CSE 482 FINAL PROJECT (Cover Page)

Project Title: Urban Crime Rate Prediction

Summary of Team Member Participation:

Name	**Participate in data collection**	**Participate in preprocessing**	**Participate in data analysis/ experiment**	**Participate in writing the final report**	**Completed assigned tasks**
Parker Goodrich	Yes	Yes	Yes	Yes	Yes
Matt Adomshick	Yes	Yes	No	No	Yes

Team Member Roles and Contributions:

Name	**Roles and Contributions**
Parker Goodrich	Main Contributor on Project Proposal and project selection. Sole contributor of Data Analysis/experiment, and writing the final report
Matt Adomshick	Main contributor for Project intermediate report and Data Collection. He has failed the class midway through so he was not able to help me with the rest of the data preprocessing, processing, analysis, or final report.

I approve the content of the final report

Parker Goodrich: Parker Goodrich

Urban Crime Rate Prediction

Parker Goodrich
Project URL: <INSERT URL HERE>

ABSTRACT

We wanted to study Urban Crime rate prediction, more specifically, predicting shootings in the various boroughs in New York City. Our goals were to test our data management and manipulation skills in a more real-world scenario, as well as try and predict whether a shooting would happen or not given a specific location, time, day of the week, and poverty level. We are solving this problem using linear regression, and our target attribute is whether a shooting would happen at a specific time and location in New York City. Our regression model gave us an r-square value of 0.1617, which means that our model was not that great in determining whether a shooting would occur or not.

1. INTRODUCTION

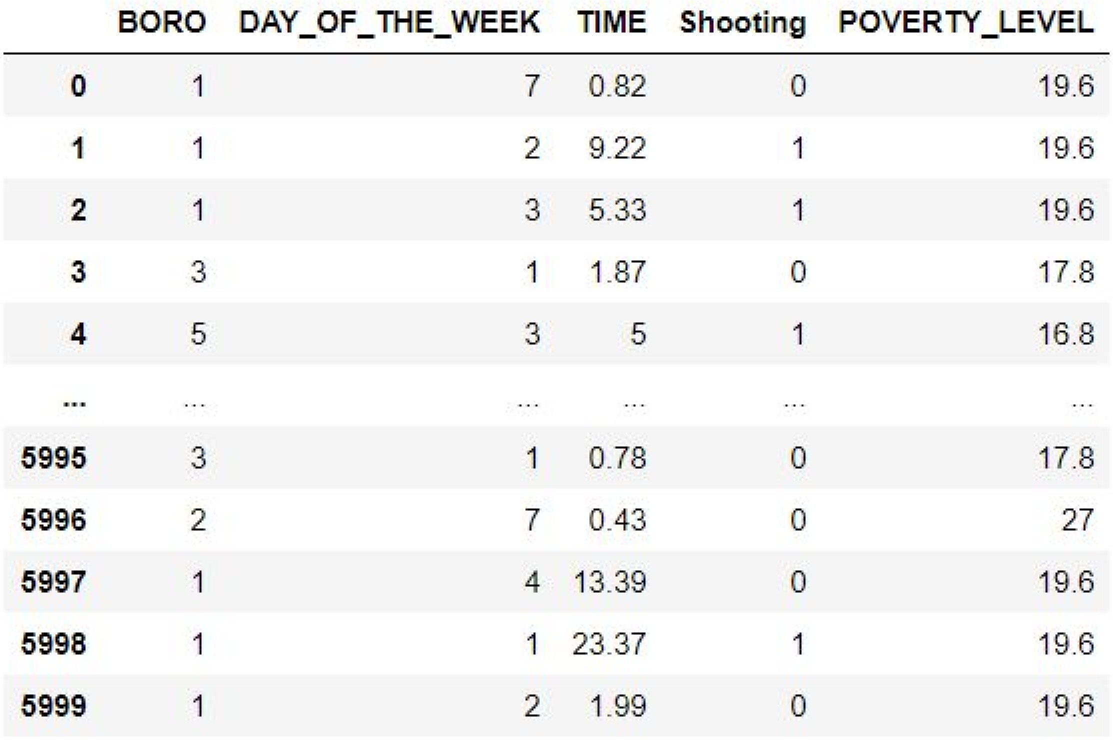
The goal of this project is to try and predict future shootings in some of the larger areas in New York. This is extremely important because if we are able to accurately predict with our test set, then this type of prediction can be used in the real world to help prevent future shootings, potentially saving lives. We used data from the NYPD between 2006 and 2019, which included 21,409 shootings. Some of the main data points we will be using are borough, day of the week, poverty level, and time of the shooting. We hope that by using these attributes, we will be able to predict future shootings in New York City.

We are going to use a Linear Regression with time, borough, day of the week, and poverty level as our predictor attributes to determine location and time of shootings in the future. We received our data directly from the NYPD. The URL for the data is linked in the References section of this report.

One challenge we faced was that we had a lot of data going all the way back to 2006. We decided to only look at the data from 2013-2019, so we removed any data points before 2013, as well as removing any missing data that was missing a time or a date associated with it. Then we had to add in the days of the week to our dataframe, as well as the poverty level for each borough, and change the time to a numeric value so that it could be used for our regression model. Once this was done, we randomly generated a set of 6000 shootings to make our training set a reasonable size. Another challenge we faced was changing over some of our data points to numeric values. To do this, we found it easiest to simply change the format of the cell in the excel file. We found that we were rarely able to predict shootings accurately. We would have liked to be able to at least predict shootings with 50% accuracy, so we did not reach our goal.

2. DATA

The only dataset we used came from data.gov. We chose to use the CSV format of the data because we have a lot of experience with that type of format from the exercises and the homeworks in class. There are 18 attributes in the table, and we are primarily focused on 5 of them; location, time, borough, poverty level, and shooting. I have pasted a few example data points with the important attributes from the table below.



	BORO	DAY_OF_THE_WEEK	TIME	Shooting	POVERTY_LEVEL
0	1	7	0.82	0	19.6
1	1	2	9.22	1	19.6
2	1	3	5.33	1	19.6
3	3	1	1.87	0	17.8
4	5	3	5	1	16.8
...	...	...	...	...	...
5995	3	1	0.78	0	17.8
5996	2	7	0.43	0	27
5997	1	4	13.39	0	19.6
5998	1	1	23.37	1	19.6
5999	1	2	1.99	0	19.6

Table 1: Example section of data from the NYPD after alterations.

The data includes larger areas of New York, including the Bronx, Brooklyn, Queens, Staten Island, and Manhattan. Most of the attributes in this dataset aren’t useful in our predictions. Some of these include incident, precinct, jurisdiction, and information about the victim and the perpetrator. We needed to discard any shootings that were missing values from any of our predictor attributes. A summary of the important attributes are attached below.

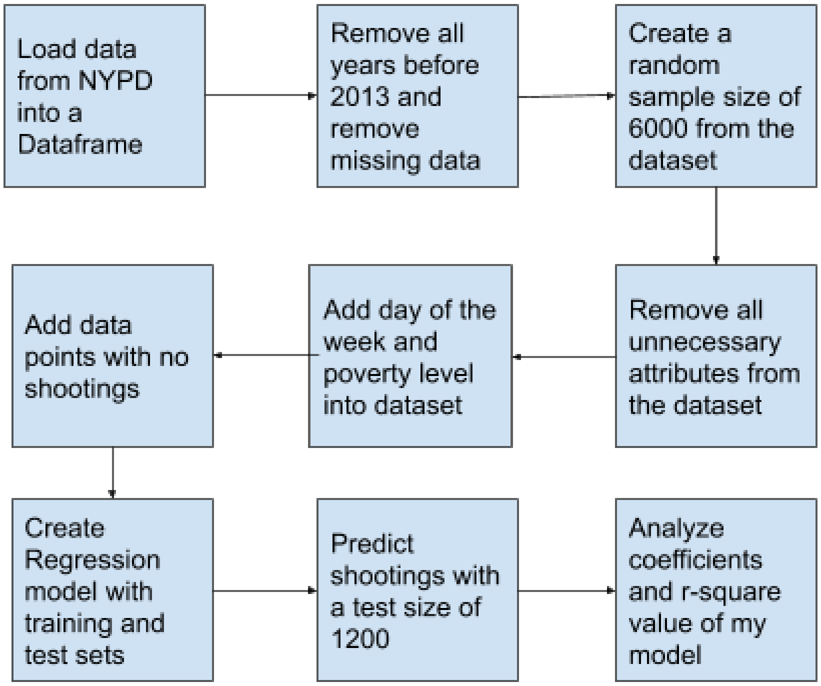
Attribute name	**Type**	**Description**
Time	Interval	Time from midnight
Poverty level	Interval	Poverty level (%)
Borough	Nominal	Name of borough
day_of_the_week	Interval	Day of the week as a number
Shooting	Nominal	Value for if the shooting happened or not

The first thing we had to do was alter the data for our regression model. To do this, we first needed to add 2 attributes to our table; poverty level and shooting flag. Once we did this, we had to change all of the values in our csv file to numeric values to be used in our regression model. This meant changing time to a floating data point, borough to an ID value, and the day of the week to a number between 1-7 depending on the day. Once we had our data all set, then we had to generate rows of data where there weren't shootings. I decided to generate 5000 data points where no shootings occurred, giving us a good amount of data for our training. The table below summarizes the data.

Number of data points	21,409
Number of attributes	18
Percentage of missing values	0%
Number of data points used for training	6000
Number of attributes used for training	4

The predictor attributes used are time, borough, day of the week, and poverty level. Our target attribute was whether the shooting would occur or not. With only 5 attributes that were used in our regression model we had to discard most of the columns in our dataset. The size of the data after I added in the times where shootings did not occur was 26,406.

3. METHODOLOGY


Load data from NYPD into a Dataframe
Remove all years before 2013 and remove missing data
Create a random sample size of 6000 from the dataset
Remove all unnecessary attributes from the dataset
Add day of the week and poverty level into dataset
Add data points with no shootings
Create Regression model with training and test sets
Predict shootings with a test size of 1200
Analyze coefficients and r-square value of my model

Flowchart 1: Flowchart of project steps

The data preprocessing portion of this project was where the bulk of the work came from. We first needed to read the data in, remove all years before 2013, and generate a random sample of the data to reduce the size down to 6000 points. Then once we did that, we needed to remove all the attributes that weren't used as predictors. Next, we had to add in 5000 rows of random data points so that we could test not data where shootings did not occur. Once this was done, then we were able to step into the predictive modeling phase of the project. We created our testing and training datasets by simply using the train_test_split method in python. This made it easy for us to test with different sizes of data.

All of our code is written in our nyc_crime_data.ipynb file in jupyter notebook. The first box reads in the data and displays the first 5 lines. The second box gets the years that we want, and adds in the poverty level to each data point. The third box generates our random sample size of 6000. The fourth box changes the day of the week and the borough to ID numbers. The fifth box drops the unnecessary columns, and the last box does the training, testing, and displays the r-squared value along with the slope coefficients and the intercept.

4. EXPERIMENTAL EVALUATION

4.1 Experimental Setup

For this project, we decided to continue to use jupyter notebook since we became quite familiar with the format of it through the exercises and the homeworks we used.

For determining how successful my test set was, I decided to just use the one dataset for training and testing. For our testing, we received an r-squared value of 0.1617. This was much lower than what I wanted it to be. I was hoping for a value around 0.5. In an article I found from BBC, a tech firm has software that can increase crime detection from 10-50%. I was hoping to detect minimally 50% accuracy with my model to be somewhat close to what they said in the article.

4.2 Experimental Results

The only experiment I performed was splitting the main dataset up into training and testing sets. This produced me the r-squared value of 0.1617. The top 2 predictor attributes were time, and borough, in that order. Time had a coefficient value of 0.025 and borough had a value of 0.015. I expected that these would be the top 2 values. Most shootings happen at night, and most of the shootings in the dataset happened in the Bronx.

The project was not successful. The main reason the results were not as good as I had hoped was that I didn’t have enough predictor attributes. I didn’t have enough time to research weather data and find a way to incorporate it into my dataset. I also could have used a couple other predictor attributes to help out. Another reason why my model could have been bad was that I was mainly looking at shooting data at a fairly broad level. I was only working with 5 boroughs, when it would have been better to break the boroughs down into many different sections. If my partner could have helped me out during this project, I think these things could have gotten done and we could have had a much better model. We also could have done research to find some more data from the NYPD to test on.

5. CONCLUSIONS

I found in this project that you need quite a lot of predictor attributes for determining shootings. I also found that choosing fairly small sections would have helped a lot as well. This would help a lot because more shootings would happen in the communities where there isn't a precinct nearby as well as the poor communities. Overall, I am happy with this project. I wish I would have taken the time to think a little more about the preprocessing steps before I dove into this project.

6. REFERENCES

[1] "Data Tool." *Data Tool - NYC Opportunity*, NYC, www1.nyc.gov/site/opportunity/poverty-in-nyc/data-tool.page.

[2] "NYPD Shooting Incident Data (Historic) - Comma Separated Values File." *Data.gov*, catalog.data.gov/dataset/nypd-shooting-incident-data-historic/resource/0aa46f63-3c4a-40ef-975a-305bdab953ab.

[3] Smith, Mark. "Can We Predict When and Where a Crime Will Take Place?" *BBC News*, BBC, 30 Oct. 2018, www.bbc.com/news/business-46017239.