Analyzing Crime Data in the Charlotte Area

Brittany Marchand, Brian Peaver, Danny Barto, Elizabeth Patterson

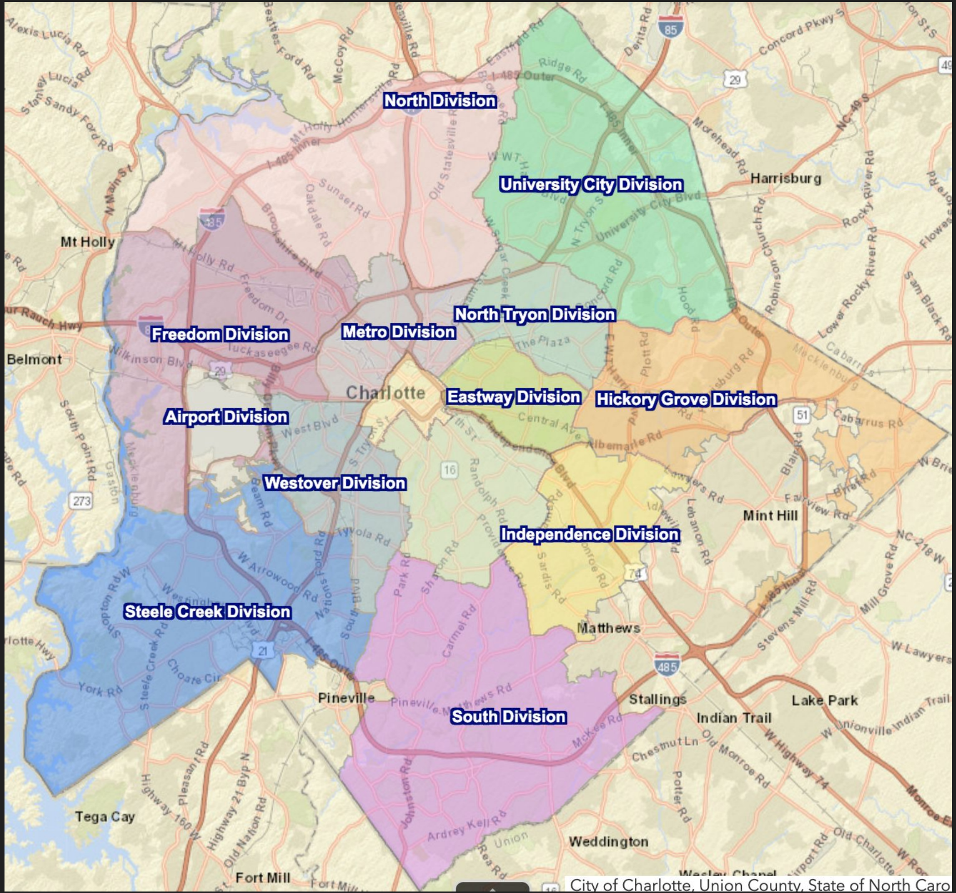
Reason for Choosing Topic and Questions to be Answered

- **Why topic was selected**
  - Team members live in the area
  - Vested interest in public safety
  - Area's steady growth
  - More effective law enforcement

- **Questions we're hoping to answer**
  - In which Division/Patrol ID are crimes occurring the most?
  - Is there a Division/Patrol ID that holds the most violent crime and does that differ from the overall number?
  - What is the most frequent type of crime occuring in the Charlotte area?
  - Is there a season in which crime is the highest?
  - Is there a type of location, in which crime is the highest. (ex. residential vs. commercial, etc)
  - How much violent vs. non-violent crime is there?

Data Used

- **City of Charlotte Open Data Portal**
  - https://data.charlottenc.gov/
- **CMPD Incident Data**
  - Within this data set, main data points of our interest were:
  - -Divison/Patrol ID
  - -NIBRS Incident Code (FBI code categorizing types of crime)
  - -Time(Year/Month) of Incidents
- **We also created our own column of data which binned each incident into Violent vs. Non-Violent Crime. This column was important for our Machine Learning Model**



Cleaning the Data

- **Remove**
  - irrelevant columns
  - Remove nulls
- **Create Consistency**
- **Changed strings to floats in order to improve ML model accuracy**

```
In [83]: clean_date_4_df.CITY.unique()

Out[83]: array(['CHARLOTTE', 'MATTHEWS', 'MECKLENBURG', 'PINEVILLE',
       'HUNTERSVILLE', 'MINT HILL', 'CHARLOTTE, NC 28211', 'CHAROLETTE',
       '28277', 'CHARLOTTE NC', 'CHARLOTE', 'HUMTERSVILLE', 'CHARLOLTTE',
       'CHARLOTTE, NC 28209', 'CHARLOTTE,', 'CHARLOTT', 'MIDLAND',
       'MATTHES', 'MATHEWS', 'CORNELIUS', 'C', 'CHARLOTTTE',
       'CHARLOTTE, NC 28206', 'CHARLOOTE', 'CHARLOTTE, NC',
       'CHARLOTTE,  28211', 'DAVIDSON', 'CHAROLTTE', 'CHRALOTTE',
       'PINEVLE', 'MINT HIL', 'FORT MILL', 'CHARROLTE', 'CHARLTOTE',
       'RT SIDE', 'CHARLOTT3215E', 'MECKLENBRUG', '28273',
       'UNKNOWN/REFUSED', 'BALLANTYNE', 'HUNTERVILLE', 'CHARTLOTTE',
       '110', 'OTHER/NOT LISTED', '28205', 'CHRLOTTE', 'MOUNT HOLLY',
       '28226', 'CHARRLOTE', 'CONCORD', 'CHAARLOTTE', '1640 DEWBERRY TER',
       'CHARLOTTE/NC/28269', 'CHARLOTTLE', 'MECKLENBERG',
       'CHARLOTTEJAVASCRIPT:VOID PT_SU', 'CHARLOTTEE', 'MINTHILL', 'H',
       '28210'], dtype=object)
```

```
In [97]: clean_date_4_df.loc[clean_date_4_df['CITY'].str.contains('CH', case=False), 'CITY_NEW'] = 'Charlotte'
clean_date_4_df.loc[clean_date_4_df['CITY'].str.contains('MAT', case=False), 'CITY_NEW'] = 'Matthews'
clean_date_4_df.loc[clean_date_4_df['CITY'].str.contains('MECK', case=False), 'CITY_NEW'] = 'Mecklenburg'
clean_date_4_df.loc[clean_date_4_df['CITY'].str.contains('HU', case=False), 'CITY_NEW'] = 'Huntersville'
clean_date_4_df.loc[clean_date_4_df['CITY'].str.contains('MINT', case=False), 'CITY_NEW'] = 'Mint Hill'
clean_date_4_df.loc[clean_date_4_df['CITY'].str.contains('PIN', case=False), 'CITY_NEW'] = 'Pineville'
clean_date_4_df.loc[clean_date_4_df['CITY'].str.contains('CORNELIUS', case=False), 'CITY_NEW'] = 'Cornelius'
clean_date_4_df.loc[clean_date_4_df['CITY'].str.contains('MIDLAND', case=False), 'CITY_NEW'] = 'Midland'
clean_date_4_df.loc[clean_date_4_df['CITY'].str.contains('DAVIDSON', case=False), 'CITY_NEW'] = 'Davidson'
clean_date_4_df.loc[clean_date_4_df['CITY'].str.contains('BALLANTYNE', case=False), 'CITY_NEW'] = 'Ballantyne'
clean_date_4_df.loc[clean_date_4_df['CITY'].str.contains('FORT MILL', case=False), 'CITY_NEW'] = 'Fort Mill'
clean_date_4_df.loc[clean_date_4_df['CITY'].str.contains('MOUNT HOLLY', case=False), 'CITY_NEW'] = 'Mount Holly'
clean_date_4_df.loc[clean_date_4_df['CITY'].str.contains('CONCORD', case=False), 'CITY_NEW'] = 'Concord'
clean_date_4_df
```

Database

- **PostgreSQL**- store tabular data
- **SQLAlchemy**- communicate databases and machine learning model
- **Schema**

```
Call_Data
-
ROW_TYPE PK string FK >- Offense_Data.ROW_TYPE
GEOGRAPHY string INDEX FK >- Offense_Data.GEOGRAPHY
CALENDAR_YEAR int FK >- Offense_Data.CALENDAR_YEAR
CALENDAR_MONTH int FK >- Offense_Data.CALENDAR_MONTH
CALL_DESCRIPTION string
CALL_COUNT int


Offense_Data
-
ROW_TYPE PK string
GEOGRAPHY string
CALENDAR_YEAR int
CALENDAR_MONTH int
OFFENSE_DESCRIPTION string
OFFENSE_COUNT int

Incident_Data
-
CMPD_PATROL_DIVISION string
HIGHEST_NIBRS_DESCRIPTION string
PLACE_TYPE_DESCRIPTION string
CLEARANCE_STATUS string
```

Machine Learning

- **Supervised Learning-Classification**
  - Logistic Regression- used to conduct analysis when the dependent variable is binary. Examples are “Yes/No” or “True/False”
- **Categorize violent versus non-violent crimes**
  - Utilize NIBRS to help determine the category

Goal- map out violent crime to see where heaviest to better allocate police resources

Results

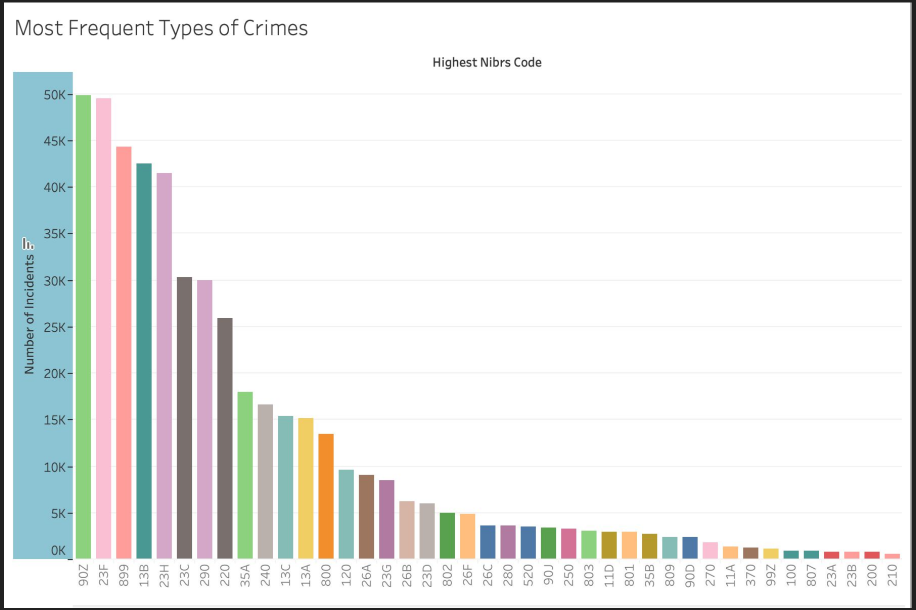
- **Incidents per Division ID/Division ID with highest crime**
  - A total of nearly 500,000 calls across all 17 divisions
  - Divisions 21 and 14 had over 50,000 calls each
    - This represents over 20% of all calls
- **Frequency and Types of Crimes**
  - Close to 50% of all calls are comprised of 5 different types
    - 23F- Theft from Motor Vehicle Property- 10%
    - 90Z- All Other Offenses Person, Property or Society: 10%
    - Other Unsolicited/Non-Criminal- 9%
    - 13B- Simple Assault Person- 9%
    - 23H- All other Larceny Property-8%

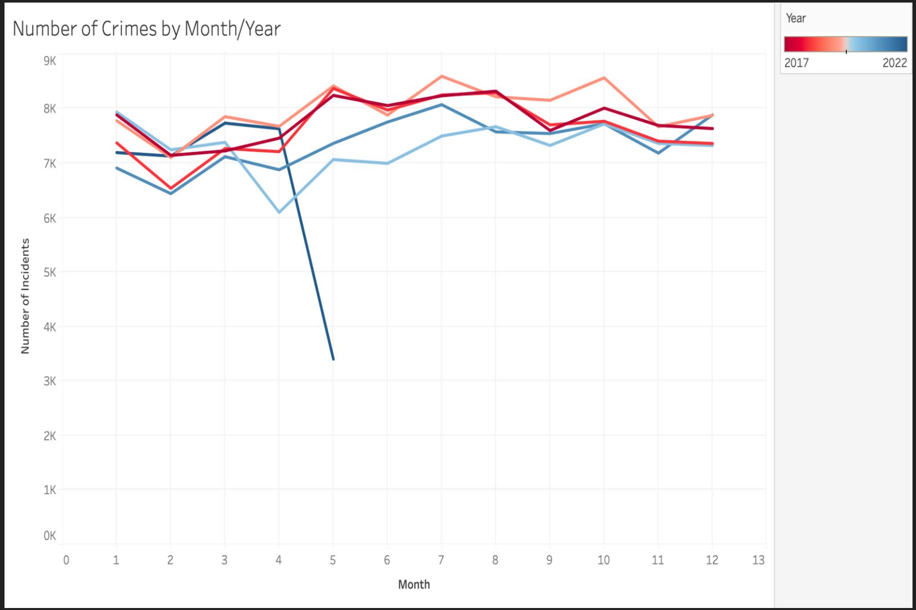
Results (continued)

- **Number of Crimes by Month/Year**
  - The number of incidents is fairly consistent across the timeline
  - We see a there is a slight tendency for there to be calls in the winter and spring months
- **Violent vs. Non Violent**
  - The data demonstrates that the number of violent crimes tends to correlate with the number of incidents in total across all districts
  - Divisions 21 and 14 are the most violent divisions in the CMPD service area
- **Map View**
  - The map view allows the user to further analyze the data using a map of the service area to view the aforementioned data
  - This provides the user with an efficient method to view the data using various geographical inputs
  - This allows for a more granular view of the data as needed

Dashboard

- [CMPD Crime Dashboard]





Summary

- This model can be used to determine the optimal deployment of various resources by looking at where and when the most dangerous incidents are happening
- We determined that out of nearly 500,000 calls across 17 divisions, two of them , divisions 21 and 14 represented over 20% of all calls.
  - We were also able to determine that the most dangerous incidents were occurring in these divisions.
- The most frequent type of crime occurring in the Charlotte area is "Theft from Motor Vehicle." The most frequent violent crime is "Simple Assault"
- A view of the data by season didn't show a significant change in reporting from season to season. The winter and spring months show a slight uptick.
- Since 2017, CMPD has logged nearly 500,000 incidents with an average of about 91,000 per year. It should be noted that there was a spike in 2019 at 95,794 while the subsequent 2 years have logged less than 90,000.
  - So far in 2022 the monthly average of 7,420 is on pace to continue the downward trend and fall somewhere around 87,000 calls
- The data shows that nearly half of all calls come from a personal residence. At these residences calls classified as "violent" are outpacing the "non-violent" calls 55% to 45%