CTA CRIME PREDICTION

First name	Last Name	
Shubham	Madke	A20445594
Yash	Agrawal	A20442647
Tanmay	Gaikwad	A20416375
Deeptanshu	Jariwala	A20448079
Megha	Shrivastava	A20450886

1. Introduction

- In this the CTA Crime prediction is done. The first of our project is the data preparation and the provision of the data. In this we have implemented the ETL that is extracted and transformed, In transformation we have executed feature selection in that we have fetched the best attributes that we will use in data visualization, secondly we have data include the incident category, in which we have replaced the names of the incidents to its severity.
- In the second transformation we have performed
- In the next step the visualization and charts are created and the interactive elements are made which shows that the dashboard implementation is done. Further the analysis is being done which describe the crime prediction according to the CTA data. The data set is being collected including the period January, February and March 2020.
- In the after the entire analysis the final prediction is been shown according to the ETL process and the predictive analytics.

2. Data Gathering

The dataset contains 603 rows and is for the month of January, February and March 2020. Also in data. The main attributes in the data of the incident report file are:
the data set we have build the crime data and fetched the weather data. The data was stored in the form of csv file. There are two files first the incident report and second including the weather

- Time of incident
- Date of incident
- Site of incident
- Incident description
- Incident
- ATM
- Restaurant
- Time category

The attributes are further classified into different types namely, Continuous, Discrete, Nominal, Binary.

The second file consisting of the weather data include the attributes:

- Humidity
- Temperature
- Cloud cover
- Heat index
- Sunrise
- Sunset
- Moonrise
- Moonset

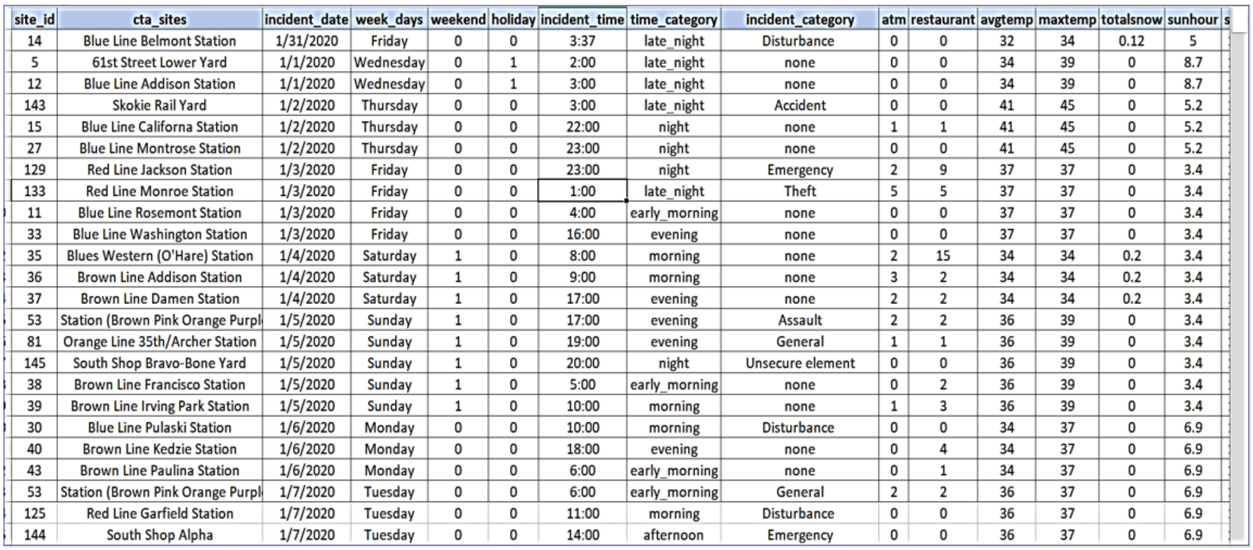
The attributes of the weather data are further classified into different types as Nominal, Continuous, Discrete.

By using all these attributes, the prediction of the crime in the Chicago state is been performed. The data and the information are collected from the website:

https://data.cityofchicago.org/Public-Safety/CTA-Crime/5xiy-qnsz

https://weatherstack.com/dashboard?loggedin=1

Overview of the dataset displaying all the collected feature:



site_id	cta_sites	incident_date	week_days	weekend	holiday	incident_time	time_category	incident_category	atm	restaurant	avgtemp	maxtemp	totalsnow	sunhour	s
14	Blue Line Belmont Station	1/31/2020	Friday	0	0	3:37	late_night	Disturbance	0	0	32	34	0.12	5	
5	61st Street Lower Yard	1/1/2020	Wednesday	0	1	2:00	late_night	none	0	0	34	39	0	8.7	
12	Blue Line Addison Station	1/1/2020	Wednesday	0	1	3:00	late_night	none	0	0	34	39	0	8.7	
143	Skokie Rail Yard	1/2/2020	Thursday	0	0	3:00	late_night	Accident	0	0	41	45	0	5.2	
15	Blue Line Californa Station	1/2/2020	Thursday	0	0	22:00	night	none	1	1	41	45	0	5.2	
27	Blue Line Montrose Station	1/2/2020	Thursday	0	0	23:00	night	none	0	0	41	45	0	5.2	
129	Red Line Jackson Station	1/3/2020	Friday	0	0	23:00	night	Emergency	2	9	37	37	0	3.4	
133	Red Line Monroe Station	1/3/2020	Friday	0	0	1:00	late_night	Theft	5	5	37	37	0	3.4	
11	Blue Line Rosemont Station	1/3/2020	Friday	0	0	4:00	early_morning	none	0	0	37	37	0	3.4	
33	Blue Line Washington Station	1/3/2020	Friday	0	0	16:00	evening	none	0	0	37	37	0	3.4	
35	Blues Western (O'Hare) Station	1/4/2020	Saturday	1	0	8:00	morning	none	2	15	34	34	0.2	3.4	
36	Brown Line Addison Station	1/4/2020	Saturday	1	0	9:00	morning	none	3	2	34	34	0.2	3.4	
37	Brown Line Damen Station	1/4/2020	Saturday	1	0	17:00	evening	none	2	2	34	34	0.2	3.4	
53	Station (Brown Pink Orange Purpl	1/5/2020	Sunday	1	0	17:00	evening	Assault	2	2	36	39	0	3.4	
81	Orange Line 35th/Archer Station	1/5/2020	Sunday	1	0	19:00	evening	General	1	1	36	39	0	3.4	
145	South Shop Bravo-Bone Yard	1/5/2020	Sunday	1	0	20:00	night	Unsecure element	0	0	36	39	0	3.4	
38	Brown Line Francisco Station	1/5/2020	Sunday	1	0	5:00	early_morning	none	0	2	36	39	0	3.4	
39	Brown Line Irving Park Station	1/5/2020	Sunday	1	0	10:00	morning	none	1	3	36	39	0	3.4	
30	Blue Line Pulaski Station	1/6/2020	Monday	0	0	10:00	morning	Disturbance	0	0	34	37	0	6.9	
40	Brown Line Kedzie Station	1/6/2020	Monday	0	0	18:00	evening	none	0	4	34	37	0	6.9	
43	Brown Line Paulina Station	1/6/2020	Monday	0	0	6:00	early_morning	none	0	1	34	37	0	6.9	
53	Station (Brown Pink Orange Purpl	1/7/2020	Tuesday	0	0	6:00	early_morning	General	2	2	36	37	0	6.9	
125	Red Line Garfield Station	1/7/2020	Tuesday	0	0	11:00	morning	Disturbance	0	0	36	37	0	6.9	
144	South Shop Alpha	1/7/2020	Tuesday	0	0	14:00	afternoon	Emergency	0	0	36	37	0	6.9	

3. **Extract Transform and Load:**
   In this Process we have fetched data and made more column named sun data which we have processed and exported this data to tableau for further visualization. We have a lot of columns in the dataset so we have selected some important attributes to it.
   From this we got two files:
   1. Crime reports file
   2. Weather Data

4. Research Problems

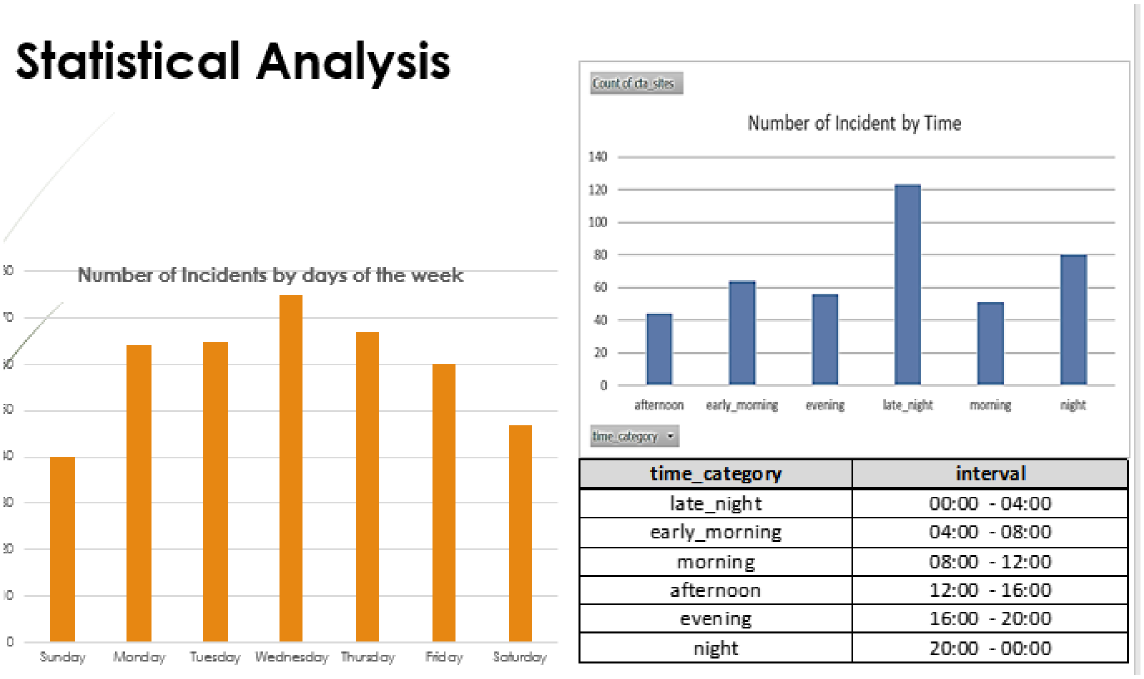
- In the Chicago area, as we see that the crime is increasing day by day with a high speed which shows that safety is the major concern. The entire police department and security agency have been working and trying hard to reduce these crimes. However, the number of crimes is not reducing and its quite difficult to control them at some regions.
- Here, we have done the prediction of the crime rate and the following question are been answered:
  1. What are the main factor that encourage crime at a given location?
  2. How to predict whether crime is likely to happen at a given location and date?
  3. What type of crime is more likely to happen at a location and date?
- By solving all these problems and providing the answer to this the prediction of the crime in the Chicago city can be easily done.

5. Evaluation

The data is collected and is being preprocessed to develop the variables. The two ETL process are being implemented. First ETL process is being implemented on the site information and the second one on the weather data then the SQL cross and join operation is being performed on the data.

The ETL process implemented on the site information shows the information from all the sites in the three months. The values which are required for data modeling are selected and then the crime category is being replaced. In the ETL process which is implemented on the weather data shows the transformation of the three months according to the weather data. Here also we have selected the data required for data modeling and then it concatenated the entire sun data with the rising and the setting time. Once the ETL is being performed and the operations are done the data consists of two new changes. The incident category is changed into severity scale and the sunrise and sunset data into concatenation of sun data. This shows that the two tables will consist of the severity in the CTA table and a sun data column in the temperature table. The SQL query is being done for combining the two datasets and creating the tables for the ETL process.

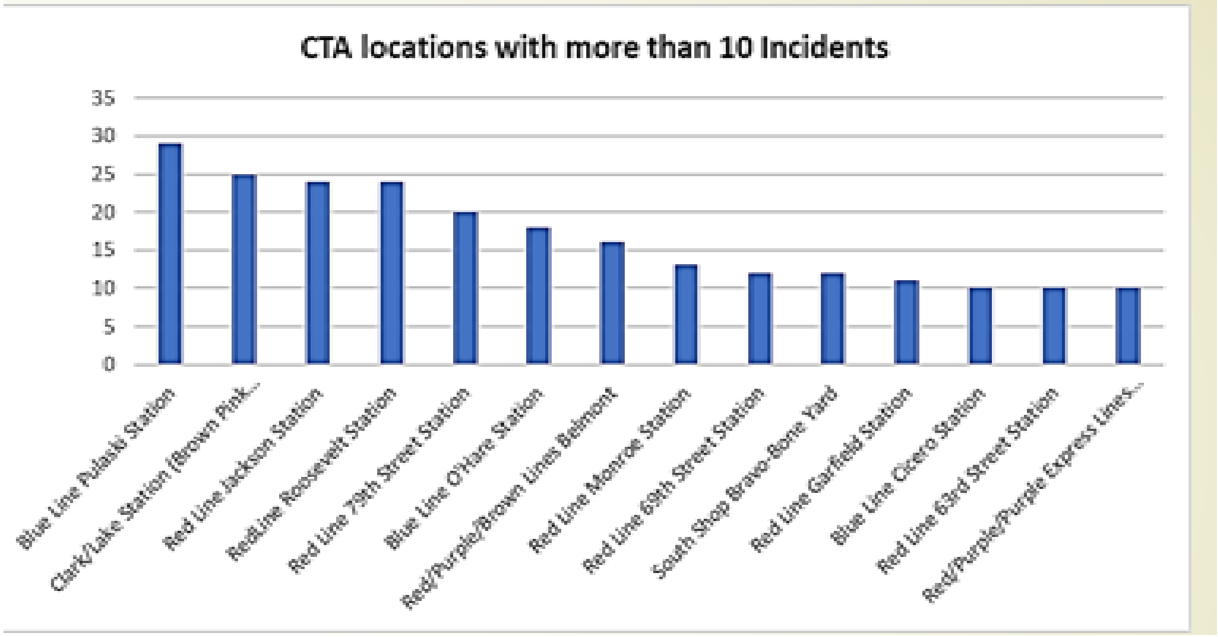
The Statistical Analysis of the data:



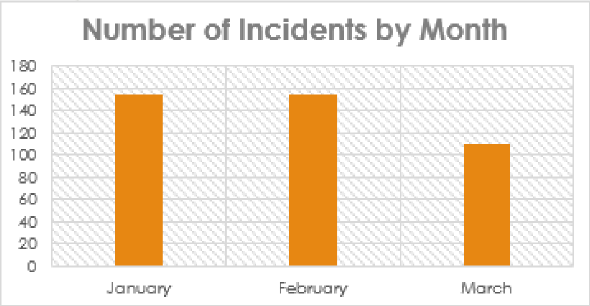
Statistical Analysis

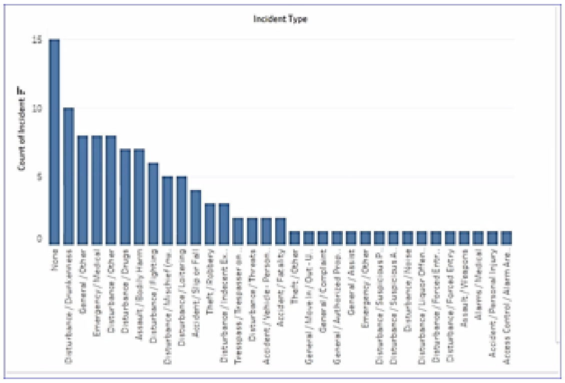
time_category	interval
late_night	00:00 - 04:00
early_morning	04:00 - 08:00
morning	08:00 - 12:00
afternoon	12:00 - 16:00
evening	16:00 - 20:00
night	20:00 - 00:00

The statistical analysis showing the number of incidents by the CTA locations:



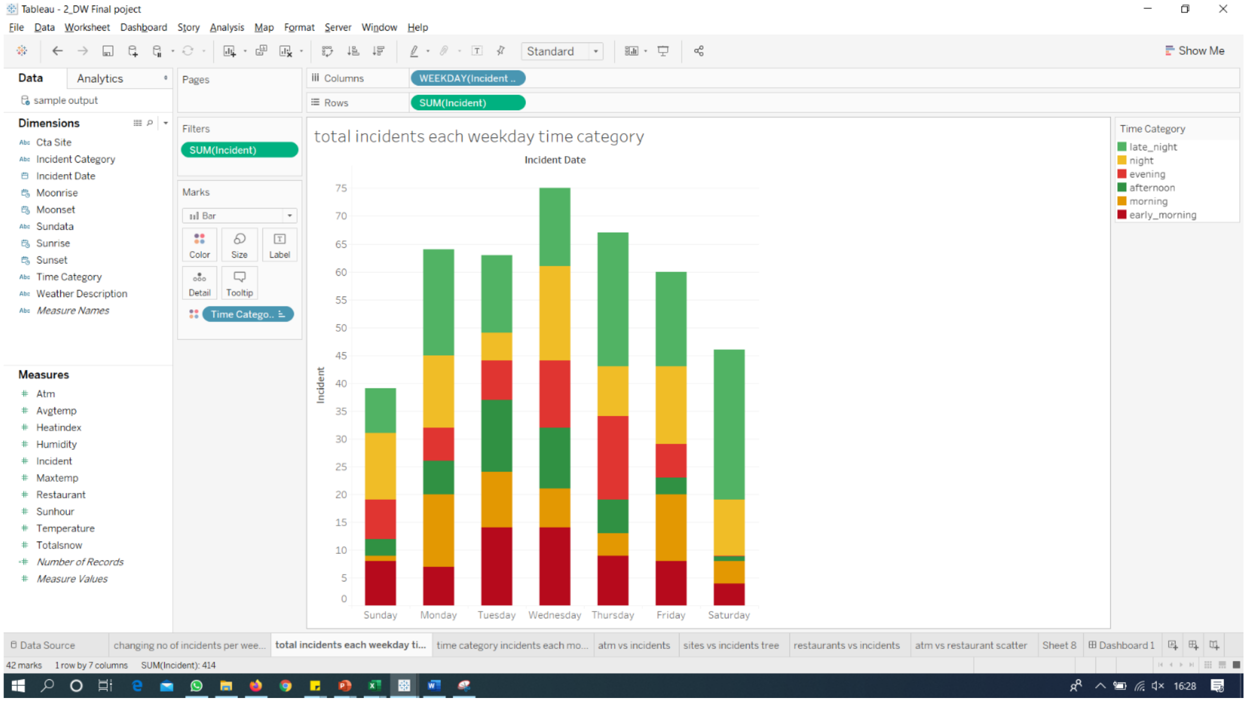
The statistical analysis according to the month and the type of incident:



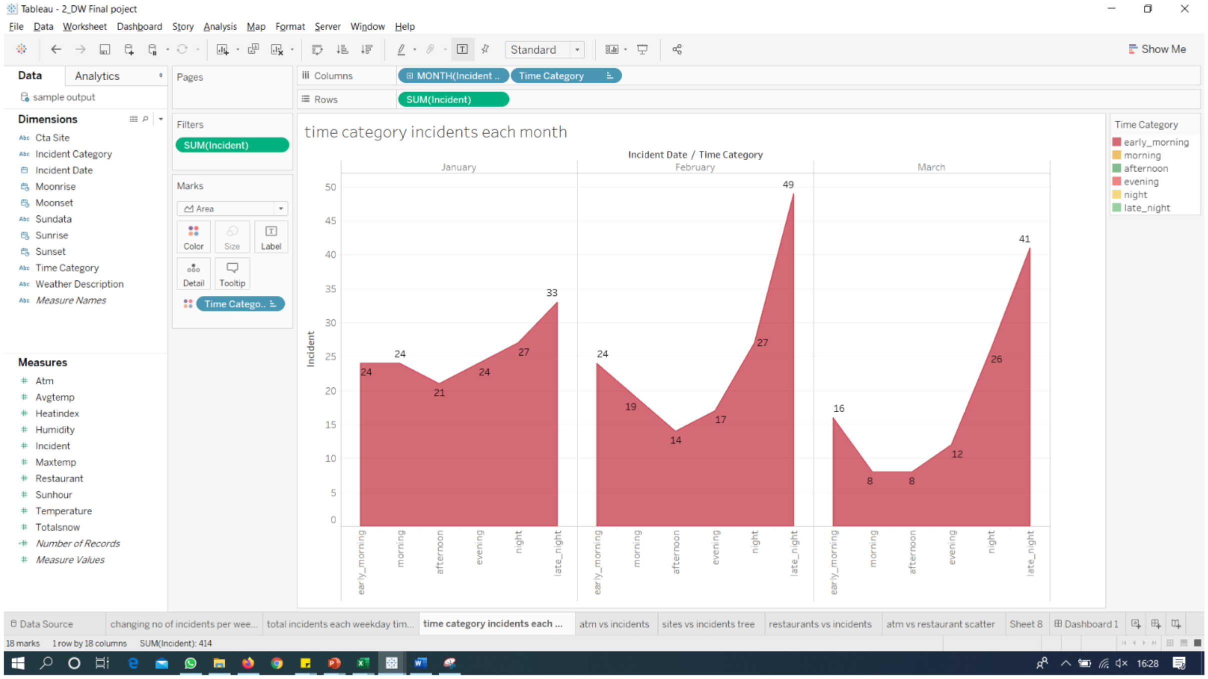


By this we can see that due to COVID 19 and the stay at home order the crime is being reduced and affected the crime rate. Also, it shows that in the three months the incidents are mostly the disturbance type.

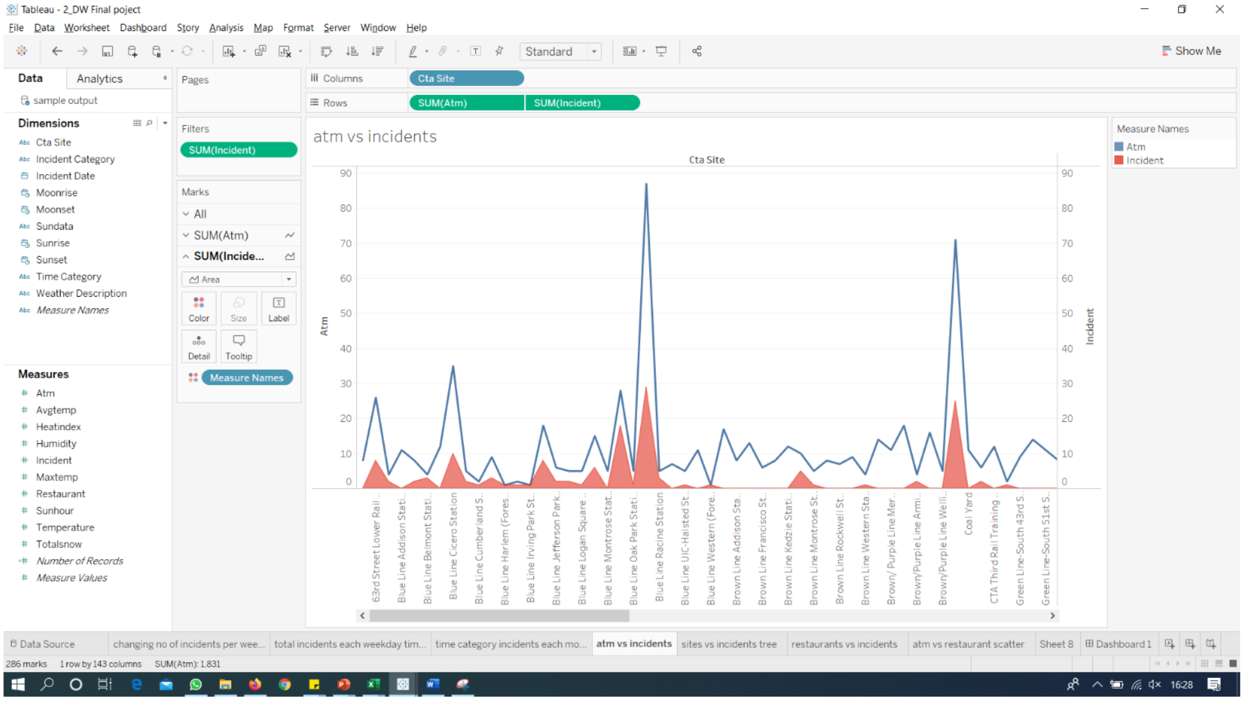
Visualization of the data in Tableau:



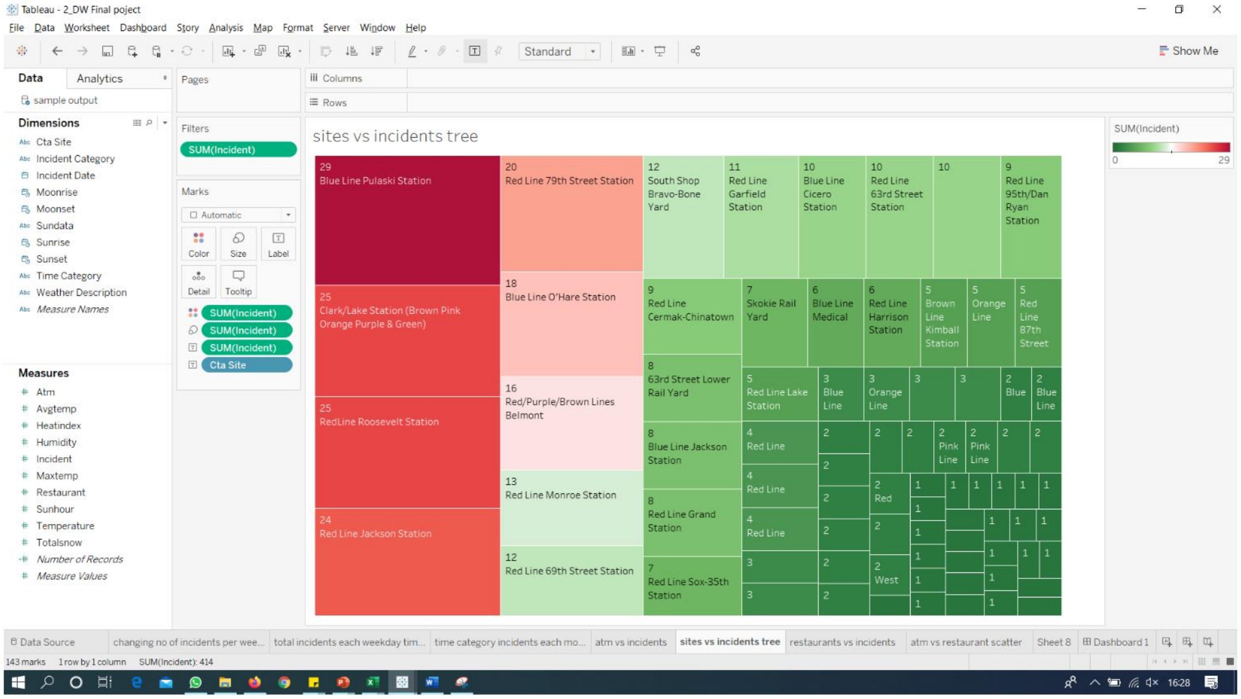
This visual graph is representing the number of incidents with weekday and also includes time category in that. We can analyze that most of the crime happened on Wednesday and most of the incident happened during night and then late_night category respectively. It is showing and telling us which are the safe day to with.



From the above graph we can understand that during the span of 3 months there are different type of incidents during the different phases of the day. In this we can also understand that during the late night most of the incidents happened in the 3 months.



This graph sketched between atm and incidents and shows that where there are more number of atm there are more probability of incidents. From this we can also note that **blue line oak park** registered most number of the incidents.



This graph is between incident and cta sites. This shows how many incidents are there happening in each cta sites**.**

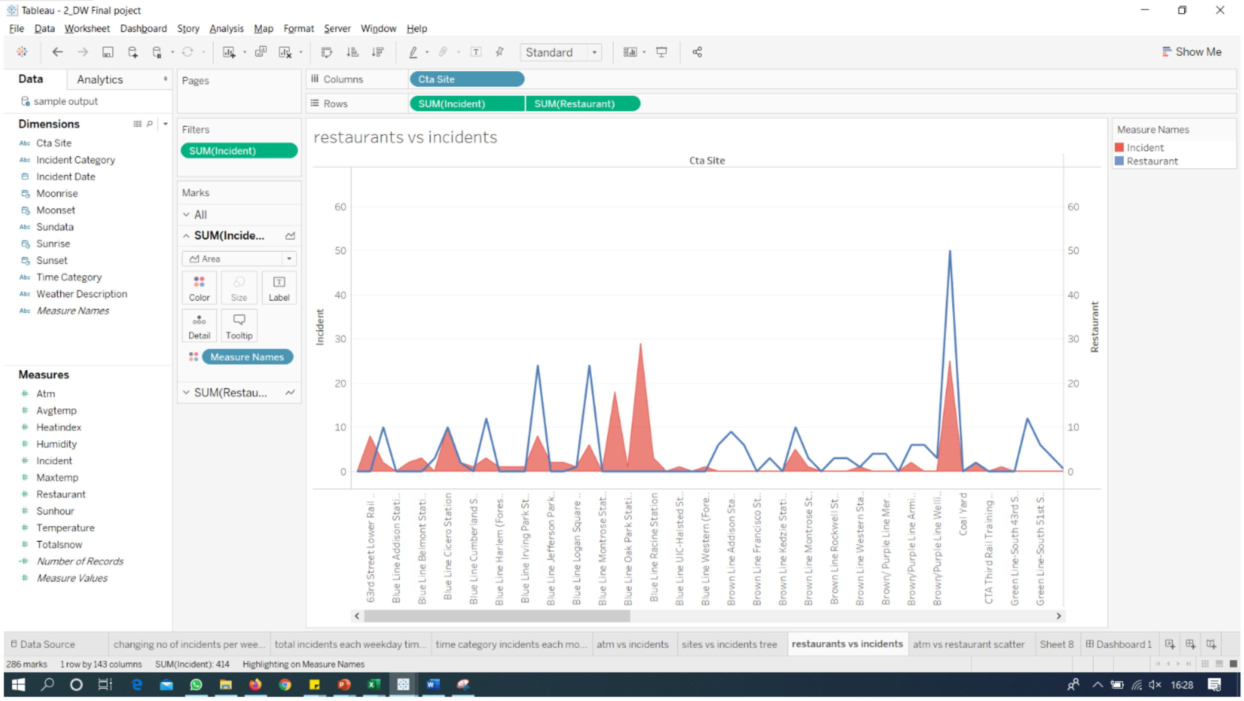
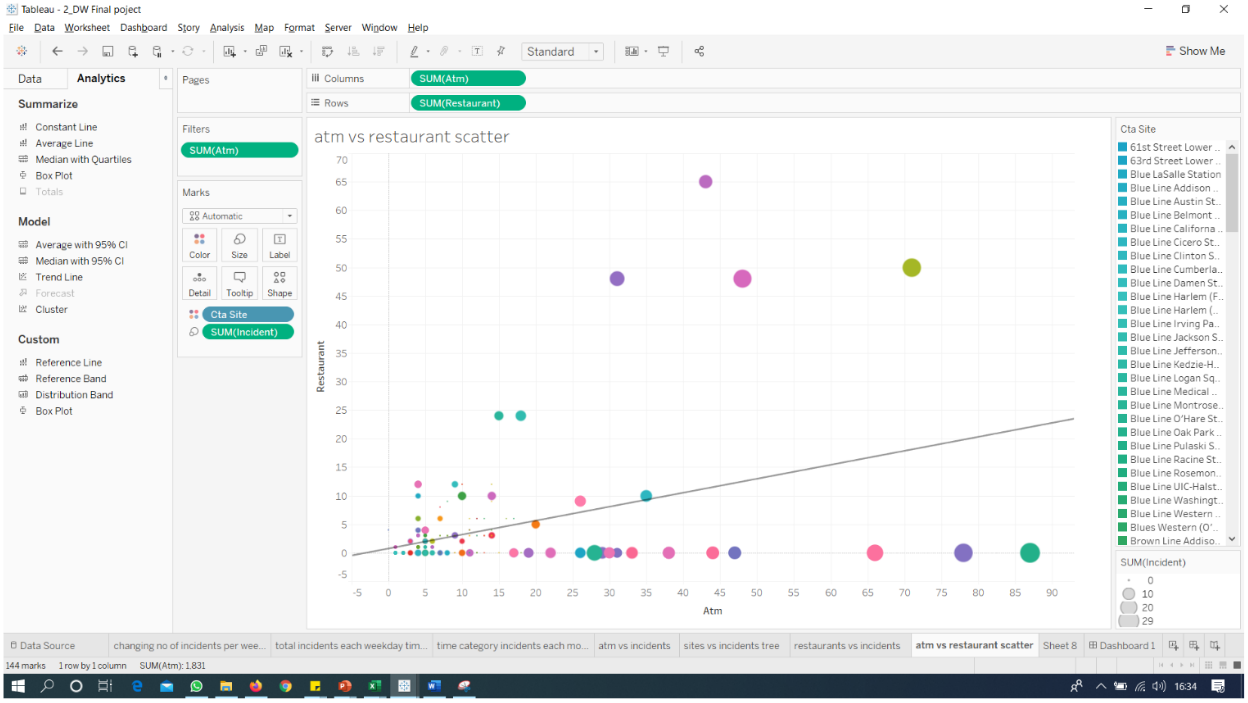


Tableau - 2_DW Final poject
File Data Worksheet Dashboard Story Analysis Map Format Server Window Help
Standard
Show Me
Data
Analytics
sample output
Dimensions
Cta Site
Incident Category
Incident Date
Moonrise
Moonset
Sundata
Sunrise
Sunset
Time Category
Weather Description
Measure Names
Measures
Atm
Avgtemp
Heatindex
Humidity
Incident
Maxtemp
Restaurant
Sunhour
Temperature
Totalsnow
Number of Records
Measure Values
Pages
Filters
SUM(Incident)
Marks
All
SUM(Incide...
Area
Color
Size
Label
Detail
Tooltip
Measure Names
SUM(Restau...
Columns
Cta Site
Rows
SUM(Incident)
SUM(Restaurant)
restaurants vs incidents
Cta Site
Incident
Restaurant
0
10
20
30
40
50
60
63rd Street Lower Rail...
Blue Line Addison Stati...
Blue Line Belmont Stati...
Blue Line Cicero Station
Blue Line Cumberland S...
Blue Line Harlem (Fores...
Blue Line Irving Park St...
Blue Line Jefferson Park...
Blue Line Logan Square ...
Blue Line Montrose Stat...
Blue Line Oak Park Stati...
Blue Line Racine Station
Blue Line UIC-Halsted St...
Blue Line Western (Fore...
Brown Line Addison Sta...
Brown Line Francisco St...
Brown Line Kedzie Stati...
Brown Line Montrose St...
Brown Line Rockwell St...
Brown Line Western Sta...
Brown/ Purple Line Mer...
Brown/Purple Line Armi...
Brown/Purple Line Welli...
Coal Yard
CTA Third Rail Training...
Green Line-South 43rd S...
Green Line-South 51st S...
Measure Names
🟥 Incident
🟦 Restaurant
Data Source
changing no of incidents per wee...
total incidents each weekday tim...
time category incidents each mo...
atm vs incidents
sites vs incidents tree
restaurants vs incidents
atm vs restaurant scatter
Sheet 8
Dashboard 1
286 marks
1 row by 143 columns
SUM(Incident): 414
Highlighting on Measure Names
16:28

This graph sketched between restaurant and incidents shows that where there are more number of restaurant there are more probability of incidents. From this we can also note that **coal yard** registered most number of the incidents.



In this we have sketched data trend between atm and restaurant and according to the size of the data and incidents.

6. Expected outcomes:

After completing the project, we will be able to say according to the month what is crime rate in the particular city by considering the main locations. The ETL implementation is done and the data is been visualized to find the final result.

We received the transformed data and all the visualized.

- From statistical analysis we can see that the Wednesday is the day having more crime in past three months.
- Also most of the crime happened during late night around(0:00 am to 4:00 am)
-

  Blue Line Pulaski has highest rate of Crime also the ATM and Restaurant count affects the number of Crimes.
- Covid-19 stay at home might create some bias in data which could affect the accuracy of these model.