PREDICTIVE ANALYSIS OF CHICAGO'S WEATHER ON VIOLENCE

Introduction to Data Analytics

IS 534

Instructor: Dr. Lynn Collen

ST. CLOUD STATE UNIVERSITY™

Submitted by,

Venkata Ayyappa D

13228200

Table of contents

PREDICTIVE ANALYSIS OF CHICAGO'S WEATHER ON VIOLENCE

Executive Summary

Chicago's extreme gun violence—762 homicides last year and more than 4,000 people wounded—has been described as an epidemic. Most importantly, south side of Chicago has been very much sensitive. Primarily gang-related, the shootings are often spontaneous and unpredictable. That's why the FBI's Chicago Division, working with the Chicago Police Department (CPD) and other agencies, has undertaken significant measures to address the problem. Looking at these numbers and the FBI's involvement in getting down to the street level to address violent crime, it is very clear that the City of Chicago needs a lot of hope and support.

This report is made to show how Chicago's crime rate trends with its climate. t's important to remember correlation does not mean causation, i.e., even if this process achieved promising trends, we cannot say that climate might be the major reason behind the crimes. Climate may or may not have a major effect on the crime's happening on the streets. But, this report brings out the possible trends of the crime rate against the climate and other analytical reports including Principle component analysis and Regression analysis on the Chicago's Crime related data.

Software Used:

- Minitab
- Microsoft SQL Server
- Tableau
- Microsoft Excel

Data Preparation:

- To perform analysis on 17 years of data, 2001 -2017. To do that, I need weather/ Climate related data from any weather station in Chicago. Similarly, 17 years of crime related data.
- City of Chicago has three main weather stations that log the weather data. But most of the weather stations has 10 years of data (2007-2017). But, the station 'CHICAGO OHARE INTERNATIONAL AIRPORT, IL US' has all the data I required.
- Visited 'NOAA' (National Oceanic and Atmospheric Administration) website for the required dataset.
  https://www.ncdc.noaa.gov/cdo-web/datasets
- Up on valid request (for educational purposes), they have accepted my proposal and allowed me to download the data set of Chicago's airport's weather log.

This is how the Dataset looks:

STATION	NAME	DATE	TAVG	TMAX	TMIN
USW00094846	CHICAGO OHARE INTERNATIONAL AIRPORT, IL US	1/1/2001	15	24	5
USW00094846	CHICAGO OHARE INTERNATIONAL AIRPORT, IL US	1/2/2001	12	19	5
USW00094846	CHICAGO OHARE INTERNATIONAL AIRPORT, IL US	1/3/2001	18	28	7
USW00094846	CHICAGO OHARE INTERNATIONAL AIRPORT, IL US	1/4/2001	25	30	19
USW00094846	CHICAGO OHARE INTERNATIONAL AIRPORT, IL US	1/5/2001	29	36	21
USW00094846	CHICAGO OHARE INTERNATIONAL AIRPORT, IL US	1/6/2001	25	33	17
USW00094846	CHICAGO OHARE INTERNATIONAL AIRPORT, IL US	1/7/2001	28	34	21
USW00094846	CHICAGO OHARE INTERNATIONAL AIRPORT, IL US	1/8/2001	19	26	12
USW00094846	CHICAGO OHARE INTERNATIONAL AIRPORT, IL US	1/9/2001	17	23	10
USW00094846	CHICAGO OHARE INTERNATIONAL AIRPORT, IL US	1/10/2001	26	34	18
USW00094846	CHICAGO OHARE INTERNATIONAL AIRPORT, IL US	1/11/2001	29	39	18
USW00094846	CHICAGO OHARE INTERNATIONAL AIRPORT, IL US	1/12/2001	30	37	23
USW00094846	CHICAGO OHARE INTERNATIONAL AIRPORT, IL US	1/13/2001	34	36	32
USW00094846	CHICAGO OHARE INTERNATIONAL AIRPORT, IL US	1/14/2001	34	36	32

Courtesy: GHCN (Global Historical Climatology Network)

Brief Description:

GHCN (Global Historical Climatology Network)-Daily is a database that addresses the critical need for historical daily temperature, precipitation, and snow records over global land areas. GHCN-Daily is a composite of climate records from numerous sources that were merged and then subjected to a suite of quality assurance reviews. The archive includes over 40 meteorological elements including temperature daily maximum/minimum, temperature at observation time, precipitation, snowfall, snow depth, evaporation, wind movement, wind maximums, soil temperature, cloudiness, and more.

Data observations (values):

- **STATION** (17 characters) is the station identification code.
- **STATION_NAME** (max 50 characters) is the name of the station (usually city/airport name). Optional output field.
- **DATE** is the year of the record (4 digits) followed by month (2 digits) and day (2 digits).
- **TMAX** = Maximum temperature (Fahrenheit or Celsius as per user preference, Fahrenheit to tenths on Daily Form pdf file)
- **TMIN** = Minimum temperature (Fahrenheit or Celsius as per user preference, Fahrenheit to tenths on Daily Form pdf file)
- **TAVG=** Average temperature (Fahrenheit or Celsius as per user preference, Fahrenheit to tenths on Daily Form pdf file)

PREDICTIVE ANALYSIS OF CHICAGO'S WEATHER ON VIOLENCE

Climate Dataset Description:

This dataset consists Climate data from 2001- 2017. And has maximum temperature (TMAX) recorded on every particular day (DATE) and similarly constitutes TMIN and TAVG (Minimum temperature and Average Temperature) recorded in Fahrenheit.

I would like to consider the maximum temperature (TMAX) field for this analysis. So, the dataset contains maximum temperature recorded on a day for 17 years.

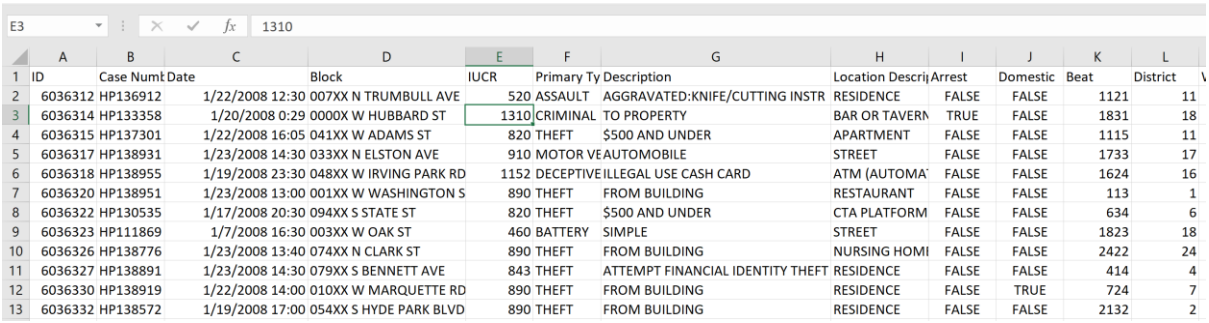
Crime Data Set:

- Chicago Daily Portal has the crime related dataset.
- This dataset reflects reported incidents of crime that occurred in the City of Chicago from 2001 to present, minus the most recent seven days. Data is extracted from the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system.

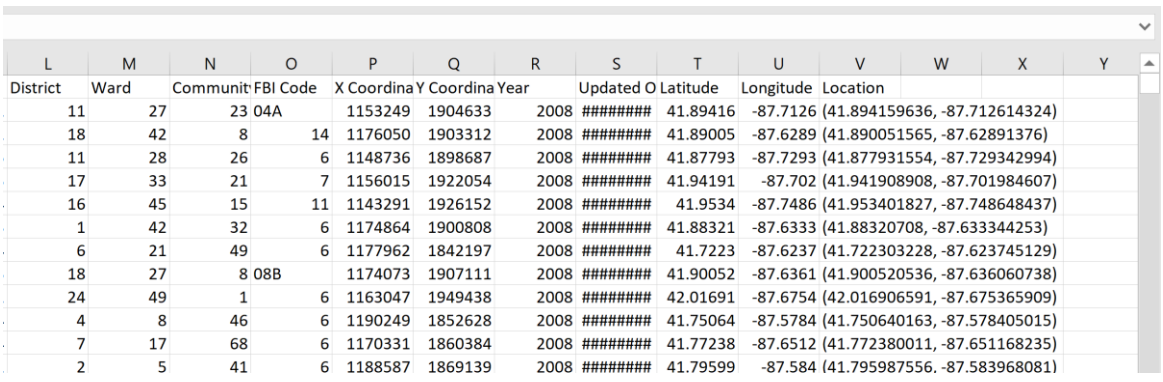
https://data.cityofchicago.org/

- This dataset that I've downloaded has full CSV dump of all 5,000,000+ reported crimes in Chicago since 2001.

This is how the data looks in the Dataset:



	A	B	C	D	E	F	G	H	I	J	K	L
1	ID	Case Numb	Date	Block	IUCR	Primary Ty	Description	Location Descri	Arrest	Domestic	Beat	District
2	6036312	HP136912	1/22/2008 12:30	007XX N TRUMBULL AVE	520	ASSAULT	AGGRAVATED:KNIFE/CUTTING INSTR	RESIDENCE	FALSE	FALSE	1121	11
3	6036314	HP133358	1/20/2008 0:29	0000X W HUBBARD ST	1310	CRIMINAL	TO PROPERTY	BAR OR TAVERN	TRUE	FALSE	1831	18
4	6036315	HP137301	1/22/2008 16:05	041XX W ADAMS ST	820	THEFT	$500 AND UNDER	APARTMENT	FALSE	FALSE	1115	11
5	6036317	HP138931	1/23/2008 14:30	033XX N ELSTON AVE	910	MOTOR VE	AUTOMOBILE	STREET	FALSE	FALSE	1733	17
6	6036318	HP138955	1/19/2008 23:30	048XX W IRVING PARK RD	1152	DECEPTIVE	ILLEGAL USE CASH CARD	ATM (AUTOMA	FALSE	FALSE	1624	16
7	6036320	HP138951	1/23/2008 13:00	001XX W WASHINGTON S	890	THEFT	FROM BUILDING	RESTAURANT	FALSE	FALSE	113	1
8	6036322	HP130535	1/17/2008 20:30	094XX S STATE ST	820	THEFT	$500 AND UNDER	CTA PLATFORM	FALSE	FALSE	634	6
9	6036323	HP111869	1/7/2008 16:30	003XX W OAK ST	460	BATTERY	SIMPLE	STREET	FALSE	FALSE	1823	18
10	6036326	HP138776	1/23/2008 13:40	074XX N CLARK ST	890	THEFT	FROM BUILDING	NURSING HOMI	FALSE	FALSE	2422	24
11	6036327	HP138891	1/23/2008 14:30	079XX S BENNETT AVE	843	THEFT	ATTEMPT FINANCIAL IDENTITY THEFT	RESIDENCE	FALSE	FALSE	414	4
12	6036330	HP138919	1/22/2008 14:00	010XX W MARQUETTE RD	890	THEFT	FROM BUILDING	RESIDENCE	FALSE	TRUE	724	7
13	6036332	HP138572	1/19/2008 17:00	054XX S HYDE PARK BLVD	890	THEFT	FROM BUILDING	RESIDENCE	FALSE	FALSE	2132	2



L	M	N	O	P	Q	R	S	T	U	V
District	Ward	Communit	FBI Code	X Coordina	Y Coordina	Year	Updated O	Latitude	Longitude	Location
11	27	23	04A	1153249	1904633	2008	########	41.89416	-87.7126	(41.894159636, -87.712614324)
18	42	8	14	1176050	1903312	2008	########	41.89005	-87.6289	(41.890051565, -87.62891376)
11	28	26	6	1148736	1898687	2008	########	41.87793	-87.7293	(41.877931554, -87.729342994)
17	33	21	7	1156015	1922054	2008	########	41.94191	-87.702	(41.941908908, -87.701984607)
16	45	15	11	1143291	1926152	2008	########	41.9534	-87.7486	(41.953401827, -87.748648437)
1	42	32	6	1174864	1900808	2008	########	41.88321	-87.6333	(41.88320708, -87.633344253)
6	21	49	6	1177962	1842197	2008	########	41.7223	-87.6237	(41.722303228, -87.623745129)
18	27	8	08B	1174073	1907111	2008	########	41.90052	-87.6361	(41.900520536, -87.636060738)
24	49	1	6	1163047	1949438	2008	########	42.01691	-87.6754	(42.016906591, -87.675365909)
4	8	46	6	1190249	1852628	2008	########	41.75064	-87.5784	(41.750640163, -87.578405015)
7	17	68	6	1170331	1860384	2008	########	41.77238	-87.6512	(41.772380011, -87.651168235)
2	5	41	6	1188587	1869139	2008	########	41.79599	-87.584	(41.795987556, -87.583968081)

PREDICTIVE ANALYSIS OF CHICAGO'S WEATHER ON VIOLENCE

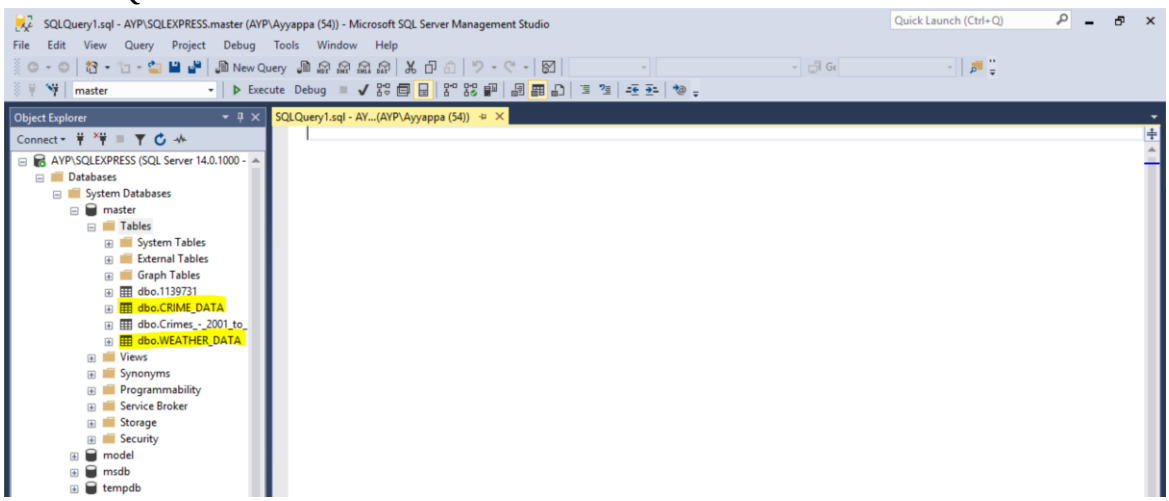
Dataset Fields:

- **ID**: (Number) Unique identifier for the record.
- **Case Number**:(Text field) The Chicago Police Department RD Number (Records Division Number), which is unique to the incident.
- **Date**: (Timestamp) Date when the incident occurred. this is sometimes a best estimate.
- **Block**: (Plain Text) The partially redacted address where the incident occurred, placing it on the same block as the actual address.
- **IUCR**: (Plain Text) The Illinois Uniform Crime Reporting code.
- **Primary Type**: (Plain Text) The primary description of the IUCR code.
- **Description**: (Plain Text) The secondary description of the IUCR code, a subcategory of the primary description.
- **Location Description**: (Plain Text) Description of the location where the incident occurred.
- **Arrest**: (Boolean) Indicates whether an arrest was made.
- **Domestic**: (Boolean) Indicates whether the incident was domestic-related as defined by the Illinois Domestic Violence Act.
- **Beat**: (Plain Text) Indicates the beat where the incident occurred. A beat is the smallest police geographic area – each beat has dedicated police beat car.
- **District**: (Plain Text) Indicates the police district where the incident occurred.
- **Ward**: (Number) The ward (City Council district) where the incident occurred.
- **Community Area**: (Plain Text) Indicates the community area where the incident occurred.
- **FBI Code**: (Plain Text) Indicates the crime classification as outlined in the FBI's National Incident-Based Reporting System (NIBRS).
- **X Coordinate**: (Number) The x coordinate of the location where the incident occurred
- **Y Coordinate**: (Number) The y coordinate of the location where the incident occurred
- **Year**: (Number) Year the incident occurred.
- **Updated On**: (Timestamp) Date and time the record was last updated.
- **Latitude**: (Number) The latitude of the location where the incident occurred
- **Longitude**: (Number)The longitude of the location where the incident occurred.
- **Location**: (Location)The location where the incident occurred in a format that allows for creation of maps and other geographic operations on this data portal.

PREDICTIVE ANALYSIS OF CHICAGO'S WEATHER ON VIOLENCE

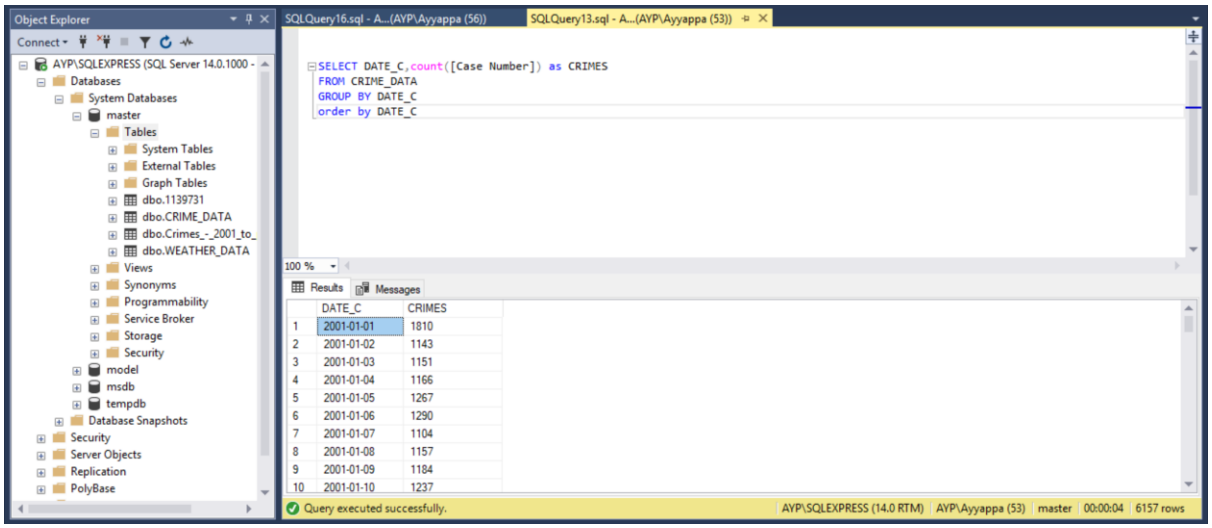
Execution:

- Using the import method in the SQL server, I have imported both the data sets in to the SQL data base.



Imported both the data sets, Crime_Data and Weather_Data (Highlighted in the picture)

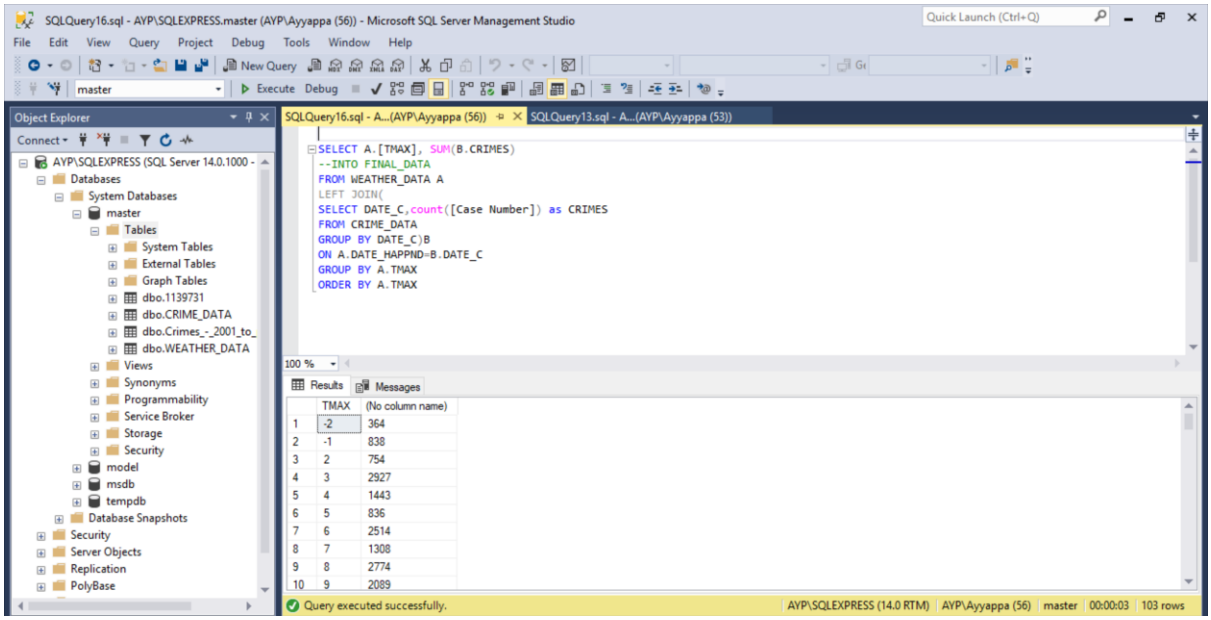
- Now, I need to group all the crimes happened in every single day. This gives me number of crimes happened in every single day.



The query written on the Crime_Data sums up all the crime happened on a single day.

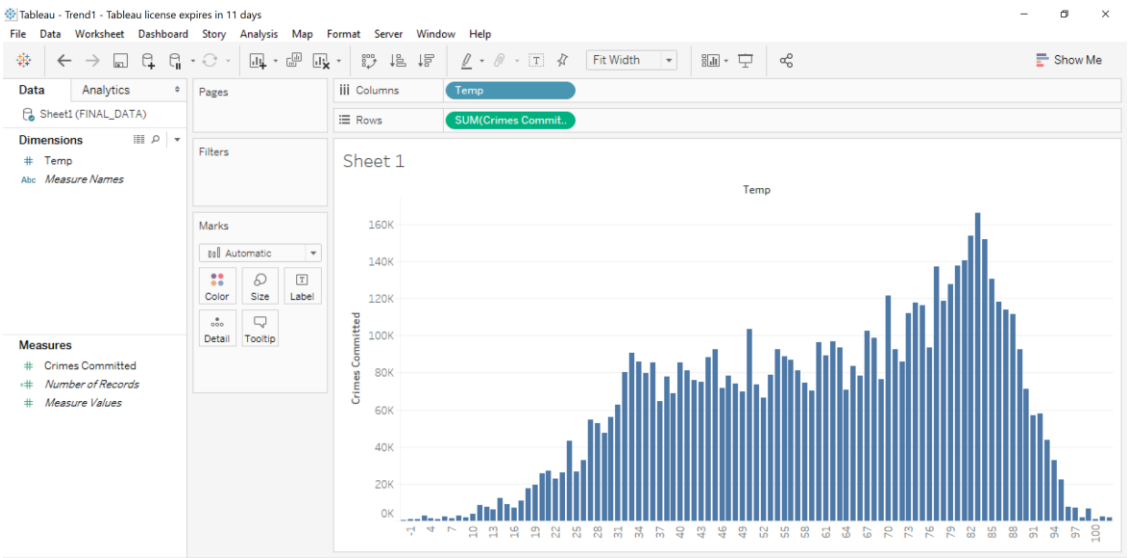
As shown above, on 2001-01-01, total of 1810 crimes happened all over the city of Chicago.

- Use this data, and Join it with Climate Data and group the result by TMAX. This query results the following. for a temperature recorded, sum of all the total crimes committed on any day, where that temperature is recorded.

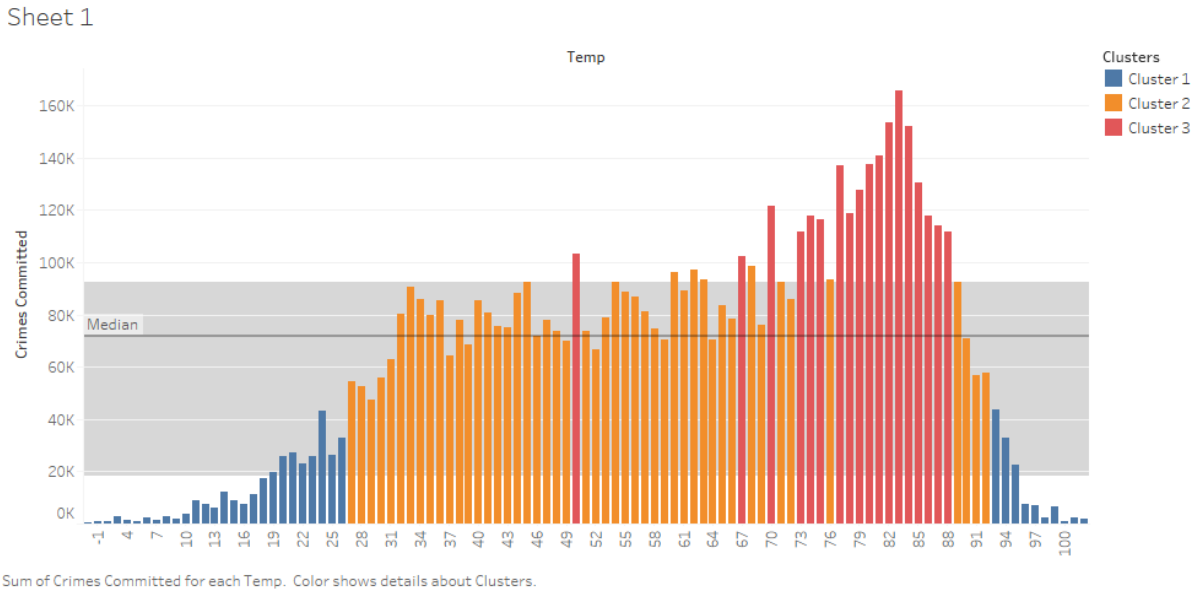


This is the important query, which helps me join the two tables based on the DATE field. This results in giving the required data as shown, Temperature vs Crimes_Committed.

- Now, I have exported the data in to Tableau and made a trend of TMAX vs Crimes.



Clustering



- In this trend, data has been grouped on to three clusters, each cluster has its range for total number of crimes committed

Cluster Number	Range
Cluster 1	0K - 50K
Cluster 2	50K-100K
Cluster 3	100K-

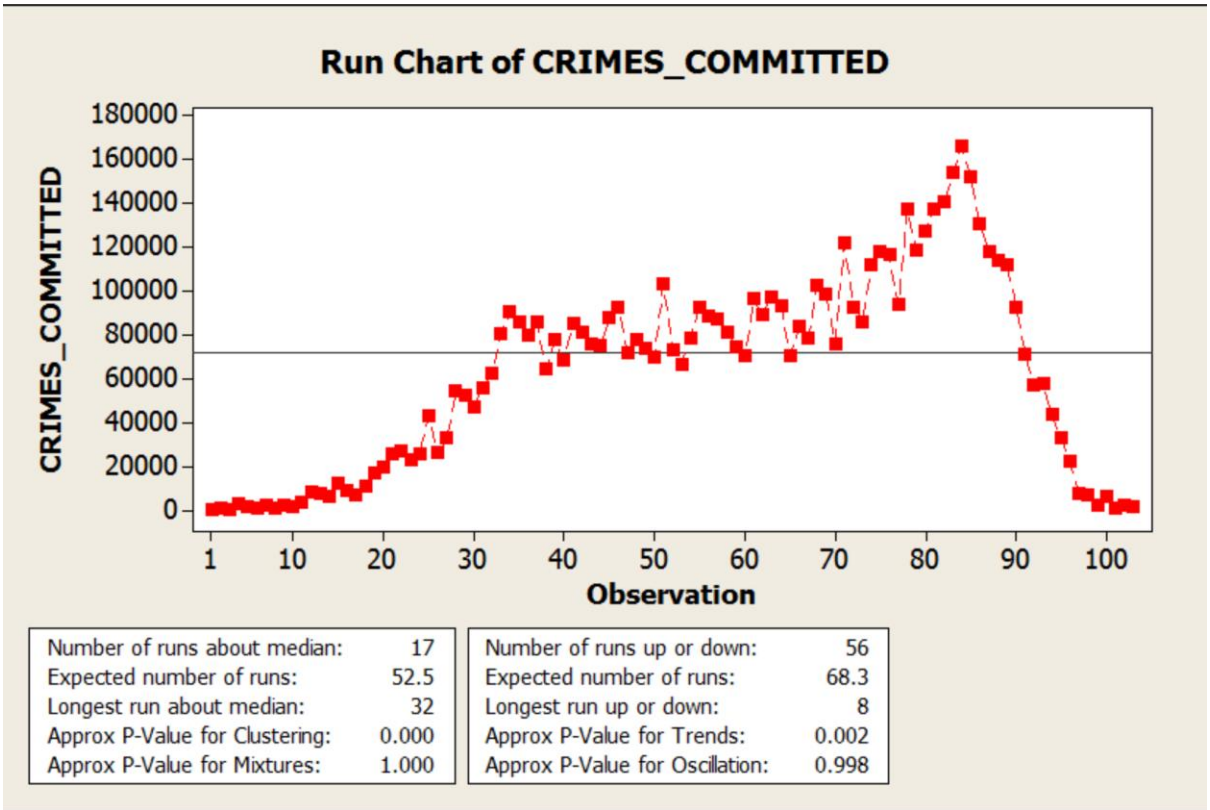
Cluster 1 with total crimes committed between (0-50K) are belonged to the temperatures at the extreme, as shown in the trend above. Temperature ranging from -3F to 26F and 93F to 103F. At these extreme temperatures, the crime count happened is relatively very low. So, it can be said that at the extreme temperatures, there is less probability of happening anything crime related. Since, at either of the extreme temperatures, people usually stay in their homes, unless there is any need to be outside.

Looking at the trend above, it was very clear that **cluster-2** with range of crimes committed in range of **50K – 100K** are belonged to **27F to 70F**. Here we can observe a constant trend of crime happening on the streets.

And coming to the last cluster, **Cluster-3** with Crime count in the range of 100K and above, belonged to **70F to 88F** temperature. At these temperatures, the crime rate was reportedly very high, since due to very comfortable temperature range, brings more people on to streets. Therefore, high probability of crime at these temperatures.

PREDICTIVE ANALYSIS OF CHICAGO'S WEATHER ON VIOLENCE

At 83F, the highest crime rate has been observed with 167K crimes committed at this temperature.



These analysis shows mostly a positive trend of Crimes Committed for a specific climate till 83F and then show a downward trend till 103F.

This analysis clearly that at relatively high temperatures, there is very high probability of violence. I would say this has been a promising trend of violence on climate.

I've always believed high temperatures contributed to violent crime in poorer neighborhoods in two ways:

1. More people on the street, since it's unbearable to be indoors if you don't have good ventilation.

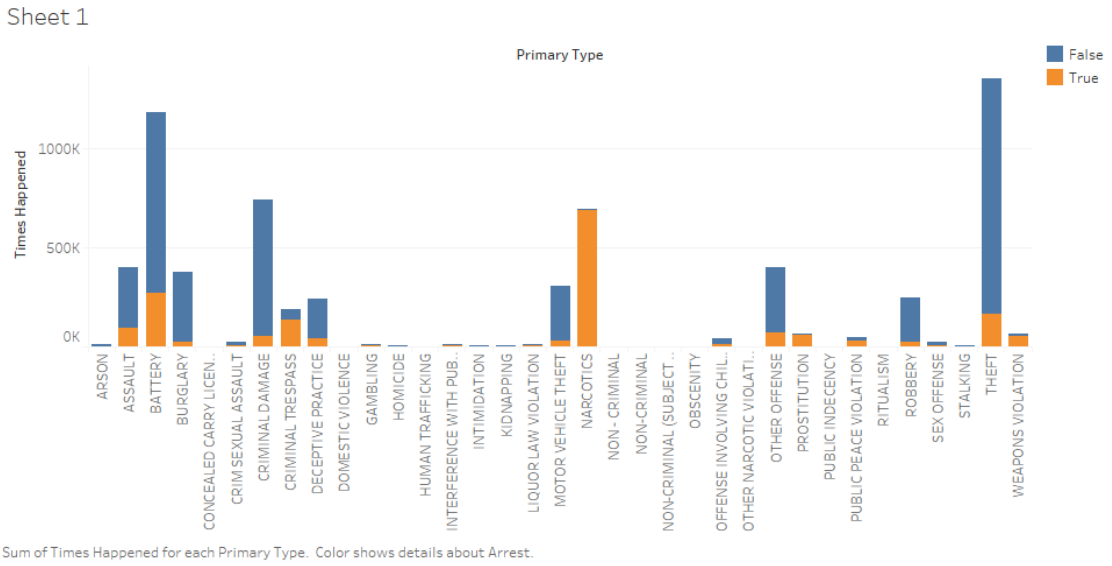
2. More frequent uncontrolled rage responses, since without proper ventilation. It's damn hard to get a good night's sleep. Angry people plus more of them outside results in more violent crime.

PREDICTIVE ANALYSIS OF CHICAGO'S WEATHER ON VIOLENCE

Violence Type vs Arrest:

This trend shows the number of times police arrested the suspects at the crime scene against to the nature of the crime committed.

As indicated, blue color shows that no one got arrested at the crime scene.

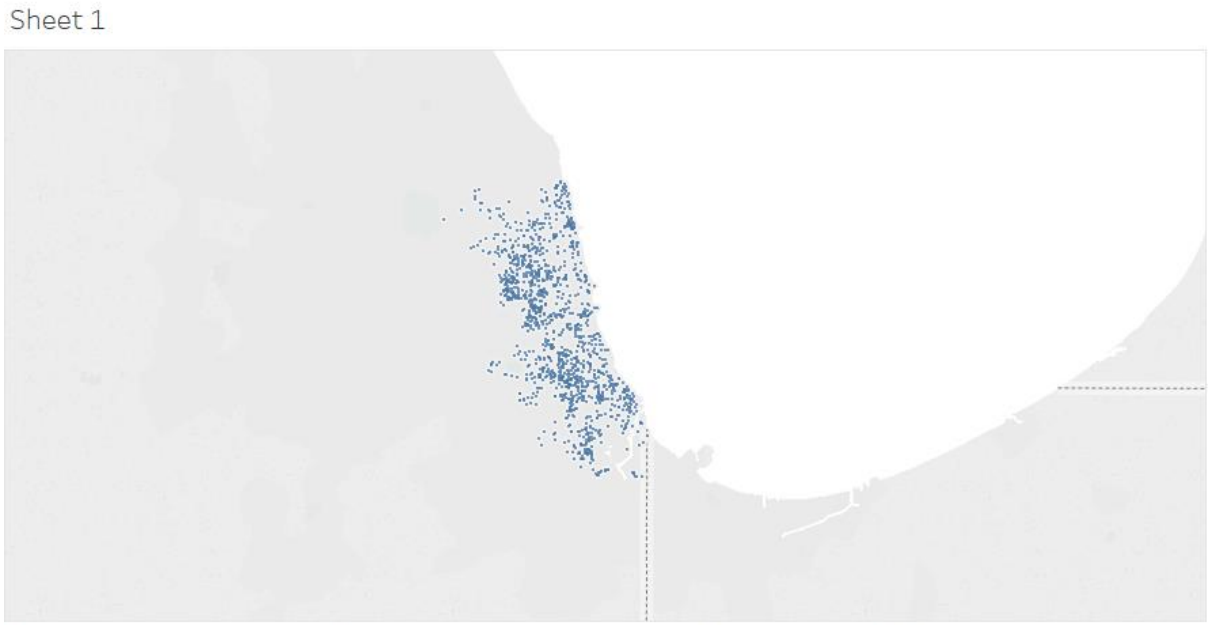


- From this trend, it can be said that when the crime committed is 'THEFT', police were unable to reach out to the situation in time and arrest the suspect.
- Same with the case of BATTERY, it was very clear that suspects might have got away from the crime scene.
- Looking at the trends of ASSAUT, BURGLARY, CRIMINAL DAMAGE and MOTOR VEHICLE THEFT, number of the suspects that got caught are very less.
- All the crimes analyzed are outdoor. May be in most of the sensitive parts of the city, more police force should be deployed to avoid this.
- More police force refers to improving the police patrol per block ratio.
- One good thing from this trend is that, in case of 'Narcotics' more than 90% of the time Police were able to arrest the suspect.

Geo Fields on Tableau

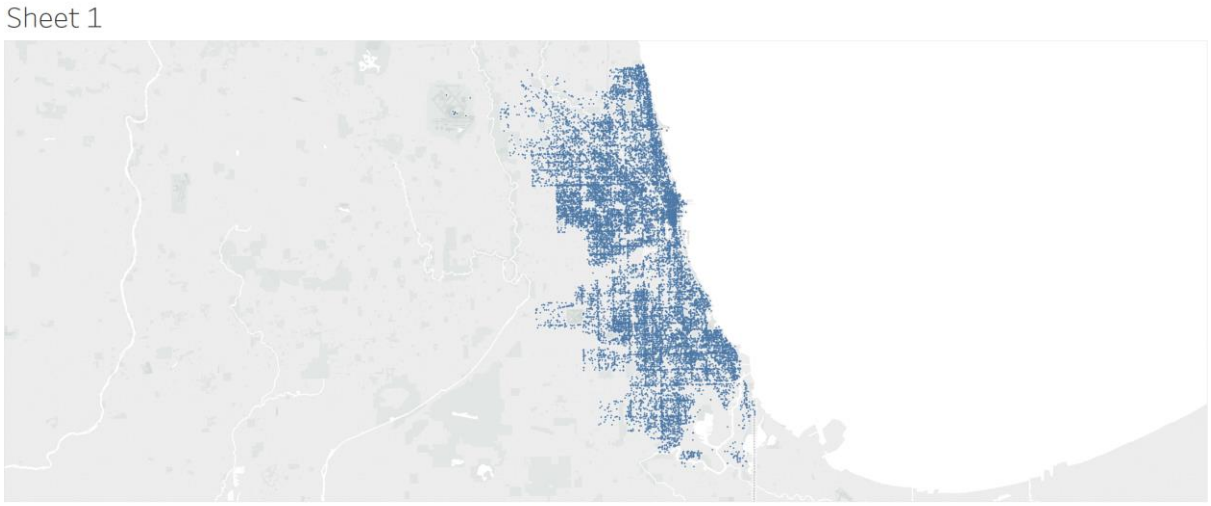
As mentioned earlier, CRIME_DATA dataset has geo fields, which are Latitude and Longitude fields for every crime record, which exactly are a pin point location in the city of Chicago.

The trend below is the locations where the arrest has been made, when ever the type of crime is ARSON.



Map based on Longitude and Latitude. The data is filtered on Arrest and Primary Type. The Arrest filter keeps True. The Primary Type filter keeps ARSON.

- Similarly, the below trend is location of the crime happened for some particular crimes, as already mentioned in the trend caption



Map based on Longitude and Latitude. The data is filtered on Description, which keeps ATTEMPT THEFT, ATTEMPT: ARMED-HANDGUN and ATTEMPT: ARMED-KNIFE/CUT INSTR.

Similarly, using the filtering technique in Tableau along the geo location fields, we can learn where any particular crime type is happening throughout the city. When the density of blue dots (crime location) is high in any of a location. Police can plan accordingly by patrol those areas more regularly and deploying more armed forces in that area.

Regression Analysis: CRIMES_COMMITTED versus TEMP

The regression equation is
CRIMES_COMMITTED = 25701 + 729 TEMP

Predictor	Coef	SE Coef	T	P
Constant	25701	7510	3.42	**0.001**
TEMP	729.2	127.1	5.74	0.000

S = 38496.2 R-Sq = 24.6% **R-Sq(adj) = 23.8%**

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	48743842067	48743842067	32.89	0.000
Residual Error	101	1.49678E+11	1481956455		
Total	102	1.98421E+11			

Unusual Observations

Obs	TEMP	CRIMES_COMMITTED	Fit	SE Fit	Residual	St Resid
84	83	165886	86225	5564	79661	2.09R
97	96	7595	95705	6867	-88110	-2.33R
98	97	6918	96434	6973	-89516	-2.36R
99	98	2117	97163	7080	-95046	-2.51R
100	99	6759	97892	7188	-91133	-2.41R
101	100	939	98622	7296	-97683	-2.58R
102	102	2329	100080	7515	-97751	-2.59R
103	103	2037	100809	7625	-98772	-2.62R

R denotes an observation with a large standardized residual.

Statistic, $\mathbf{r^2}$ measures the goodness of fit of the regression. It is also known as *coefficient of determination*, measures how well the linear approximation produced by the least-squares regression line fits the observed data.

In this case $\mathbf{r^2}$ observed to be at 24.6%.

Adjusted R2 also indicates how well terms fit a curve or line, but adjusts for the number of terms in a model. If you add more and more useless variables to a model, adjusted r-squared will decrease. If you add more useful variables, adjusted r-squared will increase.

Adjusted R2 will always be less than or equal to R2.

Adjusted R2 obtained in this case **23.8%**

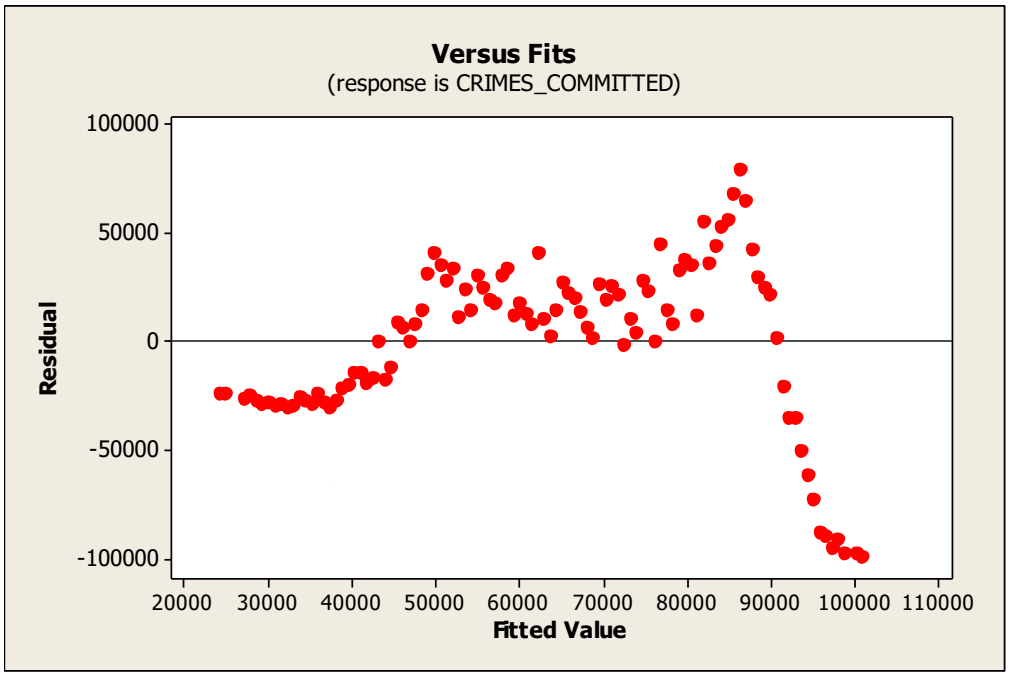
A **p value** is used in hypothesis testing to help you support or reject the null hypothesis.

The p value is the evidence against a null hypothesis. 'The smaller the p-value, the strong the evidence that you should reject the null hypothesis'.

- If $p > .10 \rightarrow$ "not significant"
- If $p \leq .10 \rightarrow$ "marginally significant"
- If $p \leq .05 \rightarrow$ "significant"
- If $p \leq .01 \rightarrow$ "highly significant."

In this case, $P < 0.01$, the results are highly significant and This means there is a very tiny(0.1%) chance your results could be random (i.e. happened by chance).

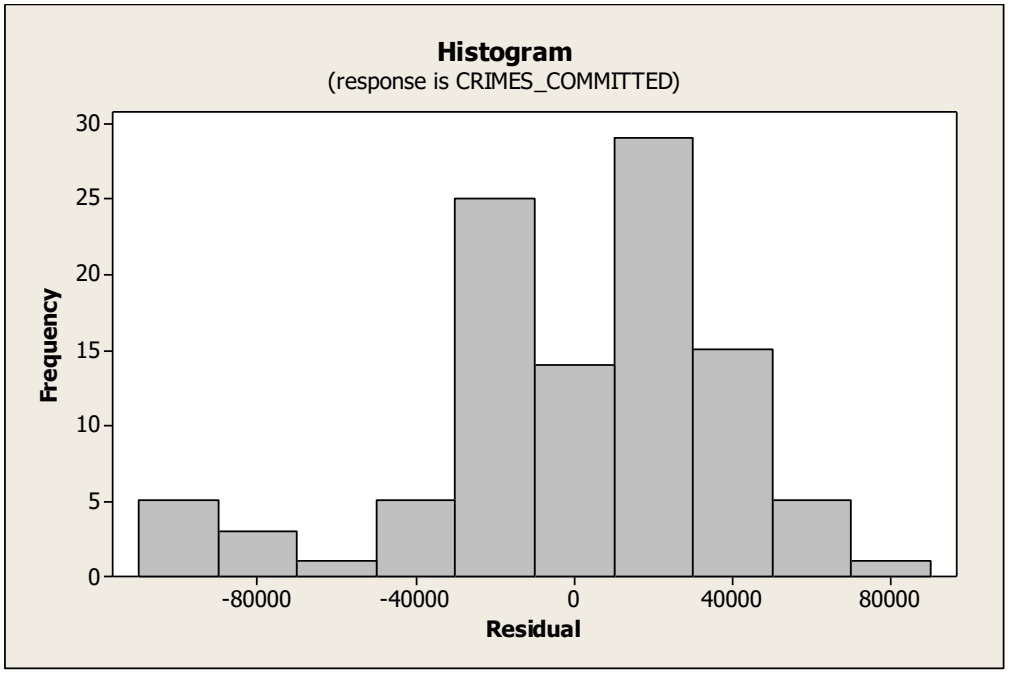
Residuals vs Fits for CRIMES_COMMITTED



A Residual vs. fits plot define the appropriateness of the simple linear regression model:

- The residuals "bounce randomly" around the 0 line. This suggests that the assumption that the relationship is linear is reasonable.
- The residuals roughly form a "horizontal band" around the 0 line. This suggests that the variances of the error terms are equal.
- No one residual "stands out" from the basic random pattern of residuals. This suggests that there are no outliers.

Residual Histogram for CRIMES_COMMITTED



(Daniel T. Larose, 2015)

References

Daniel T. Larose, C. D. (2015). *Data Mining and Predictive Analysis.* WILEY.

Programs, D. o. (n.d.). *Penn State Online course*. Retrieved from https://onlinecourses.science.psu.edu/stat501/node/36