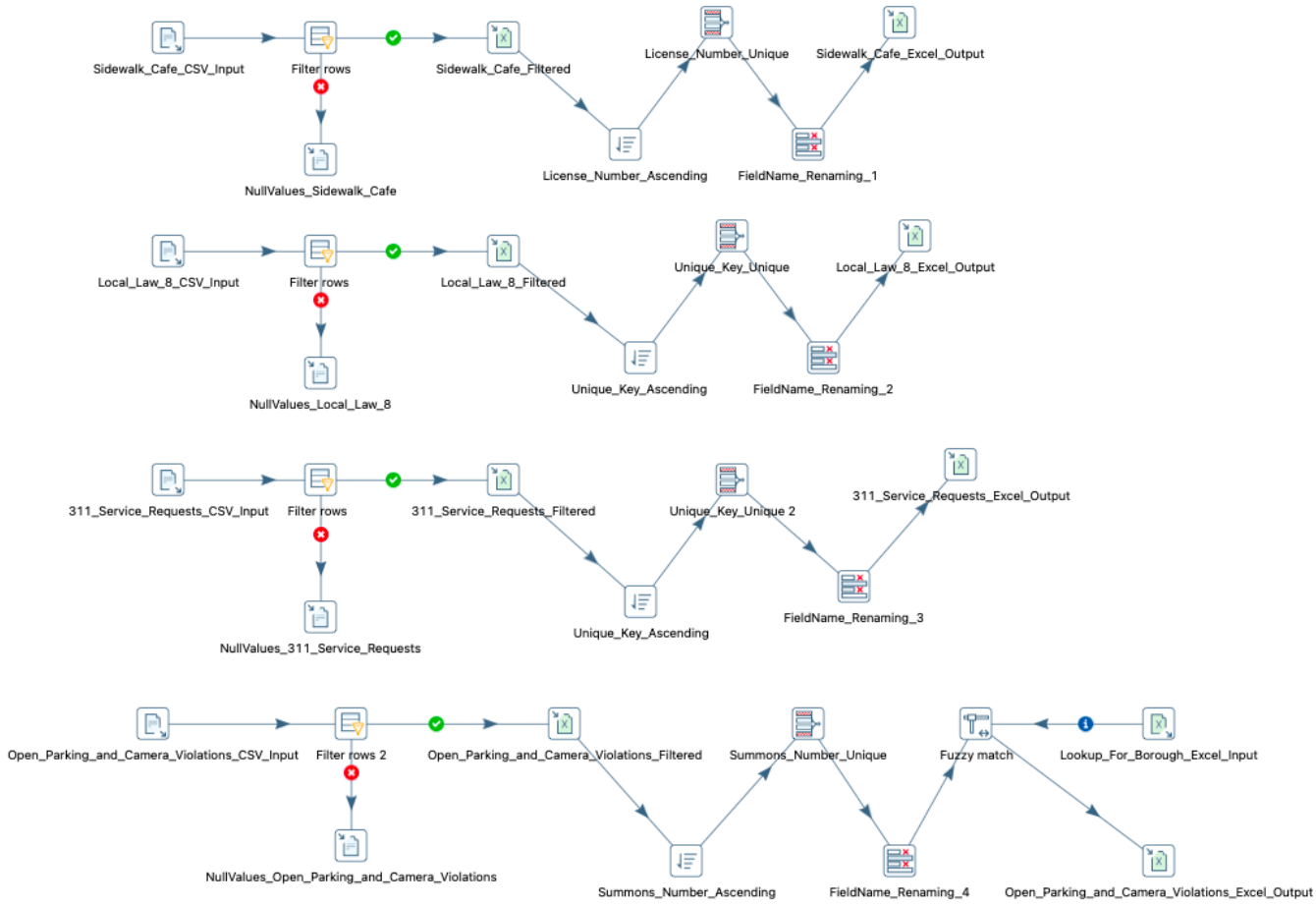


Figure 3. ETL Transformations of Primary DataSets

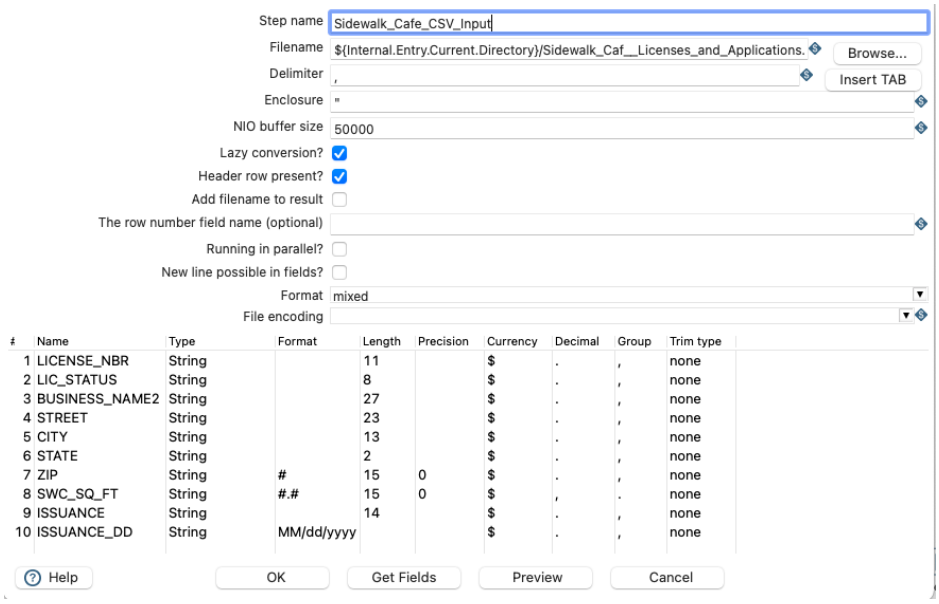


Figure 4. Transformation on Primary DataSet (Inputting and Selecting Necessary Fields)

Rows of step: Sidewalk_Cafe_CSV_Input (100 rows)

#	LICENSE_NBR	LIC_STATUS	BUSINESS_NAME2	STREET	CITY	STATE	ZIP	SWC_SQ_FT	ISSUANCE	ISSUANCE_DD
1	1420950-DCA	Active	THE SKYLARK	5TH AVE	BROOKLYN	NY	11215	236	Issued	06/21/2019
2	2019767-DCA	Active	12 CHAIRS CAFE	WYTHE AVE	BROOKLYN	NY	11249	237	Issued	03/15/2019
3	1260966-DCA	Active	LOVELY DAY	ELIZABETH ST	NEW YORK	NY	10012	53	Issued	02/07/2020
4	1448450-DCA	Inactive	<null>	E 186TH ST	BRONX	NY	10458	131	Issued	02/06/2019
5	2021276-DCA	Active	PLAYA BETTY'S	AMSTERDAM AVE	NEW YORK	NY	10023	474	Issued	10/09/2019
6	2051771-DCA	Active	<null>	5TH AVE	BROOKLYN	NY	11215	274	Issued	07/09/2019
7	2022878-DCA	Inactive	<null>	10TH AVE	NEW YORK	NY	10019	259	Issued	01/04/2019
8	0293907-DCA	Inactive	PETE'S TAVERN S/C #111	E 18TH ST	NEW YORK	NY	10003	611	Pending Review	<null>
9	1463548-DCA	Active	SERAFINA	E 61ST ST	NEW YORK	NY	10065	81	<null>	<null>
10	2056574-DCA	Inactive	AMELIE	AMSTERDAM AVE	NEW YORK	NY	10024	85	Pending Review	<null>

Figure 5. Preview of DataSet with Necessary Fields (Shows null values that will be outputs in a text file.)

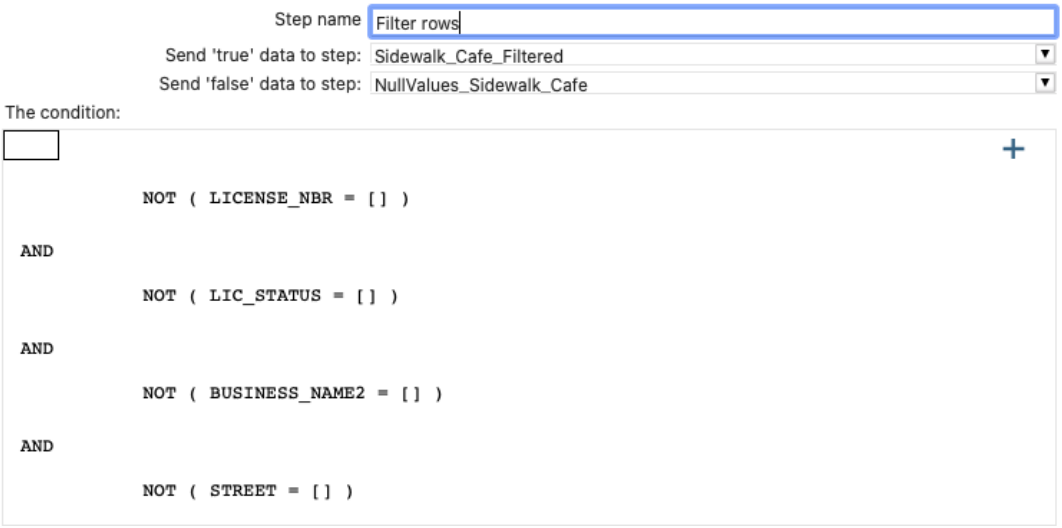
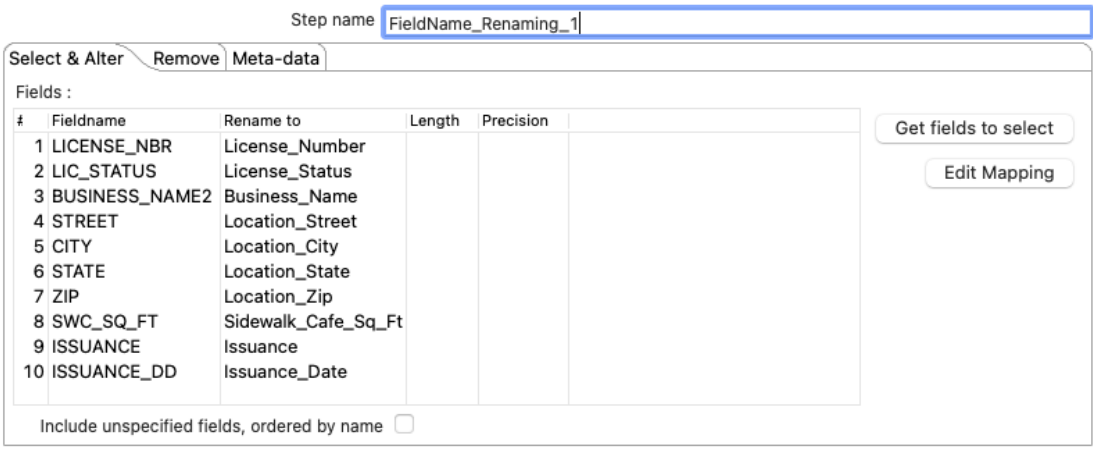


Figure 6. Filtering All Rows To Remove Null Values

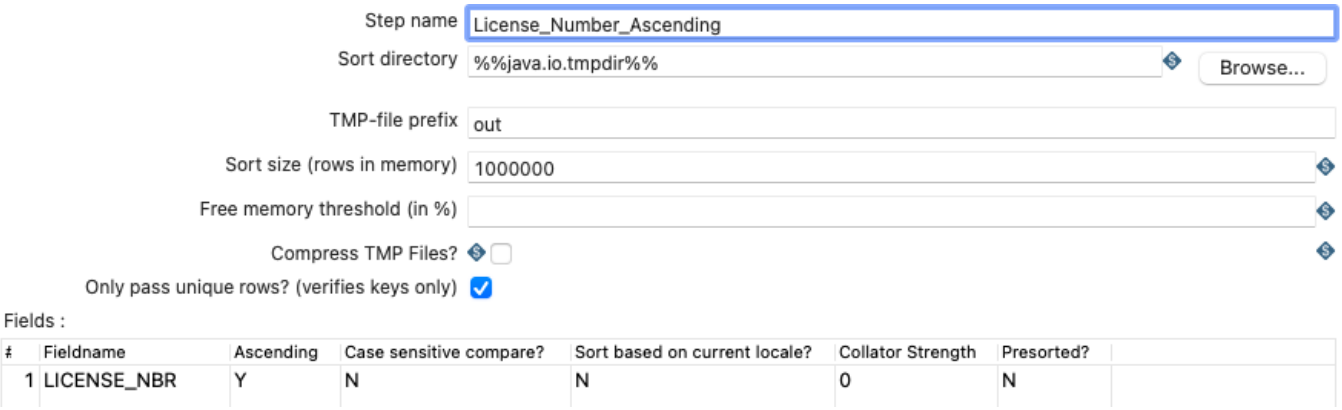
LICENSE_NBR	LIC_STATUS	BUSINESS_NAME2	STREET	CITY	STATE	ZIP	SWC_SQ_	ISSUANCE	ISSUANCE_DD
1420950-DCA	Active	THE SKYLARK	5TH AVE	BROOKLYN	NY	11215	236	Issued	06/21/2019
2019767-DCA	Active	12 CHAIRS CAFE	WYTHE AVE	BROOKLYN	NY	11249	237	Issued	03/15/2019
1260966-DCA	Active	LOVELY DAY	ELIZABETH ST	NEW YORK	NY	10012	53	Issued	02/07/2020
2021276-DCA	Active	PLAYA BETTY'S	AMSTERDAM AVE	NEW YORK	NY	10023	474	Issued	10/09/2019
0955674-DCA	Inactive	THE STUMBLE INN	2ND AVE	NEW YORK	NY	10021	212	Issued	05/02/2019
2078943-DCA	Inactive	P.J. HORGAN'S	SKILLMAN AVE	WOODSIDE	NY	11377	352	Issued	03/19/2019
2077328-DCA	Inactive	BUTLER	S 5TH ST	BROOKLYN	NY	11249	174	Issued	11/16/2018
2070370-DCA	Inactive	BELLINI	COLUMBUS AVE	NEW YORK	NY	10024	73	Issued	08/08/2018
2022715-DCA	Active	DIG INN SEASONAL MARKET	BROADWAY	NEW YORK	NY	10025	159	Issued	04/25/2019
0918521-DCA	Active	CABANA RESTAURANT AND BAR	70TH RD	FOREST HILLS	NY	11375	106	Issued	08/29/2019
2071748-DCA	Inactive	CAFEINE	FREDERICK DOUGLASS BLVD	NEW YORK	NY	10026	137	Issued	08/14/2018
2059065-DCA	Active	SAL'S PIZZA	BROOME ST	NEW YORK	NY	10013	98	Issued	08/07/2019
2050276-DCA	Active	DALTON'S	ASTORIA BLVD	EAST ELMHURST	NY	11370	232	Issued	01/27/2020
2006616-DCA	Inactive	ARIA HELLS' KITCHEN	W 51ST ST	NEW YORK	NY	10019	144	Issued	08/17/2018
2009249-DCA	Inactive	ELI'S ESSENTIALS	MADISON AVE	NEW YORK	NY	10021	20	Issued	09/18/2018
1355808-DCA	Inactive	The 13th Step	2ND AVE	NEW YORK	NY	10003	215	Issued	05/06/2019

Figure 7. Filtered DataSet with No Null Values



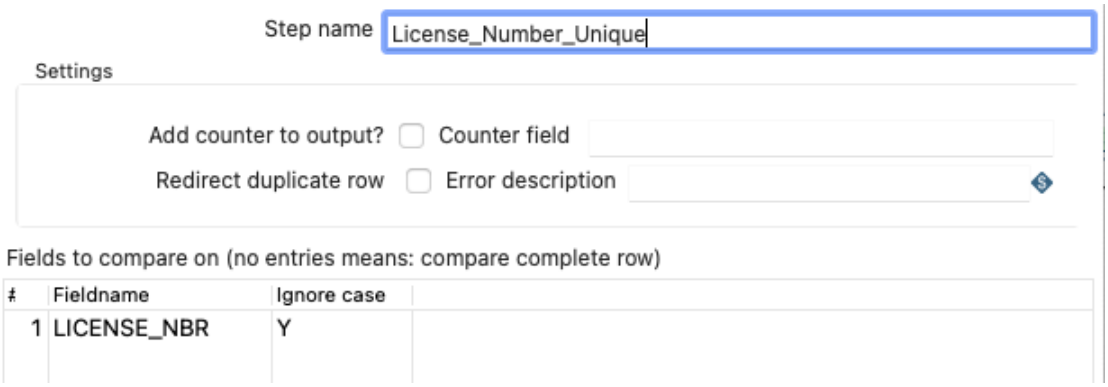
#	Fieldname	Rename to	Length	Precision
1	LICENSE_NBR	License_Number		
2	LIC_STATUS	License_Status		
3	BUSINESS_NAME2	Business_Name		
4	STREET	Location_Street		
5	CITY	Location_City		
6	STATE	Location_State		
7	ZIP	Location_Zip		
8	SWC_SQ_FT	Sidewalk_Cafe_Sq_Ft		
9	ISSUANCE	Issuance		
10	ISSUANCE_DD	Issuance_Date		

Figure 8. Renaming Fields to Coincide with Dimension Model



#	Fieldname	Ascending	Case sensitive compare?	Sort based on current locale?	Collator Strength	Presorted?
1	LICENSE_NBR	Y	N	N	0	N

Figure 9. Primary Key Ascending



#	Fieldname	Ignore case
1	LICENSE_NBR	Y

Figure 10. Primary Key as Unique Field

License_Number	License_Status	Business_Name	Location_Street	Location_City	Location_State	Location_Zip	Sidewalk_Cafe_Sq_Ft	Issuance	Issuance_Date
0670525-DCA	Active	CALIENTE CAB CO	7TH AVE S	NEW YORK	NY	10014	752	Issued	07/25/2019
0697987-DCA	Active	E/S/C #255	2ND AVE	NEW YORK	NY	10022	429	Issued	06/12/2019
0767745-DCA	Active	JUNIORS RESTAURANT	FLATBUSH AVENUE EXT	BROOKLYN	NY	11201	123	Issued	05/01/2019
0804523-DCA	Inactive	DALLAS BBQ	2ND AVE	NEW YORK	NY	10003	535	Issued	06/13/2019
0835262-DCA	Inactive	DUE	3RD AVE	NEW YORK	NY	10075	84	Issued	06/27/2019
0883095-DCA	Active	HI LIFE BAR & GRILL	AMSTERDAM AVE	NEW YORK	NY	10024	302	Issued	03/11/2020
0885881-DCA	Active	HI-LIFE BAR & GRILL	AMSTERDAM AVE	NEW YORK	NY	10024	212	Issued	05/09/2019
0890521-DCA	Active	BEYOGLU	3RD AVE	NEW YORK	NY	10028	253	Issued	07/29/2019
0895505-DCA	Active	FIREHOUSE RESTAURANT	COLUMBUS AVE	NEW YORK	NY	10024	255	Issued	09/18/2019
0895586-DCA	Active	THE STUMBLE INN	2ND AVE	NEW YORK	NY	10021	265	Issued	12/04/2019
0918358-DCA	Inactive	BUS STOP	HUDSON ST	NEW YORK	NY	10014	297	Issued	04/02/2019
0918521-DCA	Active	CABANA RESTAURANT AND BAR	70TH RD	FOREST HILLS	NY	11375	106	Issued	08/29/2019
0919619-DCA	Inactive	BAR 6	AVENUE OF THE AMERICAS	NEW YORK	NY	10011	89	Issued	05/15/2019
0919917-DCA	Active	COWGIRL AND BAR K	HUDSON ST	NEW YORK	NY	10014	206	Issued	08/29/2019
0920440-DCA	Inactive	Caffe Palermo	MULBERRY ST	NEW YORK	NY	10013	125	Issued	01/11/2018
0920983-DCA	Inactive	SAVORE RISTORANTE	SPRING ST	NEW YORK	NY	10012	293	Issued	05/02/2019

Figure 11. Preview of Finalized DataSet with Transformations

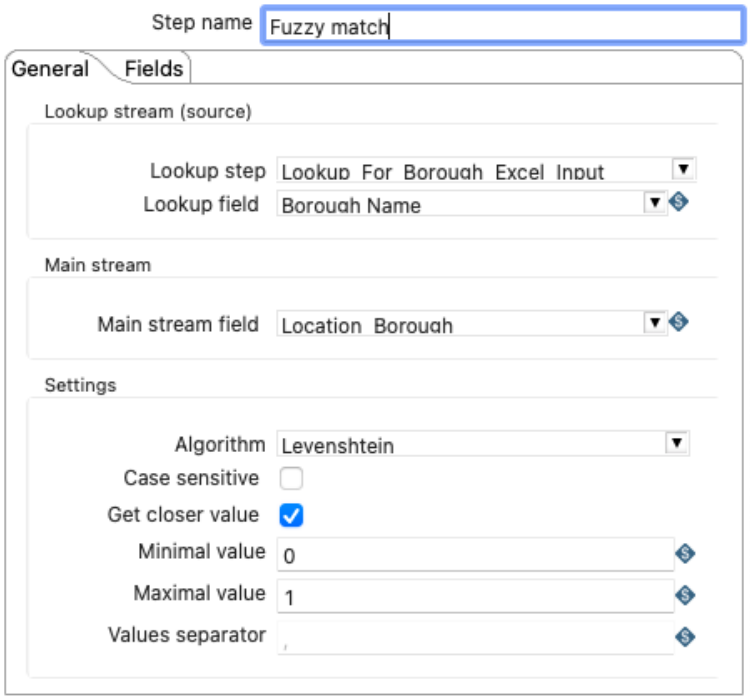


Figure 12. Matching A Specific Field Name Data with an Accurate Excel File

Rows of step: Open_Parking_and_Camera_Violations_CSV_Input (100 rows)

#	Summons Number	Issue Date	Violation Time	Violation	Fine Amount	County
1	8842345003	07/31/2021	12:10P	FAIL TO DSPLY MUNI METER RECPT	35.00	R
2	8686206669	08/28/2018	09:24A	NO STANDING-BUS STOP	115.00	NY
3	8686206670	08/28/2018	09:24A	PLTFRM LFTS LWRD POS COMM VEH	65.00	NY
4	8686207182	08/29/2018	12:36P	NO STANDING-DAY/TIME LIMITS	115.00	NY
5	8686207236	08/29/2018	01:02P	SAFETY ZONE	115.00	NY
6	8686207248	08/29/2018	01:04P	NO STANDING-BUS STOP	115.00	NY
7	8686207261	08/29/2018	01:10P	NO STANDING-DAY/TIME LIMITS	115.00	NY
8	8686207297	08/29/2018	01:16P	NO STANDING-DAY/TIME LIMITS	115.00	NY
9	8686207510	08/30/2018	01:04P	NO STANDING-DAY/TIME LIMITS	115.00	NY
10	8686207984	08/31/2018	08:14A	DOUBLE PARKING	115.00	NY
11	8686208186	09/01/2018	12:32P	FIRE HYDRANT	115.00	NY
12	8686208393	09/01/2018	02:07P	NO PARKING-DAY/TIME LIMITS	65.00	NY
13	8686208400	09/01/2018	02:15P	NO STANDING-DAY/TIME LIMITS	115.00	NY
14	8686208915	09/04/2018	08:19A	INSP. STICKER-EXPIRED/MISSING	65.00	NY
15	8686209002	09/04/2018	08:55A	NO STANDING-BUS STOP	115.00	NY

Figure 13. Example of Field Name "County" for Transformation of Matching

Borough Name
Bronx
Queens
Manhattan
Staten Island
Brooklyn

Figure 14. Example of Accurate Excel File for Matching Field Name "County"

ETL Transformations Summary

We used an "Excel Input" to enter our datasets and auto-filled fields. We then removed unnecessary fields in our dataset, represented in figure 4. In our next step of our cleaning process we removed null entries into a text file and kept valid entries in an Excel sheet. We accomplished this by using the "Unique Row" function and provided the SQL code as 'not null values', as shown in figure 6, and results are shown in figure 7. We changed the metadata field names to match the ones in our dimensional model for easier interpretation. Furthermore, in our efforts to clean data we removed duplicate records and sorted our primary keys in ascending order. Sample of final preview shown in figure 11. We repeated these steps for all datasets. Excluding data set "Open Parking and Camera Violations" which includes a transformation to match field "County" with accurate metadata. This was completed using the tool "Fuzzy Match" and connecting with an outside Excel file to match the metadata respectively to their best match. This transformation step is shown in figueres 12, 13, and 14. We waited several hours for Pentaho to run its transformation and decided that writing SQL code would be faster. It was fairly simple code, I selected a distinct Complaint_ID/Summons number, renamed my fields to match our dimensional model, and made sure there were no null values figure 15. The only real issue we encountered was trying to convert 'county' into 'boroughs' to have a consistent location type through all our datasets.

openparkjing NNV.xls_195	11/21/2021 12:28 AM	Microsoft Excel Worksheet	7,708 KB
openparkjing NNV.xls_196	11/21/2021 12:30 AM	Microsoft Excel Worksheet	7,705 KB
openparkjing NNV.xls_197	11/21/2021 12:32 AM	Microsoft Excel Worksheet	7,703 KB
openparkjing NNV.xls_198	11/21/2021 12:34 AM	Microsoft Excel Worksheet	7,706 KB
openparkjing NNV.xls_199	11/21/2021 12:36 AM	Microsoft Excel Worksheet	7,705 KB
openparkjing NNV.xls_200	11/21/2021 12:38 AM	Microsoft Excel Worksheet	7,701 KB
openparkjing NNV.xls_201	11/21/2021 12:40 AM	Microsoft Excel Worksheet	7,704 KB
openparkjing NNV.xls_202	11/21/2021 12:42 AM	Microsoft Excel Worksheet	7,699 KB
openparkjing NNV.xls_203	11/21/2021 12:43 AM	Microsoft Excel Worksheet	6,146 KB

Figure 14.5. Filtering out null values for the open parking dataset yielded 203 excel sheets with 70,000 lines each and took several hours. (this was only one step)

```
SELECT distinct LICENSE_NBR as License_Number,
LIC_STATUS as License_Status,
BUSINESS_NAME2 as Business_Name,
STREET as Location_Street,
CITY as Location_City,
STATE as Location_State,
ZIP as Location_Zip,
SWC_SQ_FT as Sidewalk_Cafe_Sq_Ft,
ISSUANCE as Issuance,
ISSUANCE_DD as Issuance_Date
FROM `cis-4400-homework-classwork.Group6Project.Sidewalk_Cafe`

where BUSINESS_NAME2 is not null
and LICENSE_NBR is not null
and ISSUANCE is not null
and ISSUANCE_DD is not null)
```

Figure 15. SQL code used for ETL on Sidewalk-cafe dataset

```
1  SELECT count(Summons_Number) as Num_of_violation, violation, county
2  FROM `cis-4400-homework-classwork.Group6Project.OpenParking Cleaned`
3  where violation is not null
4  group by  violation, county
5  order by Num_of_violation desc
```

Figure 16. Sql code used to find popular violations by county

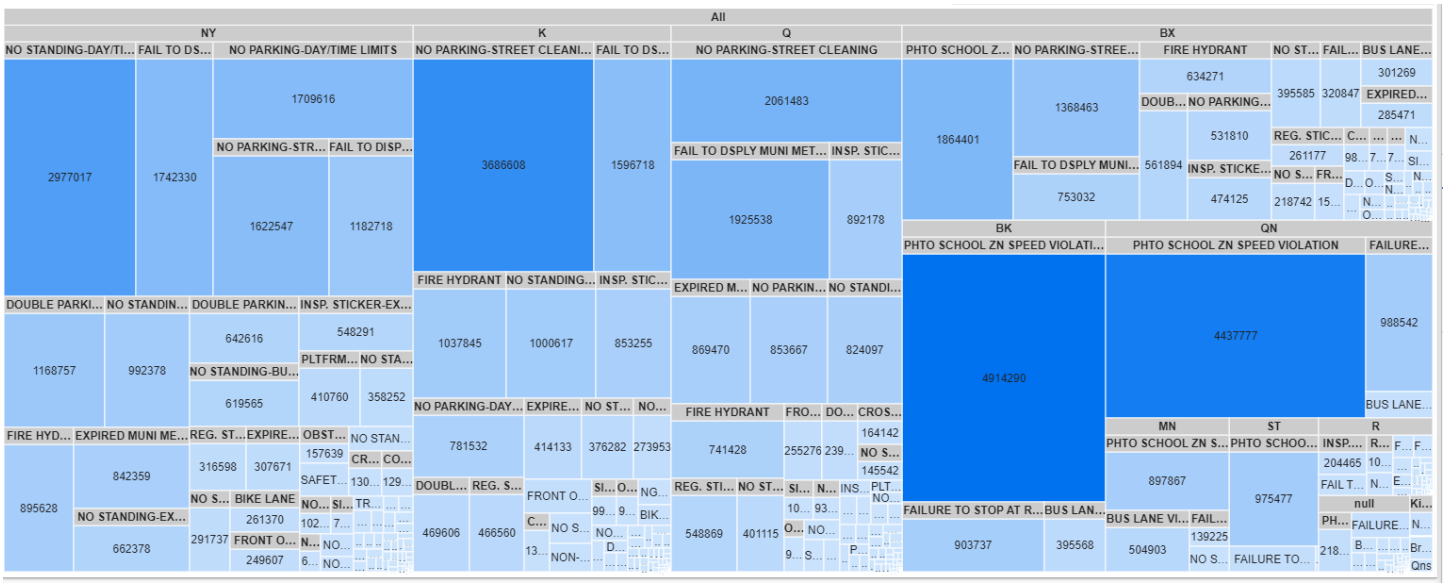
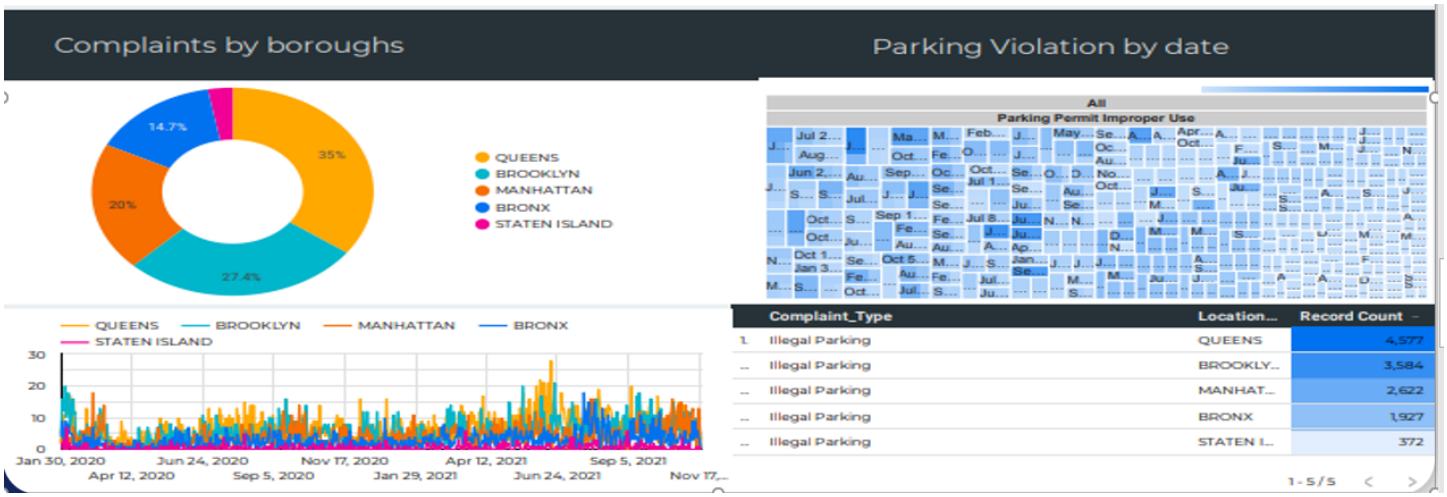


Figure 17. Popular violations by borough

The most prominent violation is speeding in a school zone in Brooklyn, Bronx and Queens while people tend to get fined more often for parking during street cleaning in Manhattan and Kings county



Here we see that Queens has the highest number of complaints as well as Illegal parking according to the donut chart. We can also conclude that June/July have the highest count of violation by month which is most noticeable in queens.

```
SELECT avg(Sidewalk_Cafe_Sq_Ft) as Avg_sidewalk_sqft
 FROM `cis-4400-homework-classwork.Group6Project.sidewalk` LIMIT 1000
```

Row	Avg_sidewalk_sqft
1	251.983870967742

We found out that the average space sidewalk-cafe's occupy is 251.98 square feet. To put that into perspective, the minimum space for a parking space is 9ft wide and 18ft long which is about 162 square ft

Final thoughts on the 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?)

This project helped us put into perspective how large a dataset can really be. The most difficult part was the wait time needed to run a transformation in Pentaho. This was difficult because one step took more than five hours and we had to terminate it because it was no longer a feasable idea. Instead we turned to google BigQuery for the ETL process which was significantly faster than Pentaho. However, there were data type issues that we've come across such as DateTime. One of the easiest parts of this project was setting up Pentaho for ETL. It proves itself to be extremely user friendly and maps out the ETL process. If we had to do this project all over again, we would definitely choose less datasets as adding two more became twice as confusing. One thing I learned that was unexpected was creating a dashboard from my current dataset/queries. It's a painless transition to the final step of this project that didn't require uploading all the datasets again.

a) the software and database tools the group used to coordinate and manage the project as well as carry out the programming tasks (list of bullet points with software or service and one sentence of what it was used for)

- Google docs- shared view/edit allowed all of us to contribute to a different part of the project proposal at once
- Whatsapp- used for communicating availabilities and issues with our project
- Lucid chart- Used for Dimensional modeling.
- Pentaho- initially used for ETL.
- Google BigQuery- Running SQL code for ETL.
- Data Studio- used for data visualization.

Meeting Logs:

Date & Time: 09/26 (10:30 PM - 12 AM)

Attendees: Melissa Ramirez, Prakhyat Rana

Discussion: Project ideas, choosing between the issue of animal abuse and illegal parking.

Date & Time: 09/27 (10:30 PM - 11:30 PM)

Attendees: Melissa Ramirez, Prakhyat Rana

Discussion: Finalization of the project proposal.

Date & Time: 10/10 (10:00 PM - 11:30 PM)

Attendees: William Zhu

Discussion: Dimensional modeling

Date & Time: 10/17 (4:00 PM - 5:30 PM)

Attendees: Melissa Ramirez, William Zhu

Discussion: Editing project proposal.

Date & Time: 11/03 (11:00 PM - 12:30 AM)

Attendees: Melissa Ramirez, William Zhu, Prakhyat Rana

Discussion: Editing dimensional modeling and deciding DBMS for data warehousing.

Date & Time: 11/06 (6:00 PM - 7:00 PM)

Attendees: Melissa Ramirez, William Zhu, Prakhyat Rana

Discussion: Finalizing dimensional model

Date & Time: 11/18 (5:00 PM - 7:00 PM)

Attendees: Melissa Ramirez, William Zhu, Prakhyat Rana

Discussion: Starting ETL

Date & Time: 11/20 (4:30 PM - 7:00 PM, 11:00 PM - 1:00 AM)

Attendees: Melissa Ramirez, William Zhu, Prakhyat Rana

Discussion: Finalizing ETL

Date & Time: 12/5

Attendees: William Zhu, Prakhyat Rana

Discussion: Looking for an alternative ETL process since pentaho takes a lot of time.

Date & Time: 12/6 - 12/7

Attendees: William Zhu, Prakhyat Rana

Discussion:Data visualiztion and presentation preparation.