Project 1

Charlotte Mecklenburg's Crime Over Time

If you are an avid or even an occasional watcher of the local news, it seems as if crime is rampant throughout our community and no place is safe. As one of those consumers of news I am interested in exploring the data to determine how “bad” crime is. During this project I want to explore the data and answer these first initial questions: 1) overall how much has crime changed from year to year? 2) Are we seeing changes in crime in certain district (neighborhoods)? 3) What types of crimes are on the rise? 4) Are there locations (residents, terminals, parking deck, etc. where it’s more likely to happen? 5) Lastly, as one of the fastest growing regions in the country, are the changes in crime just a reflection of population growth (as the population grows, we should expect for crime to increase as well)? This last question is simplistic, and you really can not say there is a correlation without much more research but it will be interesting to see.

https://data.charlottenc.gov/datasets/charlotte::cmpd-incidents-1/about

Dataset

The first dataset is available from the city of Charlotte's open data portal. It was downloaded in a CSV format. This link provides additional information about the CMPD incident data used in this project. https://data.charlottenc.gov/datasets/charlotte::cmpd-incidents-1/about It data contains all CMPD incident reports from 2017 through 2024. The contains 688,973 observations (records) and 29 features (attributes) which are 'X', 'Y', 'YEAR', 'INCIDENT_REPORT_ID', 'LOCATION', 'CITY', 'STATE', 'ZIP', 'X_COORD_PUBLIC', 'Y_COORD_PUBLIC', 'LATITUDE_PUBLIC', 'LONGITUDE_PUBLIC', 'DIVISION_ID', 'CMPD_PATROL_DIVISION', 'NPA', 'DATE_REPORTED', 'DATE_INCIDENT_BEGAN', 'DATE_INCIDENT_END', 'ADDRESS_DESCRIPTION', 'LOCATION_TYPE_DESCRIPTION', 'PLACE_TYPE_DESCRIPTION', 'PLACE_DETAIL_DESCRIPTION', 'CLEARANCE_STATUS', 'CLEARANCE_DETAIL_STATUS', 'CLEARANCE_DATE', 'HIGHEST_NIBRS_CODE', 'HIGHEST_NIBRS_DESCRIPTION', 'OBJECTID', and 'GlobalID'

The second data is available from FRED. It is the Resident Population in Charlotte-concord-Gastonia, NC-SC. The link to the data is provide here Resident Population in Charlotte-Concord-Gastonia, NC-SC (MSA) (CGRPOP) | FRED | St. Louis Fed (stlouisfed.org) As the name suggests, the population data including residents from outside of the Charlotte Mecklenburg from 2000 – 2023.

Show code

```
# import data set
data_file = '/content/sample_data/CMPD_Incidents.csv';

CLT_crime = pd.read_csv(data_file, on_bad_lines = 'skip')
```

```
<ipython-input-2-88b79778accc>:4: DtypeWarning: Columns (7) have mixed types. Specify dtype option on import or set low_memory=False.
  CLT_crime = pd.read_csv(data_file, on_bad_lines = 'skip')
```

```
CLT_crime.shape
```

```
(688973, 29)
```

Pre-Processing

Pre-processing is one of the most important steps. By thoroughly cleaning the data, we will improve the accuracy of our model. In addition, it will save us time by removing all errors in advance. The pre-processing began with importing the csv dataset and immediately it created a “ParserError: Error tokenizing data. C error” which means the python process senses some rows have more data than expected; however, no observations were removed. Its shape is (688973, 29)). Then, irrelevant columns were removed. 'X', and 'Y' were removed because they reflected as the decimal version of 'X_COORD_PUBLIC', 'Y_COORD_PUBLIC'. In addition, INCIDENT_REPORT_ID, OBJECTID, and GlobalID. The INCIDENT_REPORT_ID and GlobalID do not provide us with any helpful information. OBJECTID will be replaced with an index during this section. Next, I checked for missing values. As you can see, 4 columns were missing values – ZIP, CMPD_PATROL-DIVISION, DATE_INCIDENT_END and CLEARANCE DATE. The decision was made to drop only ZIP which would have the smallest impact to the project. CMPD_PATROL_DIVISION is a more descriptive version of DIVISION, so we can impute the missing information based on the current values in DIVISION. DATE_INCIDENT_END indicates the date that the incident or cases was resolved. I will impute those missing dates with today’s current date. This will provide an accurate measure of the number of days that a case has been open. For the same reason we will retain CLEARANCE DATE. Next, I checked the data types. All variables were of type “object”, except for YEAR, X_COORD_PUBLIC, Y_COORD_PUBLIC, LATITUDE_PUBLIC,

LATITUDE_PUBLIC, and NPA. For now, I will maintain the current types. Next, I check the value_count() for YEAR. It only contained values from 2017 – 2024. Lastly, the data set was indexed, so the first observation would be row 1 rather than row 0 and named ID.

Next, I checked for missing values. As you can see, 4 columns were missing values – ZIP, CMPD_PATROL-DIVISION, DATE_INCIDENT_END and CLEARANCE DATE. The decision was made to drop column only ZIP which would have the smallest impact to the project. CMPD_PATROL-DIVISION is a more descriptive Division, so we imputed the missing information based on the current values in DIVISION. DATE_INCIDENT_END indicated the incidents or cases that have not been resolved. I will impute those missing dates with today's current date. This will provide an accurate measure of the number of days that a case has been open. For the same reason we will retain CLEARANCE DATE. Next, I checked the data types. All the of variables were of type "object", except for YEAR, X_COORD_PUBLIC, Y_COORD_PUBLIC, LATITUDE_PUBLIC, LATITUDE_PUBLIC, and NPA. For now, I will maintain the current types. Lastly, I check the value_count() for YEAR. It only contained values from 2017 - 2024.

The final shape of the dataset is 688973 observations and 23 attributes.

```
CLT_crime = CLT_crime.drop(['X','Y','INCIDENT_REPORT_ID', 'OBJECTID','GlobalID', 'ZIP'], axis=1)
```

```
CLT_crime.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 688973 entries, 0 to 688972
Data columns (total 23 columns):
 #   Column                     Non-Null Count   Dtype  
---  ------                     --------------   -----  
 0   YEAR                       688973 non-null  int64  
 1   LOCATION                   688973 non-null  object 
 2   CITY                       688973 non-null  object 
 3   STATE                      688968 non-null  object 
 4   X_COORD_PUBLIC             688973 non-null  int64  
 5   Y_COORD_PUBLIC             688973 non-null  int64  
 6   LATITUDE_PUBLIC            688973 non-null  float64
 7   LONGITUDE_PUBLIC           688973 non-null  float64
 8   DIVISION_ID                688973 non-null  object 
 9   CMPD_PATROL_DIVISION       688346 non-null  object 
 10  NPA                        688973 non-null  int64  
 11  DATE_REPORTED              688973 non-null  object 
 12  DATE_INCIDENT_BEGAN        688973 non-null  object 
 13  DATE_INCIDENT_END          520279 non-null  object 
 14  ADDRESS_DESCRIPTION        688970 non-null  object 
 15  LOCATION_TYPE_DESCRIPTION  688973 non-null  object 
 16  PLACE_TYPE_DESCRIPTION     688973 non-null  object 
 17  PLACE_DETAIL_DESCRIPTION   688973 non-null  object 
 18  CLEARANCE_STATUS           688973 non-null  object 
 19  CLEARANCE_DETAIL_STATUS    688973 non-null  object 
 20  CLEARANCE_DATE             277816 non-null  object 
 21  HIGHEST_NIBRS_CODE         688973 non-null  object 
 22  HIGHEST_NIBRS_DESCRIPTION  688973 non-null  object 
dtypes: float64(2), int64(4), object(17)
memory usage: 120.9+ MB
```

```
CLT_crime.index = [x for x in range(1, len(CLT_crime.values)+1)]

# add index field name
CLT_crime.index.name = 'id'
CLT_crime.head(3)
```

id	YEAR	LOCATION	CITY	STATE	X_COORD_PUBLIC	Y_COORD_PUBLIC	LATITUDE_PUBLIC	LONGITUDE_PUBLIC	DIVISION_ID	CMPD_PATROL_D
1	2017	10500 TURKEY POINT DR	CHARLOTTE	NC	1405570	573264	35.308755	-80.992632	11	
2	2022	1000 N CALDWELL ST	CHARLOTTE	NC	1454066	544139	35.231309	-80.828305	06	
3	2019	100 E MCCULLOUGH DR	CHARLOTTE	NC	1476909	568896	35.300454	-80.753286	14	Unive

3 rows × 23 columns

```
CLT_crime.shape
```

```
(688973, 23)
```

Visualizations

The first visualization is to see the number of incidents from 2017 – 2024. It does indicate overall there has been a steady increase in incidents within Charlotte Mecklenburg.

```
plt.hist(CLT_crime['YEAR'], bins=8,color= 'purple', edgecolor='black')

# Add labels and title
plt.xlabel('Values')
plt.ylabel('Frequency')
plt.title('Total Incidents by Year')
```

```
Text(0.5, 1.0, 'Total Incidents by Year')
```