Econ 294A Midquarter Project Christina Louie
May 13, 2017

I. Describe your data (minimum of 200 observations)

A. What is your data?
B. What is the source of the data?
C. When/how/who can access that dataset?
D. Why is your data important?

The dataset that is used for this project is a record of offenses known to law enforcement that have occurred in the state of California in the year of the 2013. Furthermore, the count of crimes is categorized by city and by different type of offenses. Posted by the Federal Bureau of Investigation under the United States Department of Justice, the data is publicly available to anyone and can be downloaded from the uniform crime reporting webpage of the Federal Bureau of Investigation. Crimes are inevitable, especially in a well-developed nation. No matter where we live there will always be some sorts of crime existing whether it is a petty or a violent crime. By analyzing this dataset, it allows government officials (for this project, we will only look into California) to determine what sorts of program or events will increase Californian resident's awareness of this sort of crimes occurring within their state.

II. Issue(s) your data could address

A. What are the general topic(s) your issue(s) are connected to?
B. Specify why answering those questions are relevant.
C. Who will benefit from answering these questions?

California is the largest population among the fifty states. More importantly, larger population is associated with higher crime rate. Like any other states, one of the several issues that state government is concerned with is crime. A common question that state or local government pondered about is: Are there programs or policies that could reduce the number of crimes? What crime, specifically, occur the most in California? These are some issues that the dataset may be able to answer. These issues may be address through descriptive statistics of crimes categorize by what type of crimes in California. Several questions that may be answered by the datasets are the following:

- On average, which kind of crimes is higher, property or violent crimes?
- Is there a relationship between violent and property crimes, regardless of city or population?
- What is the relationship between property crime and population? How about the relationship between violent crime and population? Which crime, specifically, has the highest count in California?

These questions are relevant because observing the descriptive statistics provide an intuition of what sorts of awareness program that city officials or government officials should host or set up to increase the awareness. By answering these questions and informing these results, Californian residents, especially those in less-populated area, benefits from being aware of these offenses and for city council officials to take action to reduce the number of crimes within their cities.

III. In words, describe the algorithm you plan to conduct when working on your project in Python to address your question(s)

i. Empirical analysis:

a. Describe how you would use Python to conduct preliminary analysis on your data, perhaps by using one of the following:

1. Relationships between variable
2. Model selection (Justify variable/feature selection)
3. Calculations of predictions

b. Determine your future plans on how you can expand on this project to address the issue(s) you are interested in

c. Explain what you expect to clarify from your analysis

a) Python will be used to study relationships between variables as well as performing descriptive statistics of the counts on crimes. I will use Python to study the relationship between violent and property crimes as well as relationship among different crimes at a more define level. I will also use Python to find averages of crimes as well as the lowest and highest number of crimes occurring in the California.

b) This project is at a preliminary level. The dataset that is used in this project is cross-sectional data fixed in the year of 2013 and in the state of California. The project can be expanded further to the national level where we look at all 50 states in multiple years. Instead of addressing issues at the state or local government level, we look at the crimes occurring within the United States. At this level, we can apply not the preliminary analysis, but we can also control and observe by regions and counties, and merge other data regarding characteristics of each state that may be useful to control for outside factors. We can also further expand several relationships between crimes across regions and possibly wealth of each state. Although this project is done for basic empirical analysis purpose, the idea of project can also be turned into a randomized experiment to study a possible implementation of a safety policy or safety awareness program across counties or states.

c) From the analysis, I expect to find out which type of crimes occur the most in California as well as sufficient statistical description of crimes in California. These results should be able to provide a preliminary intuition of what sort of program to institute into the lives of Californian residents.

IV. Describe your process

A. Which strategies worked for you in using Python to work through this project? For example, you may write something about

i. Websites that helped for the code
ii. Papers that relate to the project
iii. Planning strategies that helped

B. Which strategies / coding aspects posed problems?

This project does not have a process similar to that of an experiment. The goal of this project is to collect preliminary descriptive results of the data, especially on property crime and violent crimes and be able to provide an intuition or idea of what sorts of policy, programs, or awareness events to alert Californian’s residents of these crimes. Since my project is on empirical analysis of crimes, strategies that worked in using Python to work through this project are Stack Overflow and the packages main webpage to help with the coding. I had to use Google to search through several Stack Overflow page to help with plotting and counting values sum. I also had to use the main site of the Python packages (matplotlib and pandas) to figure out some code for plot editing. Since my project is focused on primarily descriptive statistics, strategies or coding posed no problems. However, the statistical analysis of the crime data brought on more questions regarding characteristics of cities and possible relationship to certain crimes. We can further expand this project by collecting data at the municipal level, regarding wealth of city, household income and size, and average wage of each household.

Some useful sites:
https://matplotlib.org/users/pyplot_tutorial.html
https://stackoverflow.com/questions/18730299/python-sum-function-with-list-parameter

V. Discuss your analyses

A. Provide details and explanations about the following (when applicable)

i. Statistical properties

ii. Relevant plots and tables

iii. Associations between variables

B. Show your predictions/simulations/benchmark values

i. Explain why each included variable is important

C. Discuss why such analyses are important

i. Which companies/groups/individuals would benefit from your analyses?

A and B:
Answering some of the questions that I have mentioned in the previous part, I observe that the average number of property crimes in California is 1883.08 and the average number of violent crimes is 269.69. This suggests that property crimes are higher in California than that of violent crimes. I also observe a plot of violent crimes against population. This plot shows a positive relationship between population and violent crimes. As population increases, violent crimes also increase. Similarly, the plot of property crimes against population also has a positive relationship. As population increases, property crimes also increase. I also plot a figure of the relationship between property crimes and violent crimes. This figure also shows a positive relationship, suggesting that more violent crimes may lead to more property crimes. I also looked at several linear regressions between violent crimes and property crimes. Regardless of city or population, we find that the estimated effect of violent crime on property crime is, on average, 5.2302. The ratio of property crime to violent crime is 5.2302. Suppose that cities of California are divided into population group. I define a large population to be greater than or equal to 60,000 people and small population to be less 60,000 people. When I run a regression of property crime on violent crime for cities within the large population, the estimated effect of property crime on violent crime is 5.0996. Likewise for the small population, the estimated effect of property crime on violent crime is 5.5298. The three estimates do not differ by much; however, I notice that the estimated effect is larger in cities with smaller population than in cities with larger population. To expand further on other specific crimes, I also looked at the relationship between rape and murder (or manslaughter) and the relationship between arson and burglary. Running regressions on these relationships, we find that the estimated effect of rape on murder and manslaughter is 0.2926 and the estimated effect of burglary on arson is 0.0604. Both of these estimates are close to zero, and are statistically significant at the 95% level. Although these

estimates are statistically significant, the estimations are close to zero and its overall significance does not seems to be a huge influence to the overall crime in California. Overall, property crime has the largest number of crimes in the state of California. The variables chosen for the analysis are important because property crime and violent crime are the two main types of crime that affects the state of California. I also look at results based on population because cities with smaller population tend to be less developed than cities with larger population and these less developed cities are less likely to have programs or awareness events on crimes. For curiosity, I also looked into murder, rape, burglary, and arson to see if there are relationships and also because these crimes have the least amount of offenses.

C:

These preliminary results have shown that property crimes contribute to most crimes in the state of California, especially in cities with lower population. It gives us a slight intuition that programs and awareness events on property crimes should be promoted by local governments or the state government, especially in towns or cities that are less populated. Setting policies that bring awareness to less populated towns and for California overall will benefit the Californian residents and may reduce the number of the crimes overall.

VI. Concluding Remarks

A. Present arguments to relevant companies/groups/individuals why your project will benefit them

i. Detail the merits of your project

B. List steps you can take to build on your project

C. If applicable, discuss steps to get funding, additional data, and resource for experiments

By setting policies that bring awareness to less populated towns and for California overall, residents will have more awareness and it may reduce the number of the crimes overall. Based from the preliminary results, the data shows that cities will smaller population have a higher estimated effect of violent crime on property crime. If there are policies that mandate setting more street surveillance or encourage local government to set up crime-watch programs, it's presence of surveillance and residents being more aware may reduce the number of crime activities within their cities. Other possible actions that can take place is notification on electronic devices or advertisement on television that promotes residents to be aware of their surrounding and belongings. Although we have these basic ideas of bringing awareness, they are costly. We need to find better ways that are more cost effective or less costly. The results presented in this project are preliminary. There was no control of outside factors or grouping by different ranges of population. These are two concerns we should look if we build the project further. Another step we should take is to collect more data for the state of California regarding to city characteristics as well as grouping by regions.

Overall, here are two steps to build on the project:

1) Collect data on characteristics of city and control for these outside factors:
    a. What ethnicity lives in that city?
    b. Are their gangs?

c. What is the overall wealth of city?

2) Instead of the looking at just the state of California, we look at the overall crimes of the United States in multiple years.
   a. Collect crime data for the past 10 years
   b. Collect crime data for all 49 states (if possible)

Note: These data are publicly available on the site of the Federal Bureau of Investigation. Having all these data, continue further on with the empirical analysis.

Figure 1: Plot of Property Crimes on Population

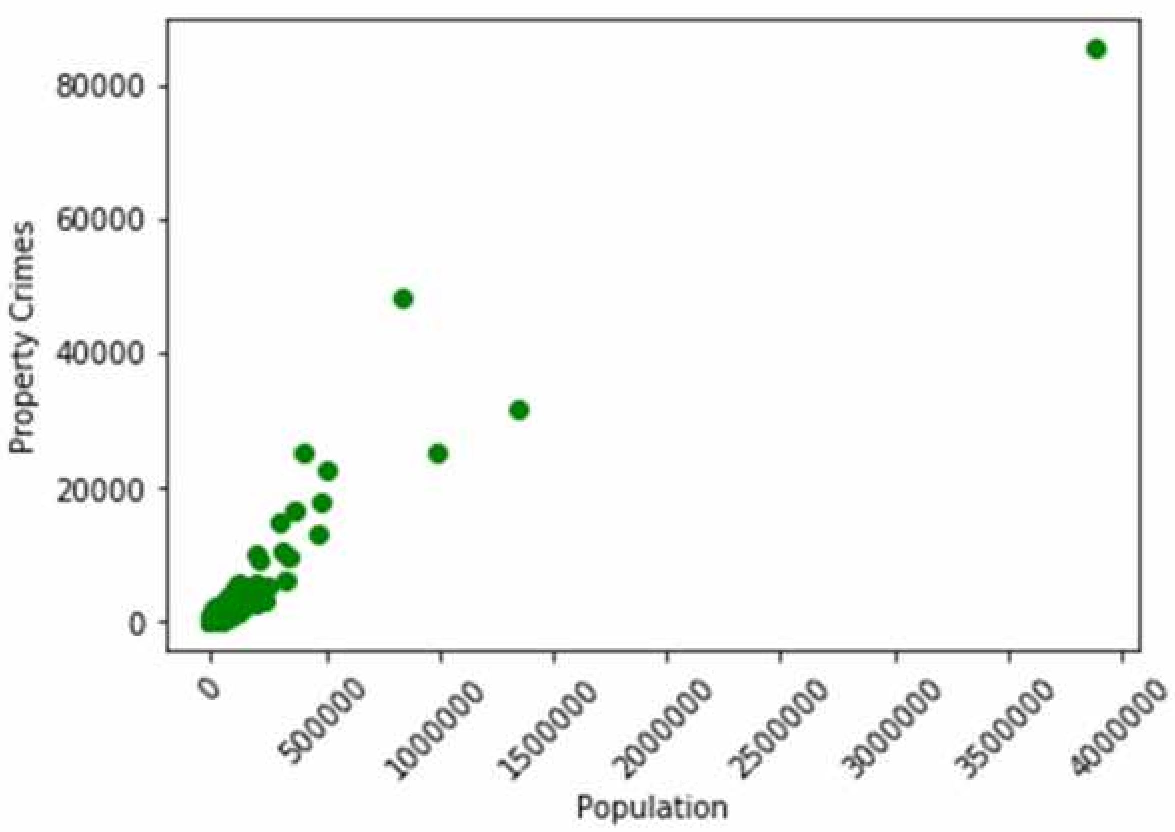


Figure 2: Plot of Violent Crimes on Population

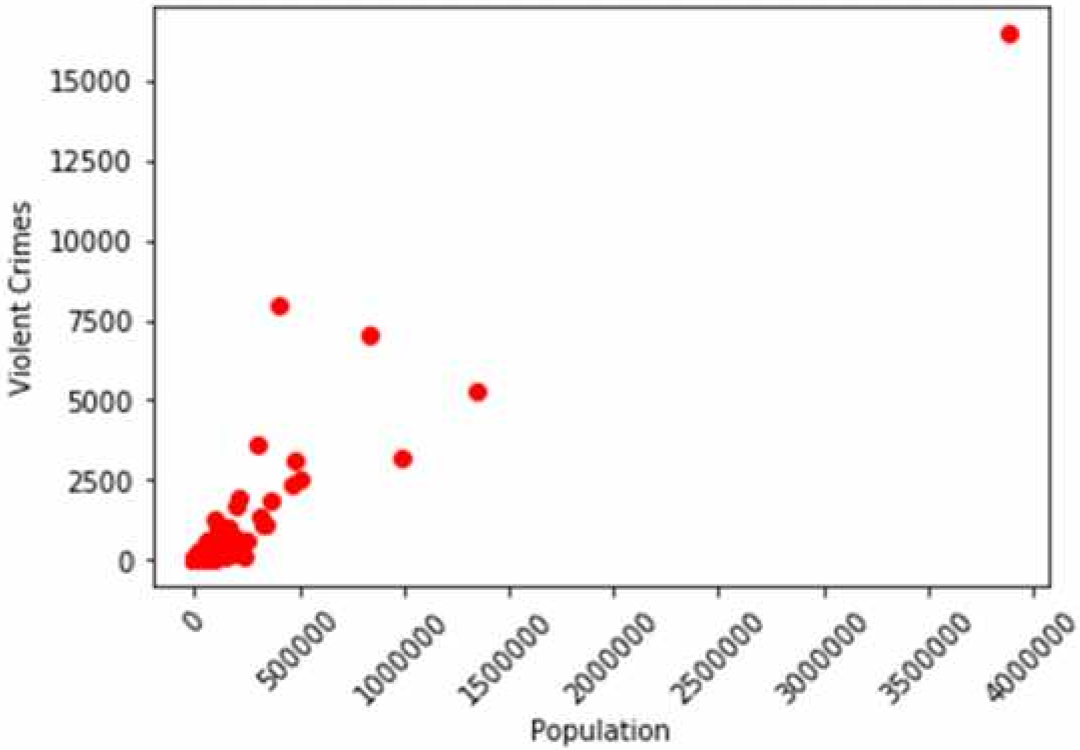
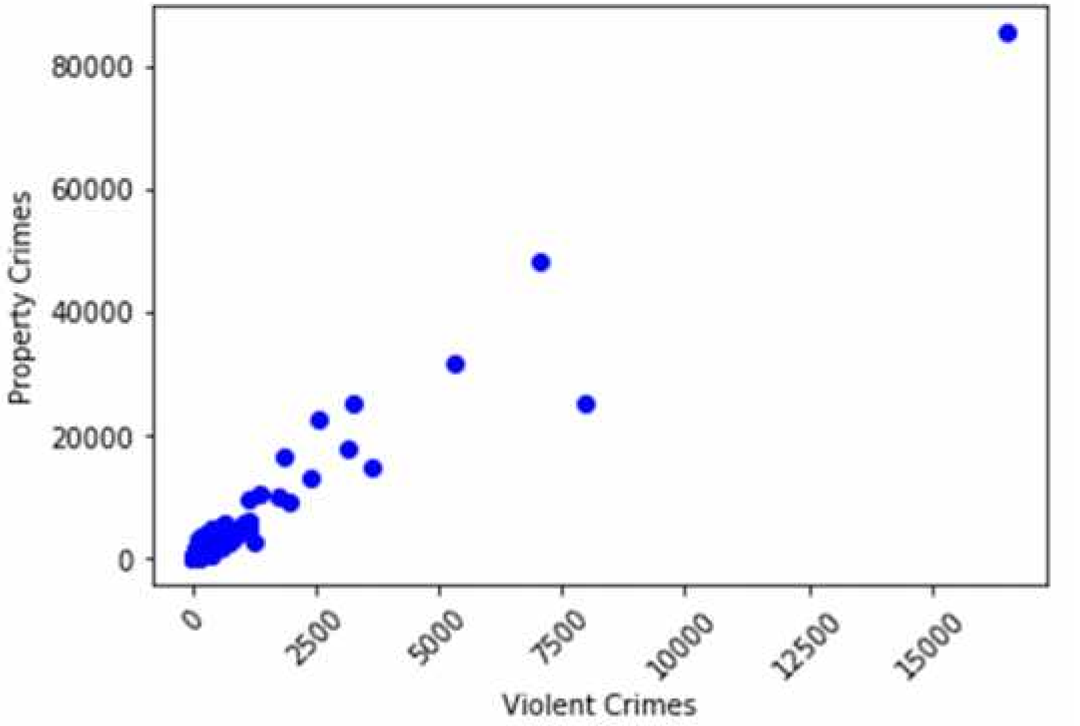
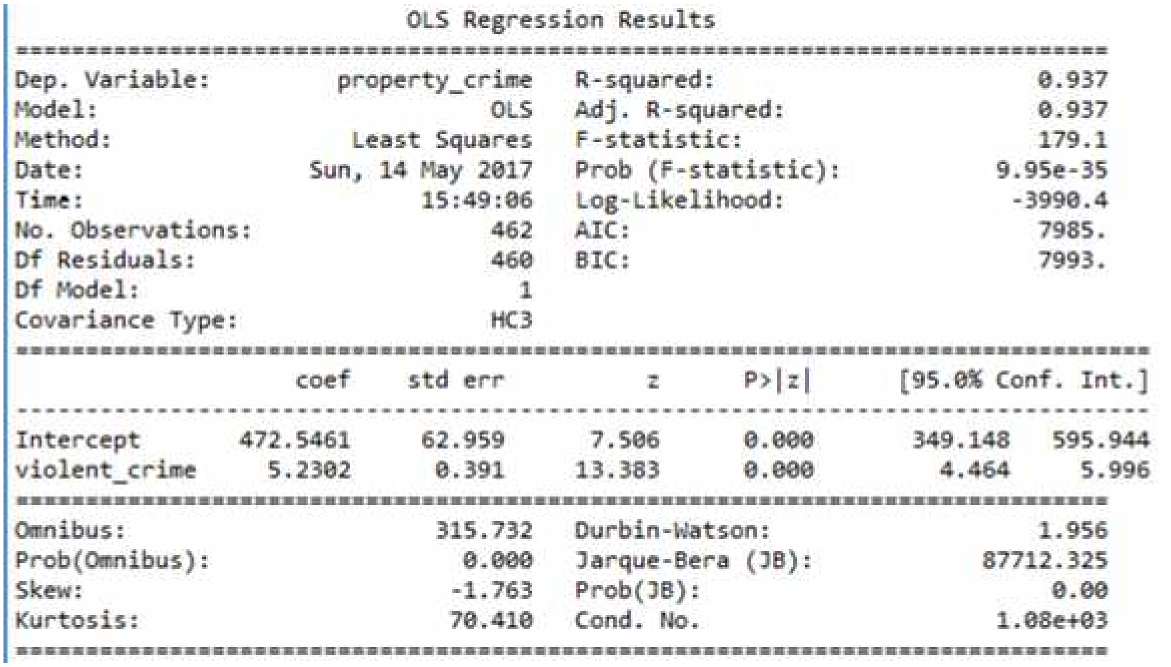


Figure 3: Plot of Property Crimes on Violent Crimes



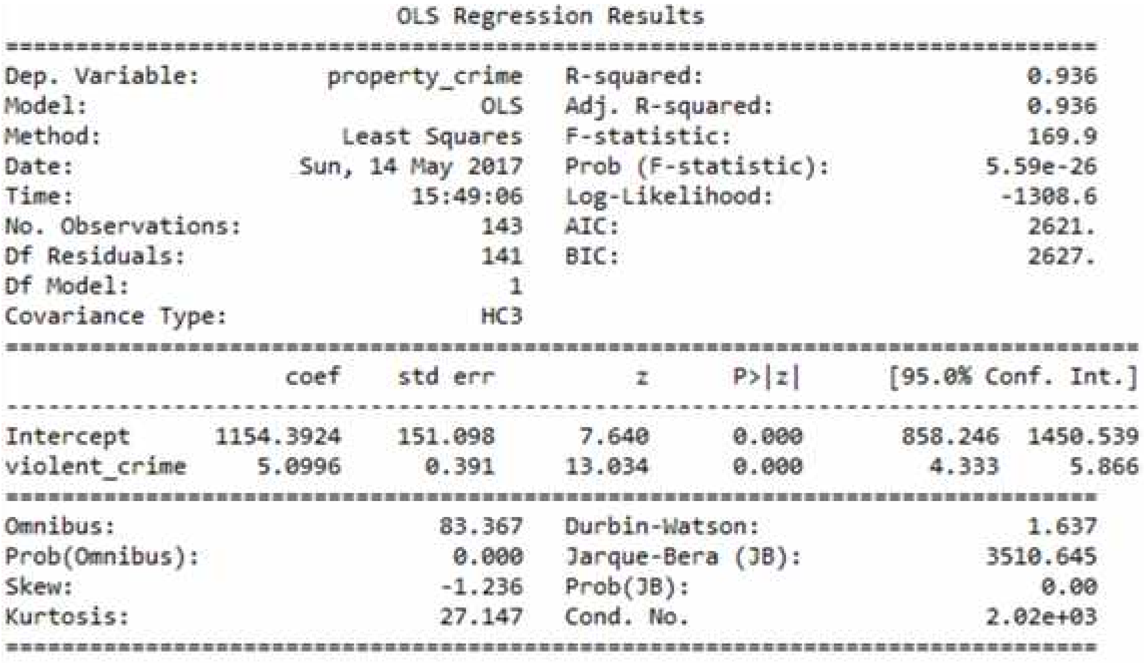
Econ 294A Midquarter Project

Regression of Property Crime on Violent Crime: All population



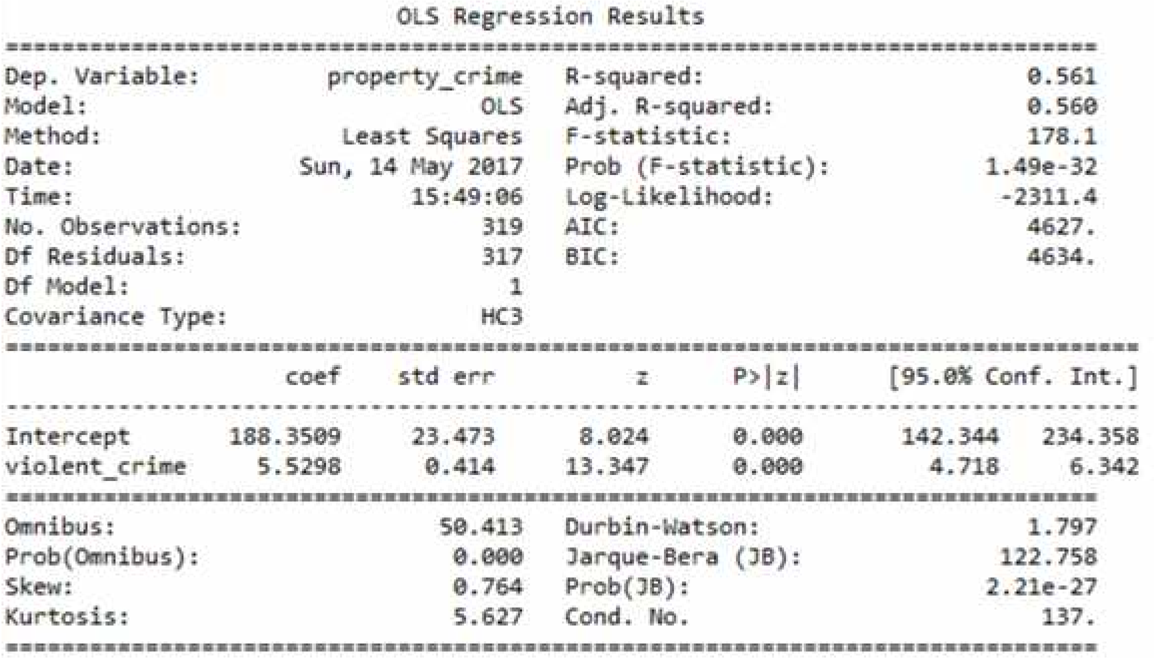
```
                            OLS Regression Results
==============================================================================
Dep. Variable:         property_crime   R-squared:                       0.937
Model:                            OLS   Adj. R-squared:                  0.937
Method:                 Least Squares   F-statistic:                     179.1
Date:                Sun, 14 May 2017   Prob (F-statistic):           9.95e-35
Time:                        15:49:06   Log-Likelihood:                -3990.4
No. Observations:                 462   AIC:                             7985.
Df Residuals:                     460   BIC:                             7993.
Df Model:                           1
Covariance Type:                  HC3
==================================================================================
                   coef    std err          z      P>|z|      [95.0% Conf. Int.]
----------------------------------------------------------------------------------
Intercept      472.5461     62.959      7.506      0.000       349.148   595.944
violent_crime    5.2302      0.391     13.383      0.000         4.464     5.996
==============================================================================
Omnibus:                      315.732   Durbin-Watson:                   1.956
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            87712.325
Skew:                          -1.763   Prob(JB):                         0.00
Kurtosis:                      70.410   Cond. No.                     1.08e+03
==============================================================================
```

For Large Population:



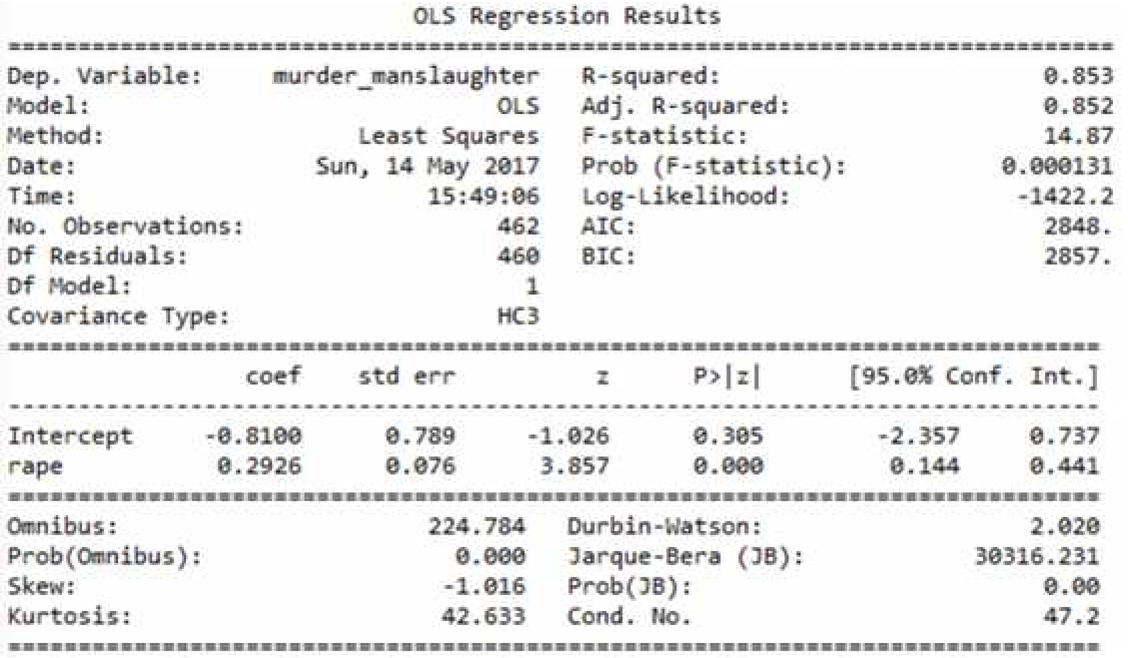
```
                            OLS Regression Results
==============================================================================
Dep. Variable:         property_crime   R-squared:                       0.936
Model:                            OLS   Adj. R-squared:                  0.936
Method:                 Least Squares   F-statistic:                     169.9
Date:                Sun, 14 May 2017   Prob (F-statistic):           5.59e-26
Time:                        15:49:06   Log-Likelihood:                -1308.6
No. Observations:                 143   AIC:                             2621.
Df Residuals:                     141   BIC:                             2627.
Df Model:                           1
Covariance Type:                  HC3
==================================================================================
                   coef    std err          z      P>|z|      [95.0% Conf. Int.]
----------------------------------------------------------------------------------
Intercept     1154.3924    151.098      7.640      0.000       858.246  1450.539
violent_crime    5.0996      0.391     13.034      0.000         4.333     5.866
==============================================================================
Omnibus:                       83.367   Durbin-Watson:                   1.637
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             3510.645
Skew:                          -1.236   Prob(JB):                         0.00
Kurtosis:                      27.147   Cond. No.                     2.02e+03
==============================================================================
```

For small population:



```
                            OLS Regression Results
==============================================================================
Dep. Variable:          property_crime   R-squared:                       0.561
Model:                             OLS   Adj. R-squared:                  0.560
Method:                  Least Squares   F-statistic:                     178.1
Date:                 Sun, 14 May 2017   Prob (F-statistic):           1.49e-32
Time:                         15:49:06   Log-Likelihood:                -2311.4
No. Observations:                  319   AIC:                             4627.
Df Residuals:                      317   BIC:                             4634.
Df Model:                            1
Covariance Type:                   HC3
=================================================================================
                    coef    std err          z      P>|z|      [95.0% Conf. Int.]
---------------------------------------------------------------------------------
Intercept       188.3509     23.473      8.024      0.000       142.344   234.358
violent_crime     5.5298      0.414     13.347      0.000         4.718     6.342
==============================================================================
Omnibus:                        50.413   Durbin-Watson:                   1.797
Prob(Omnibus):                   0.000   Jarque-Bera (JB):              122.758
Skew:                            0.764   Prob(JB):                     2.21e-27
Kurtosis:                        5.627   Cond. No.                         137.
==============================================================================
```

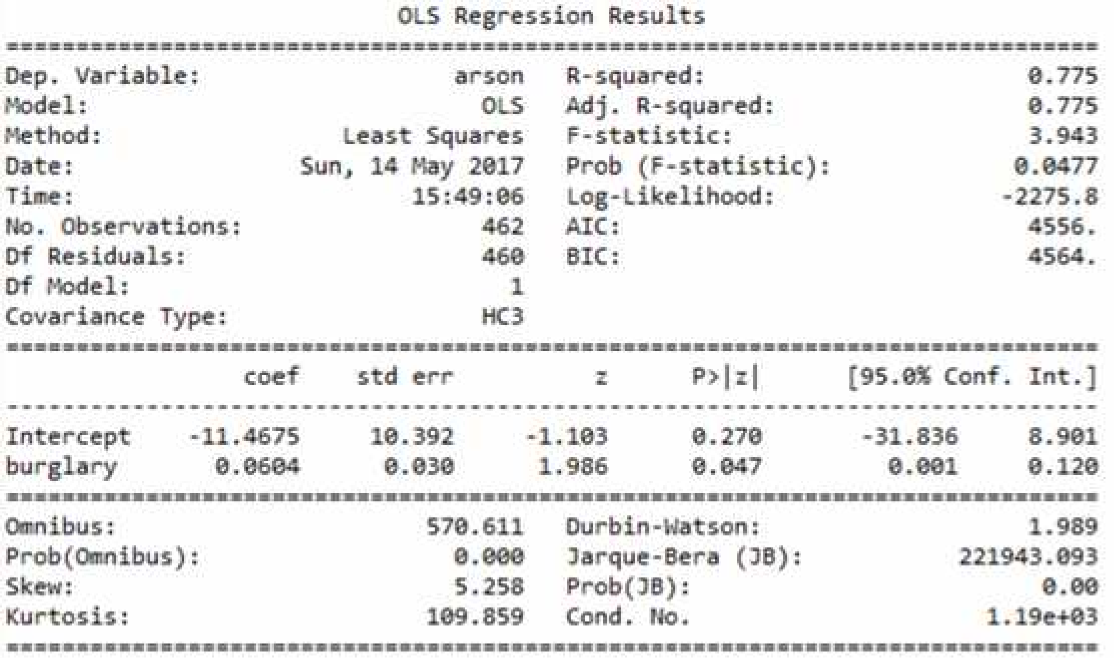
Regression of Murder and Manslaughter on Rape



```
                            OLS Regression Results
==============================================================================
Dep. Variable:     murder_manslaughter   R-squared:                       0.853
Model:                             OLS   Adj. R-squared:                  0.852
Method:                  Least Squares   F-statistic:                     14.87
Date:                 Sun, 14 May 2017   Prob (F-statistic):           0.000131
Time:                         15:49:06   Log-Likelihood:                -1422.2
No. Observations:                  462   AIC:                             2848.
Df Residuals:                      460   BIC:                             2857.
Df Model:                            1
Covariance Type:                   HC3
==============================================================================
                 coef    std err          z      P>|z|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept     -0.8100      0.789     -1.026      0.305        -2.357     0.737
rape           0.2926      0.076      3.857      0.000         0.144     0.441
==============================================================================
Omnibus:                       224.784   Durbin-Watson:                   2.020
Prob(Omnibus):                   0.000   Jarque-Bera (JB):            30316.231
Skew:                           -1.016   Prob(JB):                         0.00
Kurtosis:                       42.633   Cond. No.                         47.2
==============================================================================
```

Econ 294A Midquarter Project

Regression of Arson on Burglary



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  arson   R-squared:                       0.775
Model:                            OLS   Adj. R-squared:                  0.775
Method:                 Least Squares   F-statistic:                     3.943
Date:                Sun, 14 May 2017   Prob (F-statistic):             0.0477
Time:                        15:49:06   Log-Likelihood:                -2275.8
No. Observations:                 462   AIC:                             4556.
Df Residuals:                     460   BIC:                             4564.
Df Model:                           1
Covariance Type:                  HC3
==============================================================================
                 coef    std err          z      P>|z|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept    -11.4675     10.392     -1.103      0.270       -31.836     8.901
burglary       0.0604      0.030      1.986      0.047         0.001     0.120
==============================================================================
Omnibus:                      570.611   Durbin-Watson:                   1.989
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           221943.093
Skew:                           5.258   Prob(JB):                         0.00
Kurtosis:                     109.859   Cond. No.                     1.19e+03
==============================================================================
```

R Code:

```
"""
Created on Sat May 13 18:19:20 2017

Econ 294A Python Lab Midterm Project
@author: Christina Louie
Date: May 13, 2017
"""
# ------- libraries -------
import pandas as pd
import statsmodels.formula.api as sm
import matplotlib.pyplot as plt


# ------- Some Preliminary Analysis -----------
filelocation = r'C:\Users\ChristinaL\Documents\Econ 294A Python Lab\crime_ca_2013.csv'
crime_dataset = pd.read_csv(filelocation)
#print(crime_dataset)

# -- Property Crime --
#plot of property crimes against population
plt.plot(crime_dataset['population'],crime_dataset['property_crime'],'go')
plt.suptitle("Figure 1: Plot of Property Crimes on Population")
plt.xticks(rotation=45)
plt.xlabel('Population')
plt.ylabel('Property Crimes')
plt.show()


# Find the largest number of property crimes
sorted_property = crime_dataset.sort_values(['property_crime'], ascending=False)
print(sorted_property.head(n=10))

# Find the average number of property crimes/offenses
propertyCrimeAvg = crime_dataset['property_crime'].mean()
print(propertyCrimeAvg)

# Mean is 1883.0844155844156

# -- Violent Crime --
#plot of violent crimes against population
plt.plot(crime_dataset['population'],crime_dataset['violent_crime'],'ro')
plt.suptitle("Figure 2: Plot of Violent Crimes on Population")
plt.xticks(rotation=45)
plt.xlabel('Population')
plt.ylabel('Violent Crimes')
plt.show()


# Find the largest number of violent crimes
sorted_violent = crime_dataset.sort_values(['violent_crime'], ascending=False)
print(sorted_violent.head(n=10))

# Find the average number of violent crimes/offenses
violentCrimeAvg = crime_dataset['violent_crime'].mean()
print(violentCrimeAvg)

# Mean is 269.6926406926407

# -- Relationship between property crime and violent crime --
# plot
plt.plot(crime_dataset['violent_crime'],crime_dataset['property_crime'],'bo')
plt.suptitle("Figure 3: Plot of Property Crimes on Violent Crimes")
plt.xlabel('Violent Crimes')
plt.ylabel('Property Crimes')
plt.xticks(rotation=45)
plt.show()
```

```
# linear regression
result1 = sm.ols(formula="property_crime ~
violent_crime",data=crime_dataset).fit(cov_type='HC3')
print(result1.params)
print(result1.summary())

# --- subset of data ---
# Define large population to be greater than or equal to 60,000 people
largePop_df = crime_dataset.loc[crime_dataset['population']>=60000]

# linear regression
result2 = sm.ols(formula="property_crime ~
violent_crime",data=largePop_df).fit(cov_type='HC3')
print(result2.params)
print(result2.summary())

# Define small population to be less than 60,000 people
smallPop_df = crime_dataset.loc[crime_dataset['population']<60000]

result3 = sm.ols(formula="property_crime ~
violent_crime",data=smallPop_df).fit(cov_type='HC3')
print(result3.params)
print(result3.summary())

# --- sum of each variable to get the total number of crimes (each type) ---
total_violent = sum(list(crime_dataset['violent_crime']))
print("Total Violent Crimes: %d"  %(total_violent))

total_murder = sum(list(crime_dataset['murder_manslaughter']))
print("Total Murder and Manslaughter: %d"  %(total_murder))

total_rape = sum(list(crime_dataset['rape']))
print("Total Rape: %d"  %(total_rape))

total_robbery = sum(list(crime_dataset['robbery']))
print("Total Robbery: %d"  %(total_robbery))

total_assault = sum(list(crime_dataset['aggravated_assault']))
print("Total Aggravated Assault: %d"  %(total_assault))

total_property = sum(list(crime_dataset['property_crime']))
print("Total Property Crimes: %d"  %(total_property))

total_burglary = sum(list(crime_dataset['burglary']))
print("Total Burglary: %d"  %(total_burglary))
```

```
total_theft = sum(list(crime_dataset['larceny_theft']))
print("Total Larceny Theft: %d"  %(total_theft))

total_mvTheft = sum(list(crime_dataset['motor_vehicle_theft']))
print("Total Motor Vehicle Theft: %d"  %(total_mvTheft))

total_arson = sum(list(crime_dataset['arson']))
print("Total Arson: %d"  %(total_arson))

# Output:
#Total Violent Crimes: 124598
#Total Murder and Manslaughter: 1400
#Total Rape: 6064
#Total Robbery: 48035
#Total Aggravated Assault: 69099
#Total Property Crimes: 869985
#Total Burglary: 190417
#Total Larceny Theft: 539803
#Total Motor Vehicle Theft: 139765
#Total Arson: 6203

# Some possible correlation...
# --- relationship between rape and murder and manslaughter ---
result1 = sm.ols(formula="murder_manslaughter ~
rape",data=crime_dataset).fit(cov_type='HC3')
print(result1.params)
print(result1.summary())

# --- relationship between arson and burglary ---
result1 = sm.ols(formula="arson ~ burglary",data=crime_dataset).fit(cov_type='HC3')
print(result1.params)
print(result1.summary())
```