CRIME FORECASTING FOR CITY OF PORTLAND

BY

DEEPAK SHANMUGAM

MEGHANA POCHIRAJU

NITESH SRIVATSAV

SUNEESHA KUDIPUDI

VIGNESH VIJAYAKUMAR

VINAYAK RAJU GOPAL

PROJECT DEFINITION

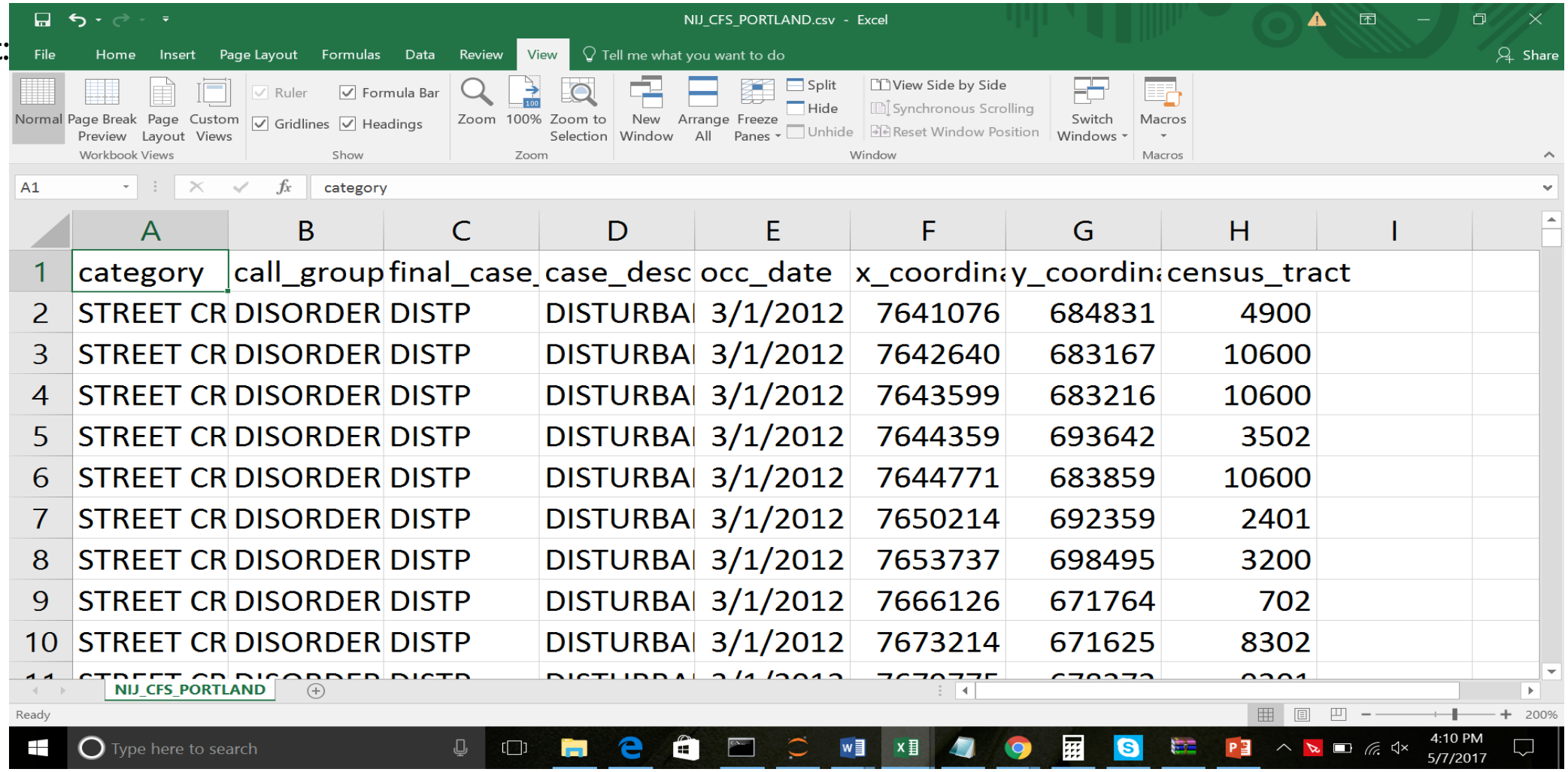
- Over the course of the last decade, Crime has seen a steady increase across all major cities in America, and this creates an uneasiness among the public. With the help of cutting edge computing technologies, we now have access to a tool that uses machine learning techniques to warn us of the number of crimes which may possibly occur for a specific instance of time. In this project, we develop an unsupervised machine learning technique with an external knowledge base, as our input to the Forecasting system which provides us with a specific number for a given location. We use the Spark MLIB for clustering the dataset for creating a structured input for our forecasting system. The forecasting system puts the ARIMA model to use.
- Our project aims to predict the number of crimes for every county in the city of Portland in Oregon for a future instance of time so that it would benefit the PPD(Portland Police Department) thereby reducing or preventing loss.

MOTIVATION

- Crime has always been difficult to predict even with increase in technology over the years . Increase in technology has not resulted in a decrease in the number of crimes .With increase in population over the years in the city of Portland crime has also been increasing at a rapid rate
- Research indicates that crimes usually occur in counties where the average per capita income is lower than the average for the entire state.These counties demonstrate a higher tendency for the number of crimes for a given instance of time. Our idea is to use the latest technologies to forecast a number(crime rate in a future instance of time) for every county(large data set) in the city of Portland so that it would benefit the local Police Department in allocating members in advance thereby preventing or minimizing damage.

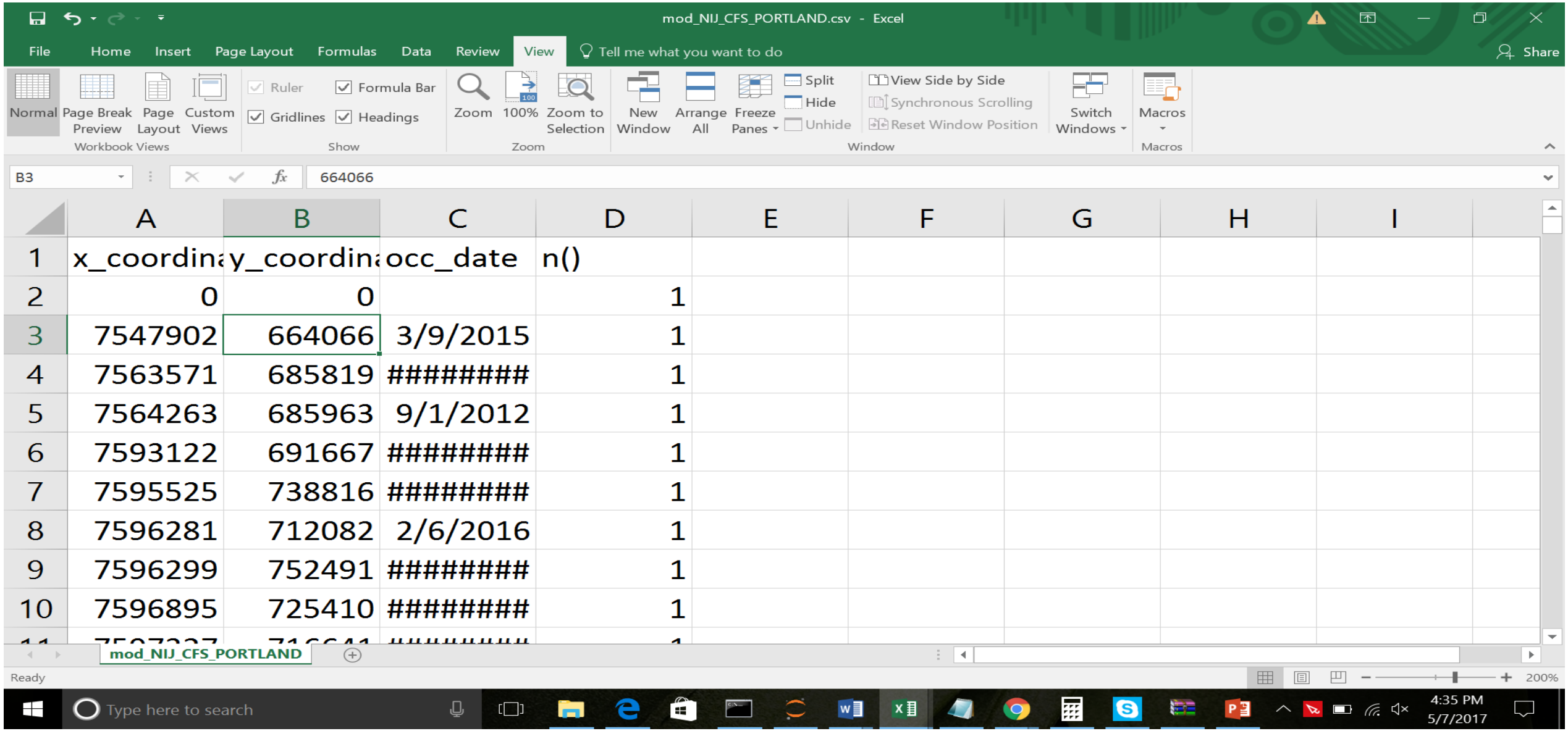
DATASETS

Initial dataset:



	A	B	C	D	E	F	G	H	I
1	category	call_group	final_case_	case_desc	occ_date	x_coordina	y_coordina	census_tract	
2	STREET CR	DISORDER	DISTP	DISTURBA	3/1/2012	7641076	684831	4900	
3	STREET CR	DISORDER	DISTP	DISTURBA	3/1/2012	7642640	683167	10600	
4	STREET CR	DISORDER	DISTP	DISTURBA	3/1/2012	7643599	683216	10600	
5	STREET CR	DISORDER	DISTP	DISTURBA	3/1/2012	7644359	693642	3502	
6	STREET CR	DISORDER	DISTP	DISTURBA	3/1/2012	7644771	683859	10600	
7	STREET CR	DISORDER	DISTP	DISTURBA	3/1/2012	7650214	692359	2401	
8	STREET CR	DISORDER	DISTP	DISTURBA	3/1/2012	7653737	698495	3200	
9	STREET CR	DISORDER	DISTP	DISTURBA	3/1/2012	7666126	671764	702	
10	STREET CR	DISORDER	DISTP	DISTURBA	3/1/2012	7673214	671625	8302	

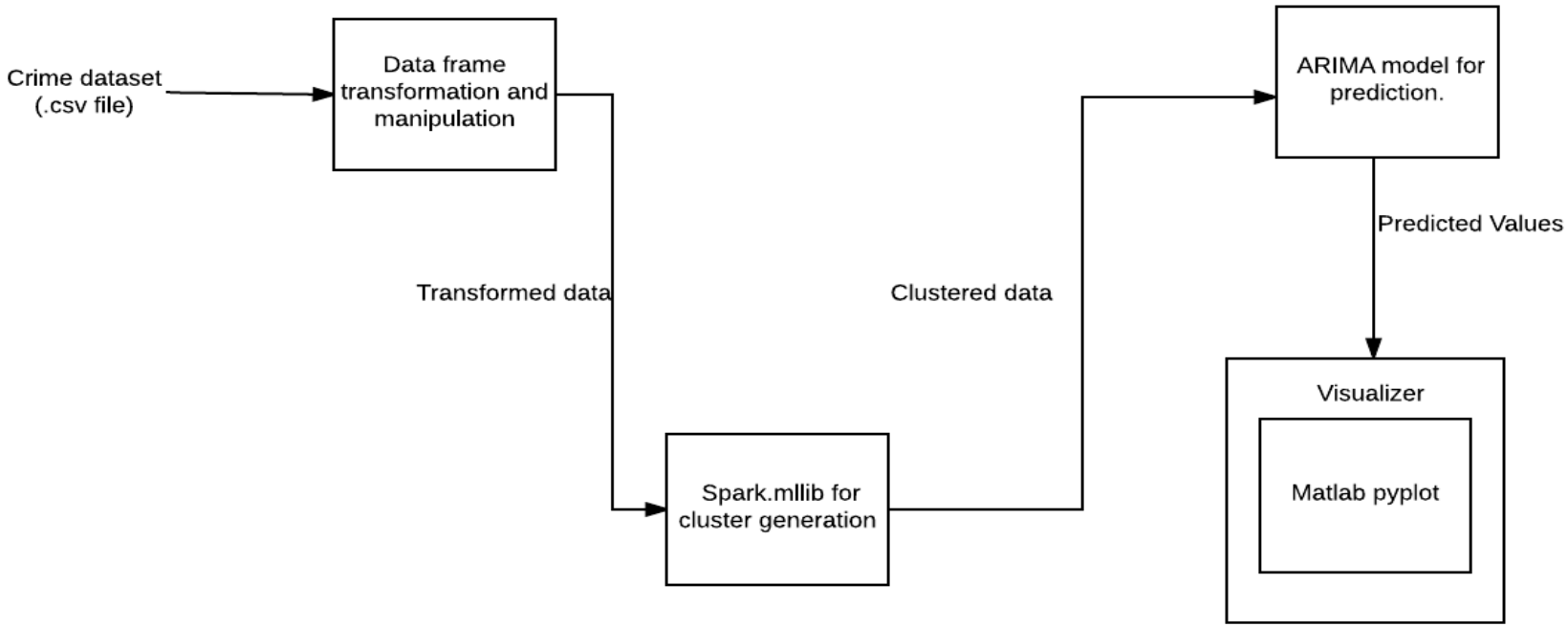
Transformed dataset(Grouped based on the location attribute and filtered the unnecessary columns)



	A	B	C	D	E	F	G	H	I
1	x_coordina	y_coordina	occ_date	n()					
2	0	0		1					
3	7547902	664066	3/9/2015	1					
4	7563571	685819	#########	1					
5	7564263	685963	9/1/2012	1					
6	7593122	691667	#########	1					
7	7595525	738816	#########	1					
8	7596281	712082	2/6/2016	1					
9	7596299	752491	#########	1					
10	7596895	725410	#########	1					

EXPERIMENTS

Project architecture:



Tools used:

Data Transformer is one of the entities of the framework which is responsible for grouping data based on location parameters to prepare the data for clustering.

Apache Spark is an open-source distributed framework for data analytics. It avoids the I/O bottleneck of the conventional two-stage MapReduce programs by providing in-memory cluster computing. Also, it supports both batch processing and streaming data

Spark MLib is Spark's scalable machine learning library consisting of common learning algorithms and utilities, including classification, regression, clustering, collaborative filtering, dimensionality reduction

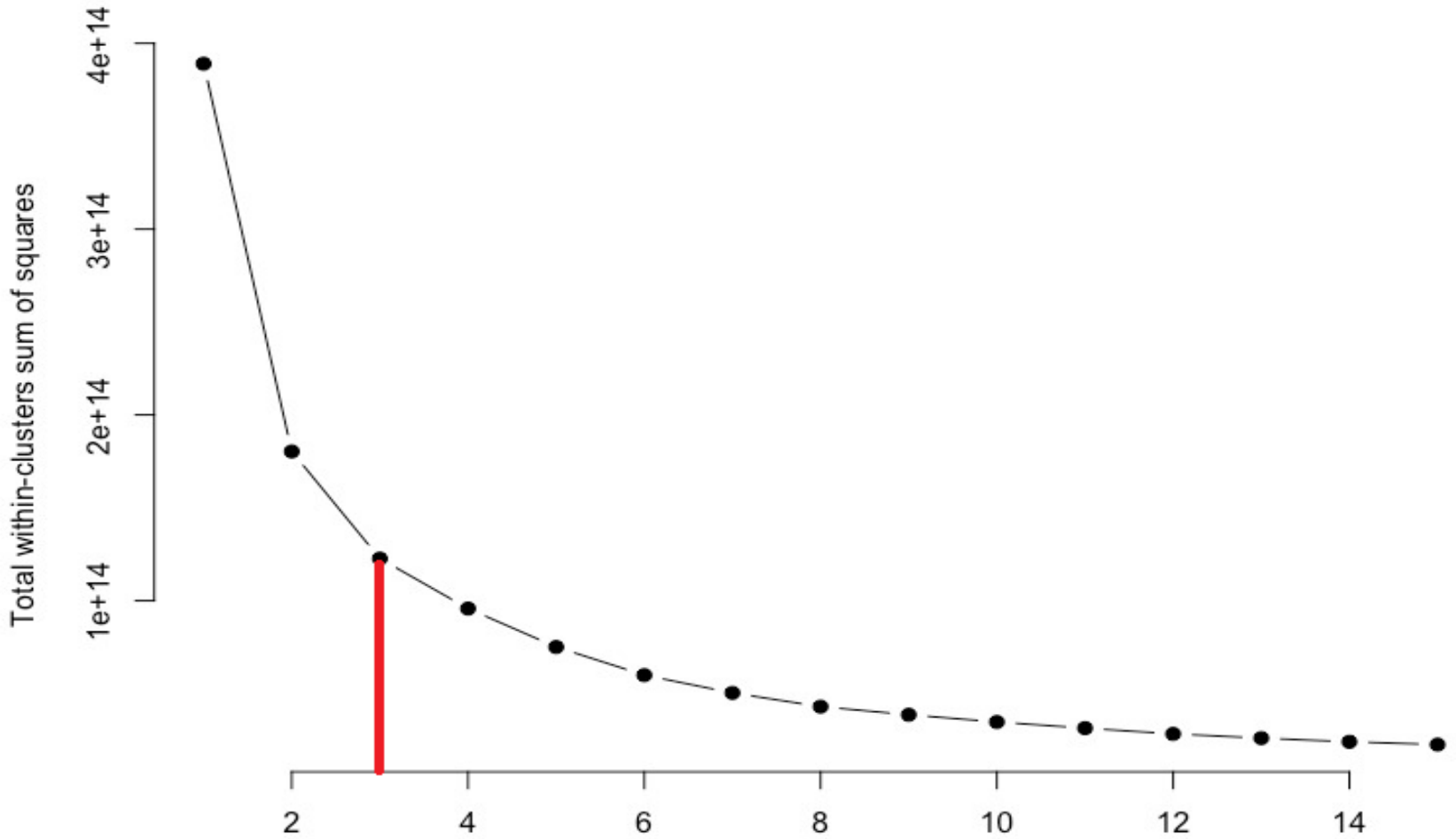
ARIMA - In the statistical analysis of time series, autoregressive–moving-average (ARIMA) models provide a parsimonious description of a (daily) stationary stochastic process in terms of two polynomials, one for the auto-regression and the second for the moving average.

Kibana is an open source data visualization plugin for Elasticsearch. It provides visualization capabilities on top of the content indexed on an Elasticsearch cluster. Users can create bar, line and scatter plots, or pie charts and maps on top of large volumes of data.

ELBOW METHOD

The **Elbow method** is a method of interpretation and validation of consistency within cluster analysis designed to help finding the appropriate number of clusters in a dataset.
This method looks at the percentage of variance explained as a function of the number of clusters: One should choose a number of clusters so that adding another cluster doesn't give much better modeling of the data. More precisely, if one plots the percentage of variance explained by the clusters against the number of clusters, the first clusters will add much information (explain a lot of variance), but at some point the marginal gain will drop, giving an angle in the graph. The number of clusters is chosen at this point, hence the "elbow criterion". This "elbow" cannot always be unambiguously identified.[1] Percentage of variance explained is the ratio of the between-group variance to the total variance, also known as an F-test. A slight variation of this method plots the curvature of the within group variance

Before performing clustering on the modified data ,to determine the optimum k value we use the **elbow method** as shown below

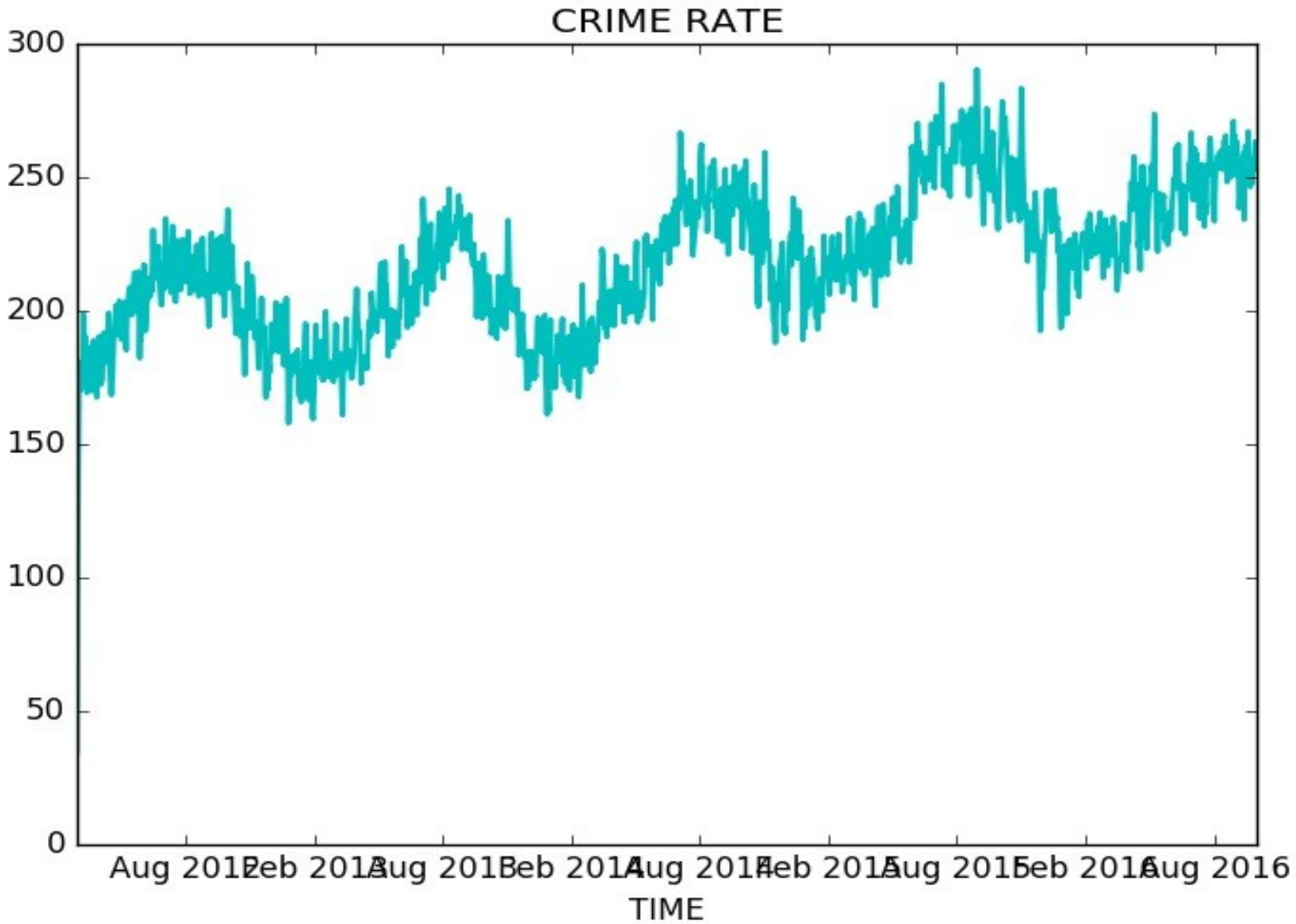


Note:Since the TSS_E sharply bends at k-3(approx.) we choose the desired number of clusters to be 3.

After performing clustering on the transformed data:

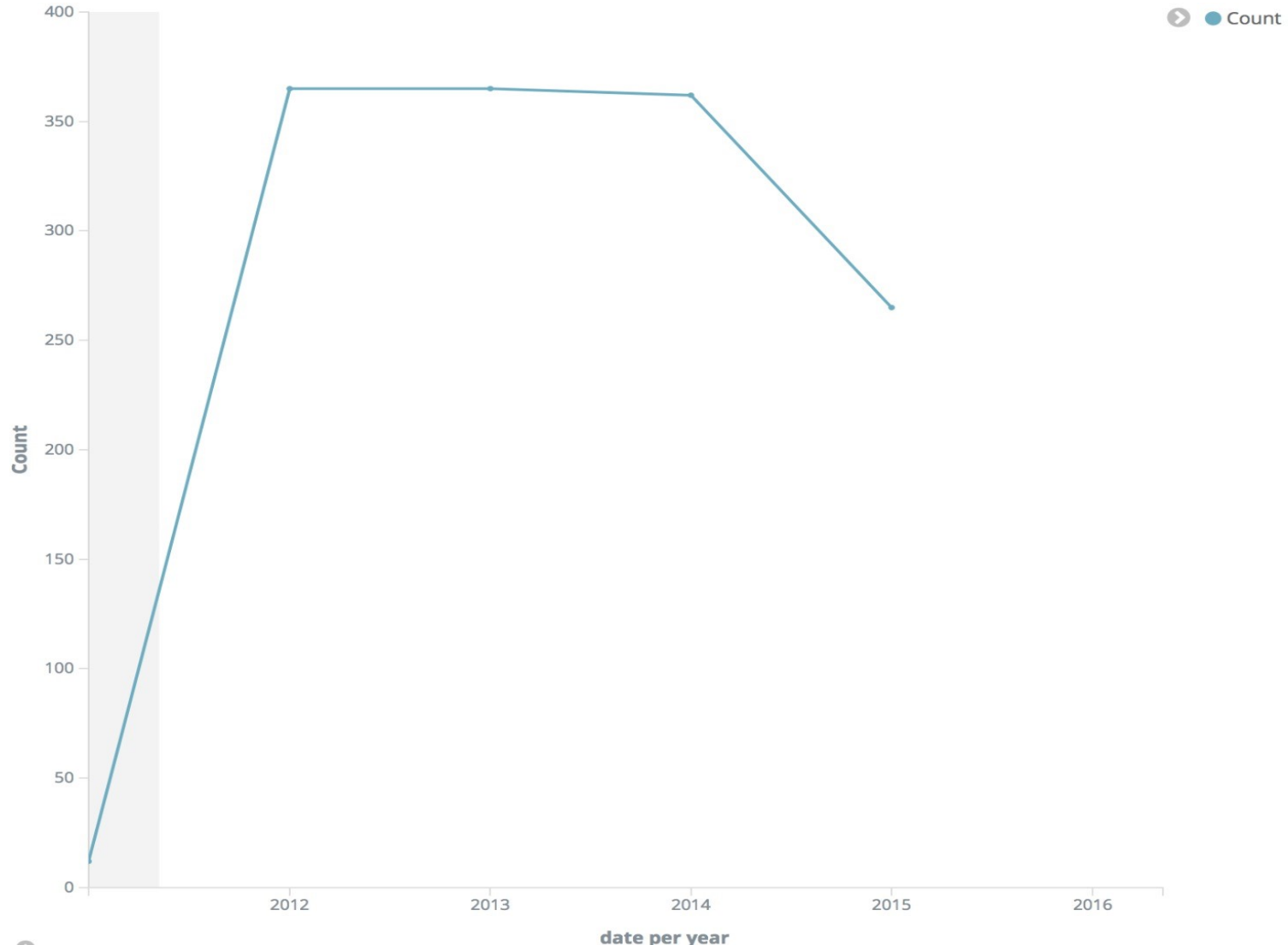
Clusters	Records Included
HIGH	345561
MEDIUM	344089
LOW	139734

Initial depiction in python matlab of the crime rate for Cluster1



Accuracy:72.16%

Initial depiction of crime rate of Cluster 1 in Kibana

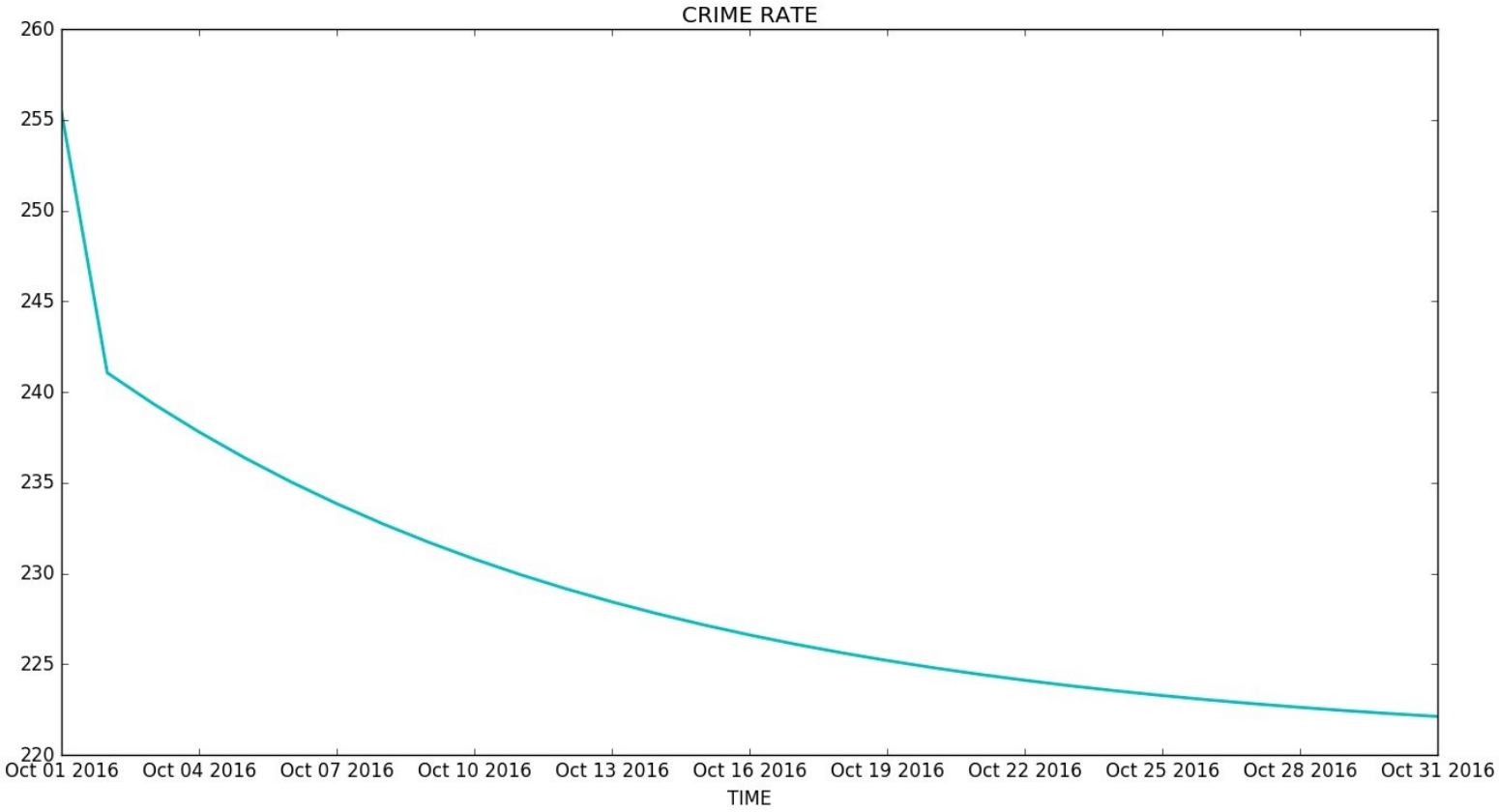


ARIMA

In statistics and econometrics, and in particular in time series analysis, an **autoregressive integrated moving average (ARIMA)** model is a generalization of an autoregressive moving average (ARMA) model. Both of these models are fitted to time series data either to better understand the data or to predict future points in the series (forecasting). ARIMA models are applied in some cases where data show evidence of non-stationarity, where an initial differencing step (corresponding to the "integrated" part of the model) can be applied one or more times to eliminate the non-stationarity.

The AR part of ARIMA indicates that the evolving variable of interest is regressed on its own lagged (i.e., prior) values. The MA part indicates that the regression error is actually a linear combination of error terms whose values occurred contemporaneously and at various times in the past. The I (for "integrated") indicates that the data values have been replaced with the difference between their values and the previous values (and this differencing process may have been performed more than once). The purpose of each of these features is to make the model fit the data as well as possible.

RESULT


CRIME RATE
260
255
250
245
240
235
230
225
220
Oct 01 2016
Oct 04 2016
Oct 07 2016
Oct 10 2016
Oct 13 2016
Oct 16 2016
Oct 19 2016
Oct 22 2016
Oct 25 2016
Oct 28 2016
Oct 31 2016
TIME

CHALLENGES

- Choosing the clustering method.
- Understanding and implementing the ARIMA model.
- Visualizing the output values(crime rates) in Kibana.

FUTURE WORK

- We can improve the model to predict locations where crimes are likely to occur on a future date.
- In accordance to the results we have obtained using this framework, we plan to make it real time by implementing the same framework using a real-time dataset.
- We would also like to tweak the algorithm used for forecasting in order to obtain results with improved accuracy.
- We can also integrate twitter data to get prior information about the crime occurring in an area and inform the PPD accordingly.

THANK YOU