Index

- Index--1
- This project---2
- Technical report---3
- Descriptive Analysis--4
- Symmetry and Kurtosis---11
- Regression---13
- Tools used in this project--14

This program has the objective of creating a statistical report from data provided by a big database.

In order to do it, it will have to order and analyze the information.

This is done by connecting to a database and running queries until the information has been filtered and tailored to the point where it can be used for this report.

Python will also give you the option to select the target of the analysis and with that in mind will create a path of commands where, depending on the quality of the data, it will decide the outcome of the analysis (for example, if a variable needs to be normalized

Technical Report

This report has been done with academical purposes only
The database information was created by a module of this program that works with the creation of tendencies in mind.

For practical purposes, we are going to define this database free of any bias that may incur while extracting the sample and its dependencies

The sample contains a list of different types of crime that were recorded in 5 different important cities during the period of one year,divided by months.This report will select one city and type of crime and perform a statistical analysis of its data.

Any extreme value will be normalized.
It is correct to assume that there might be crimes that were not recorded during this period of time. Therefore its important to define that this sample is representative of the population

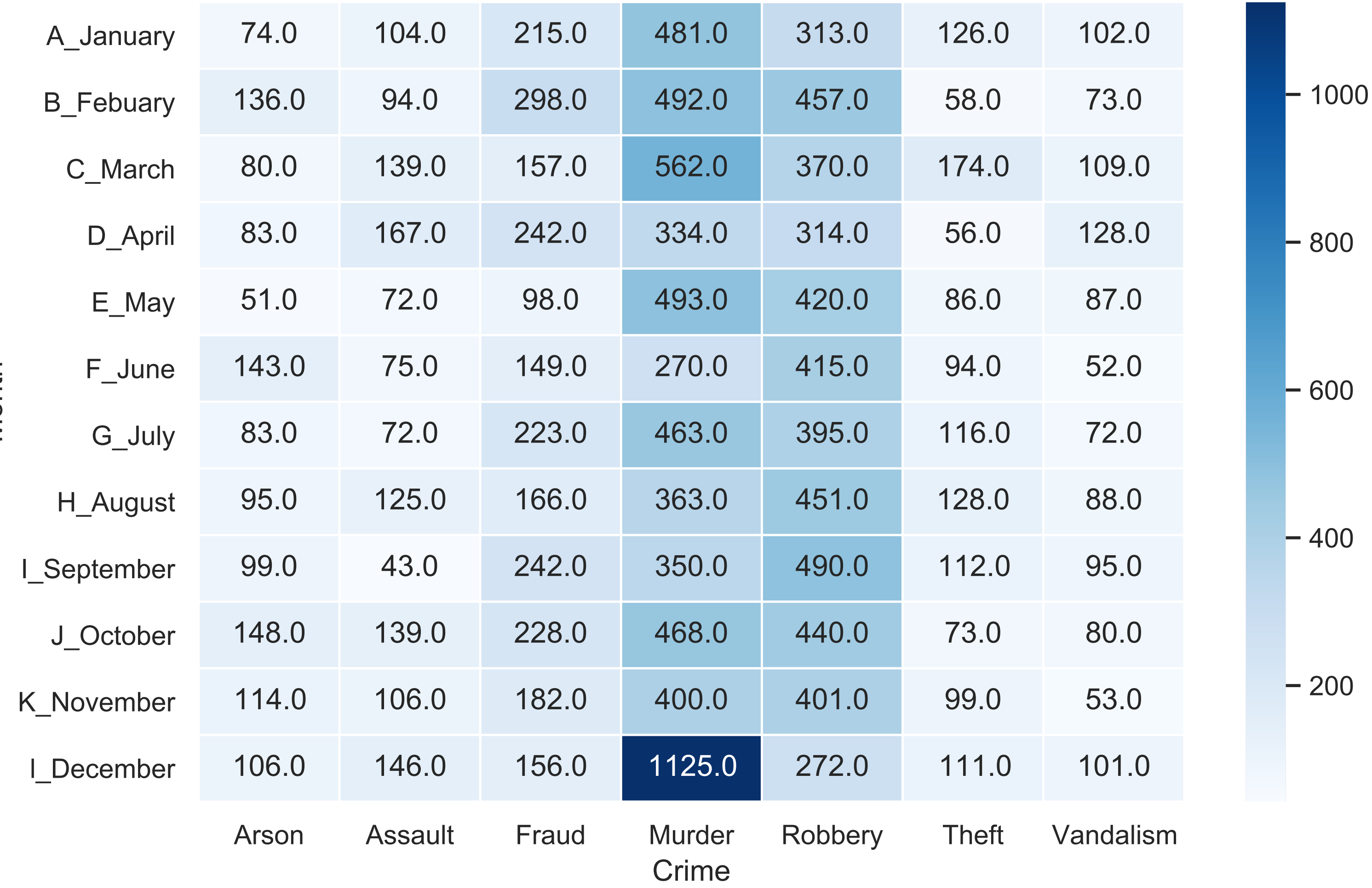
City Selected: Buenos Aires
Crime Selected: Murder
Individual entries in the sample: 15206

Descriptive analysis

The first step in this report is to evaluate the frequency te data repeats itself.

In this report, each individual case will be grouped by their month of ocurrence.

The next page will show how the data is classified by the time it happened, it will be also compare the frequency of Murders with the other crime types in the database



	Arson	Assault	Fraud	Murder	Robbery	Theft	Vandalism
A_January	74.0	104.0	215.0	481.0	313.0	126.0	102.0
B_Febuary	136.0	94.0	298.0	492.0	457.0	58.0	73.0
C_March	80.0	139.0	157.0	562.0	370.0	174.0	109.0
D_April	83.0	167.0	242.0	334.0	314.0	56.0	128.0
E_May	51.0	72.0	98.0	493.0	420.0	86.0	87.0
F_June	143.0	75.0	149.0	270.0	415.0	94.0	52.0
G_July	83.0	72.0	223.0	463.0	395.0	116.0	72.0
H_August	95.0	125.0	166.0	363.0	451.0	128.0	88.0
I_September	99.0	43.0	242.0	350.0	490.0	112.0	95.0
J_October	148.0	139.0	228.0	468.0	440.0	73.0	80.0
K_November	114.0	106.0	182.0	400.0	401.0	99.0	53.0
I_December	106.0	146.0	156.0	1125.0	272.0	111.0	101.0

As shown in the histogram, we can organize the data to make understanding it easier.

The data can be studied further, we can calculate different variables from the information that will help this study

Variables:

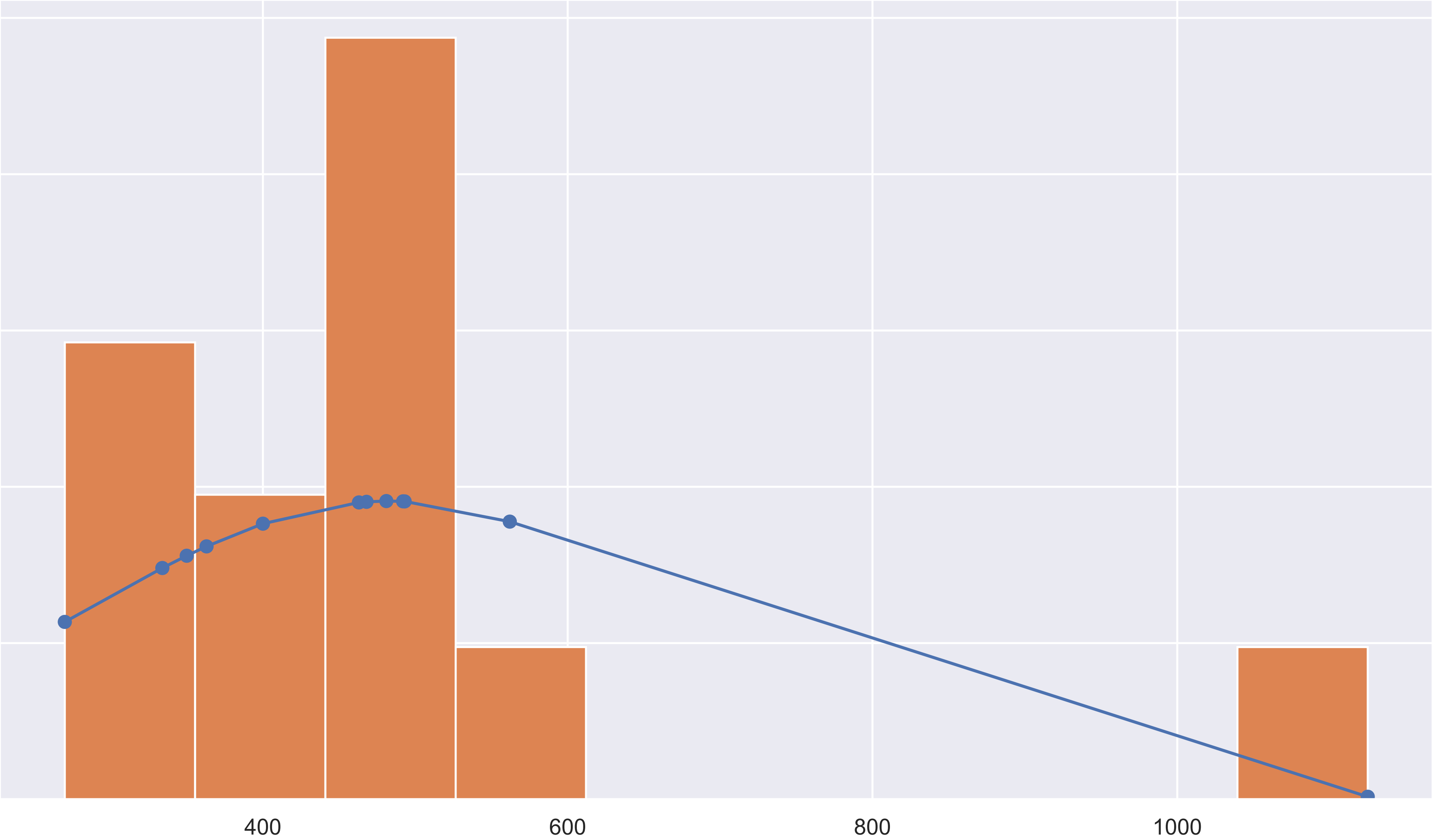
Mean: 483

Standard Deviation: 209

Highest Value: 1125

If we look at the histogram one more time, its clear that
the highes value does not help this statistical analysis
and it stands as a problem because it si not epresentative of
the population for being and isolated result

As we dont expect it to repeat again in the future we will
normalize the value, by bringing it close to the Mean
An histogram in the next page will show graphically why this
extreme value is not representative the population


400
600
800
1000

After normalizing the extreme values, we can calculate the variables again and see how the data is shown in the new plot.

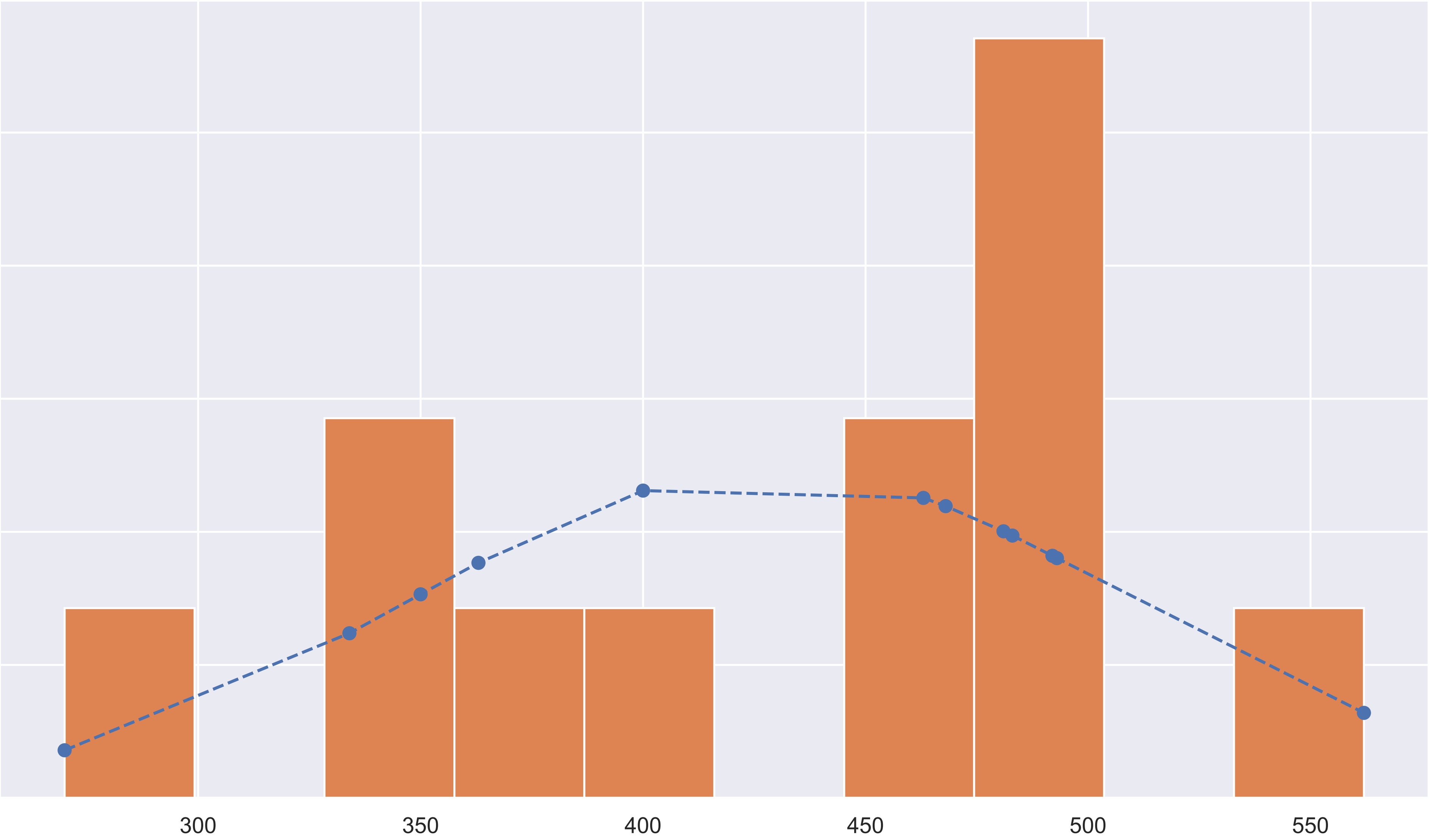
Variables:

Mean: 429

Standard Deviation: 81

Highest Value: 562

Comparing this new values with the ones we got earlier might not seem much but if we look at the next histogram, the result will be much more homogeneous.



Normal Distribution

Symmetry

If we compare this f(x) with Gauss Bell curve, our data will be skewed to the left.
This means that from all the frequency series, the mean should be the most suited for this study.

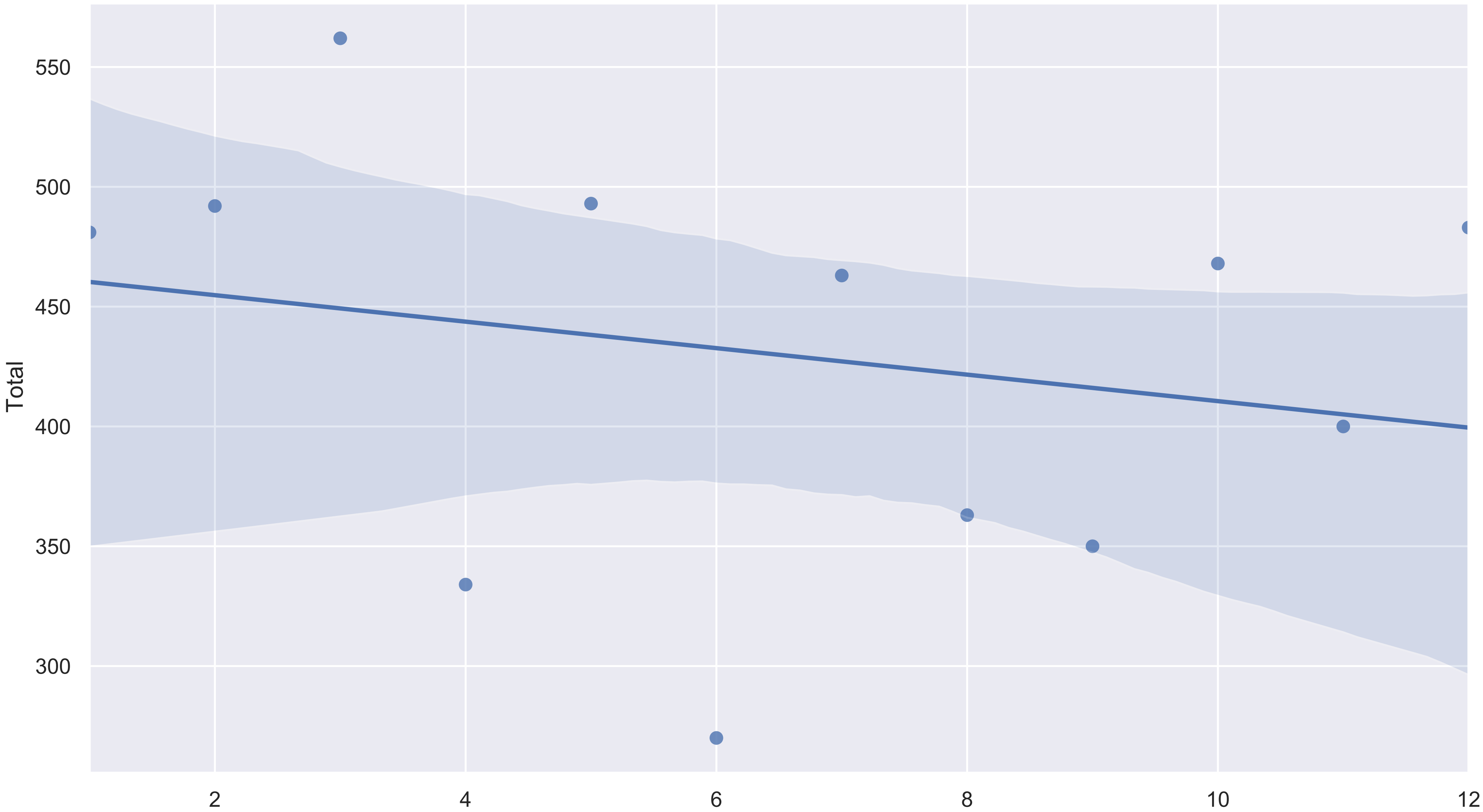
Kurtosis

The Kurtosis we get from the data is -0.83 this means that the curve is Platykurtic.
When we talk about a Platykurtic curve, it meants that
the kurtosis is flatter (less peaked) when compared with the normal distribution.
The lower the value, the less we can rely on the frequency distribution variables.

Regression

In the next page there is a plot explaining the relation between the variable time and the total of crimes in that month.

This is important not only to decision making, but it also help ups to forecast the near future. The tendency we can see is the line that connects these variables and



Regression & Tendency in the last 12 Months

Tools Used in this project

-Python
--Python Libraries
----Pymysql
----Seaborn
----Matplotlib
----Pandas
----Numpy
----Scipy.stats
----Random
----Time
----Pdfpages

-MySql

-Aws Cloud Services

-Jupyter Notebook

-Knowledge Required
--SQL Query
--ETL
--Statistics
--Programming