

IE 6200 Project - Section 9 - Group 1

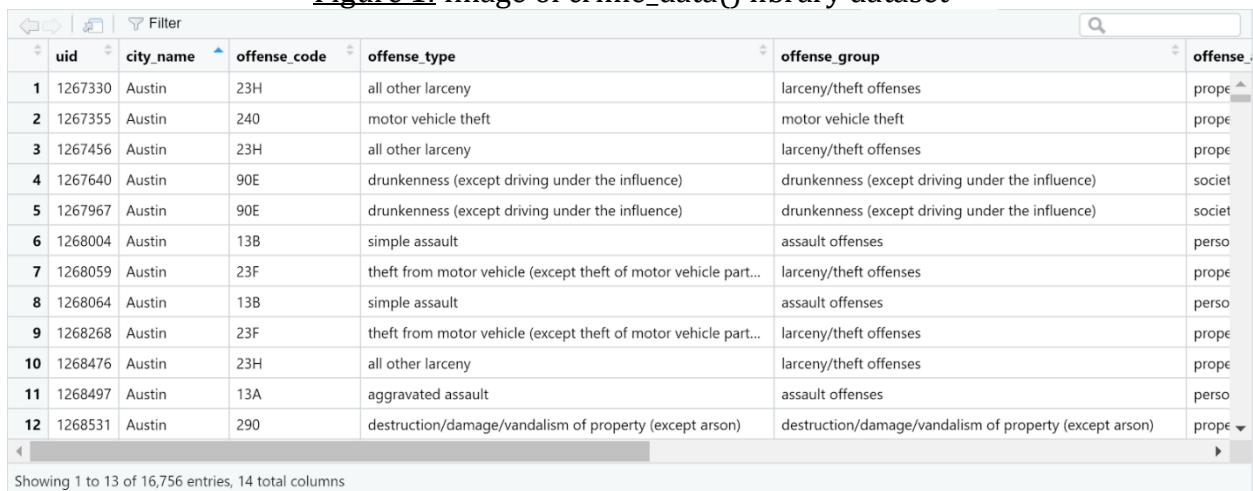
Noah Klur, Jordan Lian

09-Dec-2020

Background

For our project, we chose to look at crime rates throughout the US. CRAN has a pre-included library called `crime_data()`. From the library, we used the `get_crime_data()` function to get all crimes from 2019. Below is a screenshot of what `crime_data()` looks like:

Figure 1: Image of `crime_data()` library dataset



	uid	city_name	offense_code	offense_type	offense_group	offense_against
1	1267330	Austin	23H	all other larceny	larceny/theft offenses	property
2	1267355	Austin	240	motor vehicle theft	motor vehicle theft	property
3	1267456	Austin	23H	all other larceny	larceny/theft offenses	property
4	1267640	Austin	90E	drunkenness (except driving under the influence)	drunkenness (except driving under the influence)	society
5	1267967	Austin	90E	drunkenness (except driving under the influence)	drunkenness (except driving under the influence)	society
6	1268004	Austin	13B	simple assault	assault offenses	person
7	1268059	Austin	23F	theft from motor vehicle (except theft of motor vehicle part...)	larceny/theft offenses	property
8	1268064	Austin	13B	simple assault	assault offenses	person
9	1268268	Austin	23F	theft from motor vehicle (except theft of motor vehicle part...)	larceny/theft offenses	property
10	1268476	Austin	23H	all other larceny	larceny/theft offenses	property
11	1268497	Austin	13A	aggravated assault	assault offenses	person
12	1268531	Austin	290	destruction/damage/vandalism of property (except arson)	destruction/damage/vandalism of property (except arson)	property

Showing 1 to 13 of 16,756 entries, 14 total columns

This library dataset includes the reported crime rates for 15 major cities throughout the US in 2019, which include:

Table 1: List of 15 cities from crimedata library dataset

Austin	Boston	Chicago	Detroit	Fort Worth
Kansas City	Los Angeles	Louisville	Mesa	Nashville
New York	San Francisco	Seattle	St Louis	Tucson

Each of the 16,756 crimes were reported and organized using the following classifications:

Table 2: List of 14 organizational classifications in crimedata library dataset

uid	city_name	offense_code	offense_type	offense_group
offense_against	date_single	longitude	latitude	location_type
location_categories	census_block	start_date	end_date	

We imported this dataset to R as a dataframe, and then used it to run various statistical tests. The goal of this project was not only to determine which city has the highest crime rates overall, but we also wanted to understand if there was any correlation between crime rate and average income and crime rates and population.

Data Collection

We chose to work with the crimedata library in R. To conduct our analysis, we used dplyr, ggplot2, lubridate, tidyr, and fitdistrplus.

Variables in dataset:

uid: each crime report has a unique uid. This acts as the primary key for the entire dataset

city_name: each crime report was reported in one of the 15 cities listed above. Throughout our work we found this variable very useful for manipulating the data. We were able to make subsets

offense_code: this is an abbreviation for the offense type, enables easy searches

offense_type: this is a classification structure which groups crimes

date_single: this is the date of each crime, allows us to group/filter crimes per month

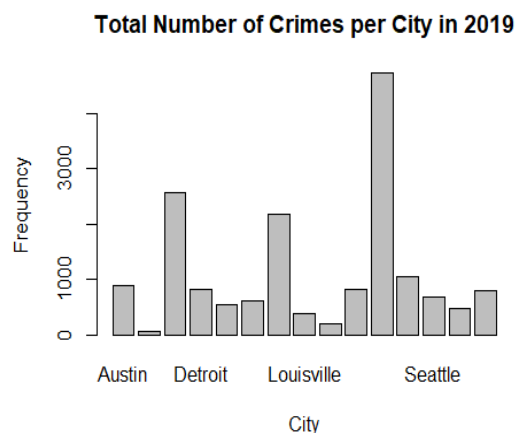
Figure 2: Setup libraries and original dataframe

```
library(crimedata)
library(dplyr)
library(ggplot2)
library(lubridate)
library(tidyr)
library(fitdistrplus)
```

Data Visualization

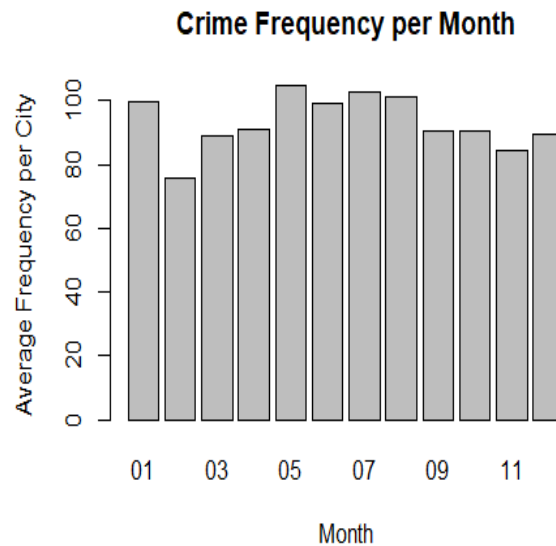
We first looked at crimes per city before picking a few cities to look at.

Figure 3: Barplot showing total crimes per city in 2019



We then decided to look at crimes per month. We had to manipulate the data and then filter out the month only from each crime. We modified the date_single variable to help us do that. Then we created a separate data frame from these results.

Figure 4: Crime frequency per month in 2019



From these results, we decided to look at New York, Chicago, Los Angeles, and Austin.

New York

Figure 5: Total crimes in New York per offence code in 2019

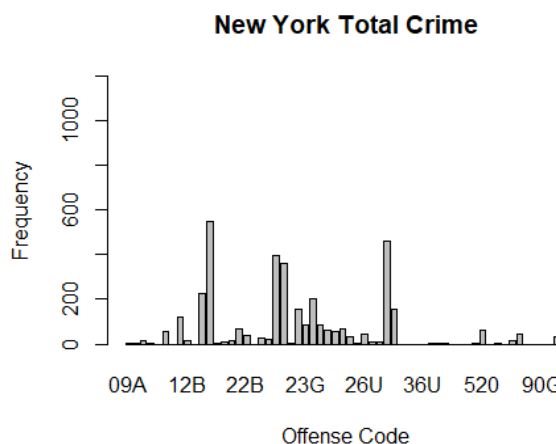
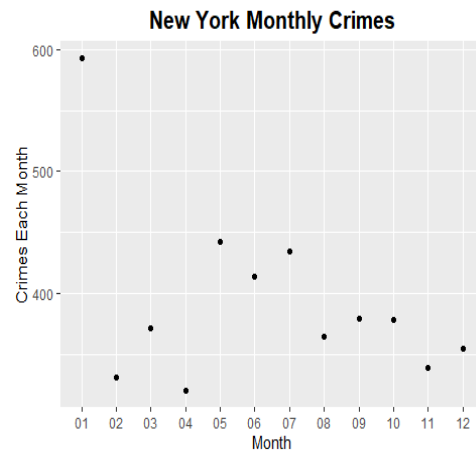


Figure 6: Monthly crime in New York in 2019



Chicago

Figure 7: Total crimes in Chicago per offence code in 2019

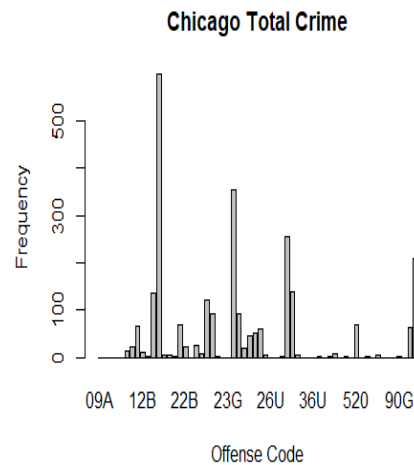
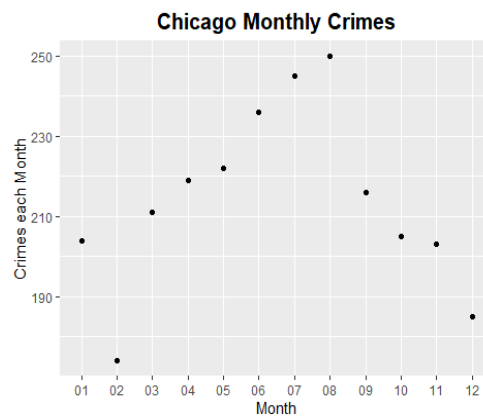


Figure 8: Monthly crimes in Chicago per month in 2019



Los Angeles

Figure 9: Total crime in Los Angeles per offense code in 2019

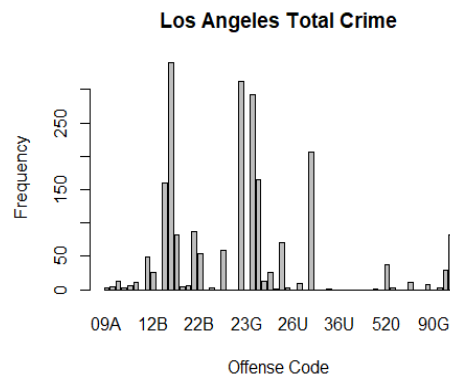
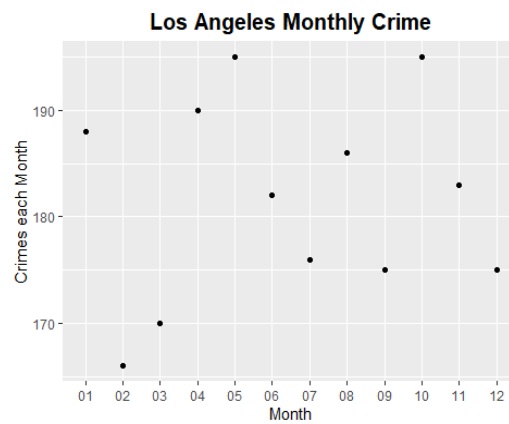


Figure 10: Monthly crime in Los Angeles per month in 2019



Austin

Figure 11: Total crime in Austin per offense code in 2019

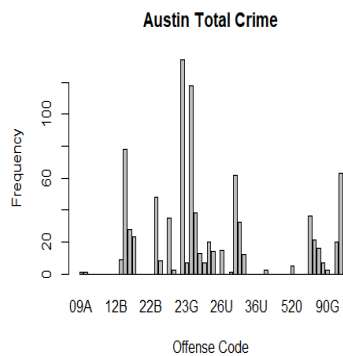


Figure 12: Monthly crimes in Austin per month in 2019

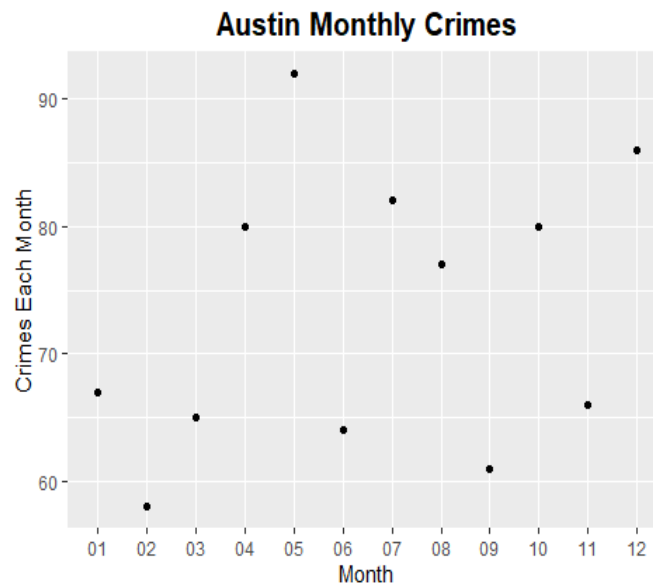
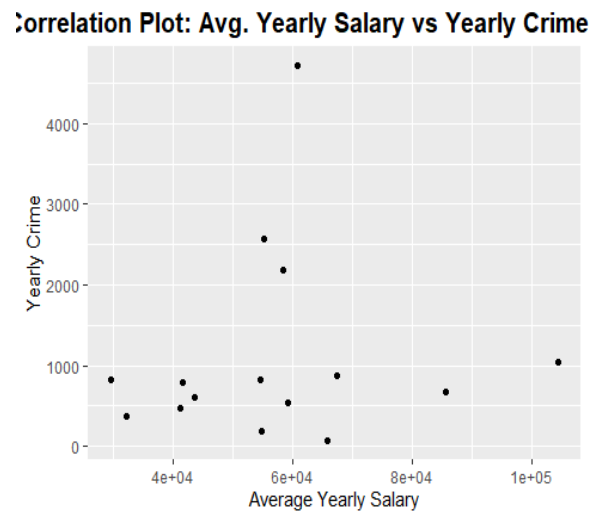


Figure 13: Correlation plot for average yearly salary vs yearly crime



```
## [1] 0.1050813
```

Figure 14: Heatmap for average yearly salary vs yearly crime

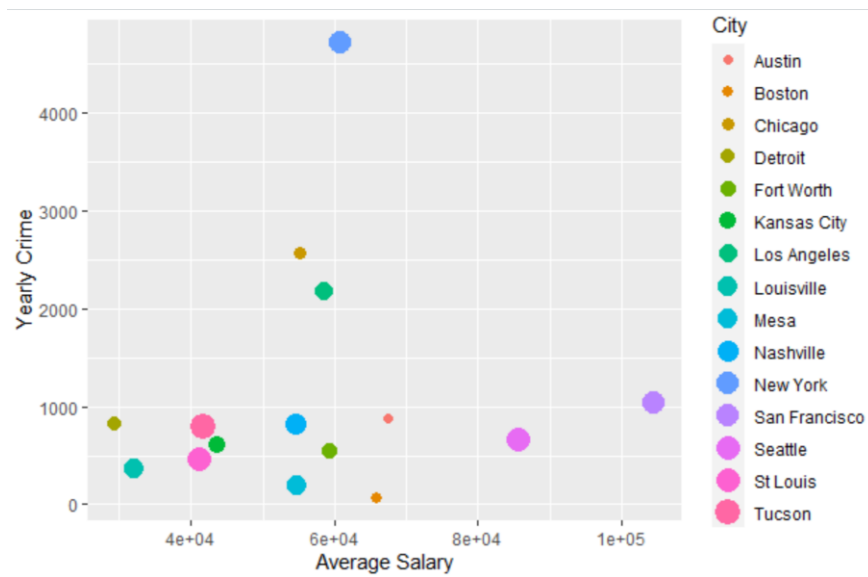
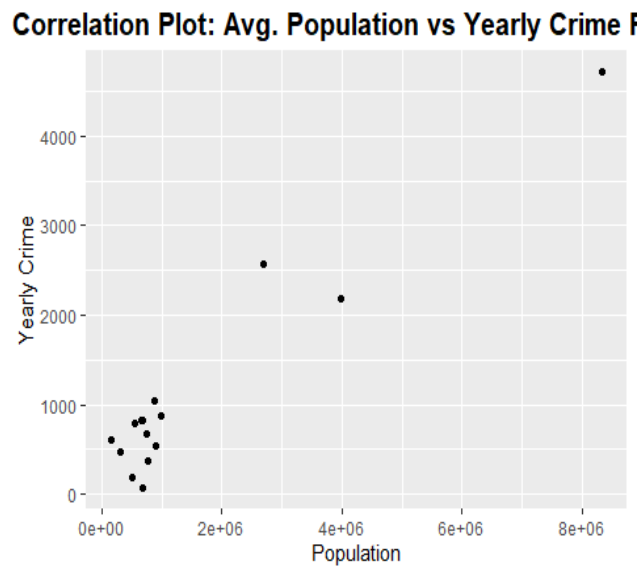
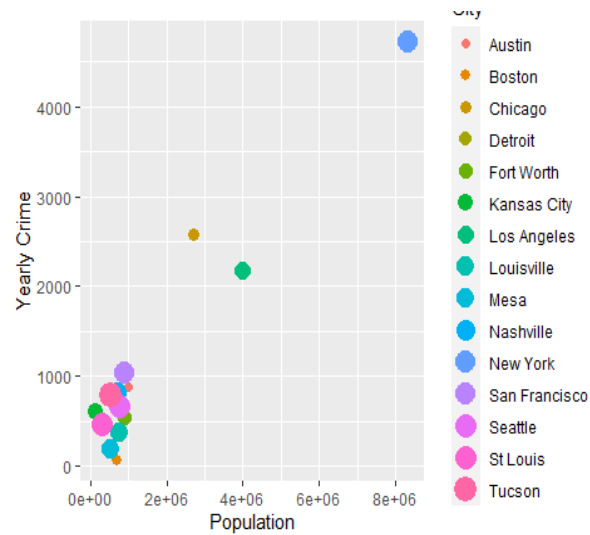


Figure 15: Correlation plot for average population vs yearly crime



```
## [1] 0.9585188
```

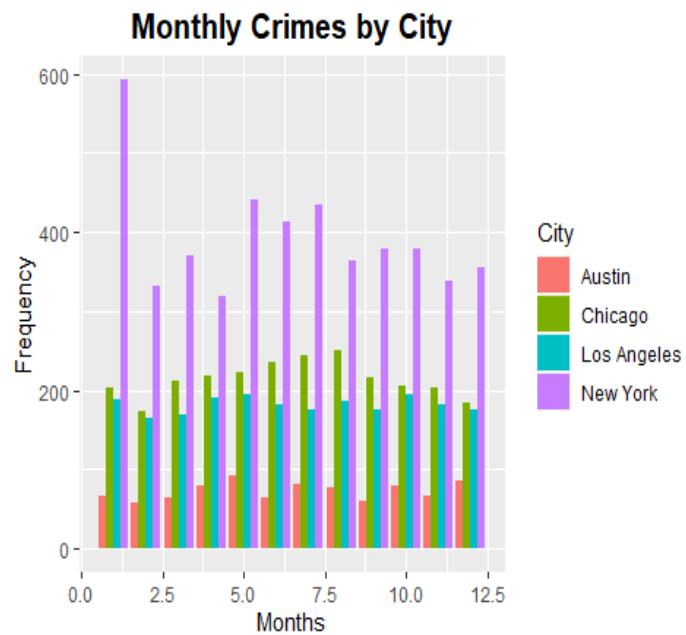
Figure 16: Heatmap for average population vs yearly crime



After getting visuals from all the cities, we compared monthly crime from each of the 4 cities. We created a new data frame to plot it on the same graph.

Grouped Bar Plot

Figure 17: Group Barplot for monthly crime per city in 2019



Statistical Analysis

We conducted 3 tests.

1. Hypothesis Test: For our hypothesis test, we chose to do two tests of unequal variances - looking at New York and then Austin. For the first hypothesis test of unequal variances, we look at New York as the sample from the population being the United States.

- New York

- Step 1: Hypothesis:
- Null Hypothesis: $H_0: \sigma^2 \leq \sigma_0^2$
- Alternate Hypothesis: $H_1: \sigma^2 > \sigma_0^2$
- Step 2: Test Statistic
- $\chi_{calc}^2 = \frac{(n-1)s^2}{\sigma_0^2}$
- For New York: $n = 4720, s^2 = 5436.424, \sigma_0^2 = 73.02411$
- $\chi_{calc}^2 = \frac{(4720-1)*5,436.424}{73.02411} = 351,315.269$
- Step 3: Rejection Region
- $\chi_{\alpha, n-1}^2 = \chi_{0.05, 4719}^2 = 4560.349$
- Step 4: Decision
- Since the calculated statistic is greater than the rejection region, we decide to reject the null hypothesis. There is enough evidence to conclude that the variance of monthly crimes in New York is greater than the average monthly crime variance.

- Austin

- Step 1: Hypothesis:
- Null Hypothesis: $H_0: \sigma^2 \leq \sigma_0^2$
- Alternate Hypothesis: $H_1: \sigma^2 > \sigma_0^2$
- Step 2: Test Statistic
- $\chi_{calc}^2 = \frac{(n-1)s^2}{\sigma_0^2}$
- For Austin: $n = 878, s^2 = 120.3333, \sigma_0^2 = 73.02411$
- $\chi_{calc}^2 = \frac{(878-1)120.333}{73.02411} = 1445.167$
- Step 3: Rejection Region
- $\chi_{\alpha, n-1}^2 = \chi_{0.05, 877}^2 = 809.2679$
- Step 4 : Decision
- Since the calculated statistic is less than the rejection region, we fail to reject the null hypothesis. There is enough evidence to conclude that the variance in monthly crimes in Austin is less than or equal to the average monthly crime variance.

2. Confidence Interval for New York
- $\alpha = 0.05, n = 4720, s^2 = 5436.424$

- $\frac{(n-1)s^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}$
- $\chi^2_{\alpha/2} = 4530.492$
- $\chi^2_{1-\alpha/2} = 4911.296$
- $\frac{(4720-1)*5436.242}{4530.492} < \sigma^2 < \frac{(4720-1)*5436.242}{4911.296}$
- $5223.567 < \sigma^2 < 5662.627$
- We can be 95% confident that the true variance for monthly crimes in New York is between 5223.567 and 5662.627.

3. Probability Sampling

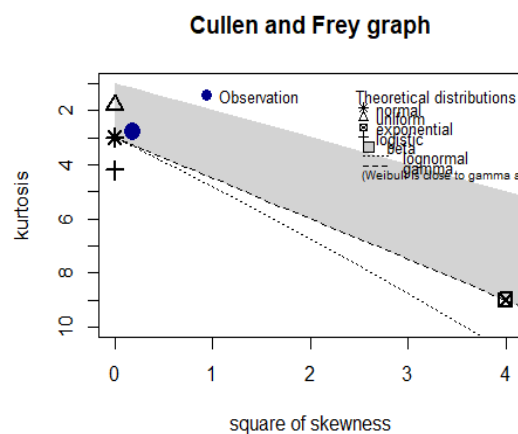
- $\alpha = 0.05$ $\bar{X} = 95$, $\mu = 93$, $\sigma = 8.545$, $n = 15$
- $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{95 - 93}{\frac{8.545}{\sqrt{15}}} = 0.9065$
- $P(x > 95) = 1 - p(z < 0.9065) = 1 - 0.2645 = 0.7356 = 73.56\%$
- There is a 73.56% probability of choosing a city with a monthly crime rate greater than 95 crimes per month.

Advanced Analytics

Below are some of the advanced analytics we chose to run for our project.

We started by doing descriptive statistics on the data, and the results are as follows:

Figure 18: Skewness graph for average monthly crime frequency



```
## summary statistics
## -----
## min: 75.86667    max: 104.6
## median: 90.53333
## mean: 93.08889
## estimated sd: 8.545414
## estimated skewness: -0.4283969
## estimated kurtosis: 2.782157
```

We then decided to fit the normal and lognormal distributions to the data and analyzing which seems to be better, as show below:

```
## Fitting of the distribution ' norm ' by maximum likelihood
## Parameters :
##      estimate Std. Error
## mean 93.088889    2.361828
## sd    8.181611    1.670064
## Loglikelihood: -42.24993    AIC: 88.49986    BIC: 89.46968
## Correlation matrix:
##      mean sd
## mean    1  0
## sd      0  1

## Fitting of the distribution ' lnorm ' by maximum likelihood
## Parameters :
##      estimate Std. Error
## meanlog 4.52956862 0.02599498
## sdlog    0.09004924 0.01837103
## Loglikelihood: -42.4933    AIC: 88.98661    BIC: 89.95642
## Correlation matrix:
##      meanlog      sdlog
## meanlog 1.000000e+00 4.241535e-13
## sdlog    4.241535e-13 1.000000e+00

## Goodness-of-fit statistics
##                                     1-mle-norm 2-mle-lnorm
## Kolmogorov-Smirnov statistic    0.1966249 0.18562611
## Cramer-von Mises statistic      0.0796565 0.07667349
## Anderson-Darling statistic      0.4502507 0.46200238
##
## Goodness-of-fit criteria
##                                     1-mle-norm 2-mle-lnorm
## Akaike's Information Criterion    88.49986    88.98661
## Bayesian Information Criterion    89.46968    89.95642
```

Figure 19: Statistical Graphs for average monthly crime frequency

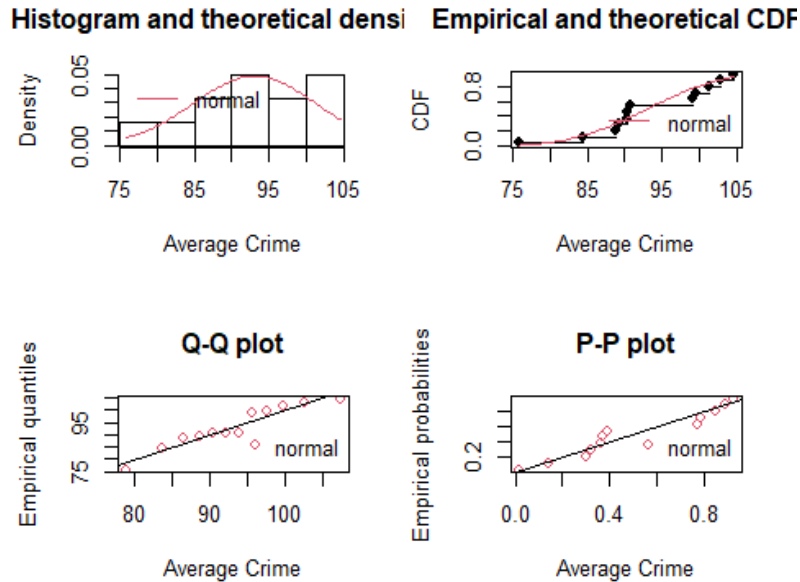
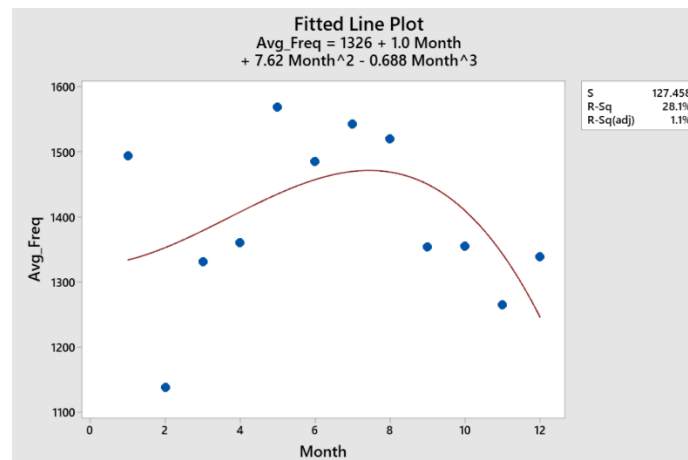


Figure 20: Minitab cubic regression



Conclusion

We found this project very interesting and generated several conclusions. Firstly, we found that New York, Chicago, and Los Angeles have significantly more yearly crimes than the other 12 cities. This led us to analyze them more and try to find correlations in these numbers. We thought of two possibilities and wanted to analyze yearly crime based on average income and population size. After manipulating the data we found extremely weak correlation between yearly crime and average yearly income. We then looked at yearly crime and population size and found a very strong positive correlation. So, we can confidently conclude that yearly crime rates are strongly correlated to population size for cities.