Crime Dynamics in Chicago

Machiraju Dheeraj Marthand

Instructor: Dr. Gahangir Hossain

Data Visualization INFO 5709

Data Visualization – Term Project

12/13/2023

Introduction:-

Crime dynamics are complex and multifaceted, influenced by a lot of factors ranging from geographical location to temporal patterns and law enforcement strategies. This project aims to unravel the intricacies of crime in Chicago in the year 2023 (till October), focusing on specific hypotheses and employing visualizations to gain insights into the correlation between crime types, arrest rates, and various contextual factors.

Our investigation is guided by several hypotheses, each focusing on different aspects of the crime landscape in Chicago. The first hypothesis delves into the temporal pattern and spatial aspects. The null hypothesis posits that there is no significant disparity in crime rates across these variables. Conversely, the alternative hypotheses propose a more detailed examination, suggesting potential variations during specific periods, times of the day, and within particular streets.

Transitioning to the second set of hypotheses, the focus shifts to evaluating the impact of street safety on crime rates. The null hypothesis asserts that the safety level of streets holds no statistical significance influence over crime rates, composition of crime, temporal patterns, or spatial distribution. In contrast, the alternative hypotheses offer deeper perspectives, exploring potential disparities in overall crime rates, the composition of crime types, temporal patterns of theft during the day, and localized occurrences of theft on streets deemed less safe.

The final set of hypotheses delves into the intersection of crime, temporal patterns, street location, and specific streets, with a null hypothesis positing that arrest counts are not influenced by different factors. The alternative hypotheses present possibilities, examining the potential influence of certain crime types on arrest counts, variations in temporal patterns of arrests for the top 3 crimes (Based on arrest count), and the prominence of specific streets with significantly higher arrest counts.

Lastly, we delve into a comparative analysis of two prominent streets—s michigan ave and n state st. Our visualization, titled "s michigan ave vs n state st," utilizes dual bar graphs to compare the crime count between the least safe street and the street with the highest arrest count, offering insights into the nuanced relationship between crime incidence and law enforcement effectiveness.

Throughout this project, we employ statistical analyses and visualizations to test these hypotheses, with a goal to identify patterns, correlations, and potential relationships within the crime data for the year 2023.

Dataset

The Dataset is collected from the city of Chicago Data Portal. This dataset reflects reported incidents of crime (with the exception of murders where data exists for each victim) that occurred in the City of Chicago in 2023. Data is extracted from the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system. In order to protect the privacy of crime victims, addresses are shown at the block level only and specific locations are not identified.

Dataset is collected from the official Chicago Data Portal Website: https://data.cityofchicago.org/Public-Safety/Crimes-2023/xguy-4ndq/about_data

Attributes

Total of 22 attributes are used to describe the mass-shooting. Below is the list:

Column Name	Description	Data Type
ID	Unique identifier for the record	Integer
Case Number	The Chicago Police Department RD Number (Records Division Number), which is unique to the incident.	Object/String
Date	Date when the incident occurred. this is sometimes a best estimate.	Object/String
Block	The partially redacted address where the incident occurred, placing it on the same block as the actual address.	Object/String
IUCR	The Illinois Uniform Crime Reporting code. This is directly linked to the Primary Type and Description	Object/String
Primary Type	The primary description of the IUCR code.	Object/String
Description	The secondary description of the IUCR code, a subcategory of the primary description	Object/String
Location Description	Description of the location where the incident occurred.	Object/String
Arrest	Indicates whether an arrest was made	Boolean
Domestic	Indicates whether the incident was domestic-related as defined by the Illinois Domestic Violence Act	Boolean
Beat	Indicates the beat where the incident occurred. A beat is the smallest police geographic area – each beat has a	Object/String

	dedicated police beat car. Three to five beats make up a police sector, and three sectors make up a police district. The Chicago Police Department has 22 police districts.	
District	Indicates the police district where the incident occurred	Object/String
Ward	The ward (City Council district) where the incident occurred.	Integer
Community Area	Indicates the community area where the incident occurred. Chicago has 77 community areas.	Object/String
FBI Code	Indicates the crime classification as outlined in the FBI's National Incident-Based Reporting System (NIBRS)	Object/String
X Coordinate	The x coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection. This location is shifted from the actual location for partial redaction but falls on the same block.	Float
Y Coordinate	The y coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection. This location is shifted from the actual location for partial redaction but falls on the same block.	Float
Year	Year the incident occurred.	Integer
Updated on	Date and time the record was last updated.	Object/String
Latitude	The latitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block.	Float
Longitude	The longitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block.	Float
Location	The location where the incident occurred in a format that allows for creation of maps and other geographic operations on this data portal. This location is shifted from the actual location for partial redaction but falls on the same block.	Object/String

Tools

- Python
- Tableau

Data Cleaning

1. Dropped the columns such as “'Case Number' ,'IUCR', 'Description', 'Beat', 'Ward', 'FBI Code', 'Updated On', 'District' ,'Location' as they do not serve much in visualization.

```
df2 = df1.drop(['Case Number','IUCR','Description',
               'Beat','Ward','FBI Code',
               'Updated On', 'District','Location'],
               axis = 1)
```

```
[ ] df2.columns
```

```
Index(['ID', 'Date', 'Block', 'Primary Type', 'Location Description', 'Arrest',
       'Domestic', 'Community Area', 'X Coordinate', 'Y Coordinate', 'Year',
       'Latitude', 'Longitude'],
      dtype='object')
```

```
[ ] print(f'Shape of the df is {df2.shape}')
```

```
Shape of the df is (216509, 13)
```

2. Changed the column names of “Primary Type” to “Crime Type” and “Location Description” to “Location Type”

```
[ ] # rename column 'Primary type' to 'Crime type'
    # rename column 'Location Description' to 'Location type' (residence, apartment, garrage, etc)
```

Renaming columns to make it easier for our understanding and also remove spaces in column names since it is not a good practice.

```
[ ] df3 = df2.rename(columns={'Primary Type':'Crime_type', 'Location Description':'Location_type',
                     'Community Area':'Community_area'})
```

```
[ ] df3.columns
```

```
Index(['ID', 'Date', 'Block', 'Crime_type', 'Location_type', 'Arrest',
       'Domestic', 'Community_area', 'X Coordinate', 'Y Coordinate', 'Year',
       'Latitude', 'Longitude'],
      dtype='object')
```

3. Converted the date from object datatype to date datatype

```python
def Time_converter_to_24hr_clock(x):
  date_time_merediem = x.split(' ')

  date = date_time_merediem[0]
  time = date_time_merediem[1]
  merediem = date_time_merediem[2]
# now we have time and merediem to work with

  split_time = time.split(':')
  hours = int(split_time[0])
  minutes = split_time[1]
  seconds = split_time[2]
# now we have hours and merediem to work with

  if(merediem == 'AM'): # we see if we have AM
    if(hours == 12):# we change only if its 12 AM and return 00
      result = '00'+':'+minutes+':'+seconds
    else:# if it is AM and not 12, we just return the same hour
      result = str(hours)+':'+minutes+':'+seconds
  elif (merediem == 'PM'): # we see if we have PM
    if(hours == 12): # if it is 12 and PM we need to return 12 and not 24
      result = '12'+':'+minutes+':'+seconds
    else:# if it is PM and not 12, we add 12 to hours and return
      result = str(hours+12)+':'+minutes+':'+seconds

  return date+' '+result # result is gonna be 'date new_hours:old_min:old_sec'
```

4. Converted the block column and extracted the street column.

'Block' column has '002XX N Wells st', so we can extract the street by removing the first token .

```python
def Block_to_street_converter(x):
  Block_split = x.split(' ')
  Street_name = ''
  for i in range(len(Block_split)-1):
    Street_name = Street_name + Block_split[i+1] + ' '
  return Street_name
```

5. Converted all street names to lower case letter to account for inconsistencies in the naming of a street. (One street had both upper case representation and lower case representation)

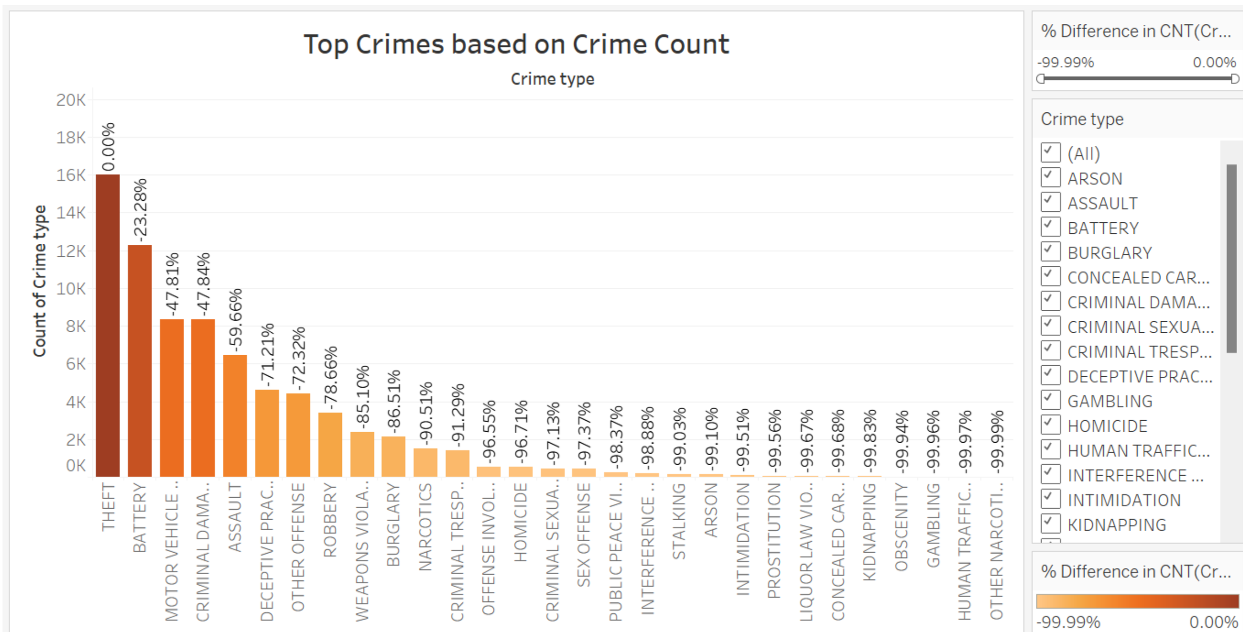
```
# there were inconsistencies with the data in "Street" column, Same street name was
# represented both in upper case letters and lower case letters.
# Converting all values in the "Street" Column into lowecase letters to deal wiht this inconsistency.

df5.Street = df5.Street.str.lower()
matching_rows = df5[df5['Street'].str.contains('s michigan ave', case=False, na=False)]
df5.shape
```

```
(216509, 13)
```

EDA

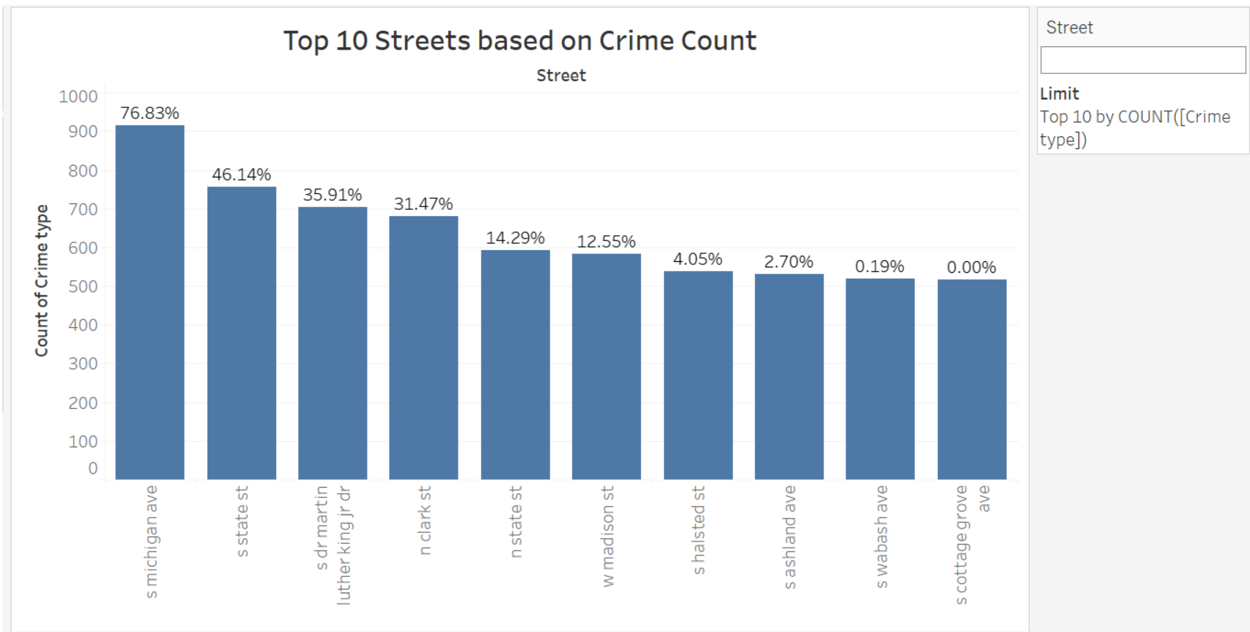
1. Which crimes are the most prevalent (with respect to crime count) what is the difference margin between crimes & how is the trend?



Result:

We can see that the Top 3 crimes in Chicago in 2023 are “Theft”, “Battery” and “Motor Vehicle Theft”, with “Theft” being the most prevalent crime(crime count about 16K). From the visualization we can infer that occurrence of “Motor Vehicle Theft” is about 48% less than that of “Theft” which makes gives us a scale to compare how prevalent “Theft” is compared to the Top three crime types.

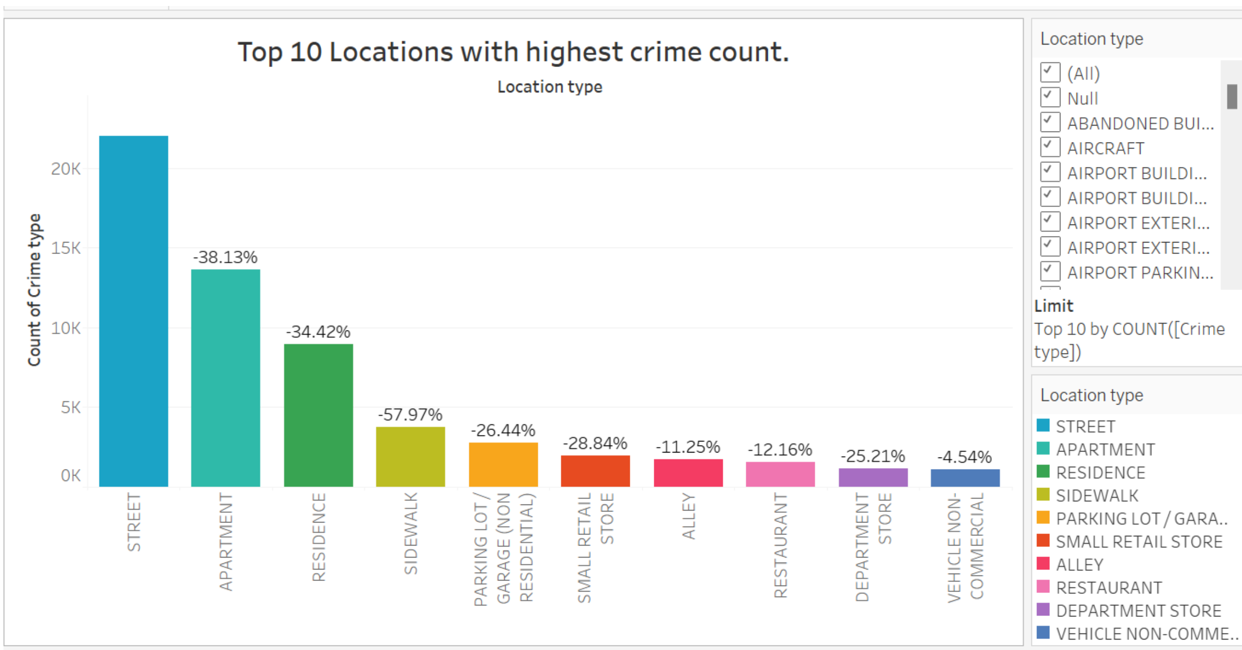
2. Which streets or blocks have the highest crime (With respect to crime count) How is the relative trend within the streets?



Result:

The visualization displays how crime count is distributed among different streets in Chicago. The street with highest crime count being "s michigan ave" and the top 10^{th} being "s cottage grove ave". And we can also see that the crime count on "s michigan ave" is 76.83% more than that of "s cottage grove ave".

3. Which crime Location has the highest crime (With respect to crime count) how are the subsequent Locations related with respect to crime count?



Result:
We can see that “Street” is the location where most of the crime occurs in Chicago, followed by “Apartment “and “Residence”.

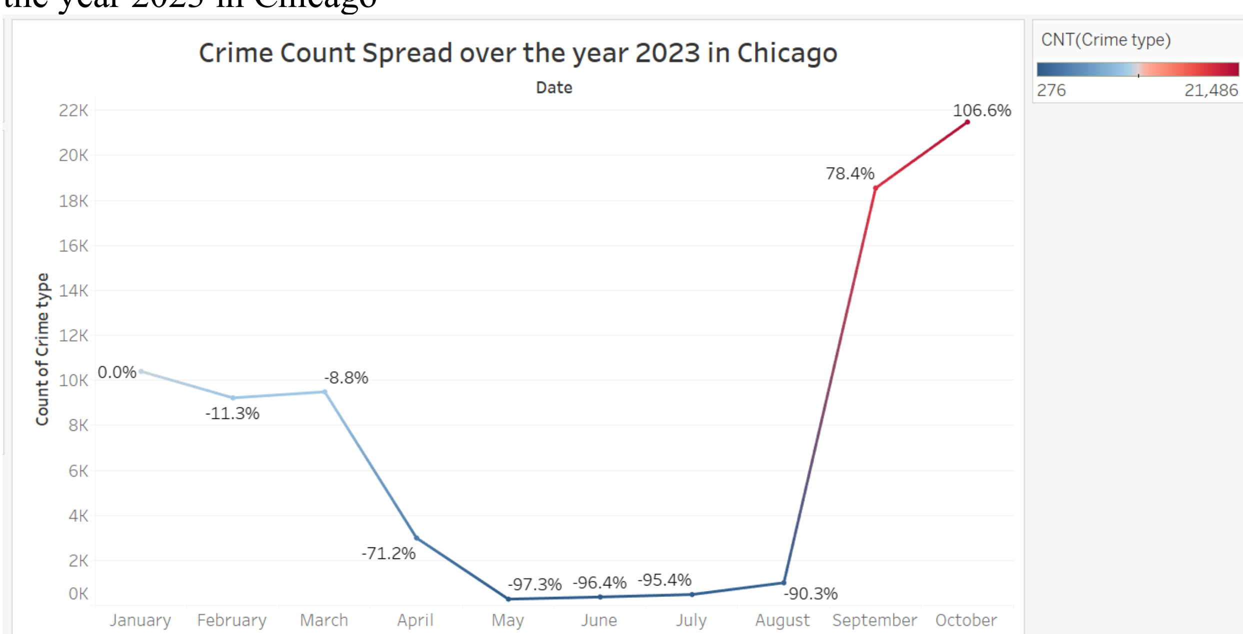
Hypothesis

1. Which <u>time of the year</u> is the safest in Chicago. Identify which <u>time of the day</u> is the safest in Chicago. Furthermore, identify the safest time in the least safe streets.

   Null Hypothesis:
   There is no significant difference in crime rates across different times of the year, times of the day, streets in Chicago in the year 2023. And there is no significant difference in crime rate in different Streets.

   Alternative Hypotheses:
   a. Certain periods of the year in 2023 exhibit significantly lower crime rates than others.
   b. Crime rates vary significantly depending on the time of day, with certain periods being safer than others.
   c. The crime rates are significantly higher in few streets compared to other streets.
   d. The overall crime rates are significantly higher on the top streets with the highest crime compared to other streets.

1. (a). **Line chart** is used to understand the distribution of crime rate over the year 2023 in Chicago

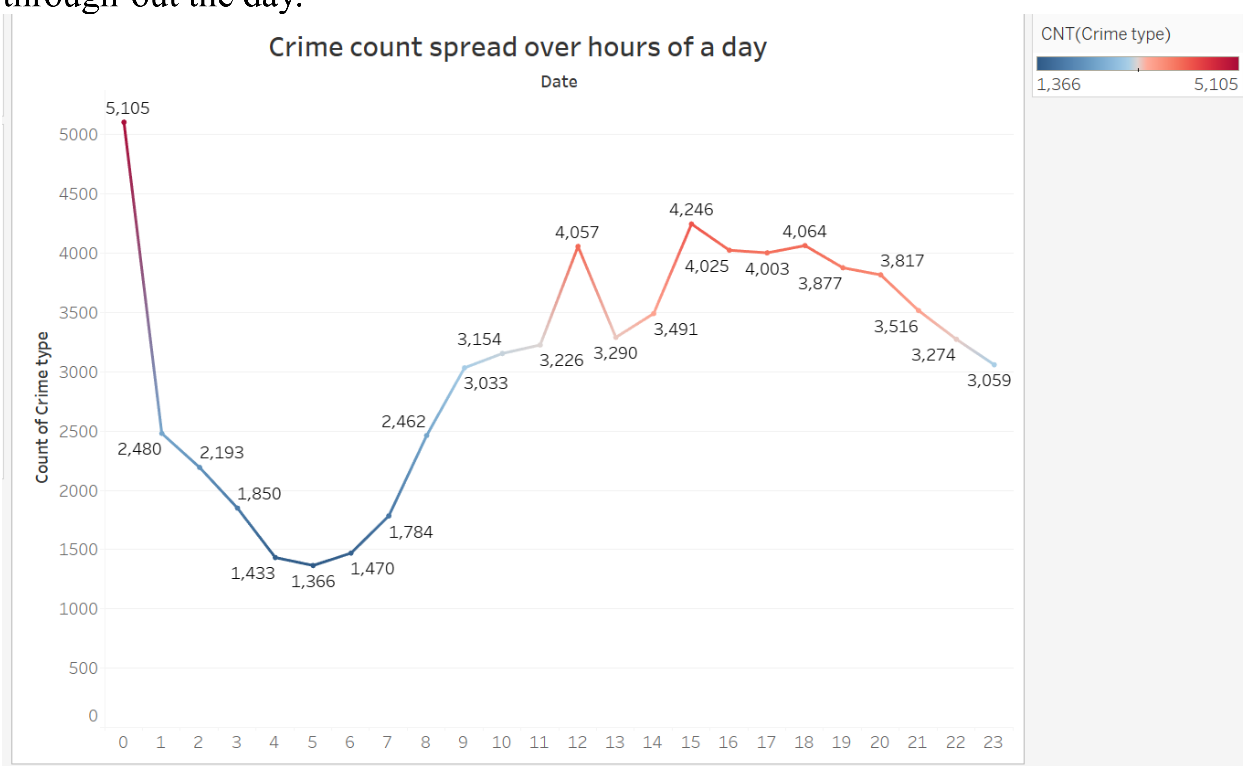


The Line Chart is achieved by adding “Month(Date)” in columns and “Count of Crime type” in rows. We further add the “Count of Crime Type” to color which we can see as a filter where we selected “Red-Blue diverging” and check “Reversed”, we also add the same attribute into “Label” and extend a “Quick table calculation” as “Percentage Difference” and select “Relative to first value.”

Result:
We can clearly see that there is a significant decline in the crime rate in the months starting from May to August (which is about 95% less that that at the start of the year) and there is a huge surge from August to October where the crime rate becomes a little above double than what it was at the beginning of the year.

So with this information we can reject our Null Hypothesis and accept the Alternative Hypothesis which is that:
“Certain periods of the year in 2023 exhibit significantly lower crime rates than others.”

1.(b). Same **Line chart** is used to visualize the distribution of crime count through-out the day.

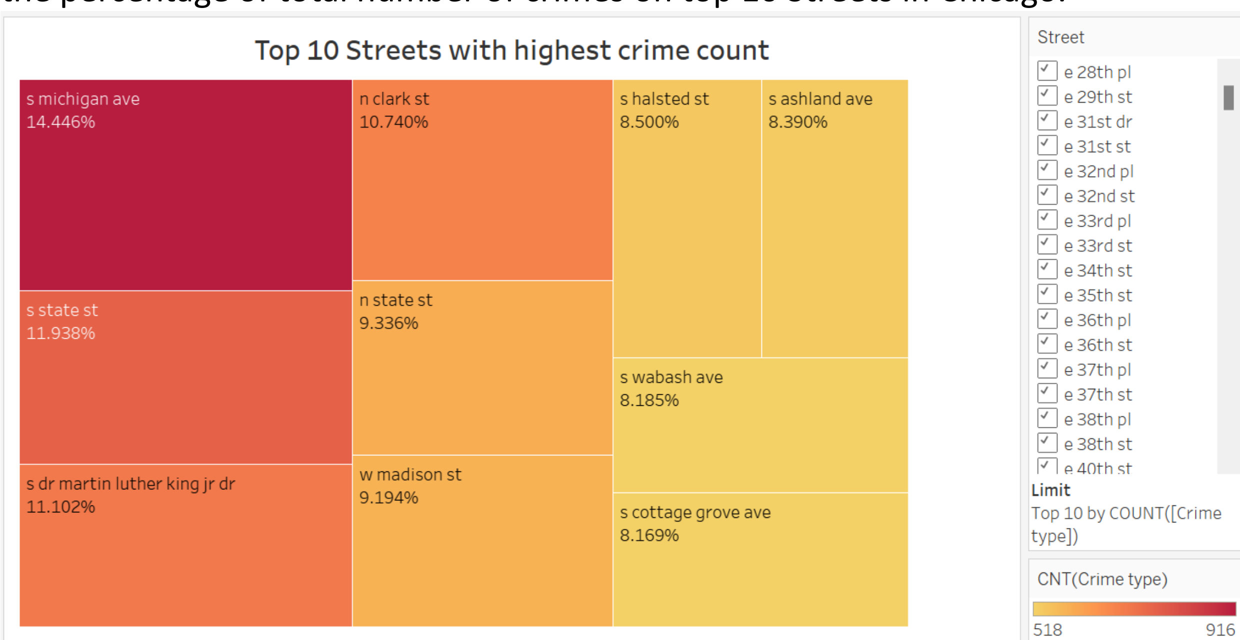


This line chart is achieved by using “Hour (Date)” in columns and “Count of Crime Type” in rows. We then add “Count of Crime Type” attribute to color, and we select “Red-Blue Diverging” as the desired color. We also add the same attribute to Label and it displays the count of crime type on each data point.

Result:
We can see that the morning hours starting from 2AM till 8AM are relatively safer times. And the safest hour is 5AM with crime count associated to that time being 1,366.
We can also infer from the above line chart that most crimes are occurring during the hour 12 AM (represented as 0 in the visualization) and the crime count associated with it being 5,105.

From this information we can confidently accept the Alternative hypothesis which is :
“Crime rates vary significantly depending on the time of day, with certain periods being safer than others.”

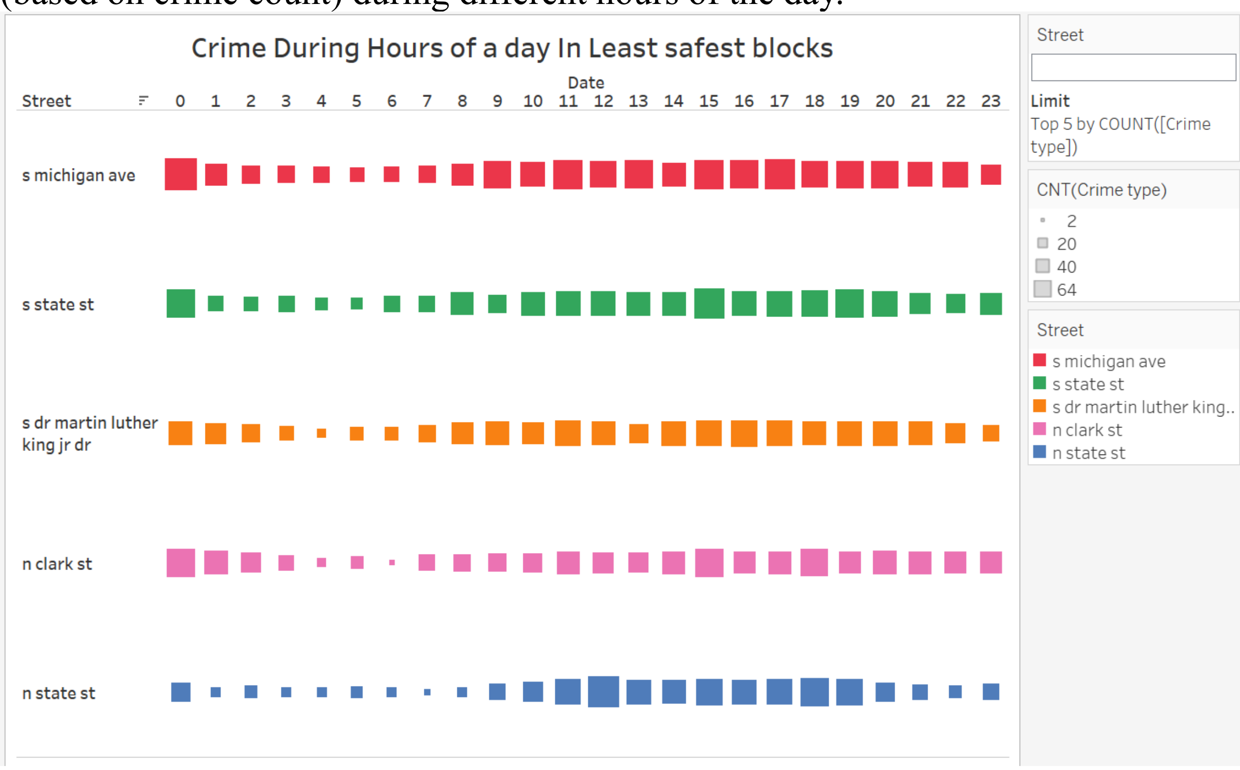
1.(c). A **Tree Map** is used to create the visualization, it is used to represent the percentage of total number of crimes on top 10 Streets in Chicago.



The Tree Map is made by selecting "Street" and Count of "Crime type" attributes and using Tableau's "Show me" to create a ready-made Tree Map. We then filter the "Street" attribute using "Top" based on the field "Crime Type" and its "Count".

Result:
We can see that the street "s michigan ave" has more crime rate related to other streets, almost 14.46% of total crime happens on this street. which makes accept the alternate hypothesis which is:
"The crime rates are significantly higher in few blocks compared to other blocks."

1.(d). **Heat Maps** is used to represent the crime count of top 5 streets (based on crime count) during different hours of the day.



This Heat Map can be made by adding “Hour (Date)” attribute to columns and “Street” attribute to rows. We then use “Count of Crime Type” attribute for sizes and we filter “Street” attribute using “Top” based on “Crime Type” attribute’s “Count”. We also add “Street” attribute to color.

Result:
We can conclude from the above visualization that there is no significant change in crime rate in different Streets/Blocks (here we are considering the top 5 streets based on crime count) when compared to that of overall crime rate during day.

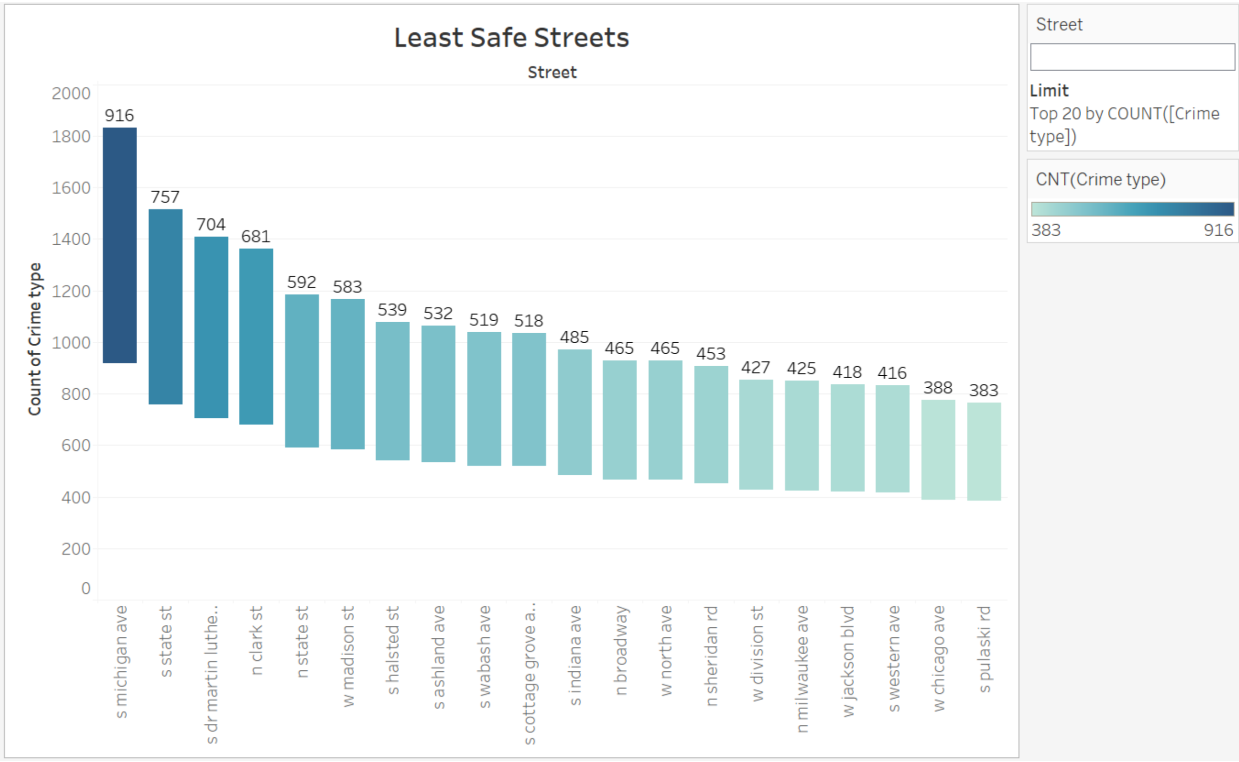
So we accept the Null Hypothesis which is:
“There is no significant difference in crime rates across different times in different Streets.”

2. Identify which street is least safe? What type of crimes are prevalent in the least safe street? Furthermore, how is the most prevalent crime occurring in the day and where is it occurring the most?

Null Hypothesis:
The safety level of streets has no significant impact on crime rates, and there is no difference in the composition, temporal patterns, or spatial distribution of crimes on the least safe street compared to the overall crime situation in Chicago.

Alternate Hypothesis:
a. There is a significant difference in overall crime rates between the least safe streets and other streets in Chicago.
b. The composition of crime types on the least safe street is significantly different from the overall crime distribution in Chicago.
c. The temporal pattern of theft on the least safe street during the day is significantly different from a uniform distribution.
d. Certain locations on the least safe street have a significantly higher occurrence of theft compared to other locations on the street.

2. (a). **Waterfall chart** is used to represent top 20 streets with highest crime count.



The waterfall chart can be created by adding "Street" to columns and "Count of Crime Type" to rows and then selecting Tableau's "Show me" and select "Gantt Chart" we then extend this by adding "Count of Crime Type" to "Sizes". We then Filter "Street" attribute based on "Crime Type" "Count" and select "Top" 20.

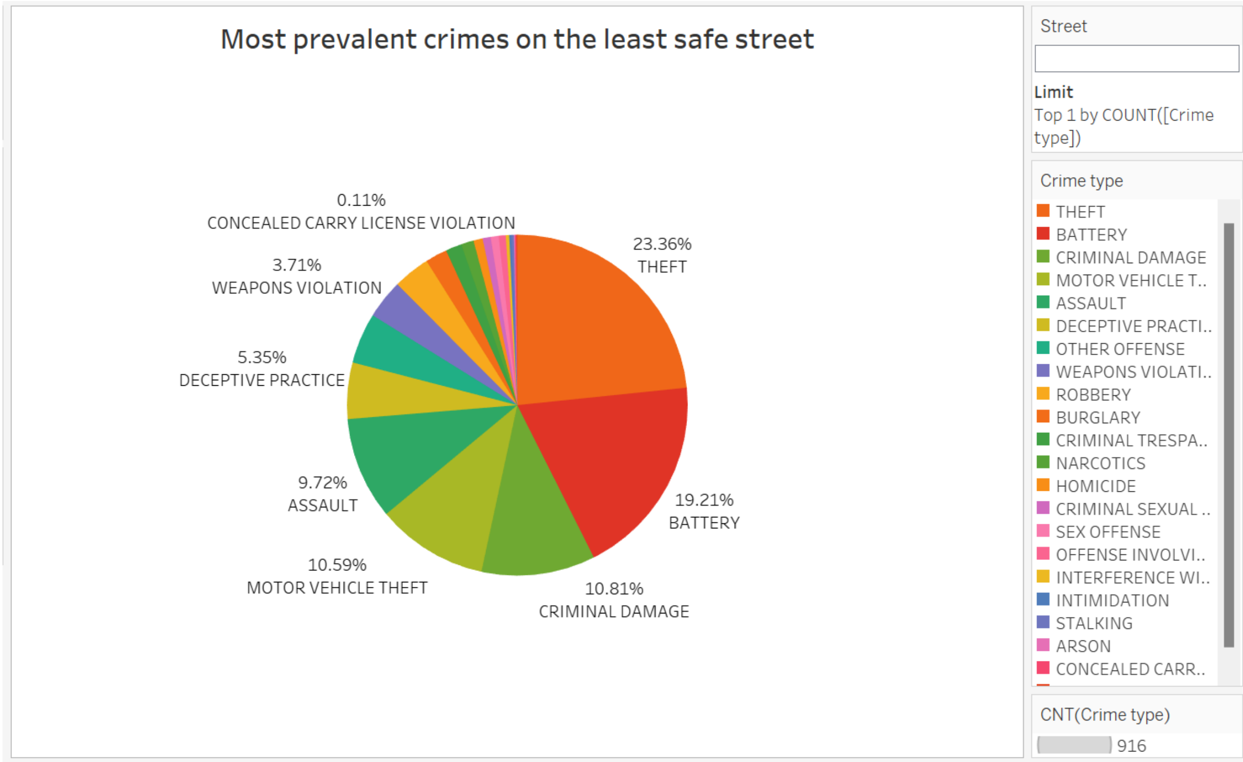
Result:

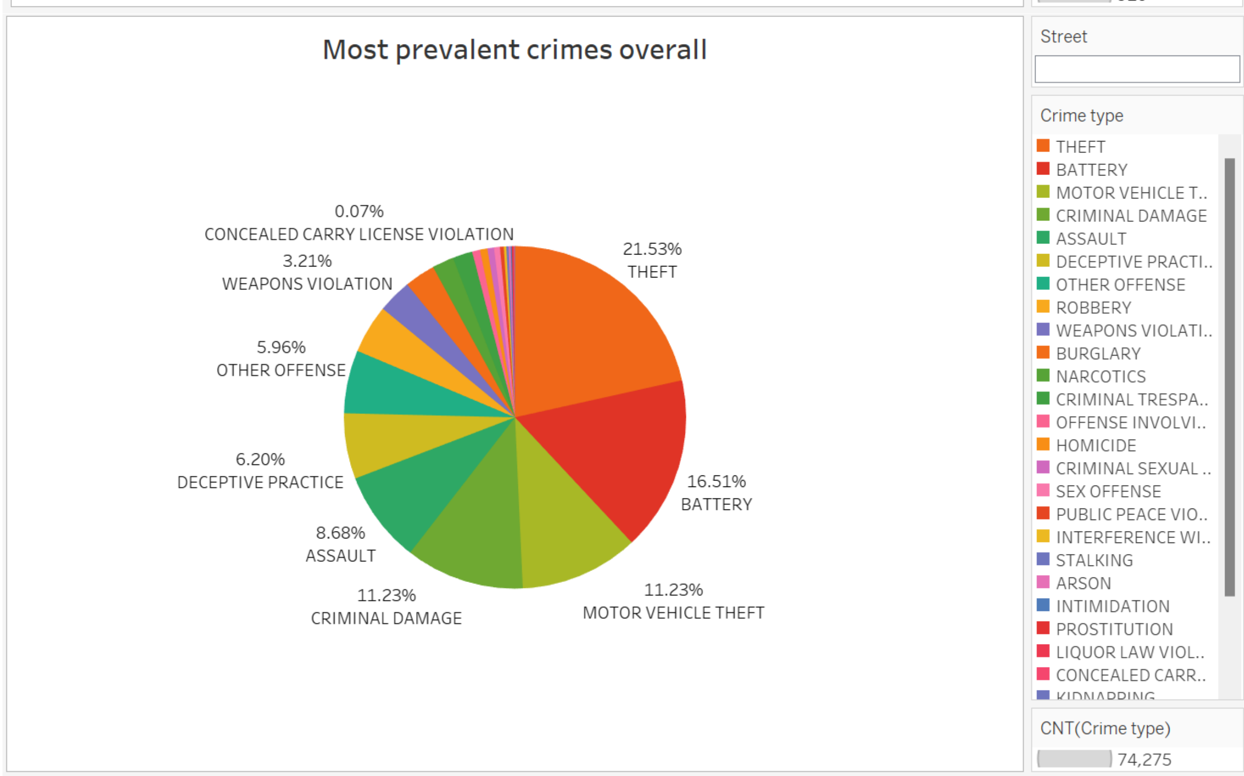
We can clearly see that the trend of crime count in different streets is as portrayed and the street which is least safe or one with highest crime count is s Michigan ave.

So with this information we can reject our null hypothesis and accept the alternative hypothesis which is:

"There is a significant difference in overall crime rates between the least safe streets and other streets in Chicago."

2.(b). **Pie charts** are used to see if the composition of crime types in the least safe street is same as that of composition of overall crime count in all streets combined.



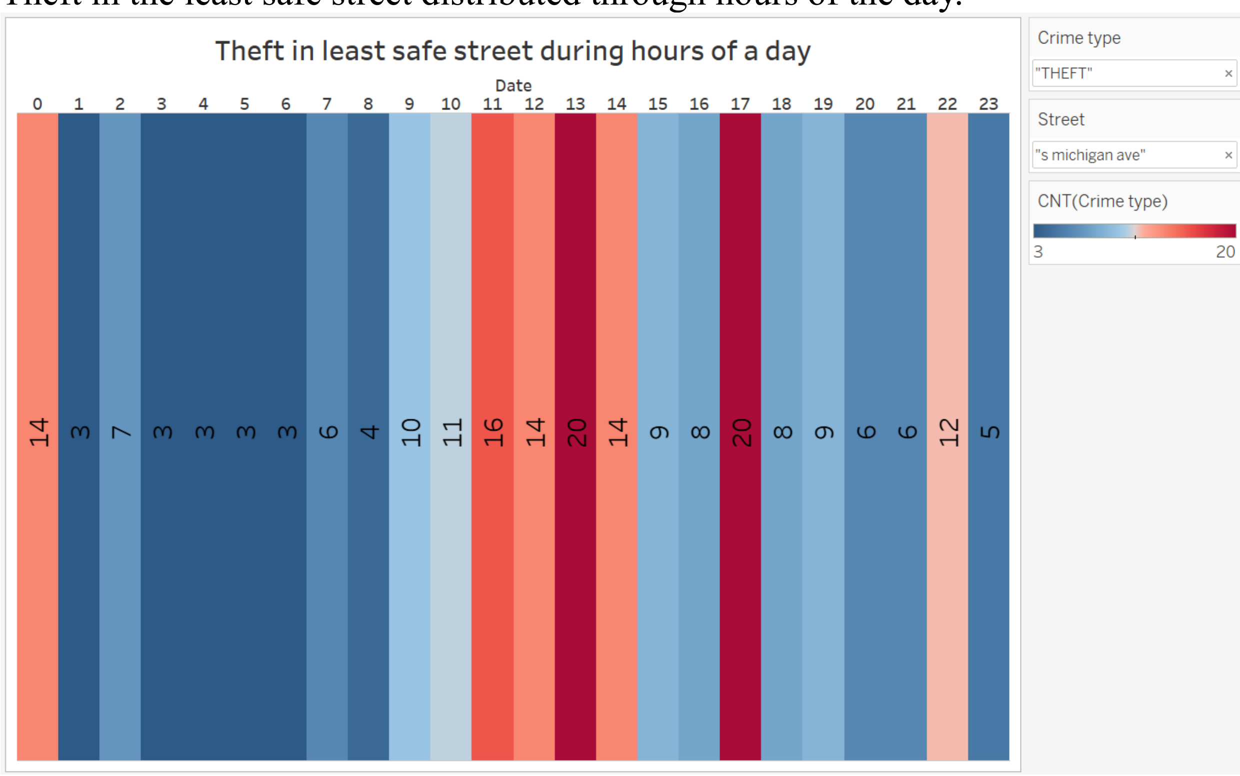


The Pie Charts can be replicated by selecting “Crime Type” and “Count of Crime Type” attributes and using Tableau’s “Show me” select Pie Chart. We then filter one chart based on “Street” to get the Top 1 based on “Crime Type” attribute’s “Count”. We also add “Crime Type” and “Count of Crime Type” attributes to labels, we then extend the “Count of Crime Type” with a “Quick Table Calculation”->” Percentage of Total.”

Result:
From the above pie charts we can see that there is not much difference (no significant difference) in the composition of crime types in the least safest street and overall, and the most prevalent crime type overall and in the least safe street is Theft.

Hence, we can accept out Null hypothesis which is:
“There is no difference in the composition of crimes on the least safe street compared to the overall crime situation in Chicago.”

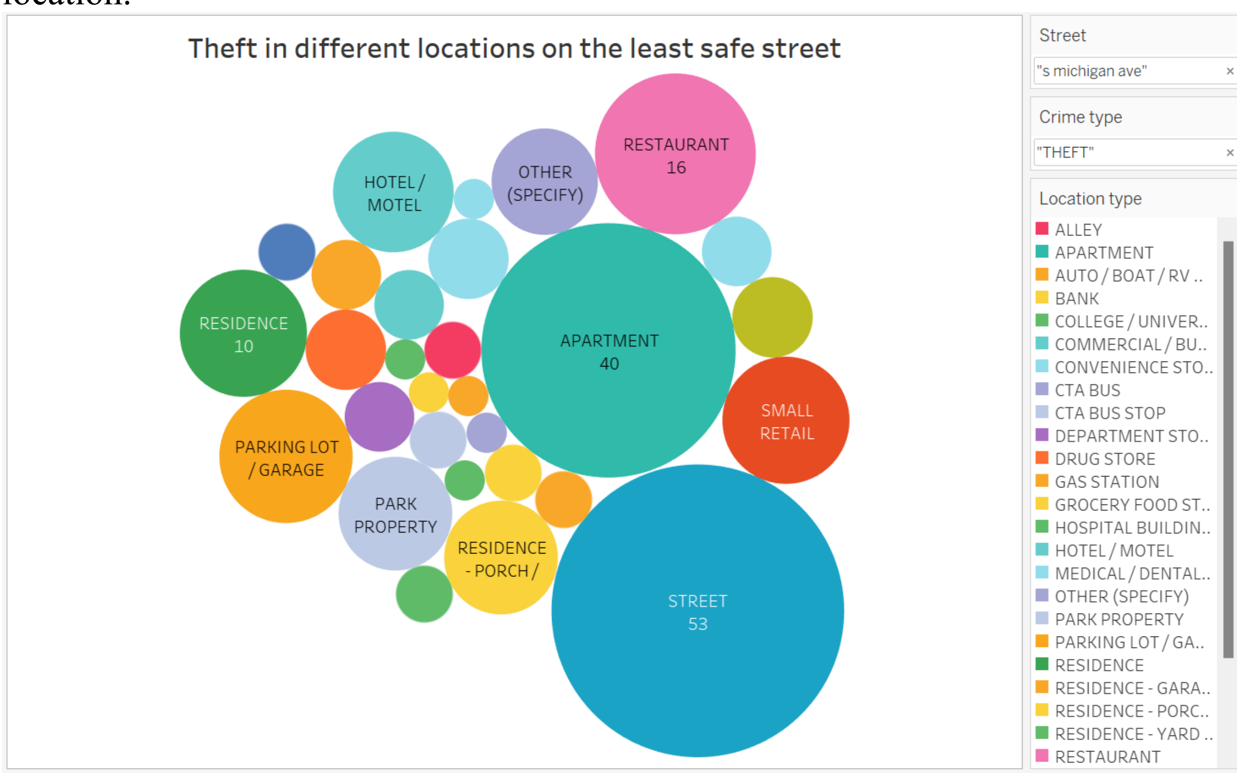
2. (c). **Heat Map (Bars) chart** is used to represent the crime count of Theft in the least safe street distributed through hours of the day.



The Bar Heat Map can be created by selecting the “Hour(Date)” and “Count of Crime Type” and using Tableau’s “Show me” to select “Heat Map”. We then change the “Marks” to “Square”. We also filter “Crime Type” check only “THEFT”, we use “wildcard” in “Street” filter and check “Exactly Contains” -> “s michigan ave”. We then add “Count of Crime Type” to Color and choose “Red-Blue Diverging” as color.

Result:
From the above visualization we can conclude that the hours 13 (1PM) and 17 (5PM) is when Theft occurs most on the least safe street. And the crime count is 20 for both.
We can confidently reject our Null Hypothesis and accept our alternate hypothesis which is:
“The temporal pattern of theft on the least safe street during the day is significantly different from a uniform distribution.”

2. (d). **Bubble chart** is used to see the locations where the most Theft occurs, the size of the bubble depends on the crime count of theft in that location.



This visualization can be replicated by selecting “Location Type” and “Count of Crime Type” attributes and use Tableau’s “Show me” and select

“Bubble Chart”. We then add “Location Type” and “Count of Crime Type” to labels. We also add “Location Type” to colors. We then Filter “Crime Type” and “Street” using “Wild card” {Crime Type: THEFT, Street : s michigan ave}

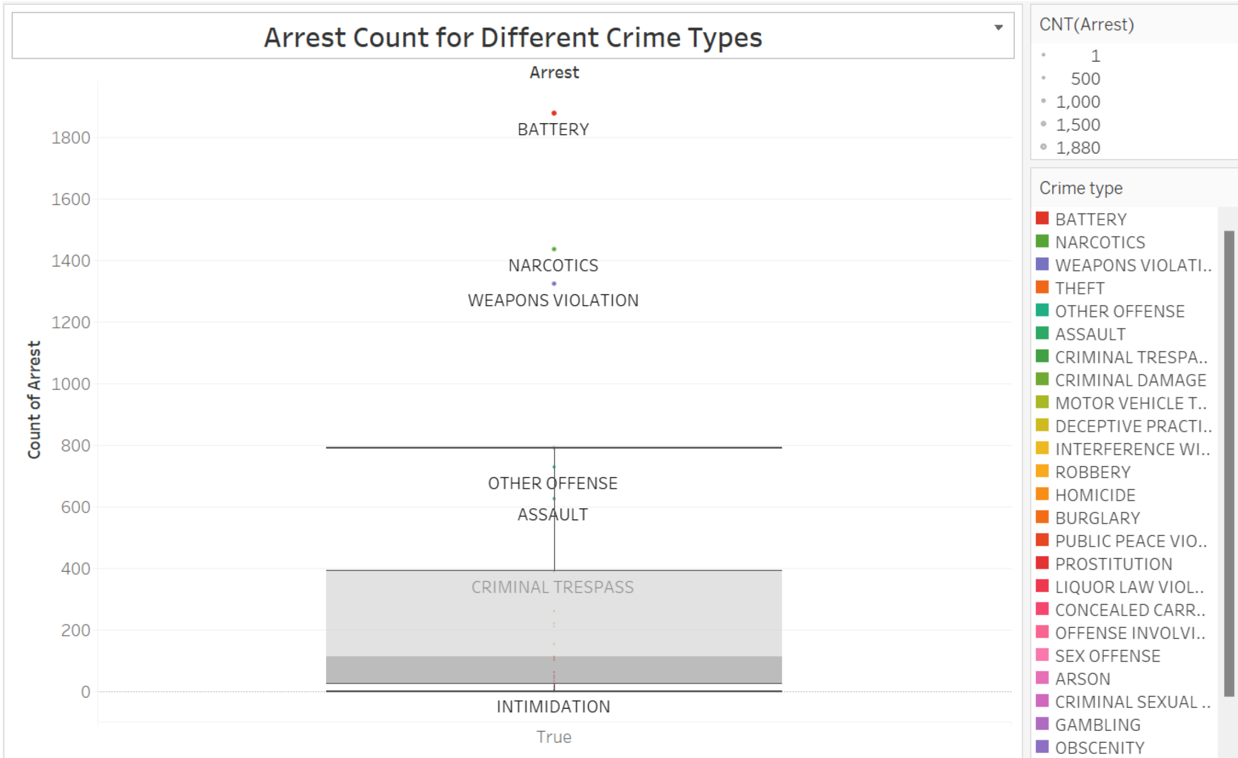
Result:
We can see that Street and Apartment are the major location types where the crime Theft is occurring most in the least safe street.
From the above visualization we can reject our null hypothesis and accept our alternative hypothesis which is:
“Certain locations on the least safe street have a significantly higher occurrence of theft compared to other locations on the street”

3. Since we now know how crimes go, let’s see how the arrests go.
Which type of crimes have the highest arrest count? How are arrests distributed over the hours of a day? Which streets have the highest Arrest count?

Null Hypothesis: The type of crime, temporal patterns, street location, and specific streets have no significant impact on the number of arrests made in Chicago.

Alternative Hypotheses: At least one of the following conditions is true:
1. Certain crime types result in significantly more arrests than others.
2. The temporal patterns of arrests differ significantly among the top 3 crimes.
3. Certain streets have a significantly higher number of arrests compared to others.

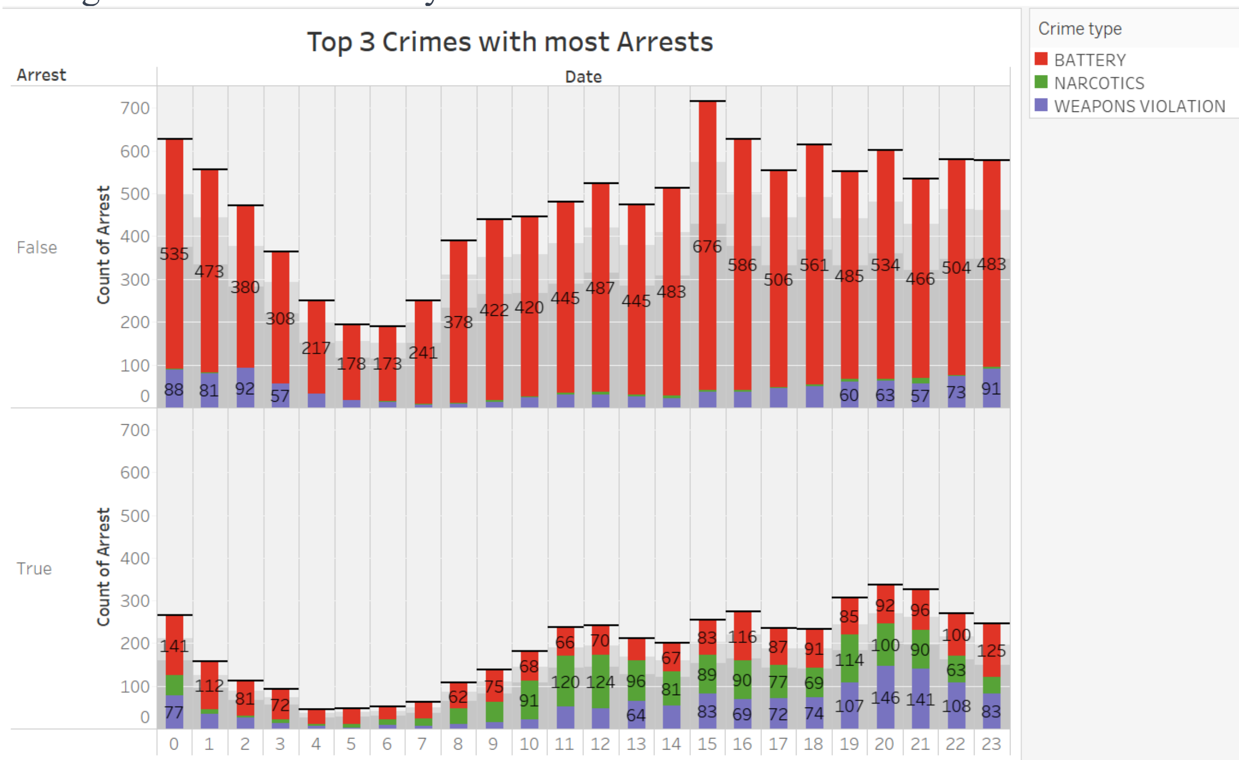
3. (a). **Box–Plot** is used to extract the top three crimes with highest arrest count.



Box plot is achieved by adding "Count of Arrest" to rows and "Crime Type" to columns and selecting "Box Plot" from Tableau's "Show me". We add "Crime Type" to colors, Labels and Detail. And we add "Count of Arrest" to "Size", we then filter "Arrest" and select "True".

Result:
From the above visualization we can see that the outliers in the box at the top are the top crime types with highest arrest count, the crimes are namely; "Arrest" "Narcotics" and "Weapon Violation".
From this we can accept out alternate hypothesis which is:
"Certain crime types result in significantly more arrests than others."

3. (b). **Vertical Bullet Graphs** are used to see how the top 3 crime types (Based on arrest count) have arrest count associated with them spread throughout the hours of a day.

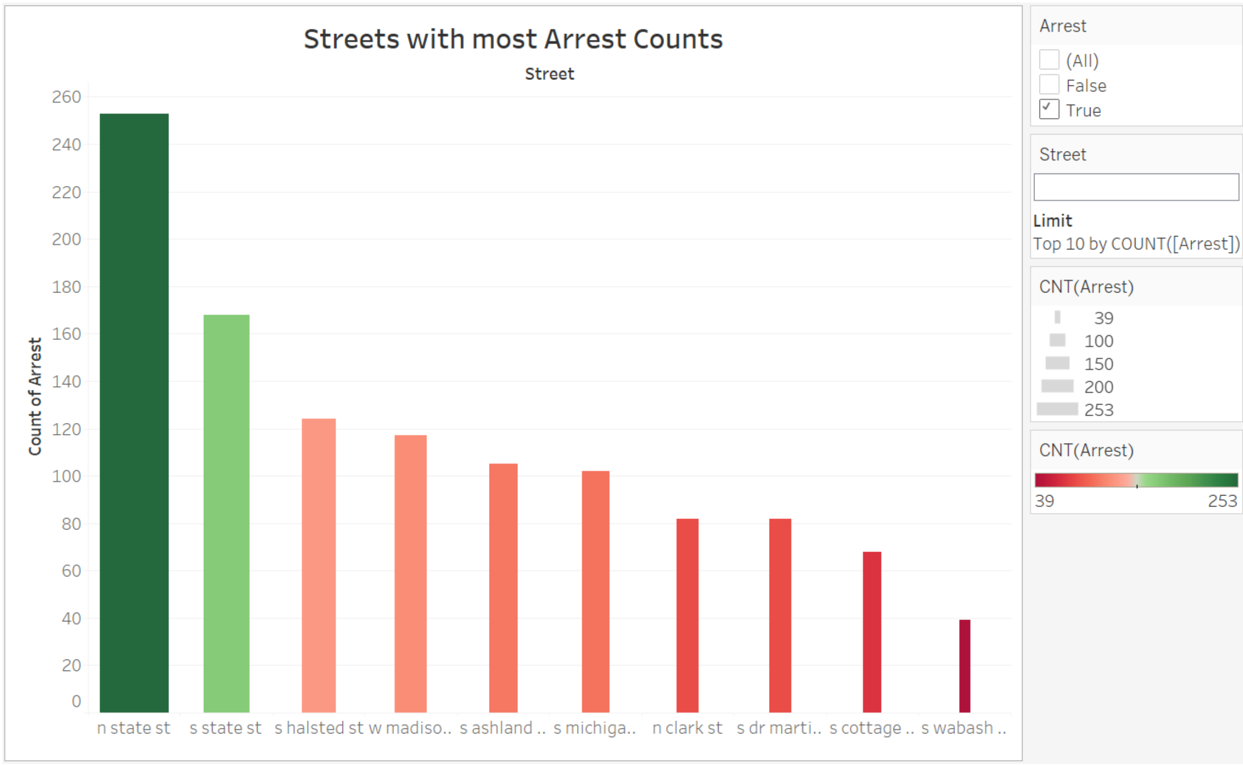


This visualization is created by selecting "Count of Arrest" and "Hour (Date)" attributes and choosing "Bullet Graphs" from Tableau's "Show me". We then add "Crime Type" to colors and filter "Crime Type" to top 3 based on "Arrest" attribute's "Count". Lastly we add "Count of Arrest" to labels.

Result:
From the visualization we can see how each hour of the day is affected by crime (Top 3 crimes bases on Arrest count) and Arrests.
We can notice that "Narcotics" has more Arrest count than Crime count, we can Confidently assume that the state's strategy(Police's Strategy) against "Narcotics" is very strict and is working as intended.
When we turn our focus onto Arrest count for "Battery" and "Weapons Violation" we can see that "Battery" crime type is very prevalent and not most Arrests are happening.

We can now confidently accept our alternative hypothesis which is:
"The temporal patterns of arrests differ significantly among the top 3 crimes."

3. (c). **Bar Graphs** are used to identify the top 10 streets with most Arrest counts.



The Sized Bar Graphs are created by adding "Street" in columns and "Count of Arrest" in rows. We then Filter "Arrest" and select "True" and "Street" attribute's "Top" 10 by "Count of Arrest". We then add "Count of Arrest" into "Size" and "color" where we choose "Red-Green Diverging".

Result:

From the Visualization we can see that the street "n state st" has the highest Arrest Count. Hence, we can accept our alternate hypothesis which is: ***"Certain streets have a significantly higher number of arrests compared to others"***

4. We use **Bar Charts** again, but this time we compare the arrest count and crime count amongst the **“Least safe street” and “Highest Arrest Count Street.”**

“s michigan ave” VS “n state ave”

Result:

We can see that the crime count on “s michigan ave” is more than that on “n state ave” but the arrest count on “s michigan ave” is less that on “n state ave” , we can infer the following possibilities from the above comparison:

1. Community Engagement:

Community engagement and cooperation with law enforcement can influence arrest rates. If the community on "N State St" is more proactive in reporting crimes or assisting law enforcement, it could contribute to a higher arrest rate despite a lower crime count.

2. Patrol Strategies:

Different patrol and crime prevention strategies may be in place on these streets. The lower arrest count on "S Michigan Ave" could be attributed to challenges in law enforcement patrolling or proactive crime prevention initiatives.

3. Data anomalies or Reporting issues:

There could be anomalies in the data or reporting issues that impact the observed patterns. It's essential to ensure data integrity and investigate if there are any discrepancies or errors in the recording of crimes and arrests.

Discussions

- We can see that crime in Chicago in 2023 has spiked in the end of the year (Comparing from January till October) and it is a little above double of what it was at the start of the year.
- The most safest hours in the city are the morning hours and the nights are unsafe with the highest crime count being at midnight.
- "s michigan ave" is the least safe street with the highest crime count with Theft being the most prevalent crime type on this street and it's count peaks during the hours 3PM and 5PM of the day.
- The most arrests happen for the crime types Battery, Narcotics and Weapons Violation.
- The most percent of arrests withe respect to crime count of a particular crime type happen for Narcotics. This means that most of the Narcotics crimes reported end up/ followed by an Arrest.This represents how strict the law enforcement authorities take Narcotics related crimes.
- "n state st" is the street with the highest arrest count.

Conclusion

In conclusion, our research unraveled intricate patterns within Chicago's crime dynamics. Visualizations and Findings supported nuanced hypotheses, showcasing temporal and spatial variations. Notably, specific streets displayed divergent crime dynamics. Insights into arrest patterns further enriched our understanding.

References

- https://www.kaggle.com/code/umeshnarayanappa/exploring-chicago-crimes-2012-2016
- https://github.com/pm831/chicago-crime-severity-modeling