

Baruch
COLLEGE

ZICKLIN
SCHOOL OF
BUSINESS

CIS 9440 Data Warehousing and Analytics

Project:

Traffic & Street Issues Data Analysis

Submission To:

Prof Richard Holowczak

By:

Xinyi Li(***)**

Immanuel Ryan Augustine (IMMANUEL.AUGUSTINE@baruchmail.cuny.edu)

Table OF Contents:

Introduction

KPIs

Dimensional Model

ETL

Tableau Analysis

Tools and Softwares Used

Challenges

Conclusion

Meeting Log

References

- **Introduction:**

A friend's high school cousin who lives in Staten Island walks to school every day. He often complains about the street conditions, claiming that he will not go back to school until those problems are solved. The sidewalks are broken and make him tumble several times; the pedestrian signals are sometimes out that he would be scared by the passing vehicles; and when he comes home late the streetlights are out which is horrible. His parents have already filed several complaints to 311, but the problems are still unsolved. We would like to use our knowledge and resources to help them, so we gather three other Data Warehousing classmates to do a project to help 311 work more effectively.

The purpose we have is to develop a working data warehouse using a commercial database management system (Oracle Autonomous Data Warehouse) and development tools. The three types of complaints we've chose are:

- Street Light Condition.
- Traffic Signal Condition.
- Sidewalk Condition.

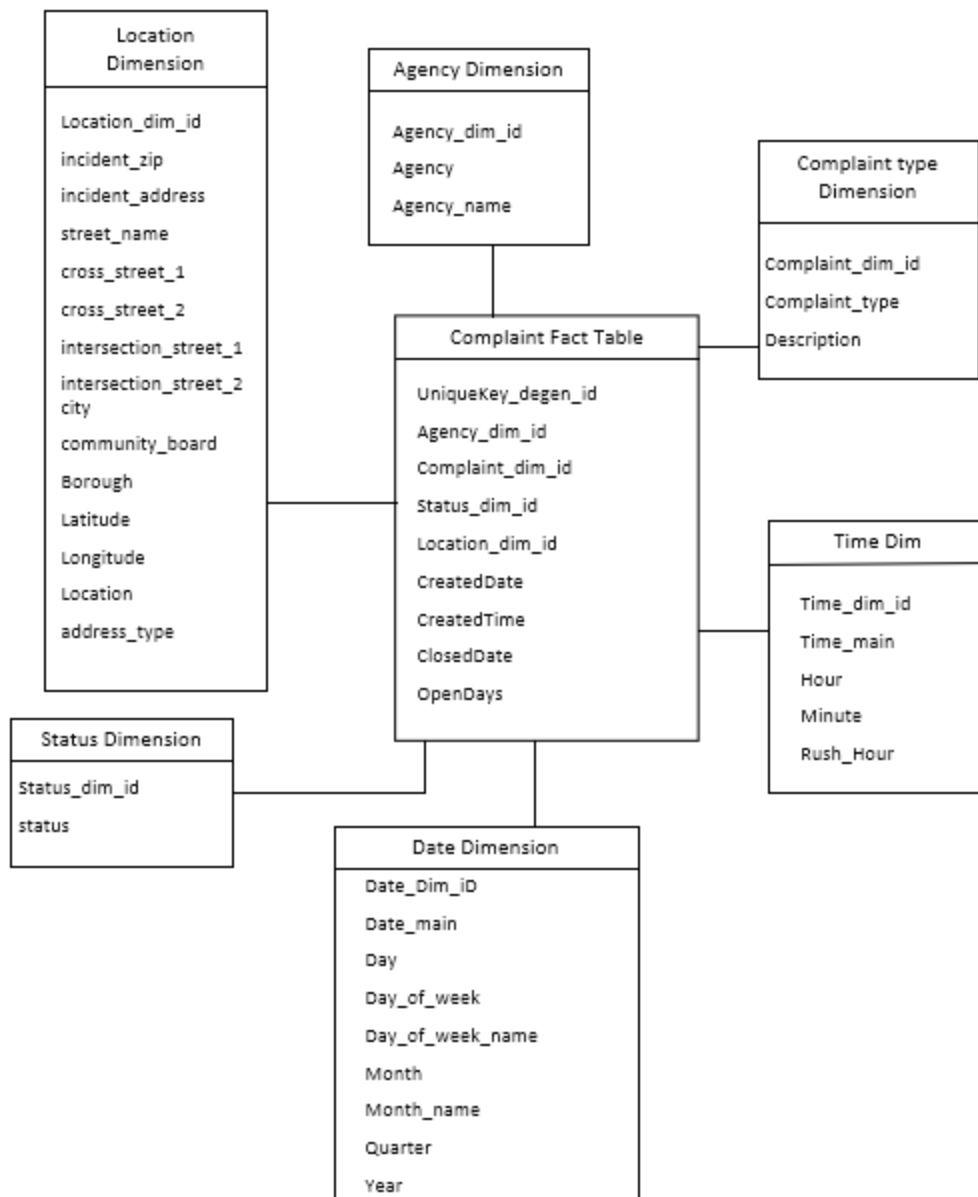
We will analyze the data over 3 years (2017-2019) to figure out which are the main problems such as Street light outage, Traffic Signal controllers not working and broken sidewalks. In addition, we need to analyze parameters such as the location, time, complaints per day/week/month/year.

- **Key Performance Indicators (KPI's):**

Based on the analysis, list of the KPIs for the complaint types we have chosen are:

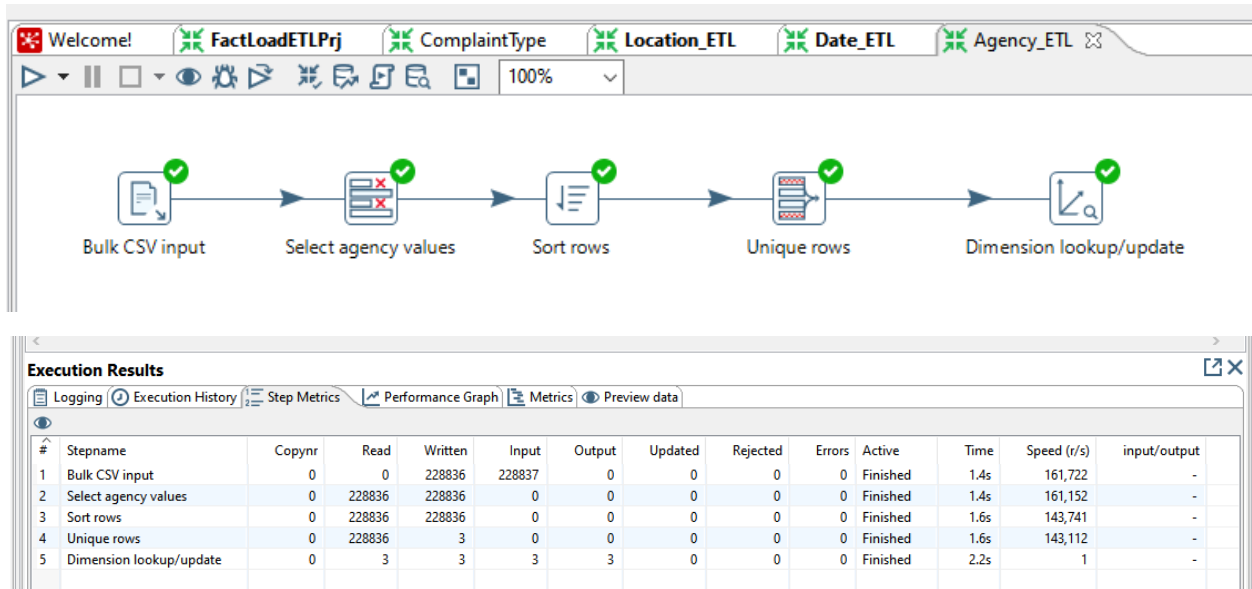
- Complaints Per Month Per Year
- Average Time of day for complaints
- Number of Complaints per weekday
- Average closure time for complaints
- Highest closure time by complaint types and description

- **Dimensional Model:**

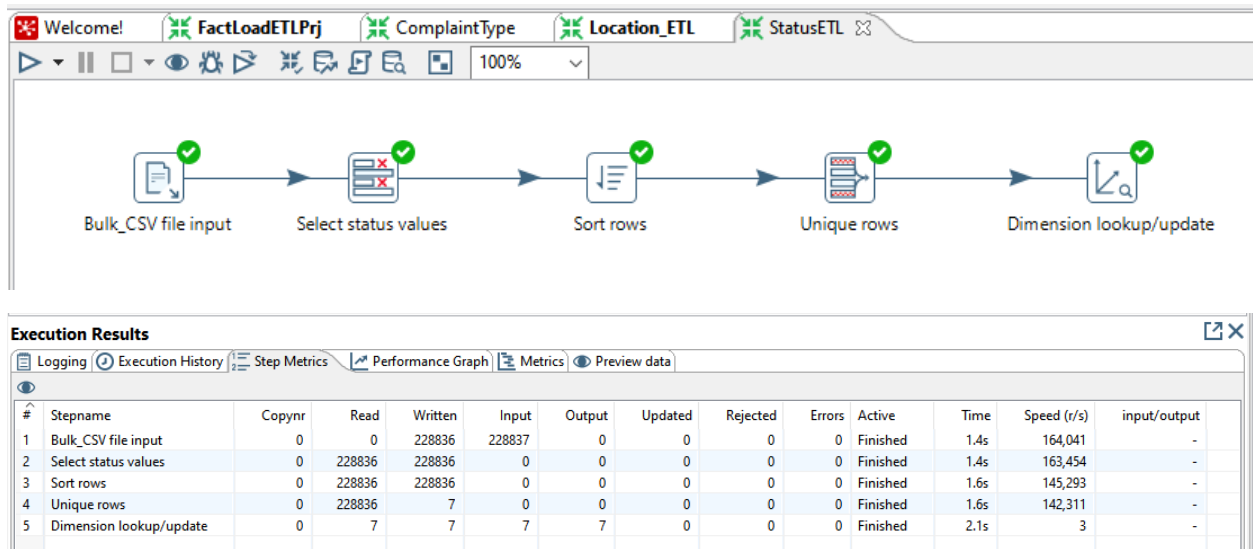


- **ETL Programming:**

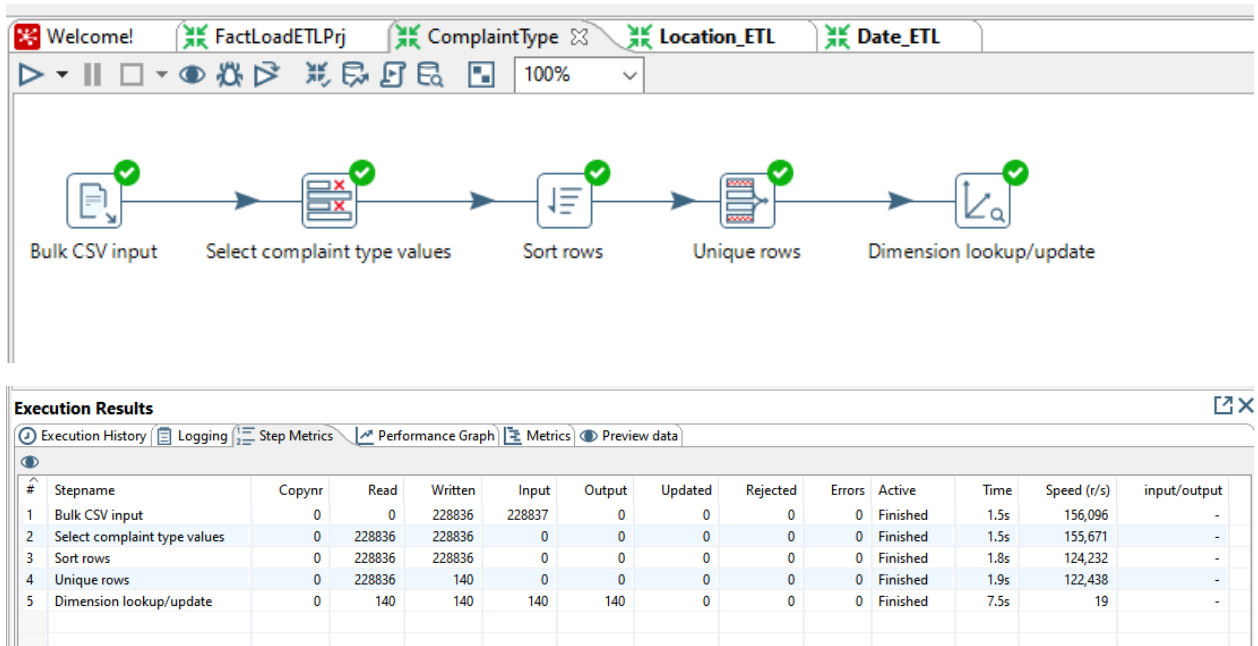
1. Agency Dimension: Input the Bulk data as CSV file and used Select Values step to pick columns related to Agency (Agency, Agency_Name). Sorted rows to filter the unique values in the Agency columns. Connected the dimensional lookup step to the Oracle Cloud Autonomous Data Warehouse and ran the transformation.



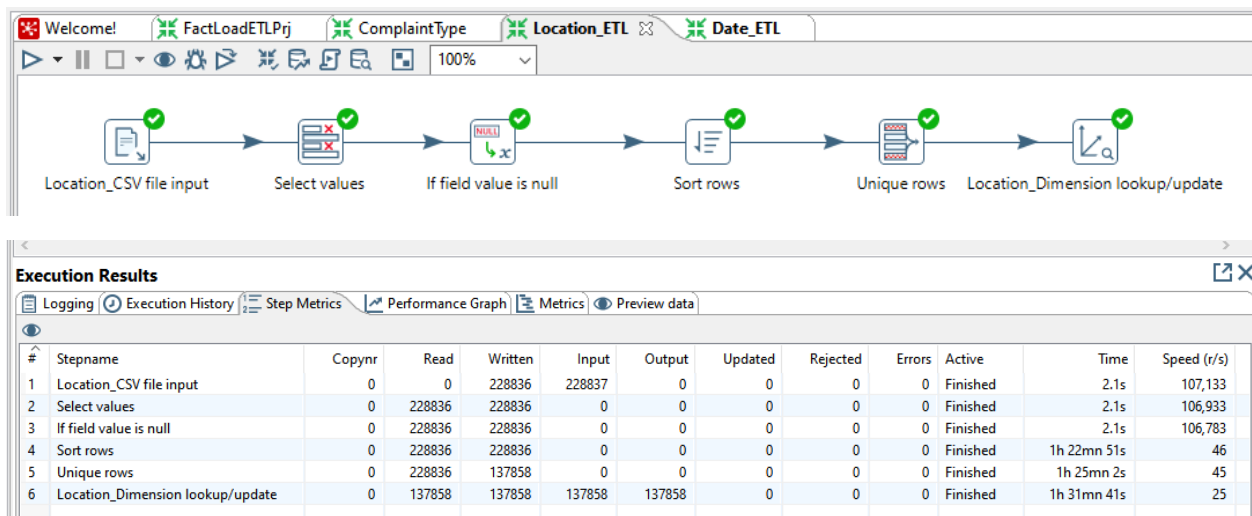
2. Status Dimension: Input the Bulk data as CSV file and used Select Values step to pick columns related to Status (Status). Sorted rows to keep the unique values in the Status columns. Connected the dimensional lookup step to the Oracle Cloud Autonomous Data Warehouse and ran the transformation.



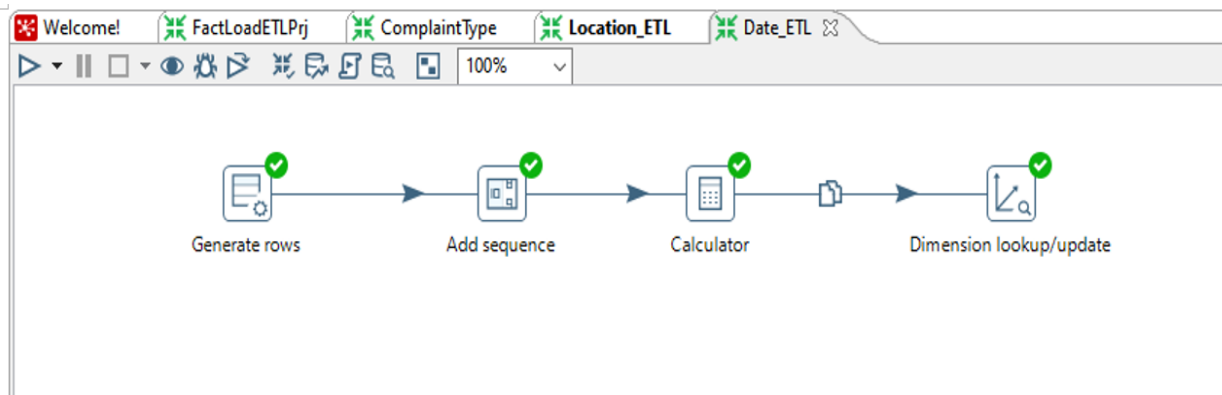
- Complaint_Type Dimension: Input the Bulk data as CSV file and used Select Values step to pick columns related to Complaint Type (Complaint_type, descriptor). Sorted rows to keep the unique values in the Complaint_Type columns. Connected the dimensional lookup step to the Oracle Cloud Autonomous Data Warehouse and ran the transformation.



4. Location Dimension: Input the Bulk data as CSV file and used Select Values step to pick columns related to Location (incident_zip, intersection_street_1, intersection_street_2, address_type, city, community_board, borough, latitude, longitude, location, incident_address, street_name, cross_street_1, cross_street_2). Since there were many rows with blank data, we ran a step to replace the NULL/blank values with a default N/A value. Sorted rows to keep the unique values in the location related columns. Connected the dimensional lookup step to the Oracle Cloud Autonomous Data Warehouse and ran the transformation.



5. Date Dimension: Generated 200 rows of a date value to which a sequence row was added with numbers incremented from 0 to 1999. Using the calculator, the 2 columns were added to get the date in ascending order from 2015 to 2020. Also, other embellishments such as quarter, day of month, etc. were added in this step. Finally, a database connection was made to the ADW and the data was loaded into a dimension using dimensional lookup/update step.



Execution Results

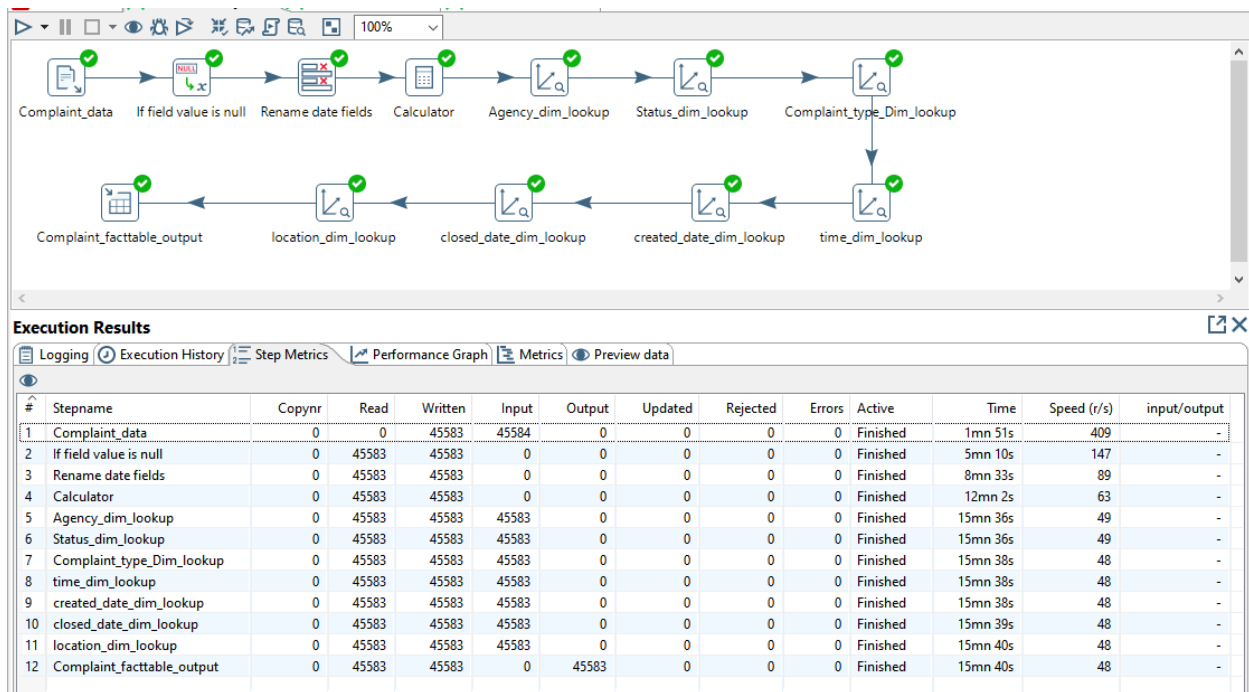
LoggingExecution HistoryStep MetricsPerformance GraphMetricsPreview data

#	Stepname	Copyn	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	input/output
1	Generate rows	0	0	2000	0	0	0	0	0	Finished	0.1s	25,316	-
2	Add sequence	0	2000	2000	0	0	0	0	0	Finished	0.1s	17,544	-
3	Calculator	0	2000	2000	0	0	0	0	0	Finished	0.7s	2,674	-
4	Dimension lookup/update	0	2000	2000	2000	2000	0	0	0	Finished	1mn 21s	24	-

6. Time Dimension: The time dimension was directly uploaded from an excel file (provided by Professor Holowczak) to the Oracle Autonomous Data Warehouse

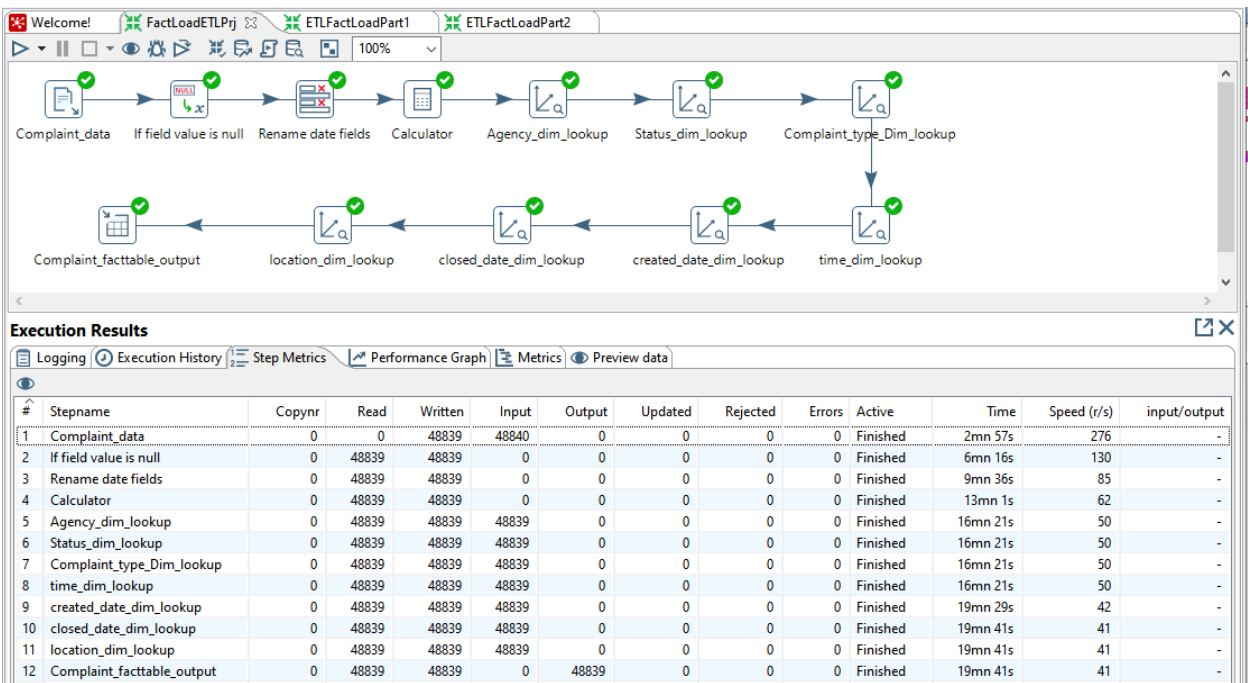
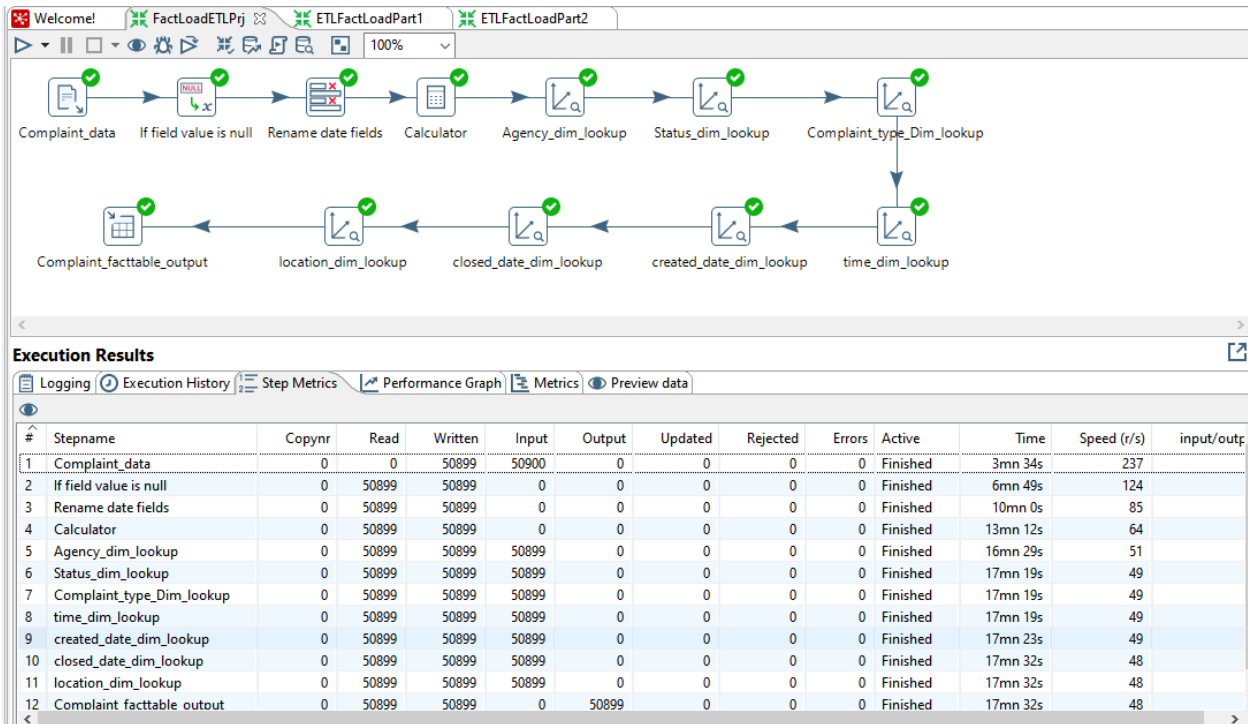
7. Loading Fact Table: Initially, we loaded the raw data in PDI. Since there were many rows with blank data, we ran a step “If field value NULL” to replace the NULL/blank values with a default value. (e.g.: NA for NULL strings, 999999 for NULL integers, etc.). We then renamed the date fields to date and time fields as they had both date and time information. In the calculator step, we split the default date value into created date, closed date and created time. After this step, we added various lookup steps for all the dimensions used in the dimensional model. Finally, on confirming that all the dimensional IDs are correctly mapped, we selected the required values in the table output step and ran the transformation to write the data to the database.

There was 1 main issue while loading the data to the fact table. PDI, after running for 25-30 minutes, would reduce the speed of the transformation (reads/second) to less than 5 reads/second for the last 3 steps in the transformation. At that speed, the data would take days to be loaded in the fact table. Hence, to get around this issue, we split the data into multiple files and ran the transformation for each file. In this way, we were able to finish loading all the fact table data in a day. This issue was understandably not observed while programming the ETL with sample data but only with the bulk data.

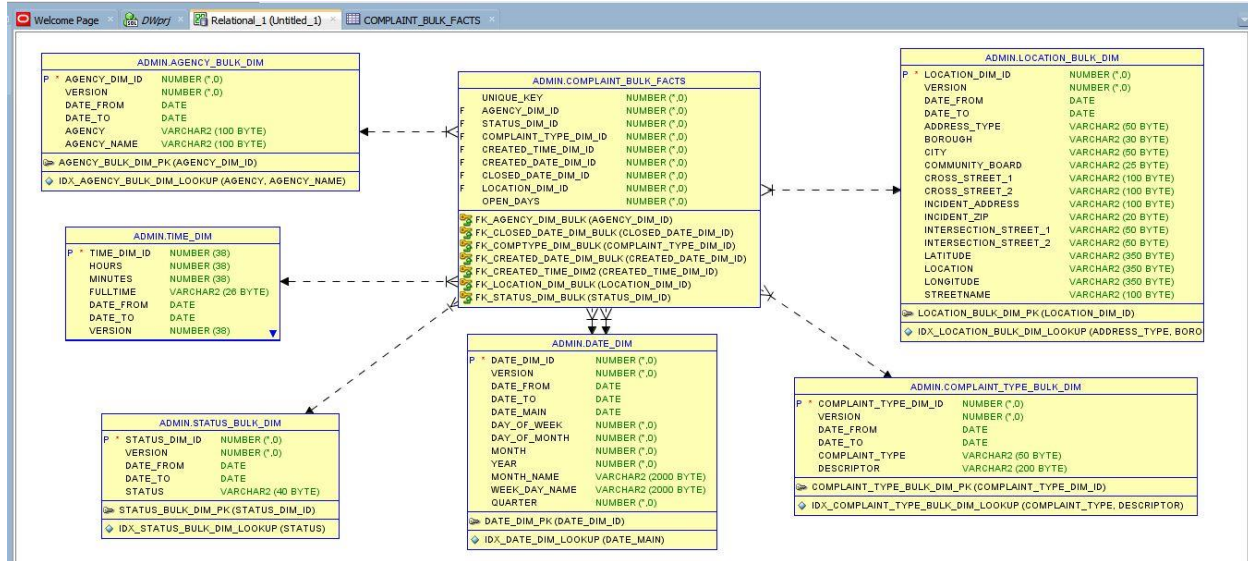


The screenshot displays the SAP Data Services Designer interface. The top section shows a job flow diagram with the following steps: Complaint_data, If field value is null, Rename date fields, Calculator, Agency_dim_lookup, Status_dim_lookup, Complaint_type_Dim_lookup, time_dim_lookup, created_date_dim_lookup, closed_date_dim_lookup, location_dim_lookup, and Complaint_facttable_output. Each step is marked with a green checkmark. The bottom section, titled 'Execution Results', provides a detailed table of the job's performance.

#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	input/output
1	Complaint_data	0	0	45583	45584	0	0	0	0	Finished	1mn 51s	409	-
2	If field value is null	0	45583	45583	0	0	0	0	0	Finished	5mn 10s	147	-
3	Rename date fields	0	45583	45583	0	0	0	0	0	Finished	8mn 33s	89	-
4	Calculator	0	45583	45583	0	0	0	0	0	Finished	12mn 2s	63	-
5	Agency_dim_lookup	0	45583	45583	45583	0	0	0	0	Finished	15mn 36s	49	-
6	Status_dim_lookup	0	45583	45583	45583	0	0	0	0	Finished	15mn 36s	49	-
7	Complaint_type_Dim_lookup	0	45583	45583	45583	0	0	0	0	Finished	15mn 38s	48	-
8	time_dim_lookup	0	45583	45583	45583	0	0	0	0	Finished	15mn 38s	48	-
9	created_date_dim_lookup	0	45583	45583	45583	0	0	0	0	Finished	15mn 38s	48	-
10	closed_date_dim_lookup	0	45583	45583	45583	0	0	0	0	Finished	15mn 39s	48	-
11	location_dim_lookup	0	45583	45583	45583	0	0	0	0	Finished	15mn 40s	48	-
12	Complaint_facttable_output	0	45583	45583	0	45583	0	0	0	Finished	15mn 40s	48	-



8. Star Schema: Finally, we added the foreign key constraints to the fact table and connected the tables to form a star schema as shown below. In the schema, each dimension is linked to the fact table by a dimensional ID key and the date dimension is linked twice, for both the created date and closed date for the complaint.



- **Analytics:**

We used Tableau to analyze the data and to visualize the KPIs. TO get the data into Tableau, we created a view in Oracle and exported it to a file which was subsequently imported in Tableau. We created a dashboard and mapped out the KPIs and made our observations.

View Creation Code:

```
CREATE VIEW factView AS

SELECT

    UNIQUE_KEY, d1.DATE_MAIN AS CREATED_DATE, d1.DAY_OF_MONTH CREATED_DAY_OF_MONTH,
    d1.DAY_OF_WEEK CREATED_DAY_OF_WEEK

    , d1.MONTH CREATED_MONTH, d1.MONTH_NAME CREATED_MONTH_NAME, d1.QUARTER
    CREATED_QUARTER, d1.WEEK_DAY_NAME CREATED_WEEKDAY_NAME

    , d1.YEAR CREATED_YEAR, OPEN_DAYS

    , d2.DATE_MAIN CLOSED_DATE, d2.DAY_OF_MONTH CLOSED_DAY_OF_MONTH,
    d2.DAY_OF_WEEK CLOSED_DAY_OF_WEEK, d2.MONTH CLOSED_MONTH

    , d2.MONTH_NAME CLOSED_MONTH_NAME, d2.QUARTER CLOSED_QUARTER, d2.WEEK_DAY_NAME
    CLOSED_WEEKDAY_NAME, d2.YEAR CLOSED_YEAR

    , FULLTIME CREATED_TIME, HOURS CREATED_HOUR, MINUTES CREATED_MIMNUTE

    , COMPLAINT_TYPE, DESCRIPTOR

    , STATUS

    , AGENCY, AGENCY_NAME

    , ADDRESS_TYPE , BOROUGH, CITY, COMMUNITY_BOARD, CROSS_STREET_1,
    CROSS_STREET_2, INCIDENT_ADDRESS, INCIDENT_ZIP

    , INTERSECTION_STREET_1, INTERSECTION_STREET_2, LATITUDE, LOCATION, LONGITUDE,
    STREETNAME

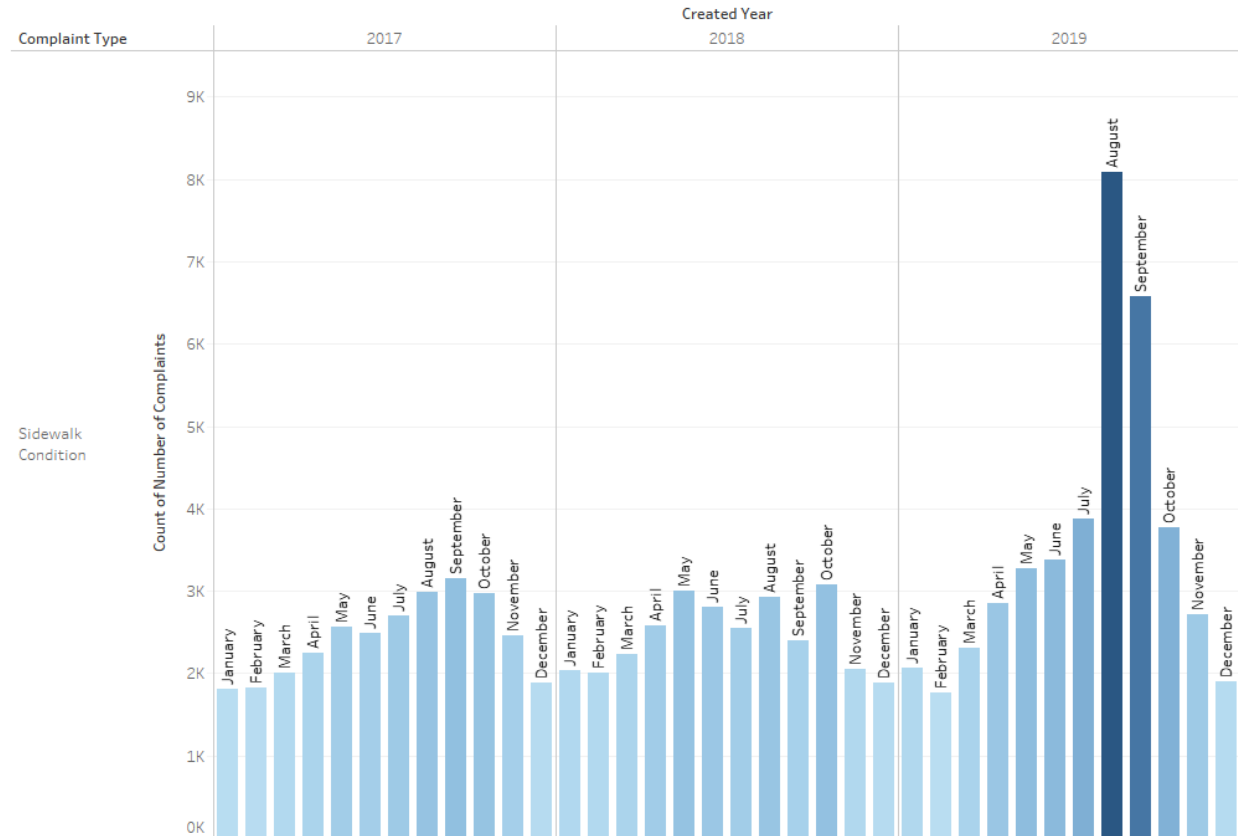
FROM COMPLAINT_BULK_FACTS ft

    INNER JOIN DATE_dim d1 ON ft.CREATED_DATE_DIM_ID = d1.date_dim_id
    INNER JOIN DATE_dim d2 ON ft.CLOSED_DATE_DIM_ID = d2.date_dim_id
    INNER JOIN time_dim t ON ft.created_time_dim_id = t.time_dim_id
    INNER JOIN agency_bulk_dim a ON ft.agency_dim_id = a.agency_dim_id
    INNER JOIN status_bulk_dim s ON ft.status_dim_id = s.status_dim_id
    INNER JOIN location_bulk_dim l ON ft.location_dim_id = l.location_dim_id
    INNER JOIN complaint_type_bulk_dim ct ON ft.complaint_type_dim_id =
    ct.complaint_type_dim_id

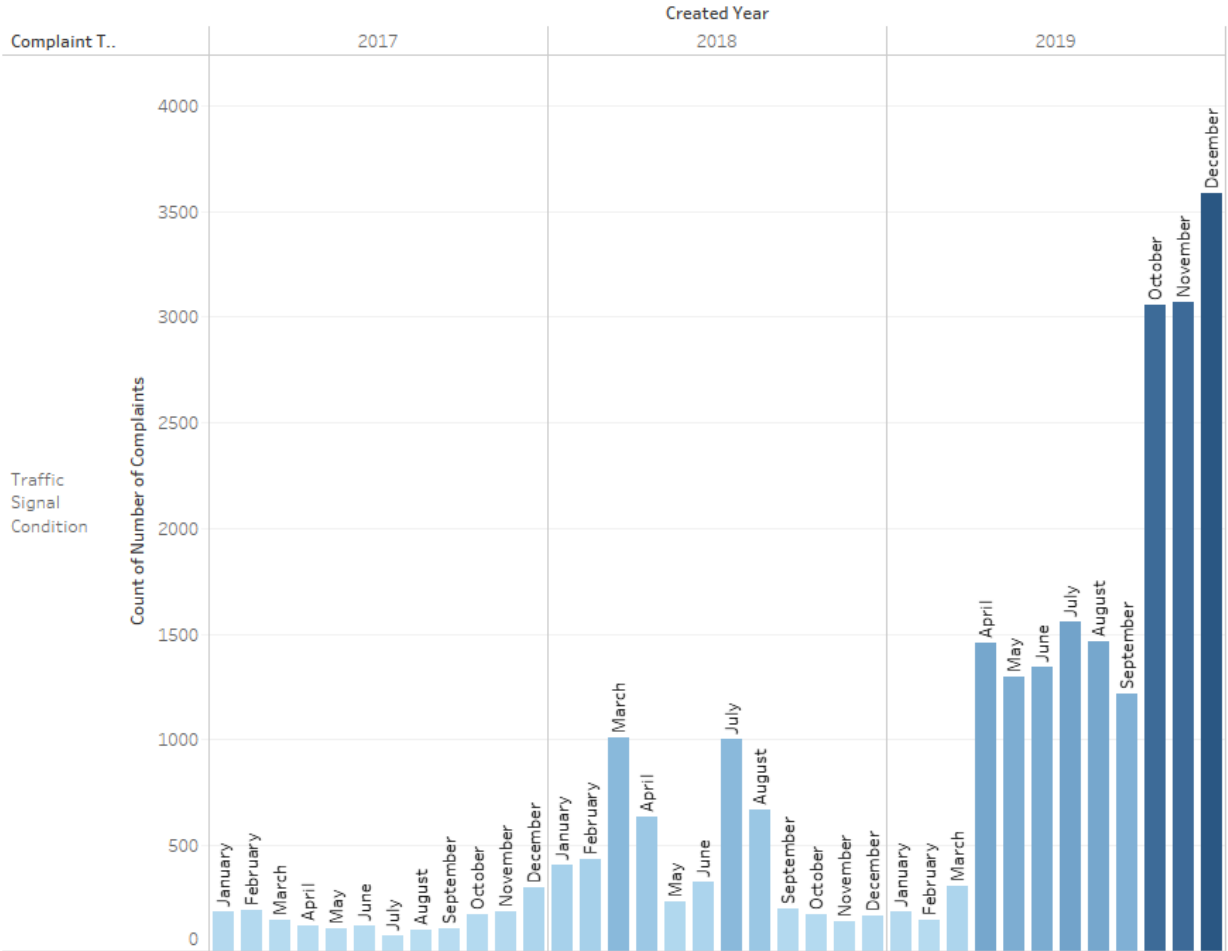
;
```

- Complaints Per Month Per Year

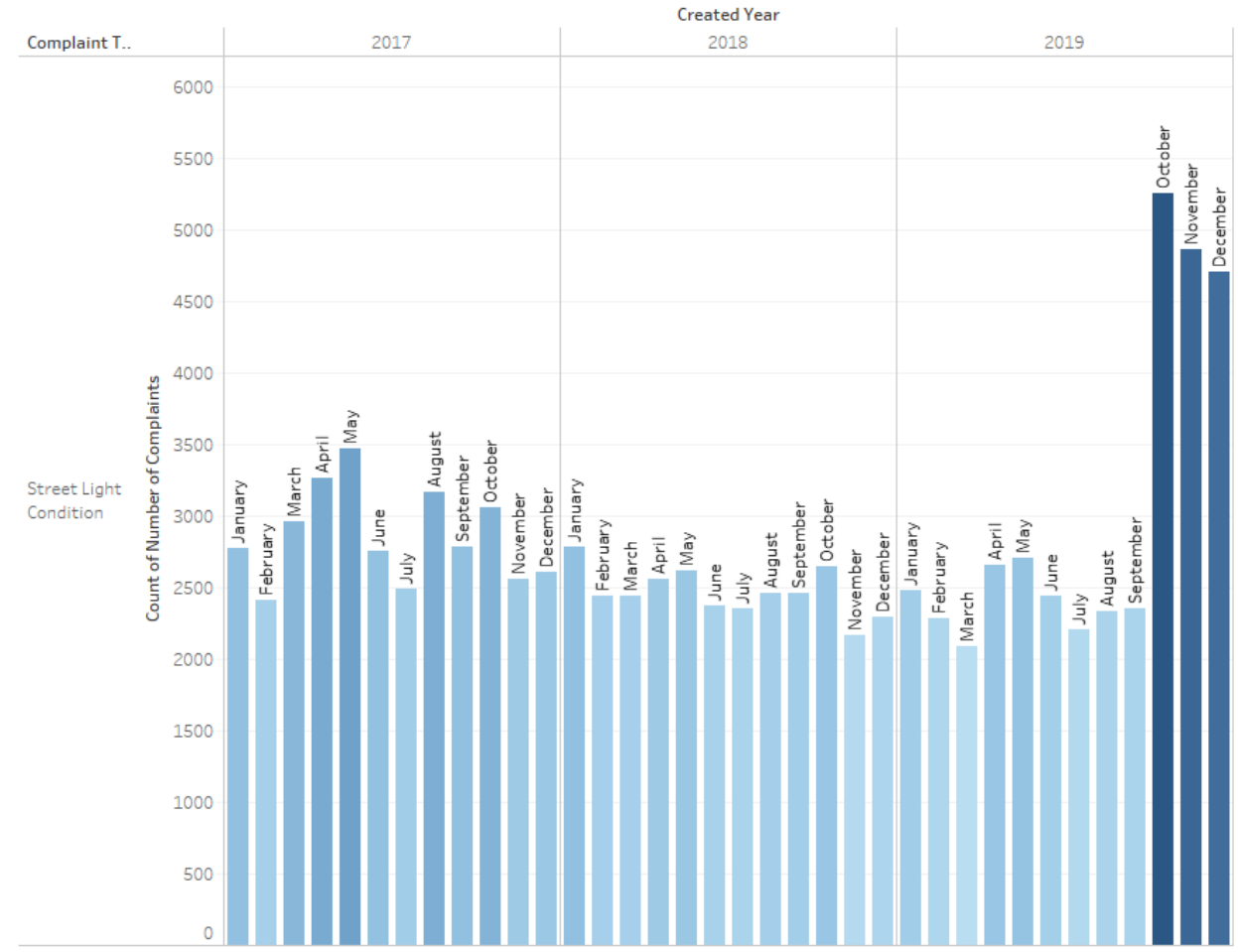
Sidewalk Condition Complaint Trend Over 3 years



Traffic Signal Complaint Trend Over 3 years

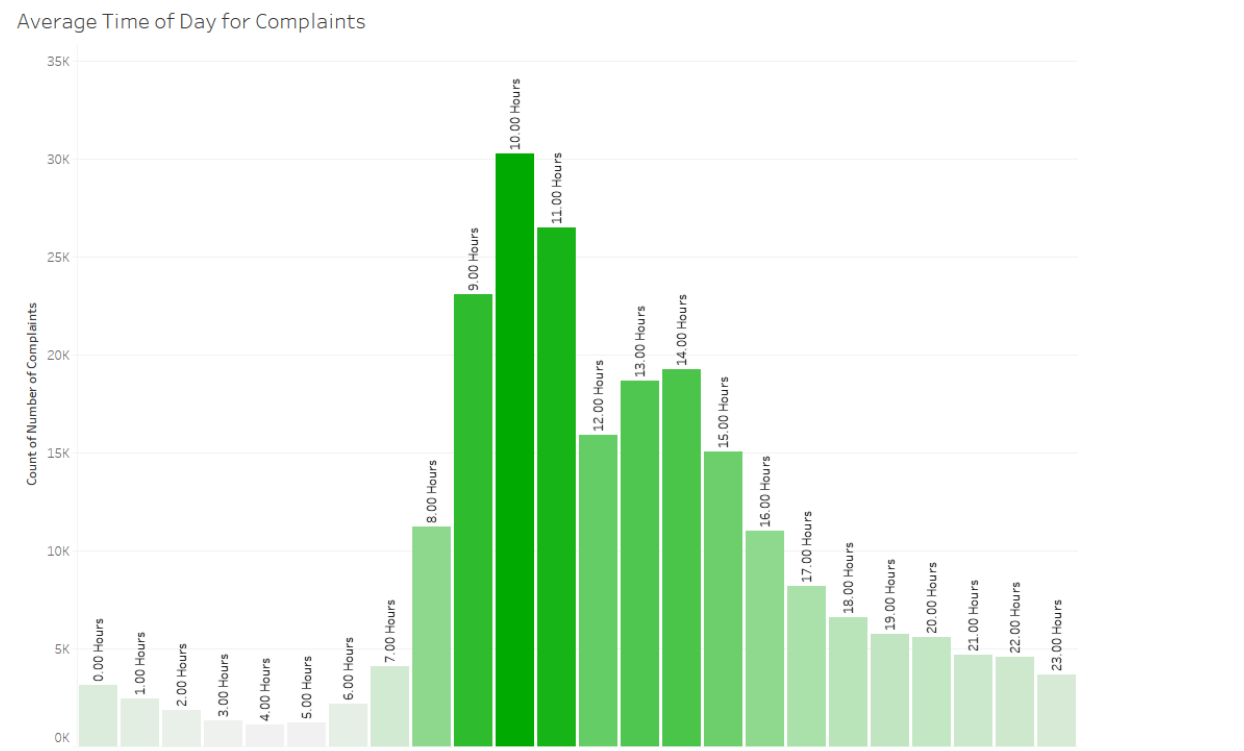


StreetLight Condition Complaint Trend Over 3 years



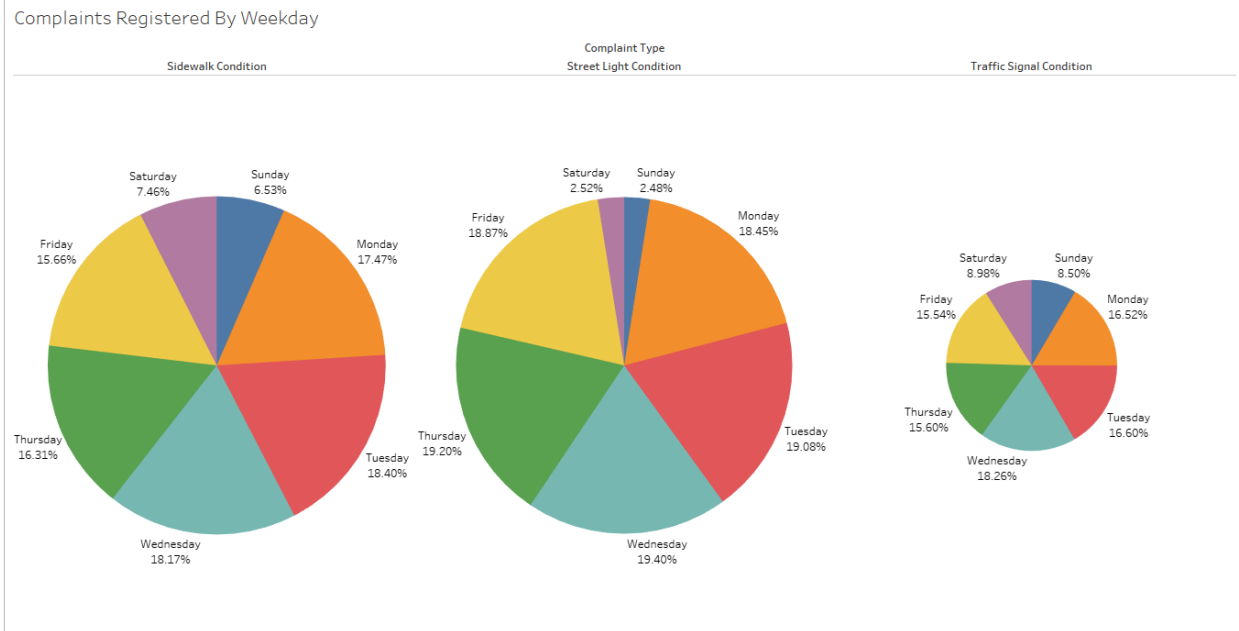
Based on the above data, the number of complaints for Sidewalk related issues has been steady over 2017 and 2018 but has seen a spike in the latter half of 2019. For traffic signal related issues, the number of complaints has increased each year with 2019 seeing the highest percentage spike in complaints. Streetlight related issues have remained steady across the 3 years except the last 3 months of 2019. Looking at all 3 graphs simultaneously, the latter half of 2019 should be investigated for factors which would have contributed to the rise in the number of cases.

- Average Time of day for complaints

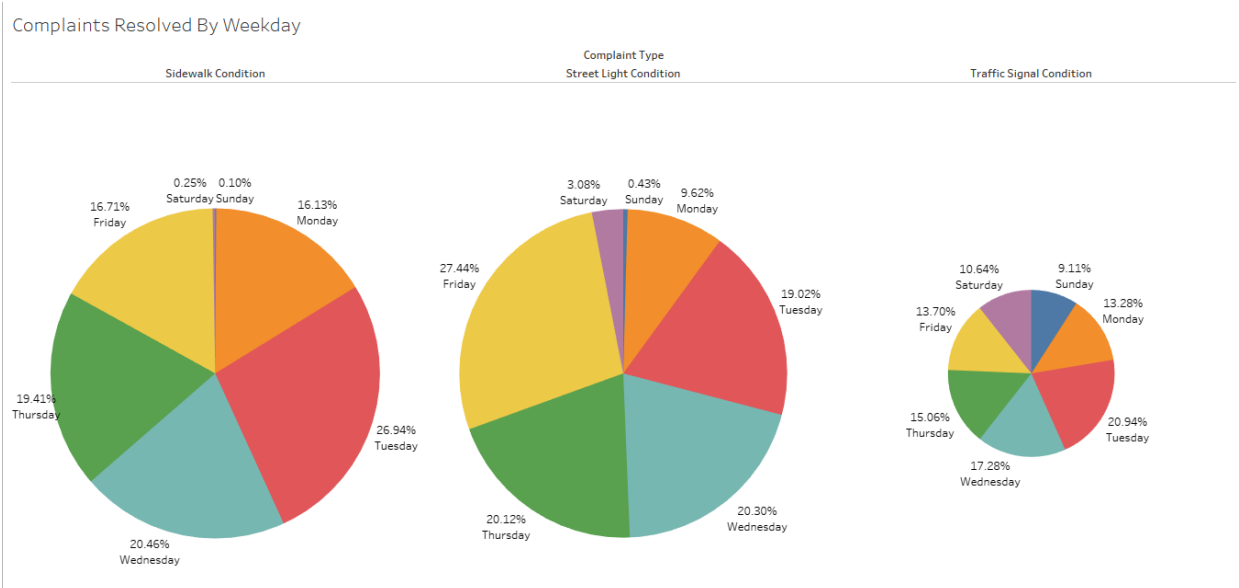


Based on the above data, the time from 9 AM to 12 PM sees the greatest number of complaints per hour of the day. This is to be expected as it is rush hour and the likelihood of bottlenecks in the transport system leading to complaints will remain high.

- Number of Complaints per weekday



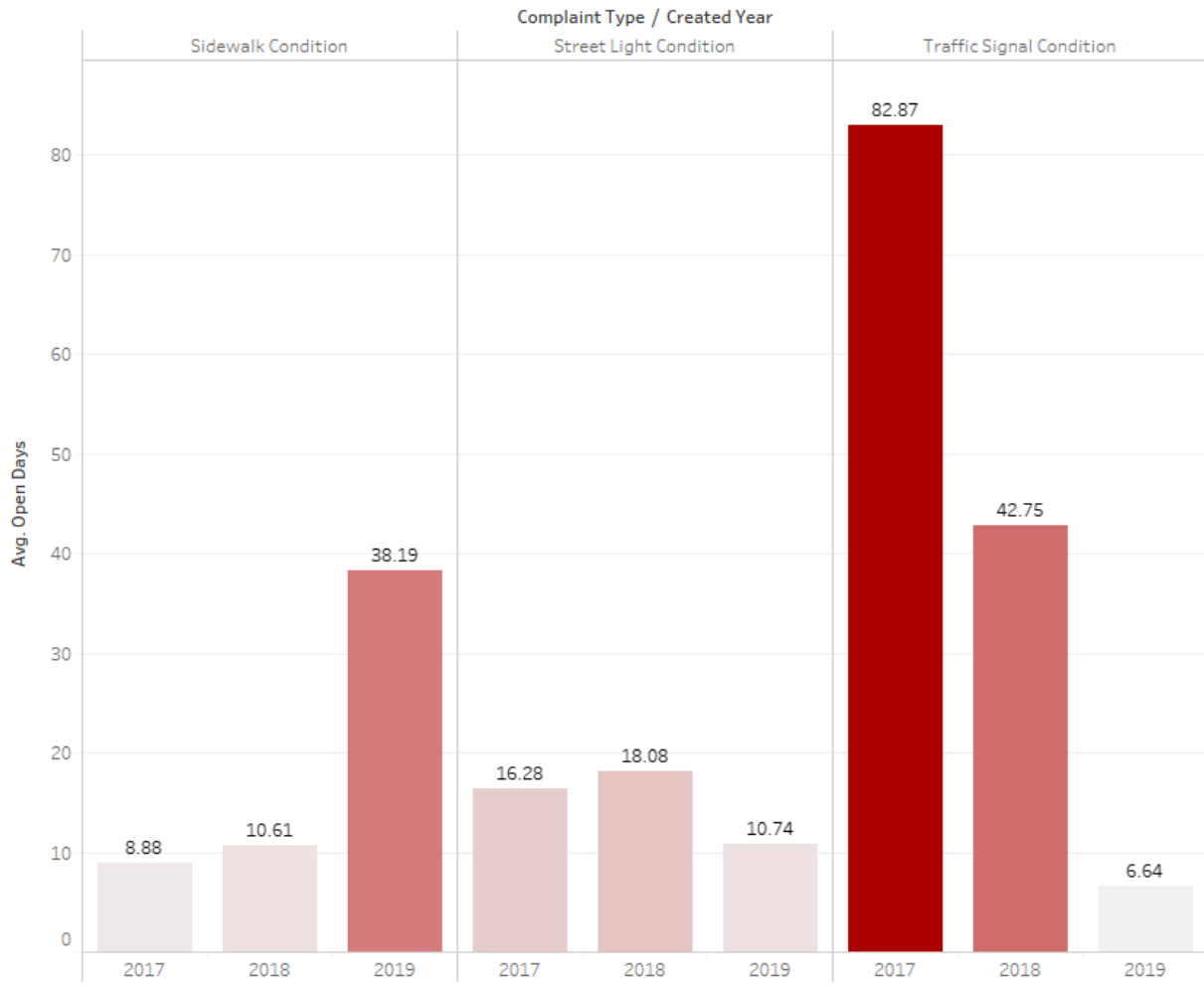
Based on the above data, most of the complaints are received during the weekdays with Streetlight related complaints the highest among the 3 complaint types under consideration. This is expected as the transport system will not operate at full capacity during the weekends and we observe no unusual patterns.



For the days on which complaints have been resolved, we see that a negligible amount of complaints are resolved on the weekends. This is to be expected with a majority of the workforce at home. Another interesting pattern to observe is that Fridays see the most cases resolved while Mondays the least (relatively).

- Average closure time for complaints

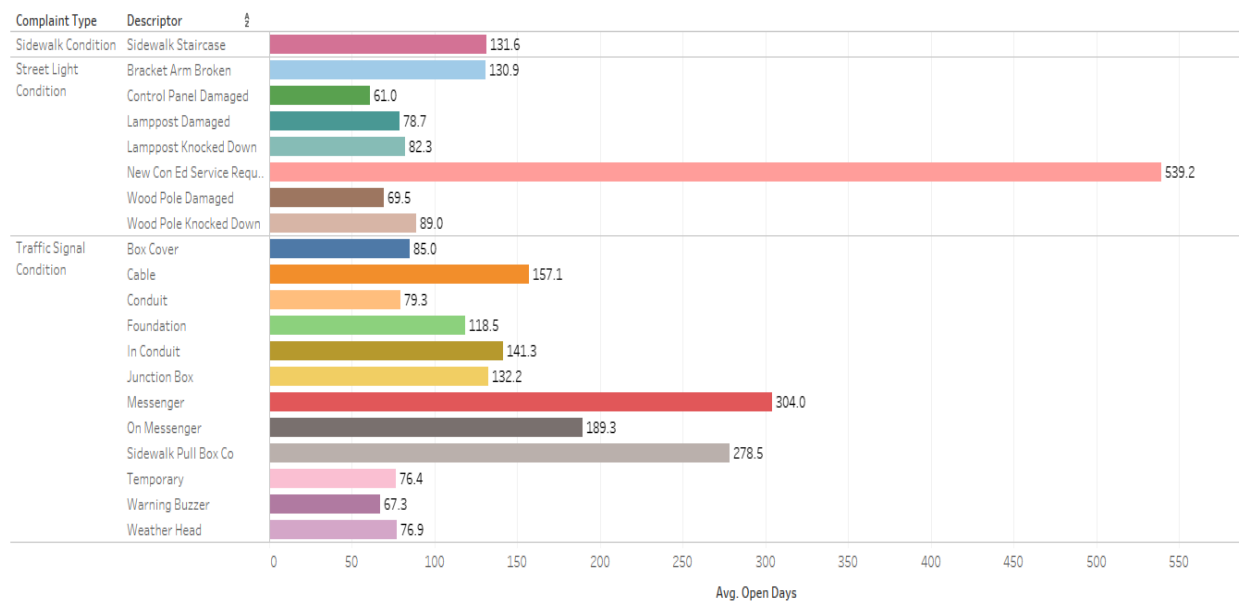
Average Days Open



For Sidewalk issues, the average closure time (time to resolve a complaint) has risen steadily from 2017-2019. This needs to be investigated. For streetlight condition, the average closure time has remained steady for 2017 and 2018 and reduced by almost 50 percent in 2019. For traffic signal complaints, the closure time was an alarming 82 days in 2017 but has recovered spectacularly in the subsequent years. The administrator responsible for this should be commended.

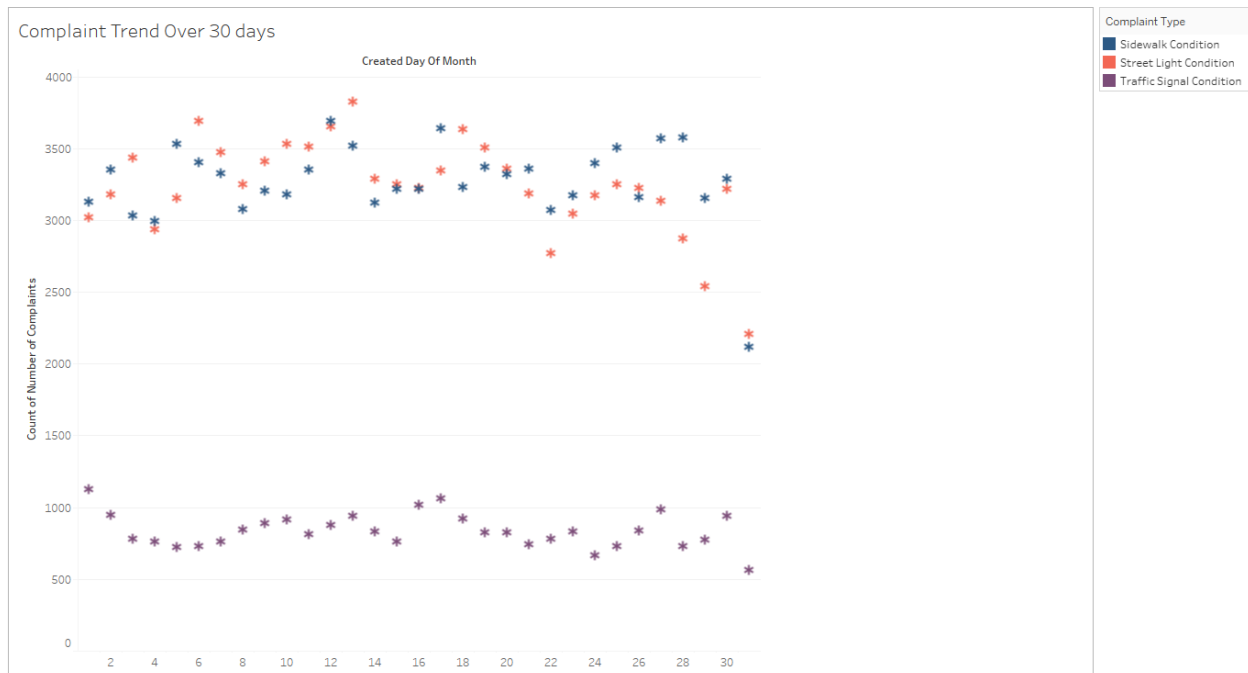
- Highest closure time by complaint types and description

Highest Average Closure Time Complaints Types (Top 20)



The above graph gives the complaint types which have taken the highest amount of time to be resolved. The winner of the prize is “New Con Ed Service Request”, a sub type of the Street Light related complaint type with an average of 540 days to resolve an issue. That is over a year and a half of an issue being open. This is a high priority matter which needs to be investigated and streamlined so that the tax dollars of the public can be saved.

- Complaints Per day of the Month



The above graph maps out the complaints received per day of the month for the 3 years. This confirms the pattern observed in the complaints per weekday graph with the complaints forming a wave pattern dipping every weekend and spiking during the week. Also, the last day of the month sees a drop in complaints across all 3 complaint types.

- Tools and Softwares Used:**

DBMS: Oracle Cloud Data Warehouse, SQL developer

ETL: Pentaho Data Integration

Data Extraction and Profiling: Python pandas and MS Excel

Analytics: Tableau

Meetings and Coordination: Slack and Zoom

We chose an Oracle Autonomous Data Warehouse to store our data. Then we used the Python Pandas library with Socrata API to extract our data into excel files which we then run through transformations in Pentaho Data Integration. Throughout this process, we used SQL developer to connect with our database to query the data and confirm if the required transformations and loads were taking place. Once the dimensions and fact tables were loaded, we created a view joining all the tables in Oracle and exported the data into tableau using an

excel file. In tableau, we finally created the visualizations according to the KPIs set at the start of the project and drew the necessary conclusions.

Our Group used Slack as our primary mode of communication and coordination. We were able to have a great turnaround time for issues and topics observed while working on the deliverables. In addition, we used Zoom to get into a conference meeting where we shared our workspace and worked together on the deliverables.

- **Challenges:**

The most challenging part of the project was loading the fact table with the final data. The write speeds for the data tanked after the first 15-20 minutes. There was more than 1 occasion where after running for an hour, an error was thrown, and the transformation was stopped. We were able to resolve the issue by splitting the data into chunks which would load in 15 to 20 minutes. Thus, we were able to load the data in the fact tables just in time and proceed with the project. The easiest part was the dimensional modelling and KPI creation. If we had to do it all over again, we would split the data in the ETL tool and not do it manually through excel as we had done initially. We had to re-do the dimensional loads by using 1 file and splitting the data in Pentaho as it easier to reuse the ETL script with new data and avoid the manual work each time.

- **Conclusion:**

It was a great experience working on the 311 data with Pentaho and Tableau. The data extraction and loading were a fun experience and gave a new perspective on databases and data analytics and we would love to work on more complex projects in the future.

- **Meeting Log:**

#	Meeting Date & Time	Attendees	Topic Discussed
1	8:15 PM, 24th June	Xinyi Li Immanuel Ryan	Team structure change Dimensional Model draft
2	11 am, 25th June	Xinyi Li Immanuel Ryan	Dimensional Model draft_v2
3	11 am, 30th June	Xinyi Li Immanuel Ryan	Final dimensional model discussion
4	11 am, 1st July	Xinyi Li Immanuel Ryan	ETL process discussion, task allocation, data profiling. ETL programming for dimensions started.
5	11 am, 4th July	Xinyi Li Immanuel Ryan	ETL programming for dimensions completed
6	8 pm, 4th July	Xinyi Li Immanuel Ryan	ETL fact table loading for the sample data started. Ran into issues, approximately 50% completed.
7	11 am, 6th July	Xinyi Li Immanuel Ryan	ETL fact table with sample data loaded.
8	8 pm, 7th July	Xinyi Li Immanuel Ryan	ETL with bulk data started. Ran into minor issues but were able to fully load the fact table. Added foreign key constraints.
9	11 am, 8th July	Xinyi Li Immanuel Ryan	Project Document update, screenshots attached and sent to Professor H.
10	11 am, 10th July	Xinyi Li Immanuel Ryan	Loading fact table with the bulk data
11	11 am, 11th July	Xinyi Li Immanuel Ryan	Tableau analysis
12	2 pm, 13th July	Xinyi Li Immanuel Ryan	Project Document final update, screenshots attached and sent to Professor H.

- **Reference List**

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