Predicting Crime Rates for travel

Damian Bocanegra

6/3/2020

Section 1: Introduction

1.1 Background:

Imagine planning to visit a city for a music festival or a work conference. While you are there you also plan to explore the city where it is occurring. You want to visit the top places to eat or other attractions while you are there but want to minimize the chances of getting robbed or stepping into a high crime area while you are there. The process of narrowing down the popular venues, then cross-referencing their locations with crime maps that may or may not be available is exhausting. It makes planning a trip even more of a hassle than it already is.

1.2 The Context/ Problem:

Step into the role of a traveler heading to Austin, TX in March of 2021 to attend the famous SXSW festival. You plan to go early so you can check out some other attractions and restaurants before all the other tourists show up. Using open crime data and data on venues around Austin from Foursquare, this project aims to predict the crime rates around the popular venues around Austin.

1.3 Interest:

This would perform all the heavy lifting in a single step, as opposed to a traveler having to look up all the various venues, then map them over a crime map by manually. Saving the traveler, a lot of time from having to spend extensive time researching. Allowing for reduced stress in planning and allowing more time to make various other preparations for their trip.

Section 2: Data Acquisition and Cleaning

2.1 Data Sources:

The first data set collected is crime data from: https://data.austintexas.gov/. This data contains information on crimes committed in the Austin area such as the zip code, highest offense committed, and the date of the crime committed. This data ranges from 2003 to May 2020.

The next data set collected from https://public.opendatasoft.com/, contains the latitudes and longitudes of all of the various zip codes in Austin, 83 in total. This data is needed to collect the next set of data from Foursquare.

The last set of data being collected from Foursquare is collected using the zip code data, so that as many venues as possible are available to be analyzed. The way it is collected is through the Foursquares API, once collected it includes the name of each venue, its latitude, longitude, the current rating on Foursquare, and a description of the venue.

2.2 Data Cleaning:

Both the Foursquare data and zip code data are collected with all their required features filled in completely, neither set requires any cleaning. Leaving only the crime data set for cleaning. The first step in cleaning the data is to remove unhelpful entries. Entries were only removed if they were missing the date of occurrence or had no zip code. Since the crime data will be broken up by zip code to make predictions, if there is no zip code it cannot be categorized. The same is true for no date of occurrence. Next the zip codes were imported as float numbers, so they were transformed into integers for clear reading and to play nice with the zip code data set. Last the time of occurrence was converted into a new timestamp format of: YYYY-MM-DD HH:MM:SS. Lastly all of the columns that were unnecessary were dropped from the data frame, leaving only the Date, Highest Offense Description, and ZipCode columns.

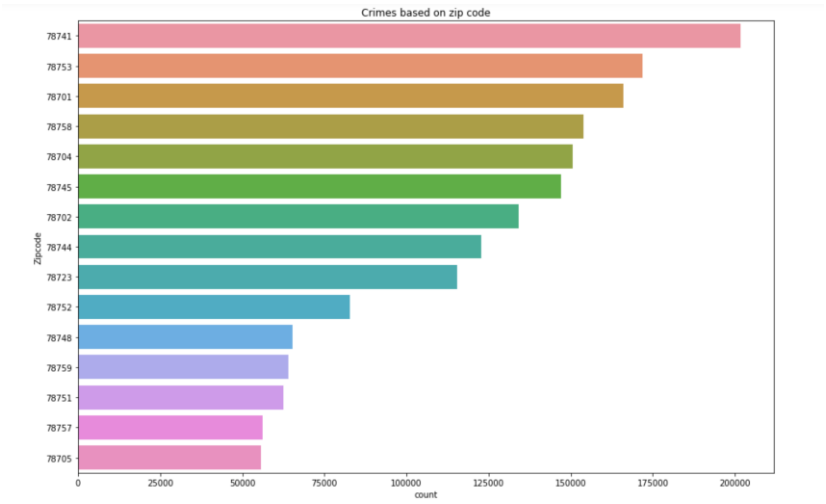
Date	Highest Offense Description	Date	Zipcode
2018-09-11 19:55:00	RAPE	2018-09-11 19:55:00	78660
2007-10-29 16:25:00	INDECENCY WITH A CHILD/CONTACT	2007-10-29 16:25:00	78719
2019-02-24 21:00:00	ASSAULT CONTACT-SEXUAL NATURE	2019-02-24 21:00:00	78749
2016-05-20 12:00:00	STATUTORY RAPE OF CHILD	2016-05-20 12:00:00	78758
2008-02-18 09:41:00	CAMPING IN PARK	2008-02-18 09:41:00	78701
...	...	...	...
2005-09-12 23:57:00	DWI	2005-09-12 23:57:00	78756
2018-03-26 23:50:00	VIOL OF EMERG PROTECTIVE ORDER	2018-03-26 23:50:00	78752
2016-07-01 21:39:00	WARRANT ARREST NON TRAFFIC	2016-07-01 21:39:00	78701
2017-05-12 03:43:00	VIOL OF EMERG PROTECTIVE ORDER	2017-05-12 03:43:00	78701
2010-06-10 02:29:00	PUBLIC INTOXICATION	2010-06-10 02:29:00	78702

Resulting data frame after cleaning and processing

Section 3: Data Exploration

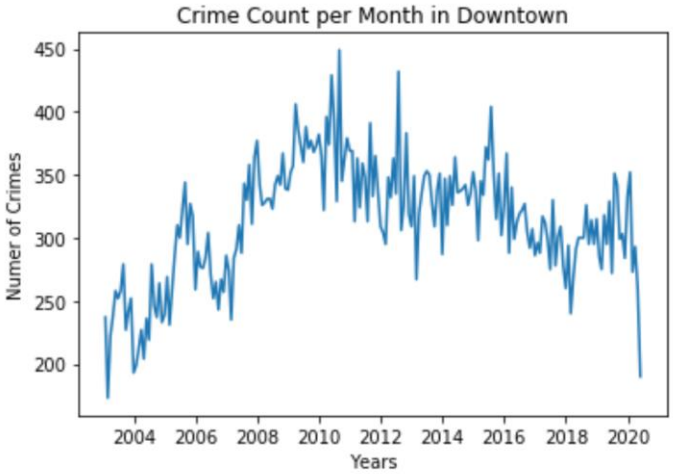
3.1 The Top Areas for crime:

It would only be natural to first locate the area with highest amount of crimes committed. So by counting the amount of times a zip code appears in the crime data frame, the zip code 78741 ends up being the area that has the highest crime count. The 78741 area is one of the main parts of downtown Austin, TX and for the rest of this report be referred to as Downtown.



With Downtown containing the highest crime count, it will be used as the example to illustrate other explorations with the data.

3.2 The Amount of Crimes Committed per Month:



In the above figure it looks like the crime count has increased in Downtown since 2004 and peaked around 2010-2011. After words taken a slight decline and has stayed steady with another slight drop around 2016. But overall, the crime count is declining in Downtown.

3.3 The Top Venues in Austin:

	Zipcode	Zipcode Latitude	Zipcode Longitude	Venue	Venue Rating	Venue Latitude	Venue Longitude	Venue Category
0	78757	30.349455	-97.733280	Tacodeli	9.3	30.348610	-97.735124	Taco Place
2	78702	30.265158	-97.718790	Lazarus Brewing Company	9.3	30.261770	-97.722267	Brewery
1	78701	30.271270	-97.741030	Texas State Capitol	9.3	30.274368	-97.740692	Capitol Building
3	78712	30.285207	-97.735394	Harry Ransom Center (HRC)	9.2	30.284245	-97.741092	Museum
4	78702	30.265158	-97.718790	Veracruz All Natural	9.2	30.263080	-97.713778	Taco Place
5	78701	30.271270	-97.741030	The Roosevelt Room	9.2	30.267842	-97.746242	Cocktail Bar
6	78748	30.172020	-97.822650	Moontower Saloon	9.2	30.169304	-97.826456	Bar
7	78704	30.246309	-97.760870	El Primo Taco Truck	9.2	30.244737	-97.757562	Food Truck
13	78756	30.320206	-97.741770	Pinthouse Pizza	9.1	30.318748	-97.739086	Pizza Place
17	78757	30.349455	-97.733280	T22 Chicken Joint	9.1	30.347976	-97.735335	Fried Chicken Joint

The traveler wants to know the top places to visit in Austin. Using the Foursquare we can see that Downtown venues are not in the list of the top 10 venues in Austin. So at least to the top venues are not in the area with the highest crime count. But it does contain venues inside from the top 5 zip codes in terms of crime count such as 78704 and 78701.

Section 4: Modeling

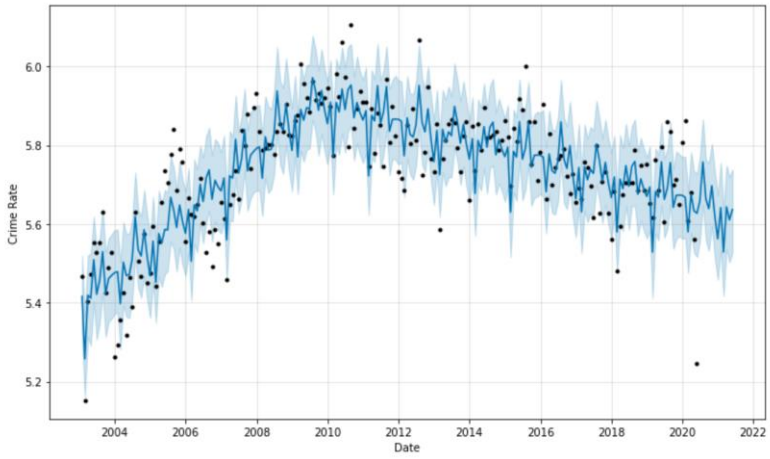
4.1 Modeling with Facebook Prophet:

To model and make prediction for this data, I will be using a tool developed by Facebook's Data Science team called Prophet. It is a Python library that will make predictions based on time series data. A time series is a set of data that is collected in regular intervals such as daily, monthly, or yearly. Since we want to know what the likely crime rate will be for a given area in March of 2021, the number of crimes per month in this case will be the time series to make predictions on.

	Date	Crime Count
0	2003-01-31	237
1	2003-02-28	173
2	2003-03-31	222
3	2003-04-30	238
4	2003-05-31	258
...	...	...
204	2020-01-31	352
205	2020-02-29	273
206	2020-03-31	293
207	2020-04-30	260
208	2020-05-31	190

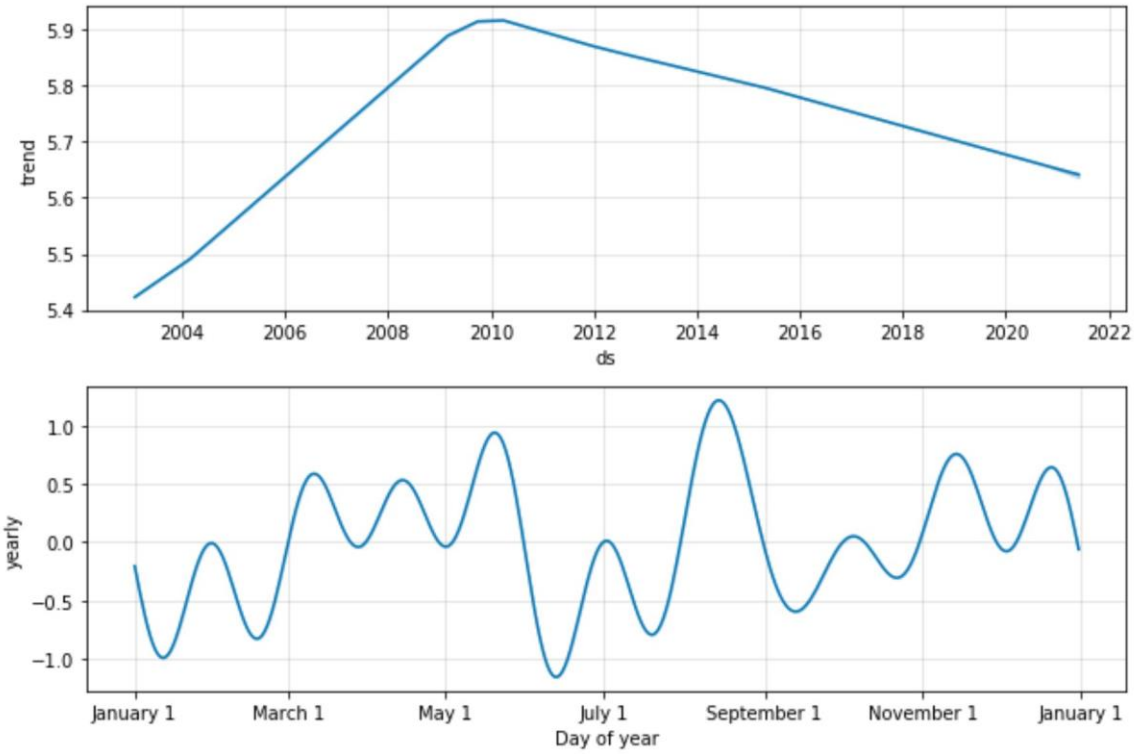
Crime data as a time series

To make the predictions work better with Prophet it is recommended to log transform the data values which is the Crime count in this case. After giving the data to Prophet it will run its calculations and return a data frame containing the predicted crime count in log transformation. This data frame also contains upper and lower bounds for this crime count. To make calculations the time series will be split into training data and test data by date. The Training data will be every entry before 2019 and the test data will be the remaining entries.



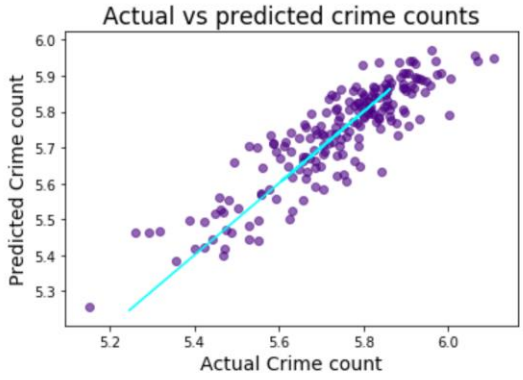
The resulting graph from the Prophet prediction, black dots are actual data points.

Splitting this data into its components shows the trends in number of crimes committed throughout the years. It will also show the trend during the year.



4.2 Evaluating the Model

With the prediction results I evaluated the mean square error between the test data and the predicted values. I got a result of 0.02 which is a very good score showing the data fits to this model very well. This would indicate a prediction using this model will be fairly accurate.



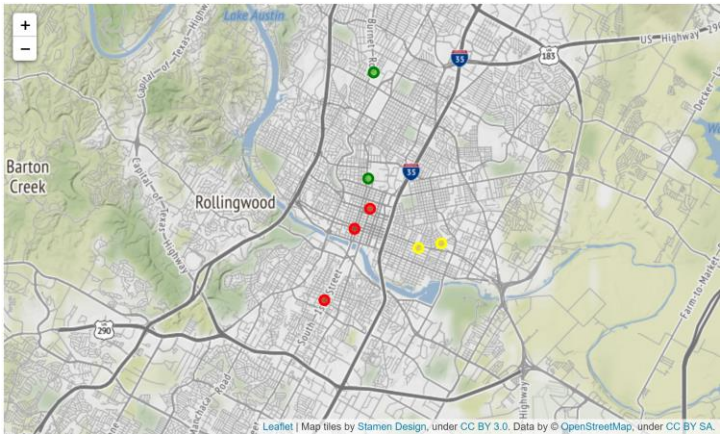
Predicted crime counts and actual crime counts(purple dots) in relation to test crime counts(cyan line).

Taking these results and re running the process on each zip code, a new data frame is created displaying the results from each model.

	Zipcode	CrimeRate_upperBound	CrimeRate	CrimeRate_lowerBound
0	78712	0.033610	0.019816	0.011539
1	78756	0.882212	0.825357	0.783616
2	78705	2.340288	2.305113	2.253464
3	78701	8.130794	7.821888	7.370643
4	78702	4.418702	4.488275	4.522790

Head of the resulting data frame

Using these results each venue from the top venues data frame is classified based on the crime in three classes of high(red) if the venue's zip code has a crime rate in the 90th percentile, moderate(yellow) if in the 40th percentile but not in the 90th, and low(green) otherwise.



Resulting map from classifications.

Section 5: Conclusions

In this report I analyzed the crime data, by identifying the areas with the highest crime count, the amount of crime committed every year, and its trends as the years have passed. Using Facebook Prophet, I can predict with high confidence the amount of crimes that will be committed at a target window of time. The model was built for every area that needed to be checked within the list of highest rated venues in the city. Now this model can be implemented into a tool to give travelers a quick and easy to digest amount of information that would normally take away from other preparations that need to be made for travel.

Section 6: Further Developments:

The output display can be further improved to show what the likely crime that a traveler would run into at a given time. This model can also be adapted to be more specific such as taking the time of day into consideration and processing data and building a model based on that specification instead of just the entire month. This project was mainly focused on a broad look at how many crimes are being committed during any given month. The next step would be to get

more specific with the window and build a new model the pursuit of achieving as precise of a result as possible, in order to give a traveler the best information to make decisions for their trip.