
AMS 325
Prof. Xiangmin Jiao
Shreyan Wankavala, Johnny Zevallos-Vila, Dongziyi He

Project Objectives

As the team was figuring out what the project topic should be, discussions were made on what the dataset should contain prior to any work or statistics being done on it. Looking through the sources provided from the slides, we came across a csv dataset of crimes reported in New York City in Kaggle, where this csv file contained the detailed reports of most crimes the NYPD profiled within their public accessible database, covering more than 10 years of data from 2006 to 2019. With our main focus now on interpreting this data in some way on python, we realized that establishing and picking apart this single dataset was something that can be done by one person, acknowledging the fact that with just this, the project could be deemed as a solo project. Of course, we then discussed ways we can elaborate the project in order to extend the work to that of three people, thinking of project end goals that would be presentable for the time allotted for the project, and to deliver results that would reflect the individual works that we committed to the project. From that discussion, we realized that there stood the importance of comparison within the city we chose, that the topic of crime stands not only in NYC, but obviously stretches to other cities as well. With this thought, the team decided to choose three cities to evaluate, study, and present results from, and in the end be able to compare the cities with one another, in hopes of finding trends between the three cities. Thus, as NYC was already decided initially, we scoped out for a dataset on the opposite side of the country, where Los Angeles was a perfect choice sitting on the west coast of the states. We then decided a city in the middle of the states would be a perfect comparison for the two, and as such Chicago was decided. With the three cities now in place, we searched for the datasets for the other two cities, and decided the project goals we wished to accomplish. Beginning with initializing the datasets, we wished to first format the data into that of a dataframe, as not only were the sets mismatched with range in years, but they also contained more than two million rows of information, each row being a single reported crime where the biggest dataset was NYC's, which had over 6 million rows upon inspection through Excel’s PivotTable feature. Upon receiving the now formatted datasets, we

decided one objective was to somehow find the correlation between weather and crimes within the cities, hopefully through that of the provided information of the months the crimes were reported in. Lastly, we wanted to predict the trends through some form of regression, in which would portray the rate of certain crimes over months and years, and as to whether or not it's on rise or fall as time progresses, and how different they might stand when compared to one another. Some objectives that ended up falling through within discussion on the project was that of the characters that are portrayed within each crime, as some data sets lacked the information of perpetrators and victims compared to that of the other cities' set of data. To go forward with the project, we decided that Victoria would work on the datasets and the slides, Shreyan would work on the datasets as well as some of the statistical procedures, and Johnny would work on the datasets and the linear regressions for all three datasets.

Techniques and Tools

To begin, we found three datasets from Kaggle, each one representing the crime rates from a different city. Even though they were all from the same source, Kaggle had acquired them from different places so this was challenging to deal with considering the focus on what was recorded varied from city to city. Originally, we'd wanted to look into factors involving the race, sex, and age of both the perpetrator of the crime as well as the victim, but this was a problem because the Chicago dataset was extremely limited compared to the New York City and Los Angeles ones when it came to person identification. It instead, was the most detailed in terms of crime classification as it contained information explaining what the crime committed actually was as well as the crime's "Primary Type," a more general classification of the crime; something the other two datasets did not do. By looking over the datasets at the start, we decided upon basing our project on three identifying factors for each crime: the month and year the crime was committed in, and the crime description.

Loading the datasets into python using pandas was tough, since the datasets were too large to work with normally. The smallest out of the three spanned more than ten years containing close to one million rows and twenty-five columns. We had to find a way to work with datasets this large that python could read in as a csv file, so we decided to add the "chunksize" parameter to our dataframe. We could read in our data as a specified number of chunks that would be concatenated together into a dataframe in the next line. Another issue was

printing out the dataframe as output. Compiling and running our code took upwards of five minutes due to the sheer size of each dataframe. This proved to be a severe limitation on our productivity as most of our time would be spent waiting for our code to print, even though the

	YEAR	MONTH	CRIME_DESCRIPTION
4	2017	3	PETIT LARCENY
5	2016	10	NARCOTICS
7	2015	10	PETIT LARCENY
10	2017	11	DANGEROUS WEAPONS
14	2017	6	BATTERY
15	2016	10	PETIT LARCENY
16	2015	12	BATTERY
17	2016	5	PETIT LARCENY
18	2015	2	NARCOTICS
19	2015	10	DANGEROUS WEAPONS

An example of our final dataframe

change made was small. We worked towards fixing this problem by making “checkpoints” of our dataframes. In other words, while working on trimming them down, we would save them to a csv file periodically. This would make the compiling and printing time much less significant with each new updated file made.

We already decided that we were going to keep the year and month the crime took place in, as well as the crime description, but how do we select the crimes we were to report? This was an extremely complicated process, as the states of New York, California, and Illinois have different laws regarding the same infractions in many different scenarios. They also have different ways to classify the same thing. For example, in New York, petit larceny is the classification for all stolen property under $1,000. In California however, this crime holds the value of $950 as its threshold, and Illinois actually divides it up into two different charges; stolen property worth $500 or less, and stolen property worth $500 or more with a corresponding charge of grand larceny. Discrepancies between the data sets were very common, and we had to go through each crime in an attempt to manually categorize it. To get a better understanding of which crimes we should eventually focus on, we used the method value_counts on our dataframes. This gave us the counts of each crime in the crime description column of our datasets. By doing this, we had a greater picture of which were the most common crimes in the

first place, and it made it a lot easier to focus on a select few. Looking at each category of crime, it was also apparent that we would have to combine crimes that eventually fell within the same general classification. For example, we wouldn't list charges for heroin and crack separately, as this detail only existed in the Chicago dataset, and they both fell under the umbrella of drugs. From this logic, they would both be renamed to 'NARCOTICS'. From this logic, we developed a system using the methods within pandas that would rename any value based on a number of factors. One of the most useful components of this was the way we could rename any crime description to its general classification if the description contained any string that we would enter, for example entering 'sex' would rename any crimes involving sex to 'SEXUAL ASSAULT'. It's worth noting that the original dataset for Los Angeles was very limited when it came to narcotics of any kind, but we felt we should include this category anyways due to the nature and prominence of the crime. This process was long and exhaustive, and eventually we found ourselves with nine crimes, chosen because of their abundance, as well as their presence in all three datasets. These crimes were sexual assault, dangerous weapons, battery, petit larceny, burglary, robbery, homicide, arson, and narcotics. In our code, the comments describe the removal or combination of at each step in the process, detailing what the crime description was labeled as originally and also what it turned into. At the very end of the file, there is a final csv saving point that creates the dataframes we'll be using for the rest of the project.

To complete our statistical analysis of the datasets, we discussed the different things we could do to best describe them. For one, we knew they were very large, so our processes had to specifically fit what we had. Our first goal was to figure out if our datasets from each city were in fact from a normal distribution. We realized this was very important, as many statistical procedures start with this first step specifically because a normally distributed dataset has many more statistical tests available to analyze it. Even though a Shapiro-Wilk test for normality is normally used, we couldn't because our sample size was so large, and opted to use the Kolmogorov-Smirnov test from python's scipy.stats package for statistical functions. This test returned test statistics and p-values for each city. Another test we wanted to perform was the Chi-Squared Test. This test is also specifically used when samples are large like what ours was, and measures if two categorical variables are independent in influencing the test statistic. We wanted to test the independence between the crimes from our dataset and the seasons in which they took place. To complete this test, we first took the twelve months that existed in the dataset

and renamed each set of three into its corresponding season. For example, the months of March, April, and May correspond to spring, June, July, and August to summer, September, October, and November to fall, and December, January, and February to winter. With the seasons and crime descriptions, we could then form a cross tabulation using pandas to make a contingency table. It was then easy to complete the chi-squared test using the chi2_contingency function from scipy.stats.

SEASON	FALL	SPRING	SUMMER	WINTER
CRIME_DESCRIPTION				
ARSON	934	963	1028	1040
BATTERY	25377	25497	29225	22356
BURGLARY	16817	14890	16875	15705
DANGEROUS WEAPONS	10458	11901	11029	10484
HOMICIDE	399	404	479	354
NARCOTICS	24004	25459	24826	22602
PETIT LARCENY	108484	101905	115503	98018
ROBBERY	19386	16609	19184	17598
SEXUAL ASSAULT	10027	10492	10683	8860

The contingency table

The most significant part of our statistical analysis on the three datasets was our linear regression on them. The purpose of linear regression is to predict the value of a variable based on the value of another variable. We wanted to see if time had an effect on the rate of crime in any way, and if we could accurately predict this relationship. Our datasets measured time using both months and years, so we decided to complete linear regressions with each as an independent variable.

For the first linear regression, we found the total number of crimes that occurred for each month throughout the five years. We then had two variables: the months ranging from 1-12, and the total crimes reported corresponding to each month. With the months as the independent variable and crimes reported as the dependent variable, we used statsmodels.formula.api.osl to regress our data. We could then fit our model and print out our parameters and summary tables which would contain key pieces of information like the equation for our line of best fit as well as the r-squared statistic, which measures how well the data fits the regression model. Matplotlib and seaborn from python were then used to provide visuals for the scatter plots and lines given by the regression analysis.

For the second linear regression, we wanted to see the effect of the year on the rates of crime, and find out if there was a positive or negative relationship amongst the two of them. By comparing the crimes individually between the three cities over the number of years as the basis of the independent factor of the graph, this part of the project was to catch the trend of certain crimes over the span of five years (2015 -2019). With this information processed through linear regression, the idea was to compare the cities with all 9 crimes, resulting in 9 graphs and the information of the cities on them. This would provide a visual opportunity in tracking the rate of these crimes, such as which ones are rising more prominently, in which from there the results would display how different these crimes are prominent between the three cities, and how they may continue to change past the five measured years.

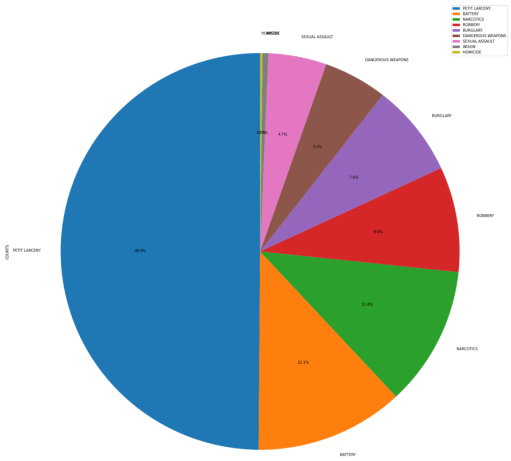
To go alongside our datasets, we used methods from pandas to form graphs and charts that would give visuals for our data. The plot function let us make a histogram for each city, letting us set parameters for which type of graph we wanted (‘bar’, ‘line’, etc…). We also used the built in plot.pie function in pandas, which turned our datasets into pie charts. Pie charts are useful because they give us a better representation of how apparent or non-apparent a specific data value is in relation to the rest of the data.

Throughout the process of our project, we collaborated using GitHub and pulled and pushed code into it often. GitHub proved to be extremely useful as we could work on code at different times and could update the repository when we were done. Our group members could then contribute on their own time by pulling and then pushing what they added. It was very important to write comments at every stage of the project, as this was the number one way we avoided any confusion and made our code transparent. We also uploaded any files that we had to the repository, like the periodic “checkpoints” we made when working on the dataframes as well as the original datasets themselves. Uploading the datasets was difficult because there is an upper limit on the size of the file you could upload. Even though we tried other methods to get all of our datasets onto the repository, only one was possible at the end.

Observations and Conclusion

The observations we have begin with the creation of bar graphs and pie charts of each city, in which we found a trend with the types of crime most reported within the three cities. According to the pie chart, from 2015-2019, nearly 50% of the crimes in New York City were

petit larceny, while narcotics, battery, robbery, burglary, dangerous weapons, and sexual assault followed closely behind. Arson and homicide was the least reported crime within the three cities. If we look at the bar chart, it is also easy to see that in New York City, the number of petit larceny crimes reported surpassed that of the other two cities. In Los Angeles, the three main majorly reported crimes were battery, petit larceny, and burglary. We were curious as to why



The NYC Pie Chart

there was an absence of drug related crimes in LA, since that portion of reported crime was not a part of the initial data set provided. In Los Angeles, burglary was the most reported crime, and it also contained a higher rate in reported crimes of dangerous weapons, battery, and robbery compared to that of New York City. Following next with Chicago, petit larceny also held its position as the highest reported crime with New York, while battery stood as the second, and narcotics the third. Similarly, battery and petit larceny were the two most common crimes in Chicago. From the value_count method and the information that it printed, we were also able to display numerically the number of reported crimes for each crime, all between the three cities. In total, New York had about four times the amount of petit larceny than that of Los Angeles and Chicago, and Los Angeles had two times the number of burglaries than that of New York City. While Chicago had the least number of burglaries among the three, it also had a very high rate in battery and narcotics.

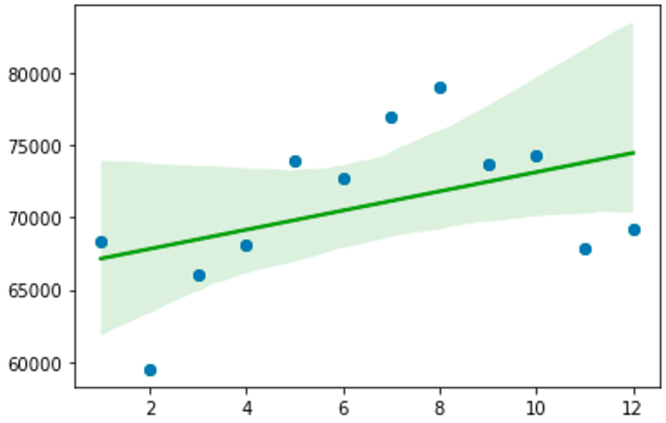
In our chi-squared test, we wanted to test the null hypothesis of whether the seasons and the type of crime were independent of one another. The p-values that we calculated were less than 0.05, so we were able to reject the null hypothesis and concluded that the seasons and types

of crime do share a relationship in some way. We can conclude that the two factors are dependent on each other, so the seasons do have an effect on crime.

From our linear regressions, we made observations on the three cities using the resulting regression graphs. The observations are that the points are less scattered on the New York City and Los Angeles graphs than the one in Chicago. From the linear regression graph from just Chicago, we can see that there's a negative relationship between each independent variable and the dependent variable. The p-values for the coefficients indicate whether these relationships are statistically significant. In Chicago, the coefficient is negative, meaning that the crime rate decreases throughout the year. "A negative coefficient suggests that as the independent variable increases, the dependent variable tends to decrease" (Statistics By Jim). However, in New York City and Los Angeles, the coefficients are positive, which means the crime rates increase throughout the year. There is a noticeable spike in all three graphs during the summer months, particularly during the months of June, July, and August. We also found the line of best fit in all three cities, and we can get the equation of each line.

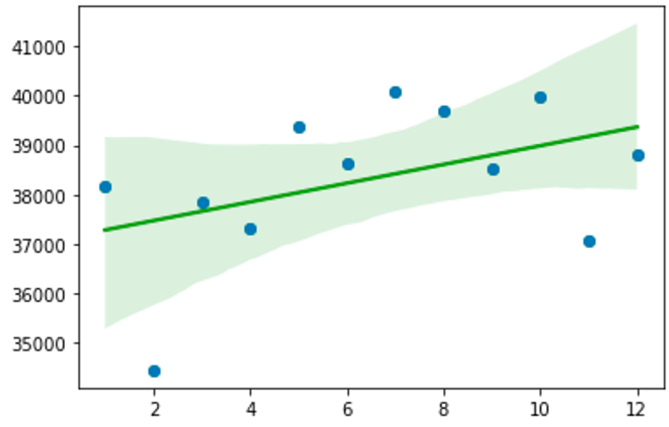
New York City:

CRIMES_REPORTED = 665.555944 * MONTH + 66495.136364.



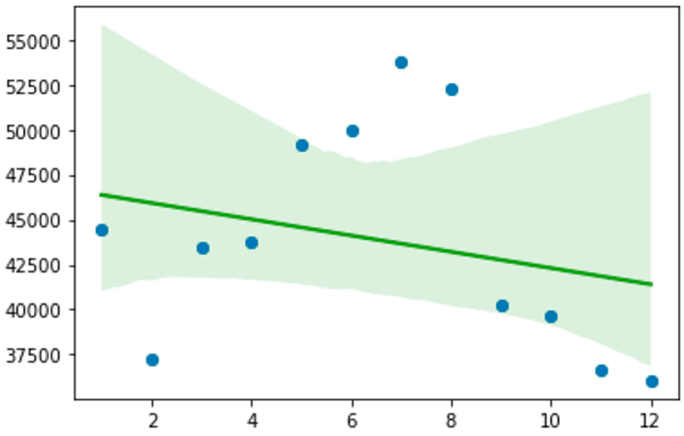
Los Angeles:

CRIMES_REPORTED = 189.5140 * MONTH + 37091.909091



Chicago:

CRIMES_REPORTED = -452.416084 * MONTH + 46825.954545



When we did the test for normality, we found that the datasets were not normal. This is because all three p-values were below 0.05, and the test statistic was close to 1, so we reject the null hypothesis at the 95% level of confidence.

At the end, the project was a really good teamwork exercise for us all, and we learned many new techniques which we used throughout the duration of it. Furthermore, the process of implementing GitHub for team-based accessibility was a first for us all, at least in practice. We ended up using it successfully, and used it to share code which helped us analyze three sets of data giving us meaningful results.

Personal Statement Contribution

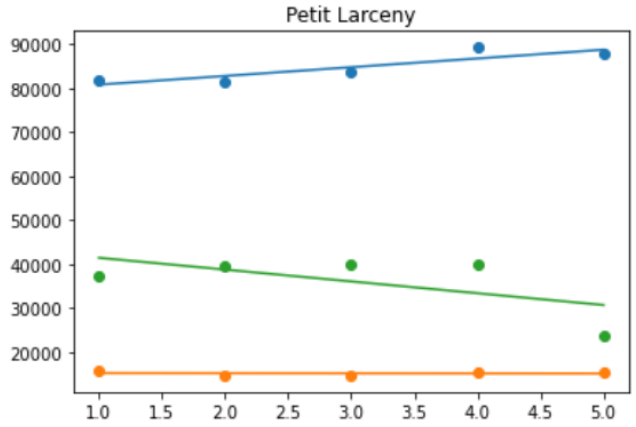
For my own personal contribution, I had the most hands-on experience with the dataframes, as I'm the one who wrote all of the code to format them into what we ended up using with help from both Victoria and Johnny. On top of this, I also implemented the chi-squared and normality test statistical procedures.

Blue : NYC

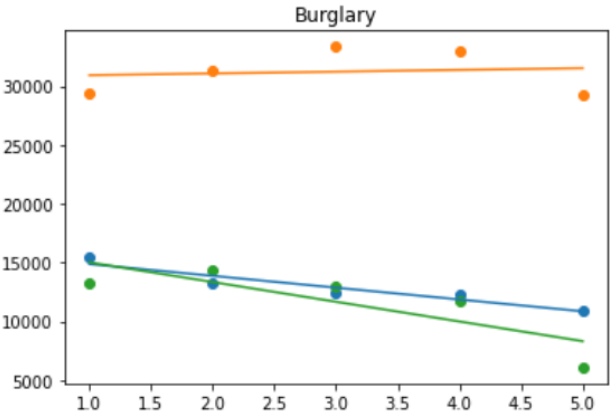
Orange: Los Angeles

Green: Chicago

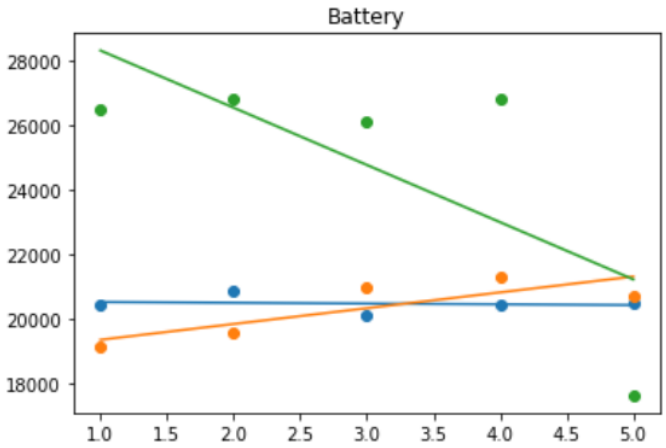
The statistics are in the same city order as the legend




```
The coefficient of determination is: 0.7762919314092773
The intercept is: [78824.]
The slope is : [[1986.]]
The predicted responses are:
[[80810.]
 [82796.]
 [84782.]
 [86768.]
 [88754.]]

<matplotlib.collections.PathCollection object at 0x00000266426844C0>
The coefficient of determination is: 0.007970558397204064
The intercept is: [15290.5]
The slope is : [[-24.5]]
The predicted responses are:
[[15266. ]
 [15241.5]
 [15217. ]
 [15192.5]
 [15168. ]]

<matplotlib.collections.PathCollection object at 0x0000026644649040>
The coefficient of determination is: 0.36348459216375106
The intercept is: [44151.4]
The slope is : [[-2682.8]]
The predicted responses are:
[[41468.6]
 [38785.8]
 [36103. ]
 [33420.2]
 [30737.4]]
```




```
The coefficient of determination is: 0.9073584508238395
The intercept is: [15871.2]
The slope is : [[-1004.6]]
The predicted responses are:
[[14866.6]
 [13862. ]
 [12857.4]
 [11852.8]
 [10848.2]]

<matplotlib.collections.PathCollection object at 0x0000026641E78F10>
The coefficient of determination is: 0.01466801877399626
The intercept is: [30784.4]
The slope is : [[149.2]]
The predicted responses are:
[[30933.6]
 [31082.8]
 [31232. ]
 [31381.2]
 [31530.4]]

<matplotlib.collections.PathCollection object at 0x0000026641EA5610>
The coefficient of determination is: 0.6657760620547652
The intercept is: [16682.7]
The slope is : [[-1674.7]]
The predicted responses are:
[[15008. ]
 [13333.3]
 [11658.6]
 [ 9983.9]
 [ 8309.2]]
```




```
The coefficient of determination is: 0.0194085710502242
The intercept is: [20561.8]
The slope is : [[-23.6]]
The predicted responses are:
[[20538.2]
 [20514.6]
 [20491. ]
 [20467.4]
 [20443.8]]

<matplotlib.collections.PathCollection object at 0x0000026641E249D0>
The coefficient of determination is: 0.6744049774390934
The intercept is: [18879.3]
The slope is : [[489.1]]
The predicted responses are:
[[19368.4]
 [19857.5]
 [20346.6]
 [20835.7]
 [21324.8]]

<matplotlib.collections.PathCollection object at 0x0000026641E244F0>
The coefficient of determination is: 0.49080069960729694
The intercept is: [30095.6]
The slope is : [[-1771.4]]
The predicted responses are:
[[28324.2]
 [26552.8]
 [24781.4]
 [23010. ]
 [21238.6]]
```

References

How to Interpret P-values and Coefficients in Regression Analysis

https://statisticsbyjim.com/regression/interpret-coefficients-p-values-regression/

NYPD Crime Complaint Data Historic (2006-2019)

https://www.kaggle.com/datasets/brunacmendes/nypd-complaint-data-historic-20062019?resource=download

Los Angeles crimes 2010-19 cleaned dataset

https://www.kaggle.com/datasets/stefanoskypritidis/los-angeles-crimes-201019-cleaned-datased

Chicago Crime 2001-2019

https://www.kaggle.com/datasets/milesius/chicago-crime-20012019

https://datatofish.com/plot-dataframe-pandas/

Chi-Squared Test Implementation

https://predictivehacks.com/how-to-run-chi-square-test-in-python/

Github

https://github.com/deathbarbiepink/final_project/settings/access?guidance_task=