NYC Crime vs Film Permits

Comparison by time and location.

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

https://data.cityofnewyork.us/Social-Services/311-Noise-Complaints/p5f6-bkga

https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data

https://data.cityofnewyork.us/City-Government/Film-Permits/tg4x-b46p

We are the ACE. Inc, a consulting firm that has been hired by the NYC government to see if there is a relationship between crime data and movie permits. Depending on our findings, the City might want to adjust permit prices and increase promotion in the areas that have more potential for being a movie location but are overlooked due to crime. Knowing which areas have the most certain types of crime might help City officials address the situation in different ways: increase or decrease their presence in the areas at a particular time of day, adjust the number of officers and type of equipment necessary to fight those crimes. Our findings might also be used by many nonprofit organizations that work with local communities in the NYC area to more precisely pinpoint areas where help is most needed to help increase the safety and bring more business to the neighborhoods.

For this project we are focusing on connecting two data sets: “NYPD Complaint Data Current Year To Year”, called from now on crime data and “Film Permits” data set. Both sets are obtained from NYC Open Data website. We had a comma issue in our data sets, so we replaced them with forward slashes.

Crime data set has 462 k rows and 35 columns. It contains all crimes in NYC 5 boroughs almost dating back to 1911 - we will focus on one year - 2019.

Columns we think will be the most helpful: name of the borough in which the incident occurred, Exact date of occurrence, Type of offense: felony, misdemeanor, violation. Specific Incident. Suspect’s: Age Group, Race, Sex. Victim’s: Age Group, Race, Sex. Then location, mainly borough and zip code. Because this data does not have neighborhoods or zip code, these columns have to be added using other data sets. Using “311 Noise Complaint Data Set” zip codes were added first based on longitude and latitude. Then based on yet more more set “NYC Airbnb Open Data” , the neighborhoods were added to the set. We also added column names Country and State to populate our location dimension.

The second data set that we were connecting was “FilmPermits”. It is featuring all information necessary while obtaining film permits in NYC in the year 2020. Data set had 14 columns: EventID, EventType, StartDateTime, EndDateTime, EnteredOn, EventAgency, ParkingHeld, Borough, Community Board(s), PolicePrecinct, Category, SubcategoryName, Country, ZipCode.

This set also had to be altered; originally all zip codes were listed in one cell because one movie permit was issued for multiple zip codes often closely located. We had to expand the rows by duplicating all information to have each row but splitting the zip code cell into multiple rows. This data set had originally 66k rows but after expanding it has 98k. All the above work, for the crime data set and for the movie set were done in Python in Jupyter notebook using Pandas DataFrame. Some data had to be deleted due to errors or missing critical information, but the amount was minuscule. Here is a link to three files that contains code for this process:

Embellishing crime data by zip codes and neighborhoods:

https://github.com/molinexx/Portfolio/blob/master/crime%20data%20merging%20with%20zip%20codes%20and%20hoods.ipynb

Cleaning crime data:

https://github.com/molinexx/Portfolio/blob/master/crime%20cleaning%20data%20and%20%20creating%20mini%20csv's.ipynb

Expanding film data with copies of rows with single zip codes:

https://github.com/molinexx/Portfolio/blob/master/expand%20rows%20using%20explode()%20Film%20Permits%20Data%20Set.ipynb

Edited Crime Data:

https://drive.google.com/open?id=1iorgj7CryG-wOemavkOs5EeMnWGWG2tF

Edited Film Permit Data:

https://drive.google.com/open?id=1YvYyznLNqlPL9No0AeuQW7JnBAJsViP3

Our KPI's:

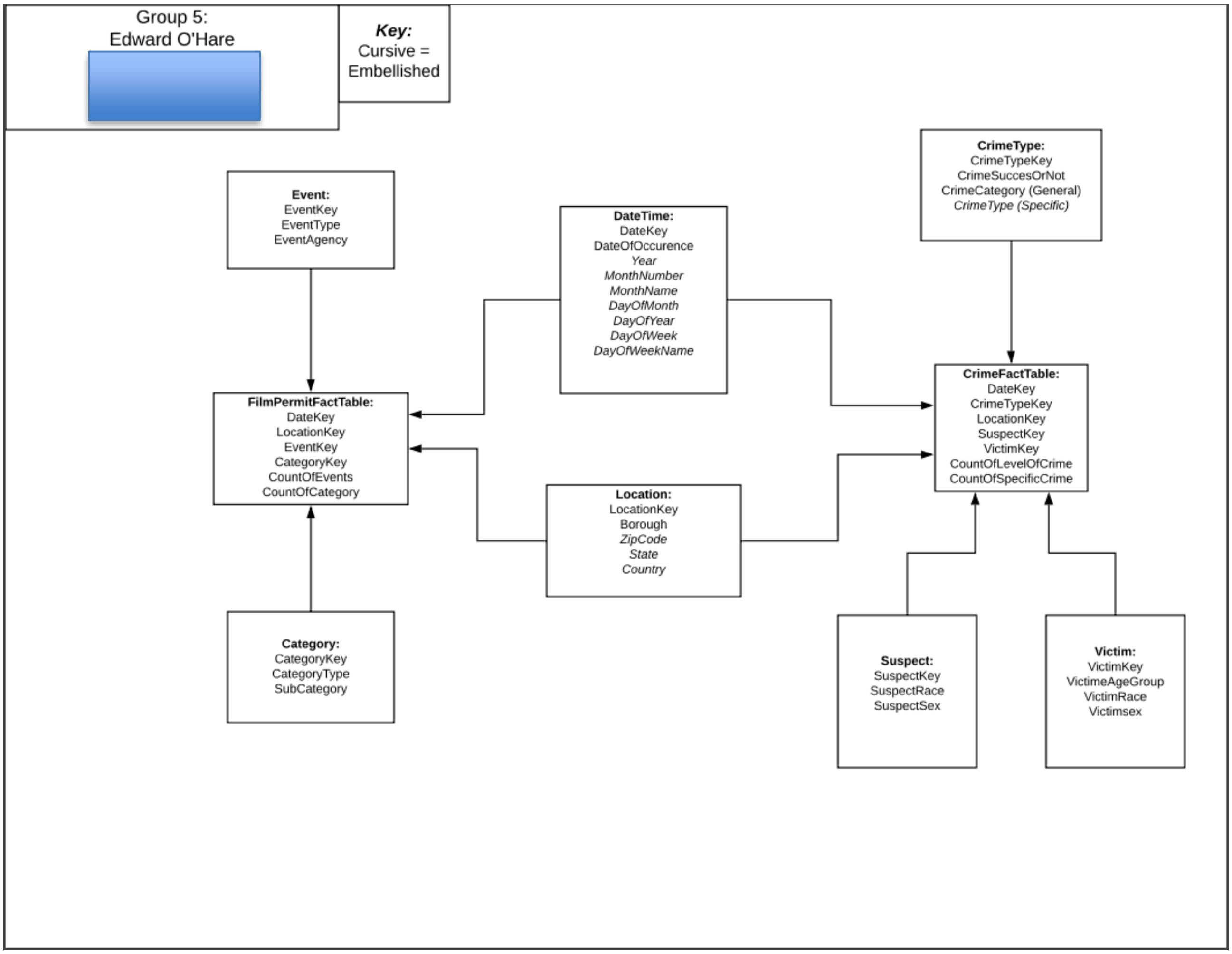
Count Of Events

Count Of Category

Count Of Level Of Crime

Count Of Specific Crime

DIMENSIONAL MODEL



There are two fact tables: Film Permit and Crime. Both are connected by location and date-time dimensions.

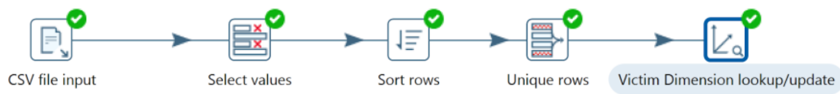
Transformations:

Using the Pentaho Data Integration tool we started creating our data warehouse. This is where we created kettle files for each dimension, and then later on we created the two fact tables.

General Transformation Rules: Take in the data, select correct values for this dimension, sort the data, find the unique rows and then populate the dimension within our database.

<u>CrimeTable: Suspect, Victim, Crime Type</u>

Victim Transformation: Pulled out the victims race, age group, and sex. Then assigned them a dim_id.



Execution Results

Logging | Execution History | Step Metrics | Performance Graph | Metrics | Preview data

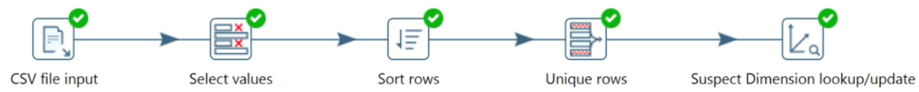
Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active
CSV file input	0	0	330517	330518	0	0	0	0	Finished
Select values	0	330517	330517	0	0	0	0	0	Finished
Sort rows	0	330517	330517	0	0	0	0	0	Finished
Unique rows	0	330517	117	0	0	0	0	0	Finished
Victim Dimension lookup/update	0	117	117	117	0	0	0	0	Finished

Examine preview data

Rows of step: victim_dim (100 rows)

#	VICTM_DIM_ID	VERSION	DATE_FROM	DATE_TO	VIC_AGE_GROUP	VIC_RACE	VIC_SEX
1	0	1	<null>	<null>	<null>	<null>	<null>
2	1	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	18-24	AMERICAN INDIAN/ALASKAN NATIVE	F
3	2	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	18-24	AMERICAN INDIAN/ALASKAN NATIVE	M
4	3	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	18-24	ASIAN / PACIFIC ISLANDER	D
5	4	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	18-24	ASIAN / PACIFIC ISLANDER	F
6	5	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	18-24	ASIAN / PACIFIC ISLANDER	M
7	6	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	18-24	BLACK	F
8	7	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	18-24	BLACK	M
9	8	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	18-24	BLACK HISPANIC	F
10	9	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	18-24	BLACK HISPANIC	M
11	10	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	18-24	UNKNOWN	D
12	11	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	18-24	UNKNOWN	F
13	12	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	18-24	UNKNOWN	M
14	13	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	18-24	WHITE	D
15	14	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	18-24	WHITE	E
16	15	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	18-24	WHITE	F
17	16	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	18-24	WHITE	M
18	17	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	18-24	WHITE HISPANIC	F
19	18	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	18-24	WHITE HISPANIC	M
20	19	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	25-44	AMERICAN INDIAN/ALASKAN NATIVE	D
21	20	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	25-44	AMERICAN INDIAN/ALASKAN NATIVE	F

Suspect Transformation: Pulled out the suspects race, age group, and sex. Then assigned them a dim_id.



Execution Results

Logging | Execution History | Step Metrics | Performance Graph | Metrics | Preview data

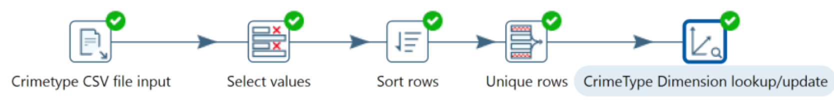
#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active
1	CSV file input	0	0	330517	330518	0	0	0	0	Finished
2	Select values	0	330517	330517	0	0	0	0	0	Finished
3	Sort rows	0	330517	330517	0	0	0	0	0	Finished
4	Unique rows	0	330517	21	0	0	0	0	0	Finished
5	Suspect Dimension lookup/update	0	21	21	21	0	0	0	0	Finished

Examine preview data

Rows of step: suspect_dim (22 rows)

#	SUSPECT_DIM_ID	VERSION	DATE_FROM	DATE_TO	SUSP_RACE	SUSP_SEX
1	0	1	<null>	<null>	<null>	<null>
2	1	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	AMERICAN INDIAN/ALASKAN NATIVE	F
3	2	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	AMERICAN INDIAN/ALASKAN NATIVE	M
4	3	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	AMERICAN INDIAN/ALASKAN NATIVE	U
5	4	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	ASIAN / PACIFIC ISLANDER	F
6	5	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	ASIAN / PACIFIC ISLANDER	M
7	6	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	ASIAN / PACIFIC ISLANDER	U
8	7	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	BLACK	F
9	8	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	BLACK	M
10	9	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	BLACK	U
11	10	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	BLACK HISPANIC	F
12	11	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	BLACK HISPANIC	M
13	12	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	BLACK HISPANIC	U
14	13	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	UNKNOWN	F
15	14	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	UNKNOWN	M
16	15	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	UNKNOWN	U
17	16	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	WHITE	F
18	17	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	WHITE	M
19	18	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	WHITE	U
20	19	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	WHITE HISPANIC	F
21	20	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	WHITE HISPANIC	M
22	21	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	WHITE HISPANIC	U

Crime Type Transformation: Pulled out the level of offense, and the exact offense description. Then assigned it a dim_id.



Execution Results

Logging | Execution History | Step Metrics | Performance Graph | Metrics | Preview data

#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active
1	Crimetype CSV file input	0	0	330517	330518	0	0	0	0	Finished
2	Select values	0	330517	330517	0	0	0	0	0	Finished
3	Sort rows	0	330517	330517	0	0	0	0	0	Finished
4	Unique rows	0	330517	117	0	0	0	0	0	Finished
5	CrimeType Dimension lookup/update	0	117	117	117	0	0	0	0	Finished

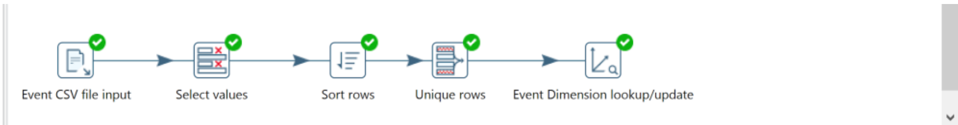
Examine preview data

Rows of step: crimetype_dim (100 rows)

#	CRIMETYPE_DIM_ID	VERSION	DATE_FROM	DATE_TO	CRM_ATPT_CPTD_CD	OFNS_DESC	LAW_CAT_CD
1	0	1	<null>	<null>	<null>	<null>	<null>
2	1	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	ATTEMPTED	ADMINISTRATIVE CODE	MISDEMEANOR
3	2	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	ATTEMPTED	ARSON	FELONY
4	3	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	ATTEMPTED	ASSAULT 3 & RELATED OFFENSES	MISDEMEANOR
5	4	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	ATTEMPTED	BURGLAR'S TOOLS	MISDEMEANOR
6	5	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	ATTEMPTED	BURGLARY	FELONY
7	6	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	ATTEMPTED	CRIMINAL MISCHIEF & RELATED OF	FELONY
8	7	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	ATTEMPTED	CRIMINAL MISCHIEF & RELATED OF	MISDEMEANOR
9	8	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	ATTEMPTED	CRIMINAL TRESPASS	MISDEMEANOR
10	9	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	ATTEMPTED	DANGEROUS DRUGS	FELONY
11	10	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	ATTEMPTED	DANGEROUS DRUGS	MISDEMEANOR
12	11	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	ATTEMPTED	DANGEROUS WEAPONS	FELONY
13	12	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	ATTEMPTED	DANGEROUS WEAPONS	MISDEMEANOR
14	13	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	ATTEMPTED	ESCAPE 3	MISDEMEANOR
15	14	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	ATTEMPTED	FELONY ASSAULT	FELONY
16	15	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	ATTEMPTED	FORGERY	FELONY
17	16	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	ATTEMPTED	FRAUDS	MISDEMEANOR
18	17	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	ATTEMPTED	FRAUDULENT ACCOSTING	MISDEMEANOR
19	18	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	ATTEMPTED	GRAND LARCENY	FELONY
20	19	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	ATTEMPTED	GRAND LARCENY OF MOTOR VEHICLE	FELONY
21	20	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	ATTEMPTED	HARRASSMENT 2	VIOLATION

Film permit table: Event, Category

Event Type Transformation: Pulled out the event type and the event agency. Then assigned it a dim_id.



Execution Results

Logging | Execution History | Step Metrics | Performance Graph | Metrics | Preview data

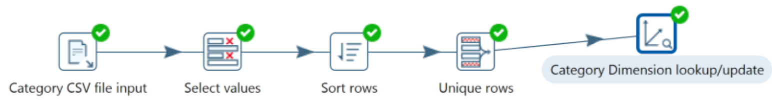
#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Tin
1	Event CSV file input	0	0	97582	97583	0	0	0	0	Finished	0.
2	Select values	0	97582	97582	0	0	0	0	0	Finished	0.
3	Sort rows	0	97582	97582	0	0	0	0	0	Finished	0.
4	Unique rows	0	97582	4	0	0	0	0	0	Finished	0.
5	Event Dimension lookup/update	0	4	4	4	0	0	0	0	Finished	0.

Examine preview data

Rows of step: event_dim (5 rows)

#	EVENT_DIM_ID	VERSION	DATE_FROM	DATE_TO	EVENTTYPE	EVENTAGENCY
1	0	1	<null>	<null>	<null>	<null>
2	1	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	DCAS Prep/Shoot/Wrap Permit	Mayor's Office of Film, Theatre & Broadcasting
3	2	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	Rigging Permit	Mayor's Office of Film, Theatre & Broadcasting
4	3	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	Shooting Permit	Mayor's Office of Film, Theatre & Broadcasting
5	4	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	Theater Load in and Load Outs	Mayor's Office of Film, Theatre & Broadcasting

Category Type Transformation: Pulled out the category type and the sub category. Then assigned it a dim_id.



Execution Results

Logging | Execution History | Step Metrics | Performance Graph | Metrics | Preview data

#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active
1	Category CSV file input	0	0	97582	97583	0	0	0	0	Finished
2	Select values	0	97582	97582	0	0	0	0	0	Finished
3	Sort rows	0	97582	97582	0	0	0	0	0	Finished
4	Unique rows	0	97582	39	0	0	0	0	0	Finished
5	category Dimension lookup/update	0	39	39	39	39	0	0	0	Finished

Examine preview data

Rows of step: CATEGORY_DIM (40 rows)

#	CATEGORY_DIM_ID	VERSION	DATE_FROM	DATE_TO	SUBCATEGORYNAME	CATEGORY
1	0	1	<null>	<null>	<null>	<null>
2	1	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	Cable-daily	Television
3	2	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	Cable-episodic	Television
4	3	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	Cable-other	Television
5	4	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	Children	Television
6	5	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	Commercial	Commercial
7	6	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	Commercial	Still Photography
8	7	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	Daytime soap	Television
9	8	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	Episodic series	Television
10	9	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	Feature	Film
11	10	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	Game show	Television
12	11	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	Independent Artist	Music Video
13	12	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	Industrial/Corporate	Commercial
14	13	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	Made for TV/mini-series	Television
15	14	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	Magazine Show	Television
16	15	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	Morning Show	Television
17	16	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	News	Television
18	17	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	Not Applicable	Commercial
19	18	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	Not Applicable	Documentary
20	19	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	Not Applicable	Film

<u>Date and Location Dimensions</u>

Date Transformation: Created dates for the year 2019, then embellished the date dimension to give an analyst more information to query on, and then assigned it a dim_id. For the embellishment we added: Day of Year, Month, Year, Quarter, Month Name, Day of Week Name, Day of Week, and Day of Month.

Execution Results

Logging | Execution History | Step Metrics | Performance Graph | Metrics | Preview data

#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time
1	Generate rows	0	0	1100	0	0	0	0	0	Finished	0.0s
2	Add date sequence	0	1100	1100	0	0	0	0	0	Finished	0.0s
3	CalculateDates	0	1100	1100	0	0	0	0	0	Finished	0.0s
4	Select values	0	1100	1100	0	0	0	0	0	Finished	0.0s
5	Calculate Additional Fields	0	1100	1100	0	0	0	0	0	Finished	0.1s
6	Date Dimension Update	0	1100	1100	1100	0	0	0	0	Finished	0.5s

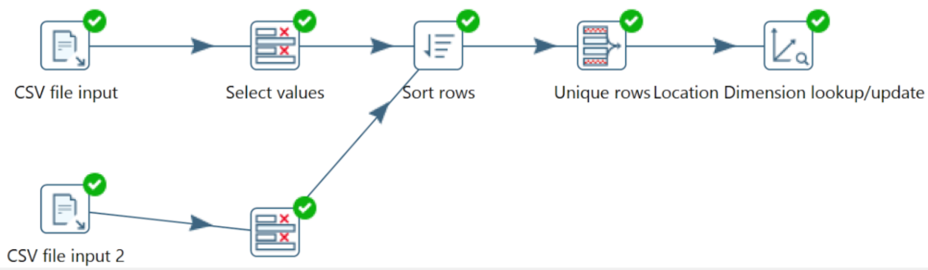
Examine preview data

Rows of step: date_dim3 (100 rows)

#	DATE_DIM_ID	VERSION	DATE_FROM	DATE_TO	DATE	DAY_OF_YEAR	MONTH	YEAR	QUARTER	MONTH_NAME
1	0	1	<null>	<null>	<null>	<null>	<null>	<null>	<null>	<null>
2	1	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	2019/01/02 00:00:00.000000000	2.0	1.0	2019.0	1.0	January
3	2	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	2019/01/03 00:00:00.000000000	3.0	1.0	2019.0	1.0	January
4	3	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	2019/01/04 00:00:00.000000000	4.0	1.0	2019.0	1.0	January
5	4	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	2019/01/05 00:00:00.000000000	5.0	1.0	2019.0	1.0	January
6	5	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	2019/01/06 00:00:00.000000000	6.0	1.0	2019.0	1.0	January
7	6	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	2019/01/07 00:00:00.000000000	7.0	1.0	2019.0	1.0	January
8	7	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	2019/01/08 00:00:00.000000000	8.0	1.0	2019.0	1.0	January
9	8	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	2019/01/09 00:00:00.000000000	9.0	1.0	2019.0	1.0	January
10	9	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	2019/01/10 00:00:00.000000000	10.0	1.0	2019.0	1.0	January
11	10	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	2019/01/11 00:00:00.000000000	11.0	1.0	2019.0	1.0	January
12	11	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	2019/01/12 00:00:00.000000000	12.0	1.0	2019.0	1.0	January
13	12	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	2019/01/13 00:00:00.000000000	13.0	1.0	2019.0	1.0	January
14	13	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	2019/01/14 00:00:00.000000000	14.0	1.0	2019.0	1.0	January
15	14	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	2019/01/15 00:00:00.000000000	15.0	1.0	2019.0	1.0	January
16	15	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	2019/01/16 00:00:00.000000000	16.0	1.0	2019.0	1.0	January
17	16	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	2019/01/17 00:00:00.000000000	17.0	1.0	2019.0	1.0	January
18	17	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	2019/01/18 00:00:00.000000000	18.0	1.0	2019.0	1.0	January
19	18	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	2019/01/19 00:00:00.000000000	19.0	1.0	2019.0	1.0	January
20	19	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	2019/01/20 00:00:00.000000000	20.0	1.0	2019.0	1.0	January
21	20	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	2019/01/21 00:00:00.000000000	21.0	1.0	2019.0	1.0	January
22	21	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	2019/01/22 00:00:00.000000000	22.0	1.0	2019.0	1.0	January
23	22	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	2019/01/23 00:00:00.000000000	23.0	1.0	2019.0	1.0	January

DAY_OF_WEEK_NAME	DAY_OF_WEEK	DAY_OF_MONTH
<null>	<null>	<null>
Wednesday	4.0	2.0
Thursday	5.0	3.0
Friday	6.0	4.0
Saturday	7.0	5.0
Sunday	1.0	6.0
Monday	2.0	7.0
Tuesday	3.0	8.0
Wednesday	4.0	9.0
Thursday	5.0	10.0
Friday	6.0	11.0
Saturday	7.0	12.0
Sunday	1.0	13.0
Monday	2.0	14.0
Tuesday	3.0	15.0
Wednesday	4.0	16.0
Thursday	5.0	17.0
Friday	6.0	18.0
Saturday	7.0	19.0
Sunday	1.0	20.0
Monday	2.0	21.0
Tuesday	3.0	22.0
Wednesday	4.0	23.0
Thursday	5.0	24.0
Friday	6.0	25.0

Location Transformation: Had to input the location from our two csv's, select the correct values from each, and then add this to our database. We mainly focused our attention on the borough, as zip code could get a little overwhelming.



Execution Results

Logging | Execution History | Step Metrics | Performance Graph | Metrics | Preview data

#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active
1	CSV file input 2	0	0	330517	330518	0	0	0	0	Finished
2	CSV file input	0	0	97582	97583	0	0	0	0	Finished
3	Select values 2	0	330517	330517	0	0	0	0	0	Finished
4	Select values	0	97582	97582	0	0	0	0	0	Finished
5	Sort rows	0	428099	428099	0	0	0	0	0	Finished
6	Unique rows	0	428099	967	0	0	0	0	0	Finished
7	Location Dimension lookup/update	0	967	967	967	0	0	0	0	Finished

Examine preview data

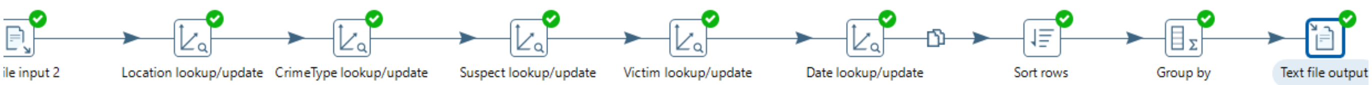
Rows of step: location_dim (100 rows)

#	LOCATION_DIM	VERSION	DATE_FROM	DATE_TO	ZIPCODE	BOROUGH	STATE	COUNTRY
1	0	1	<null>	<null>	<null>	<null>	<null>	<null>
2	1	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	10001	Bronx	New York	United States of America
3	2	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	10002	Bronx	New York	United States of America
4	3	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	10003	Bronx	New York	United States of America
5	4	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	10004	Bronx	New York	United States of America
6	5	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	10005	Bronx	New York	United States of America
7	6	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	10006	Bronx	New York	United States of America
8	7	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	10007	Bronx	New York	United States of America
9	8	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	10009	Bronx	New York	United States of America
10	9	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	10010	Bronx	New York	United States of America
11	10	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	10011	Bronx	New York	United States of America
12	11	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	10012	Bronx	New York	United States of America
13	12	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	10013	Bronx	New York	United States of America
14	13	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	10014	Bronx	New York	United States of America
15	14	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	10016	Bronx	New York	United States of America
16	15	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	10017	Bronx	New York	United States of America
17	16	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	10018	Bronx	New York	United States of America
18	17	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	10019	Bronx	New York	United States of America
19	18	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	10020	Bronx	New York	United States of America
20	19	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	10022	Bronx	New York	United States of America
21	20	1	1900/01/01 00:00:00.000000000	2199/12/31 23:59:59.999000000	10023	Bronx	New York	United States of America

<u>Fact Tables:</u>

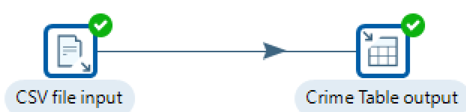
Crime Fact Table

We took the different dimensions we individually made for the crime data set, put them together, sorted them, and downloaded them to a csv. From that csv file, we uploaded that data directly into the database.



Logging | Execution History | Step Metrics | Performance Graph | Metrics | Preview data

#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	input/output
1	CSV file input 2	0	0	330517	330518	0	0	0	0	Finished	2h 8mn 56s	43	-
2	Location lookup/update	0	330517	330517	330517	0	0	0	0	Finished	2h 12mn 51s	41	-
3	CrimeType lookup/update	0	330517	330517	330517	0	0	0	0	Finished	2h 16mn 55s	40	-
4	Suspect lookup/update	0	330517	330517	330517	0	0	0	0	Finished	2h 20mn 44s	39	-
5	Victim lookup/update	0	330517	330517	330517	0	0	0	0	Finished	2h 20mn 47s	39	-
6	Date lookup/update	0	330517	330517	330517	0	0	0	0	Finished	2h 20mn 49s	39	-
7	Sort rows	0	330517	330517	0	0	0	0	0	Finished	2h 20mn 51s	39	-
8	Group by	0	330517	330517	0	0	0	0	0	Finished	2h 20mn 51s	39	-
9	Text file output	0	330517	330517	0	330518	0	0	0	Finished	2h 20mn 51s	39	-



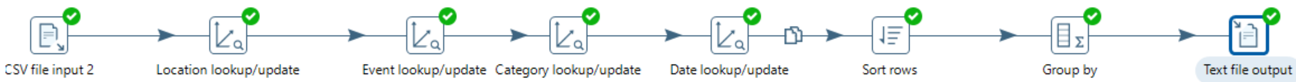
Logging | Execution History | Step Metrics | Performance Graph | Metrics | Preview data

#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	input/output
1	CSV file input	0	0	330517	330518	0	0	0	0	Finished	49.7s	6,657	-
2	Crime Table output	0	330517	330517	0	330517	0	0	0	Finished	51.5s	6,418	-

LOCATION_DIM	CRIMETYPE_DIM	SUSPECT_DIM	VICTIM_DIM	DATE_DIM_ID	CountOfLevelOfCrime	CountOfSpecific
1	1	1	1	0	4	4
1	1	1	1	0	4	4
1	1	1	1	0	4	4
1	1	1	1	0	4	4
1	1	1	1	1	2	2
1	1	1	1	1	2	2
1	1	1	1	2	3	3
1	1	1	1	2	3	3
1	1	1	1	2	3	3
1	1	1	1	3	1	1
1	1	1	1	4	6	6
1	1	1	1	4	6	6
1	1	1	1	4	6	6
1	1	1	1	4	6	6
1	1	1	1	4	6	6
1	1	1	1	4	6	6
1	1	1	1	5	1	1
1	1	1	1	6	1	1
1	1	1	1	7	3	3

Film Permit Fact Table

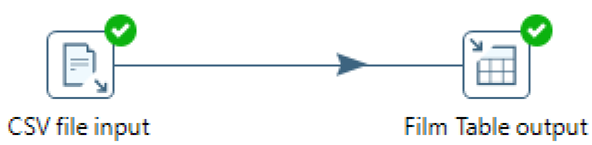
We took the different dimensions we individually made for the film permit data set, put them together, sorted them, and downloaded them to a csv. From that csv file, we uploaded that data directly into the database.



Execution Results

Logging | Execution History | Step Metrics | Performance Graph | Metrics | Preview data

#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	input/output
1	CSV file input 2	0	0	97582	97583	0	0	0	0	Finished	36mn 52s	44	-
2	Location lookup/update	0	97582	97582	97582	0	0	0	0	Finished	41mn 59s	39	-
3	Event lookup/update	0	97582	97582	97582	0	0	0	0	Finished	42mn 3s	39	-
4	Category lookup/update	0	97582	97582	97582	0	0	0	0	Finished	43mn 33s	37	-
5	Date lookup/update	0	97582	97582	97582	0	0	0	0	Finished	44mn 23s	37	-
6	Sort rows	0	97582	97582	0	0	0	0	0	Finished	44mn 23s	37	-
7	Group by	0	97582	97582	0	0	0	0	0	Finished	44mn 23s	37	-
8	Text file output	0	97582	97582	0	97583	0	0	0	Finished	44mn 23s	37	-



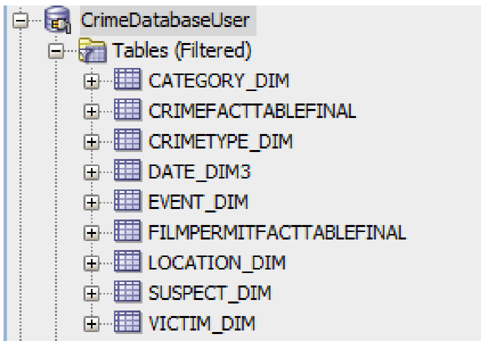
Execution Results

Logging | Execution History | Step Metrics | Performance Graph | Metrics | Preview data

#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	input/output
1	CSV file input	0	0	97582	97583	0	0	0	0	Finished	6.0s	16,237	-
2	Film Table output	0	97582	97582	0	97582	0	0	0	Finished	6.8s	14,291	-

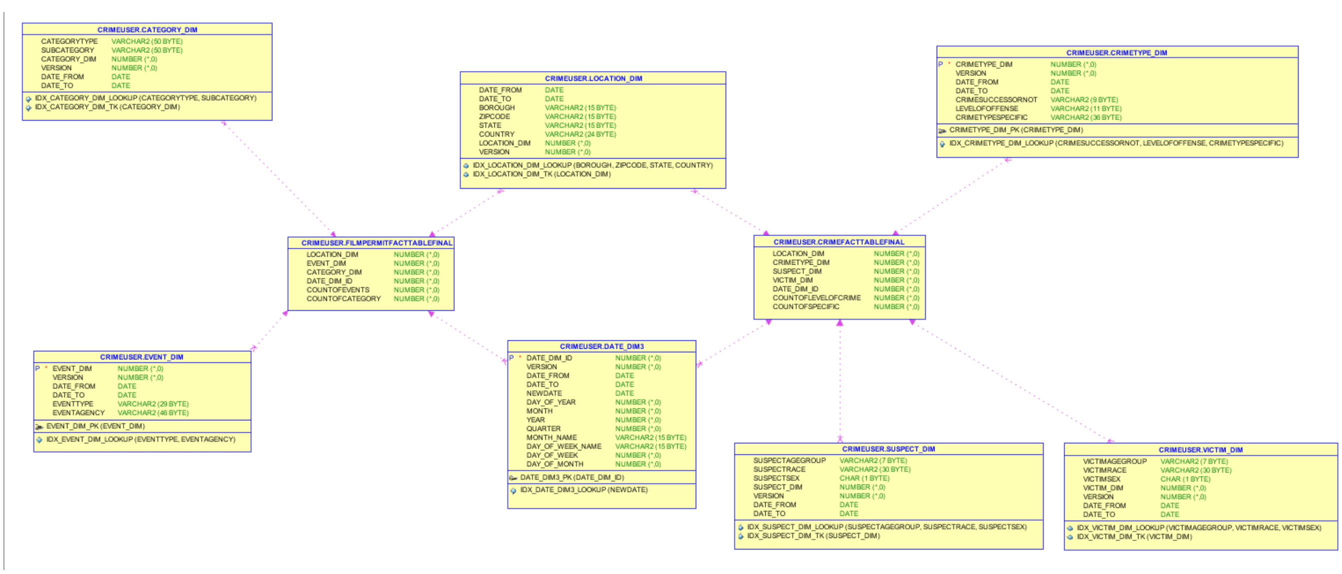
	A	B	C	D	E	F
1	LOCATION_DIM	event_dim	CATEGORY_DIM	DATE_DIM_ID	CountOfEvents	CountOfCategory
2	1	1	1	0	2	2
3	1	1	1	0	2	2
4	1	1	9	0	13	13
5	1	1	9	0	13	13
6	1	1	9	0	13	13
7	1	1	9	0	13	13
8	1	1	9	0	13	13
9	1	1	9	0	13	13
10	1	1	9	0	13	13
11	1	1	9	0	13	13
12	1	1	9	0	13	13
13	1	1	9	0	13	13
14	1	1	9	0	13	13
15	1	1	9	0	13	13
16	1	1	9	0	13	13
17	1	1	9	94	1	1
18	1	1	16	0	3	3
19	1	1	16	0	3	3
20	1	1	16	0	3	3

Data in our Oracle Database:



After assembling the tables in Pentaho, we created the link to our oracle database and uploaded them. Oracle server makes our data available for us to work with in SQL Developer using the wallet zip folder. This is where using SQL we created relationships between tables. Below you can see a physical diagram that was created after this step.

Physical diagram: SQL Developer with Foreign Keys

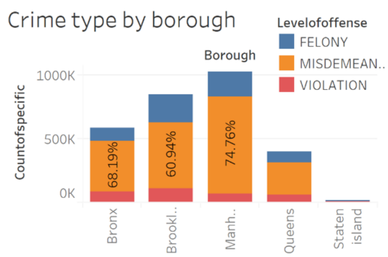


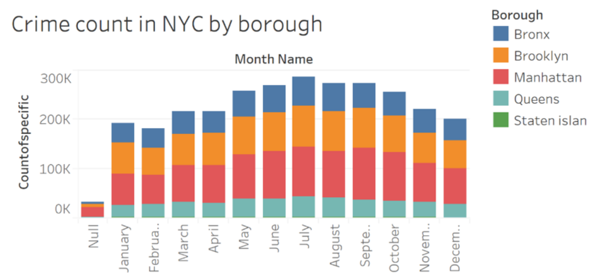
Our Final Technical Schema:

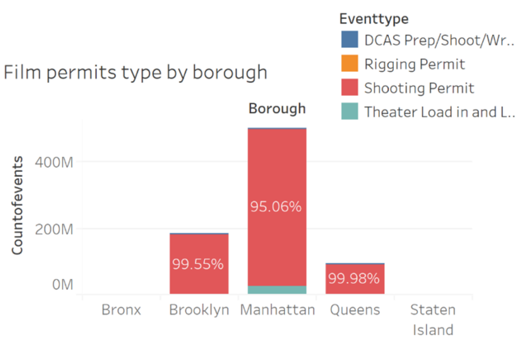
Summarizing, we used Pentaho for creating our dimensions, Oracle database for storing our data, and Tableau for creating our dashboard. We were able to connect Tableau to the Oracle server using your tutorials, where we created a schema for data visualization: for example this is the schema for Crime Fact Table.

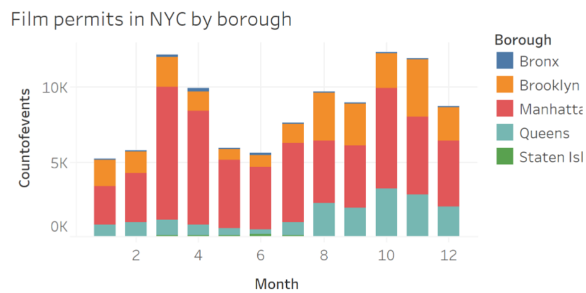
We used KPI such as CountOfEvents, CountOfCategory, CountOfLevelOfCrime, CountOfSpecificCrime to create below dashboard:

Dashboard:









From our dashboard we can easily compare the amount of crime and film permits by borough and over the year of 2019. May, June, and July are some of the most popular months for crime, but some of the least popular months for film permits. We also can see there is almost no filming done in Staten Island, but there is a very small amount of crime there as well. It may be time for more filming to take place in Staten Island.

Narrative Conclusion:

We think that even though there seems to be some slight evidence that less film permits are equal to more crime, there would have to be a lot more analysis to make any concrete predictions or recommendations. We think adding weather could be an interesting added factor into this analysis. Meaning, on extra hot days is there more crime and less film permits and vice versa?