capstone_project

July 24, 2021

Modeling and Forecasting Crime Rate in Colorado

Data Science Capstone Project, putting it all together * Student name: Elena Kazakova * Student pace: Full-time * Cohort: DS02222021 * Scheduled project review date: 07/26/2021 * Instructor name: James Irving * Application url: TBD

TABLE OF CONTENTS

1 Introduction

1.1 Outline of the Project

The goal of this project is to provide transparency, create easier access, and expand awareness of criminal data both for general and categorical crime, improve resources allocation for law enforcement, and provide a foundation to help shape public policy and preventive measures with the result of a safer state. The "Crime in Colorado application" should help to discover available data through visualizations and statistics.

1.2 Description of sub-notebooks

It is the main notebook for the Capstone Project, Crime in Colorado. It combines all the information about the project while separate notebook have code addressing sections of the project. The parts can be found in the following notebooks: 1. Part 0, creation of SQLite database with the original data. 2. Part I, preprocessing the data in the databases' tables and building DataFrames, SQL part. 3. Part II, preprocessing of the data in DataFrames and EDA. 4. Part III, modeling of the General Crime rate. 5. [Part IV]((capstone_project_part4.ipynb), modeling categorical crime rates.

If you are running this notebook without restarting the kernel replace '%load_ext autoreload' in imports with '%reload_ext autoreload'

2 Obtain

The code for processing the original data and creating the database is in part ZERO notebook.

2.1 Data Source

Data is from FBI Crime Data Explorer NIBRS data for Colorado from 2009-2019 The data dictionary is and a record descriptiopn are available.

The description of the main and reference tables is in data/README.md file.

2.2 SQLite Database

2.2.1 Changes needed

The FBI implemented some changes to the files structure in 2016 and removed the sqlite create and load scripts from the zip directories. Another fact worth mentioning is that files 'nibrs_property_desc.csv' from 2014 and 2015 have duplicated nibrs_property_desc_ids (unique identifier in the nibrs_property_desc table) which complicated the loading of the data.

All 2016-2019 files need to be cleaned up because FBI changed the file format. There is a YEAR column that needs to be removed as well as the legacy columns from the previous years need to be added up. It's a tedious job and it needs to be done once and the files need to be backed up.

In order to clean the tables up the following needs to be done

1. Remove all **DATA_YEAR** columns from each file, it's the first column

2. Files that do not need any changes beyond **DATA_YEAR** column removal

   nibrs_arrestee_weapon.csv nibrs_bias_motivation.csv nibrs_criminal_act.csv nibrs_property_desc.csv nibrs_suspect_using.csv nibrs_suspected_drug.csv nibrs_victim_circumstances.csv nibrs_victim_injury.csv nibrs_victim_offender_rel.csv nibrs_victim_offense.csv nibrs_weapon.csv

3. in **nibrs_arestee.csv file**:

a. between **ARRESTEE_SEQ_NUM** and **ARREST_DATE** there should be an **arrest_num column**
b. Between **CLEARANCE_IND** and **AGE_RANGE_LOW_NUM** should be a **ff_line_number** column.

4. in **nibrs_incident** file: a.between **NIBRS_MONTH_ID** and **CARGO_THEFT_FLAG** column **incident_number** b.between **DATA_HOME** and **ORIG_FORMAT** column **ddocname** c.between **ORIG_FORMAT** and **DID** column **ff_line_number**

5. in **nibrs_month.csv** file: a.between **REPORT_DATE** and **UPDATE_FLAG** add **prepared_date** column b.between **ORIG_FORMAT** and **DATA_HOME** column **ff_line_number** c.column **MONTH_PUB_STATUS** removed

6. in **nibrs_offender.csv** file: a.between **ETHNICITY_ID** and **AGE_RANGE_LOW_NUM** column **ff_line_number**

7. in **nibrs_offense.csv** file:

   a. the last column **ff_line_number** should be added

8. in **nibrs_property.csv** file:

   a. the last column **ff_line_number** should be added

9. in **nibrs_victim.csv** file:

   a. between **RESIDENT_STATUS_CODE** and **AGE_RANGE_LOW_NUM** two columns **agency_data_year** and **ff_line_number** (in that order) should be added

2.2.2 Database

All the tables from 2009-2019 incidents in Colorado is in production1 database in data/sqlite/db folder. The database can be used moving forward.

3 Scrub

3.1 Scrubbing data in the database tables

Notebook with Part I of the project. Its' goal is to pre-process data in the SQLite database in order to use it for building DataFrames in the modeling part of the project.

1. There were 44 table uploaded to the database

   agencies agency_participation cde_agencies nibrs_activity_type nibrs_age nibrs_arrest_type nibrs_assignment_type nibrs_bias_list nibrs_location_type nibrs_offense_type nibrs_prop_desc_type nibrs_victim_type nibrs_circumstances nibrs_cleared_except nibrs_criminal_act nibrs_criminal_act_type nibrs_drug_measure_type nibrs_ethnicity nibrs_injury nibrs_justifiable_force nibrs_prop_loss_type nibrs_relationship nibrs_suspected_drug_type nibrs_using_list nibrs_weapon_type ref_race ref_state nibrs_arrestee nibrs_arrestee_weapon nibrs_bias_motivation nibrs_month nibrs_incident nibrs_offender nibrs_offense nibrs_property nibrs_property_desc nibrs_suspect_using nibrs_suspected_drug nibrs_victim nibrs_victim_circumstances nibrs_victim_injury nibrs_victim_offender_rel nibrs_victim_offense nibrs_weapon

2. The following 24 tables were dropped right away as irrelevant:

   nibrs_month nibrs_justifiable_force nibrs_arrest_type nibrs_drug_measure_type nibrs_injury nibrs_suspect_using nibrs_suspected_drug nibrs_suspected_drug_type nibrs_using_list nibrs_arrestee nibrs_arrestee_weapon nibrs_activity_type nibrs_assignment_type nibrs_property nibrs_property_desc nibrs_prop_loss_type nibrs_victim_injury nibrs_prop_desc_type nibrs_circumstances nibrs_victim_circumstances ref_state nibrs_criminal_act nibrs_criminal_act_type nibrs_victim_offense

3. Processing of separate tables
   a. tables **agencies** and **cde_agencies** compared:
      i. only **cde_agencies** is left
      ii. table **agencies** is a subset of cde_agencies and does not have any location information
   b. **nibrs_incident table:** i. table renamed to **incident_main** ii. 4 fields were left in the table incident_id offense_id agency_id incident_hour iii. 2 fields were added based on agency_id from cde_agencies table primary_county icpsr_zip iv. None value in incident_hour replaced with 25 to be able to separate these records from the rest
   c. **nibrs_offense table:** i. table renamed to **offense_main** ii. 4 fields were left in the table offense_id incident_id offense_type_id location_id iii. location_id was replaced with location_name from nibrs_location table iv. offense_type_id was replaced with 3 based on nibrs_offense_type table values offense_name crime_against offense_category_name
   d. **nibrs_offender table:** i. table renamed to **offender_main** ii. 7 fields were left in the table offender_id incident_id age_id age_num sex_code race_id ethnicity_id iii. Replaced with values from the reference tables age_id with age_name from nibrs_age table race_id with race_desc from ref_race table ethnicity_id with ethnicity_name from nibrs_ethnicity table sex_code was spelled out to ‘Female’, ‘Male’, and ‘Unknown’
   e. **nibrs_victim table:** i. table renamed to **victim_main** ii. 9 fields were left in the table victim_id incident_id victim_type_id age_id age_num sex_code race_id ethnicity_id residence_status_code iii. Replaced with values from the reference tables: victim_type_id with victim_type_name from nibrs_victim_type age_id with age_name from nibrs_age table race_id with race_desc from ref_race table ethnicity_id with ethnicity_name from nibrs_ethnicity table sex_code was spelled out to ‘Female’, ‘Male’, and ‘Unknown’ residence_status_code was spelled out to ‘Resident’,‘Non-Resident’, and ‘Unknown’
   f. **nibrs_weapon table**: i.table renamed to **weapon_main**
      ii. 2 fields were left in the table offense_id weapon_id
      iii. Replaced with values from the reference table weapon_id with weapon_name from nibrs_weapon table mapped ‘Firearm’ ‘Handgun’ ‘Rifle’ ‘Shotgun’ ‘Personal Weapons’ ‘Other Firearm’ to ‘Non-automatic firearm’<br mapped ‘Unarmed’ ‘None’

to ‘Unarmed’ mapped anything with ‘Automatic’ to ‘Automatic Firearm’

g. **nibrs_bias_motivation**
   i. renamed to **bias_main**
   ii. replaced with values from reference table: bias_id with bias_name from nibrs_bias

h. **nibrs_victim_offender_rel**
   i. table renamed to **victim_offender_rel**
   ii. 2 fields were left offender_id victim_id
   iii. replaced with values from the reference tables: relationship_id with relationship_name from nibrs_victim_type from nibrs_relationship

4. The following tables were dropped because the information from them was used in incident, offender, victim and weapon tables and they were no longer needed:

   agencies agency_participation nibrs_age nibrs_victim_type nibrs_ethnicity ref_race nibrs_weapon_type ’nibrs_bias_list nibrs_location_type nibrs_offense_type nibrs_cleared_except nibrs_relationship nibrs_bias_motivation

5. The remaining tables were:

   incident_main offender_main victim_main weapon_main cde_agencies bias_main offense_main victim_offender_rel

6. These tables were made into DataFrames and saved as pickle files in /data/pickled_dataframes as:

   incident.pickle offender.pickle victim.pickle weapon.pickle cde_agencies bias.pickle offense.picklen relationship.pickle

3.2 Scrubbing data in the dataframes

Notebook with Part II of the project. Its’ goal is to pre-process data in the dataframes created in Part I in order to use the data in the modeling part of the project.

Dataframes processed:

1. **df_incident** a. timestamp column turned from string to datetime b. 548 duplicate incident_id found, records from the earlier date retained. Duplicate incident_ids is most probably a human error when the system got switched to another format in 2016. c. There are no NaN values but ”‘(empty string) values are present in primary_county and icpsr_zipcode fields i. Due to the fact that all primary_county missing values are associated with 80215 zip code, which belongs to Jefferson county. I am filling in these records primary county with ’Jefferson’ string ii. The missing zip codes belong to the following agencies: >agency_id=1982: Fort Lewis College, located in 81301 zip code agency_id=23131: South Metro Drug Task Force, located in 80160 zip code agency_id=25314: Gypsum Police Department, located in 81637 zip code

2. **df_offense** a. No duplicate IDs, NaN values or empty strings

3. **df_victim** a. The same person can be a victim in several incidents therefore I was only checking for duplicates with victim_ids AND incident_ids; no duplicates were found b. There are empty strings in the **age_num** column i. Empty string values in the age_num column of victims with types ‘Society/Public’, ‘Business’, ‘Government’,

‘Other’,‘Unknown’, Financial Institution’, and ‘Religious Organization’ were replaced with 999. ii. Empty string values in the age_num column of victims with types ‘Law Enforcement Officer’, ‘Individual’ AND age_group equal ‘Unknown’ were replaced with 999. iii. Empty string values in the age_num column of victims with of type ‘Individual’ AND age_group in (‘7-364 Days Old’,‘Under 24 Hours’,‘1-6 Days Old’) were replaced with 0. iv. Empty string values in the age_num column of victims with of types ‘Law Enforcement Officer’, ‘Individual’ AND age_group ‘Over 98 Years Old’ were replaced with 99. c. There are empty strings in the **sex_code** column i. Empty string values in the sex_code column of victims with of types ‘Society/Public’, ‘Business’, ‘Government’, ‘Other’,‘Unknown’, Financial Institution’, and ‘Religious Organization’ were replaced with ‘NA’ value. d. There are empty strings in the **resident_status_code** column i. The empty string values in the resident_status_code column of victims with of types ‘Society/Public’, ‘Business’, ‘Government’, ‘Other’,‘Unknown’, Financial Institution’, and ‘Religious Organization’ were replaced with ‘NA’ value. ii. The empty string values in the resident_status_code column of victims with of types ‘Law Enforcement Officer’, ‘Individual’ were replaced with ‘Unknown’ value. e. There are NaN values in the **race** column i. The NAN values in the race column of victims with of types ‘Society/Public’, ‘Business’, ‘Government’, ‘Other’,‘Unknown’, ‘Financial Institution’, and ‘Religious Organization’ were replaced with ‘NA’ value. f. There are NaN values in the **ethnicity** column i. The NaN values in the ethnicity column of victims with of types ‘Society/Public’, ‘Business’, ‘Government’, ‘Other’,‘Unknown’, Financial Institution’, and ‘Religious Organization’ were replaced with ‘NA’ value. ii. The NaN values in the ethnicity column of victims with of types ‘Law Enforcement Officer’ & ‘Individual’ were replaced with ‘Unknown’ value. g. There are NaN values in the **age_group** column i. The NAN values in the age_group column of victims with of types ‘Society/Public’, ‘Business’, ‘Government’, ‘Other’,‘Unknown’, ‘Financial Institution’, and ‘Religious Organization’ were replaced with ‘NA’ value. h. The following columns were renamed >age_num to victim_age sex_code to victim_sex resident_status_code to victim_resident_status race to victim_race age_group to victim_age_group ethnicity to victim_ethnicity

4. **df_offender** a. The same person can be an offender in several incidents therefore I was only checking for duplicates with offender_ids AND incident_ids; no duplicates were found. b. There are empty strings in the **age_num column** i. Empty string in the age_num of offender table with age_group values equal ‘Over 98 Years Old’ were replaced with 99 value. ii. Empty string in the age_num of offender table with age_group values equal ‘Unknown’ were replaced with 999 value. c. There are empty strings in the **sex_code column** i. Empty string values in the sex_code column of offender were replaced with ‘Unknown’ value. d. There are NaN values in the **race** column i. The NaN value in the race column of offender table will be replaced with Unknown value. e. There are NaN values in the **ethnicity** column i. The NaN value in the ethnicity column of offender table will be replaced with ‘Unknown’ value. f. There are NaN values in the **age_group column** i. The NaN value in the age_group column of offender table will be replaced with Unknown value. Spot checking the records did not generate any insights. All those offenders are simply not known, never got identified. g. The following columns were renamed >age_num to offender_age_age sex_code to offender_sex race to offender_race age_group to offender_age_group ethnicity to offender_ethnicity

5. **df_weapon** a. There can be several types of weapons used in one offense. For the sake of simplicity I will drop duplicates from the table. b. No duplicates, empty strings or NaN

values

6. **df__bias** a. There can be several types of biases associated with one offense. The number of duplicates is low, 15. For the sake of simplicity the duplicates were dropped from the table.
7. **df__rel** a. There are 2289 NaN values in the relationship column b. NaN values in the relationship column were replaced with 'Relationship Unknown'.

Dataframes were saved to data/pickled__dataframes as >incident__clean.pickle offense__clean.pickle victim__clean.pickle offender__clean.pickle weapon__clean.pickle bias__clean.pickle rel__clean.pickle

1. Offense, incident, bias and weapon DataFrames were combined into one for the Times-series analysis 2. Offender, victim, and relationship DataFrames were set aside for the dashboard application.

4 Explore

4.1 General exploratory analysis of the data

There were 3201143 records of offenses in Colorado between 2009 and 2019

[1]:
```
import pandas as pd
import numpy as np
import matplotlib
import matplotlib.pyplot as plt

import pickle
import os
import json

import warnings
warnings.filterwarnings(action='ignore', category=FutureWarning)

from functions_all import *

%reload_ext autoreload
%autoreload 2
%matplotlib inline
```

[2]:
```
with open('images/pickled_figs/crime_cat.pickle', 'rb') as f:
    fig=pickle.load(f)

fig.show()
```

The plot above indicates that Larceny/Theft crime category is the most abundant among all crime categories, followed by the Destruction/Damage/Vandalism of Property. It is quite vivid that almost all of the crime categories have a seasonal component to them. Most of the crime categories had a downturn during 2019—all but the Weapon Law Violations.

[3]:
```python
with open('images/pickled_figs/weapons.pickle', 'rb') as f:
    fig=pickle.load(f)

fig.show()
```

The plot above indicates that most of the committed offenses involve non-automatic firearms. The plot does not include any of the offenses which by their nature could not have any weapon associated with them (like Fraud or Prostitution).

[4]:
```python
with open('images/pickled_figs/counties.pickle', 'rb') as f:
    fig=pickle.load(f)

fig.show()
```

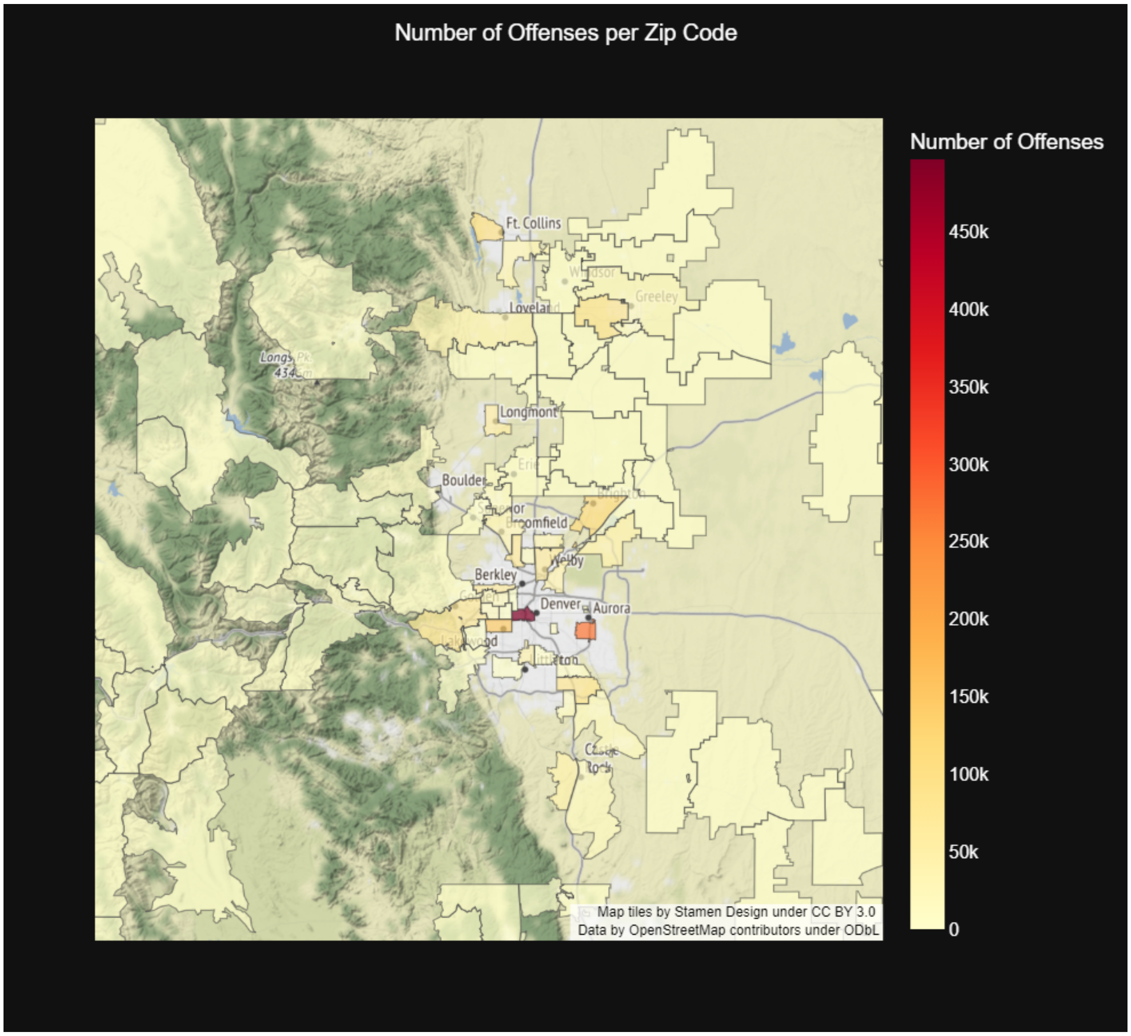
The plot above indicates that the counties with the highest number of commited offenses are 1. Denver 2. El Paso 3. Arapahoe 4. Adams 5. Jefferson The rest of the counties are trailing significantly behind. It is worth mentioning that the plot reflects the overall number of offenses over ten years. The distribution of the number of offenses per country per each year could be slightly different than on the plot above.

[5]:
```python
with open('images/pickled_figs/county_map.pickle', 'rb') as f:
    fig=pickle.load(f)
fig.show()
```

The same information is being conveyed by the map above. All counties with high crime rates are colored darker; the continuous color scale is included. One can see that the order of the sequence is the same. The main advantage of presenting similar information on a map is giving a viewer better spatial perception.

[6]:
```python
with open('images/pickled_figs/zips.pickle', 'rb') as f:
    fig=pickle.load(f)
fig.show()
```

The next plot displays the zip codes with the most offenses reported over the decade (2009-2019). The zip codes are ordered by the crime rate. However, it is worth mentioning that this information might be misleading because of the way law enforcement agencies report their data to the FBI. Zipcodes are associated not with the geographic location of an incident but rather with a reporting agency's geographic location, which accumulates the offense information for several zip codes where offenses occur.




```
[7]: with open('images/pickled_figs/hours.pickle', 'rb') as f:
         fig=pickle.load(f)

     fig.show()
```

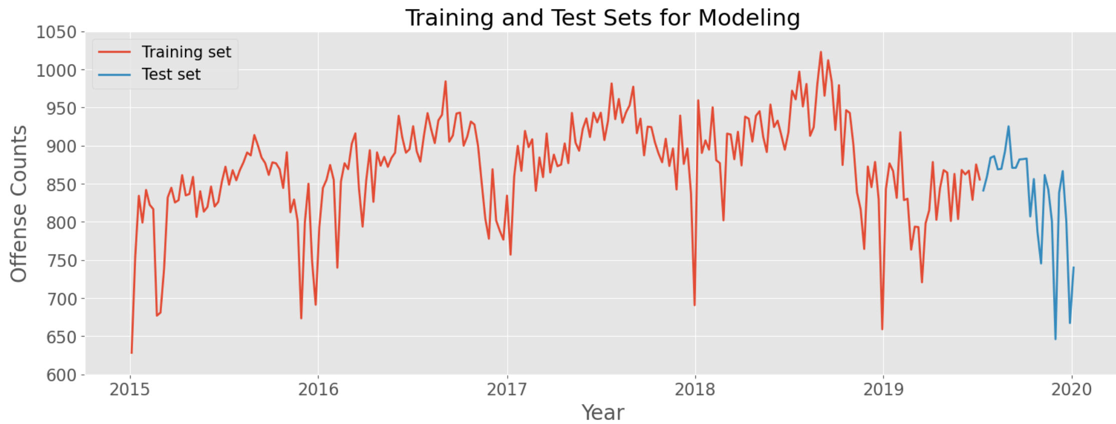
Another interesting plot presents information on what time of a day the most offenses occur. While one might expect that "dark deeds occur in the middle of a night," and they sure do, midnight has a very prominent association with the most offenses committed. However, the other three peaks happen at 8 AM, 12 PM, and 5 PM. One can speculate that these are the rush hours when the public transportation is most crowded, providing the best opportunities for theft and larceny, the category that outpaces all other categories of offenses.

4.2 Exploratory analysis of the data specific to modeling

Due to the sheer number of the records in the dataset it proved to be very time-consuming to model it. Therefore, I decided to limit the records to the last five years (2015-2019). Another reason to limit the dataset came from the fact that in 2019 the crime took a downturn. The only way to include this tendency into a training set of records was to limit the overall number of records and to use only 10% of them as a test dataset. Also, the original dataset with a count of daily offenses was resampled to a weekly count. As a result, the training set had 262 weeks between 2015 and the middle of 2019, while the test set included 26 weeks of the second part of 2019.

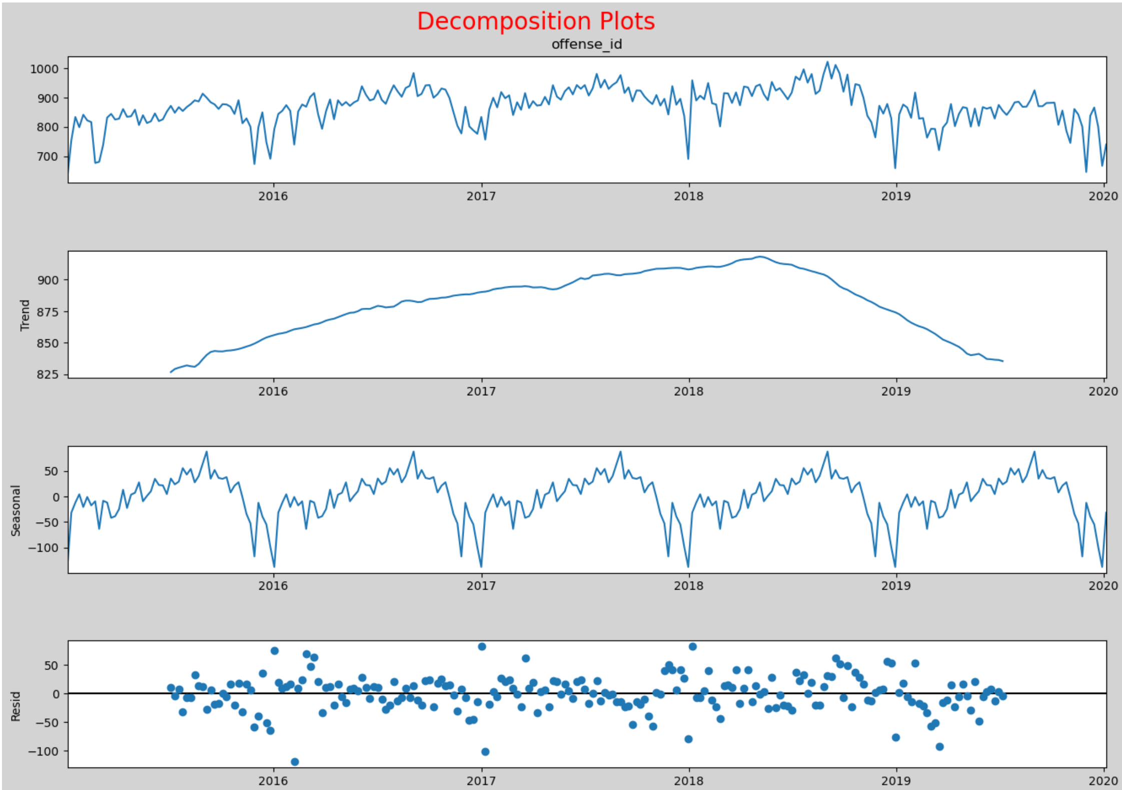
[8]:
```
with open('images/pickled_figs/ts_weekly_train_test.pickle', 'rb') as f:
    fig=pickle.load(f)
fig
```

[8]:



[9]:
```
with open('images/pickled_figs/decomposition_plot_ts_weekly.pickle', 'rb') as f:
    fig=pickle.load(f)

fig;
```



The time-series clearly displays a trend and a seasonality components. The crime rate increases in the middle of a year and drops in cold months of a year especially noticeably around Thanksgiving and Christmas holidays.

5 Model&iNterpret

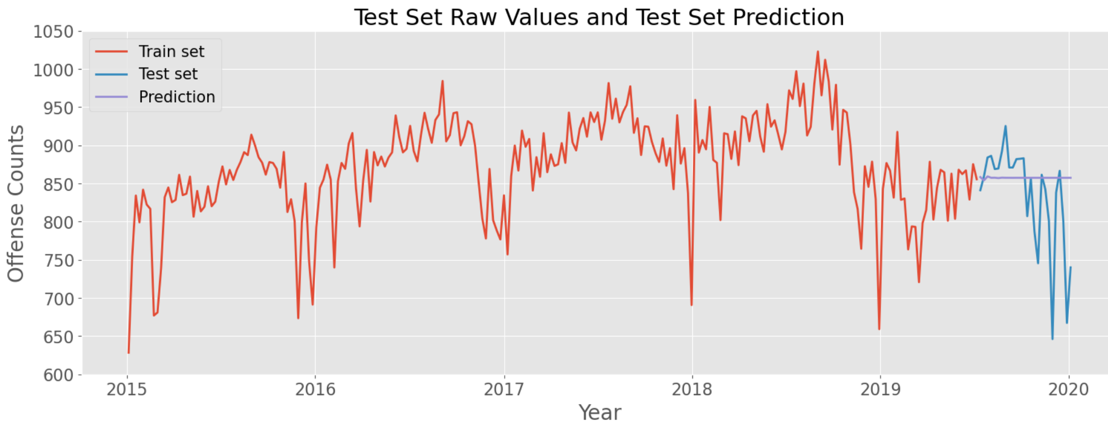
5.1 Modeling of General Crime Rate

I have decided to build n ARIMA(3, 1, 0) model with only trend autocorrelation components and first differencing as a Baseline model. As expected, the model performed relatively well, picking up an average trend for the last year but failed to account for seasonality.

[10]:
```
with open('images/pickled_figs/arima_train_test.pickle', 'rb') as f:
    fig=pickle.load(f)

fig
```

[10]:

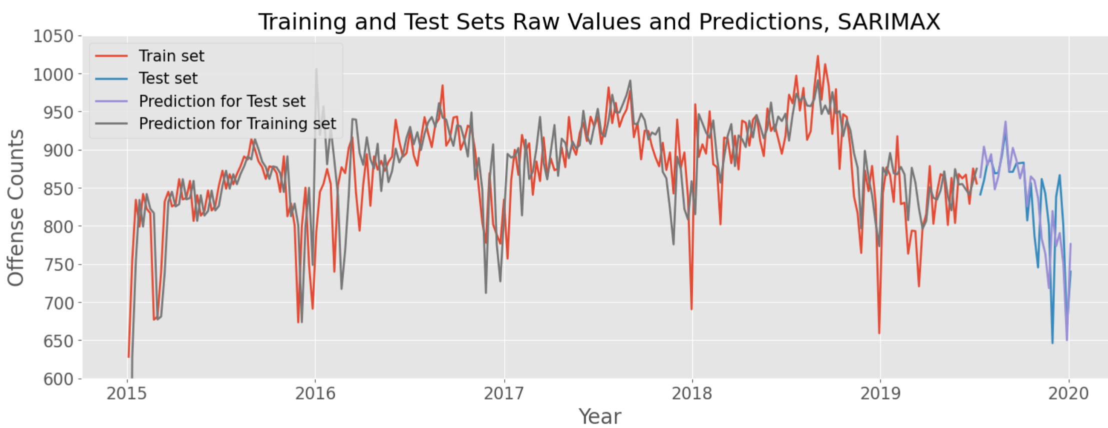


Gridsearch for the best SARIMAX model suggested ARIMA(3, 1, 0)x(3, 1, 0, 52) combination that generated relatively good test results and a reasonable forecast for two years forward.

[11]:
```python
with open('images/pickled_figs/sarimax_mod1_train_test.pickle', 'rb') as f:
    fig=pickle.load(f)

fig
```

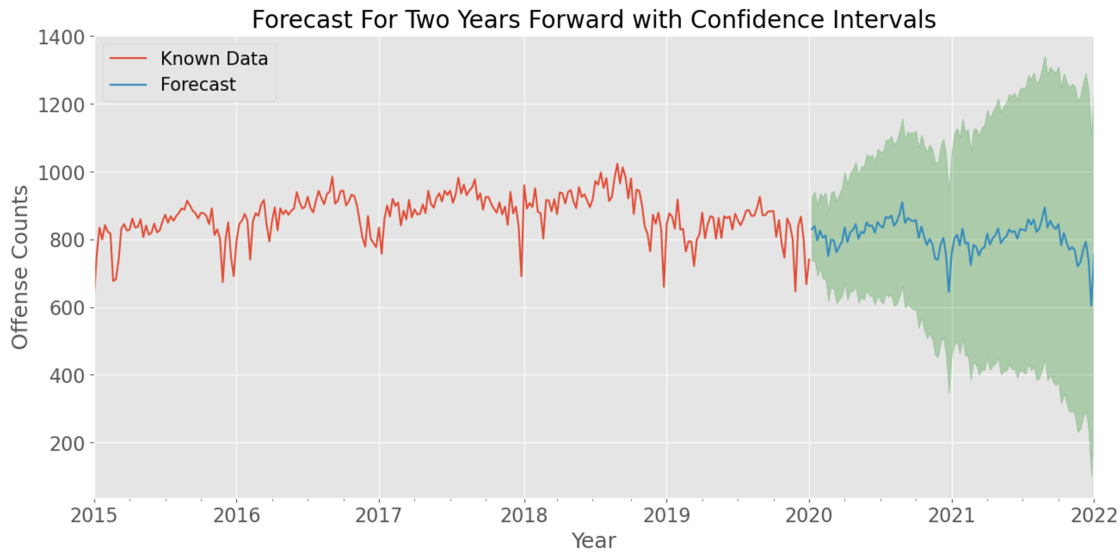
[11]:



[12]:
```python
with open('images/pickled_figs/sarimax_mod1_forecast.pickle', 'rb') as f:
    fig=pickle.load(f)

fig
```

[12]:



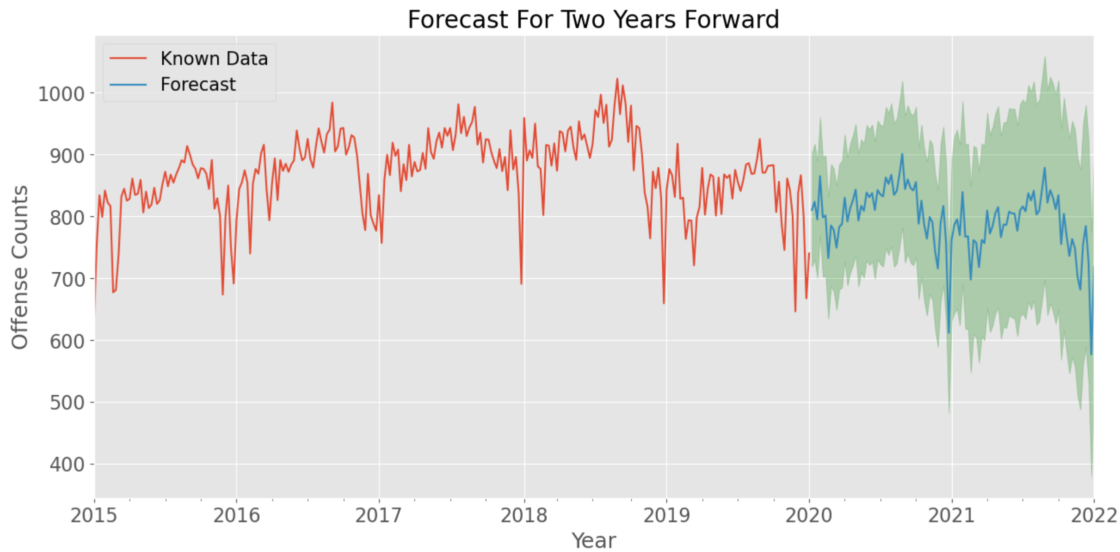
Based on the fact that the crime rate seemingly had downturns around holiday season I included exogenous predictors (US holiday time-series) into the mode. However, I used the same best combination from the previous gridsearch of ARIMA(3, 1, 0)x(3, 1, 0, 52). The results of the testing and forecasting were virtually the same.

The last step in the modeling of the general crime rate was using an auto-arima approach to search for the best combination of p,d,q and P,D,Q,d for ten and seasonal components of the time-series. The best model generated was a SARIMAX(1, 1, 1)x(2, 1, 0, 52) model. It displayed a relatively good fit with the test data and a

[13]:
```
with open('images/pickled_figs/auto_arima_forecast.pickle', 'rb') as f:
    fig=pickle.load(f)

fig
```

[13]:



5.2 Modeling of Crime Rate per Category

There are 23 separate crime categories in the original dataset. Some of them have sub-categories. But it has been decided to limit the analysis only to categories level.

Various categories demonstrated significantly different stationarity, trend and seasonality characteristics. All categories were modeled using auto-arima grid search, tested, and two-year forecast for each category generated. The final result can be found in the saved pickle files at data/pickled_models/RESULTS1 and RESULTS2. Each file contains a category model along with a forecast plot. Due to the sheer volume of the data it's better to see the corresponding notebook

6 Conclusions

1. The project demonstrated that the real crime data could be analyzed and modeled with significant accuracy. 2. It also demonstrated that the generated models could well predict future general crime rates and categories crime rates. 3. The Exploratory Data Analysis depicted which geographic areas show higher crime rates and require more resources and preventive programs. 4. The dataset, results of EDA, and the models are the base for a web-based dashboard built with dash python package.

7 Recommendations

1. The first recommendation is to obtain current data; it is difficult to forecast future trends with data almost two years old. 2. Suppose dynamic data becomes available, build and an API. This approach would be the most helpful to the general public. 3. Add exogenous predictors to the time-series

to improve modeling performance. The most helpful predictors would be the socio-economic features of the geographic areas. 4. Add geographic locations of committed offenses to improve knowledge of most crime-prone areas to plan for resources and preventive measures. 5. Improve performance and quality of the web-based `Crime in Colorado' application.