**Using Pharmaceutical Locations to Predict Crime**

October 12, 2024

Patrick Adeosun

Padeos1@lsu.edu

**Table of Contents**

**Introduction**

                        Motivation

                        Project Description

**System Design**

                        Chosen Big Data Frameworks

                        Chosen Datasets

**Detailed Description of Components**

                        Each Component Description

**Evaluation & Test**

Software Manual

                        Test Environment

                        Test Results

**Conclusion**

**Appendix**

**Introduction**

The city of Baton Rouge has one of the highest crime rates in the US. In fact, it is ranked as the 6th deadliest city as it is estimated to have 54 crimes per 1000 residents. This is a cause for concern for current and future residents, as most would prefer to live in an area with little or no crime. For this reason, it is necessary to identify areas of high crime and predict where the next crime could potentially occur, so the crime can be prevented. Since traditional crime prediction methods are ineffective, a non-traditional method must be used. Hence, this project involves the use of pharmaceutical locations to predict the location of the next crime.

**System Design**

In order to predict the next crime location in Baton Rouge, the dataset of crimes in Baton Rouge from 2001 to the present was used. The pharmaceutical locations in Baton Rouge were also used. To analyze the datasets properly, they were uploaded into the Databricks online IDE and queried with Apache Spark.

**Detailed Description of Components**

The first component of the experiment involved preprocessing the data. First, the columns containing the x and y coordinates were isolated in each dataset, Latitude and Longitude for the crime dataset and New Georeferenced Column for the pharmacy dataset. Next, a floating-point conversion was attempted for each data point. If the point did not convert, it was replaced with a zero and filtered out of the dataset. Next, the x and y coordinates of the data were zipped together and used to create a data frame using Apache Spark. The next step involved using k-means clustering, a machine learning algorithm used to group data points based on their similarity. The goal is to create a clustered heatmap based on the dataset to identify areas of high crime and use this to aid in the prediction. To achieve this, the Euclidean distance was calculated between the pharmaceutical locations and crime locations, and the spark\_min function was used to identify the minimum distance between the two locations. The crime dataset is then combined with the minimum distance and pharmaceutical data frames. Spark’s vector assemble was used to combine the columns into one, and then it was further preprocessed by Spark’s standard scaler. The data was passed into a model which used k-means clustering to identify patterns within the dataset.

**Evaluation & Test**

To test the model, the code was ran in the Databricks environment. The model identified that \_\_\_ degrees on the x-axis and \_\_ degrees on the y-axis is the likely are of the next crime.

**Conclusion**

In conclusion, it has been identified that areas from x = -87.8 to -87.5, y=41.65 to 42.05 are the areas where crime occurs, and are therefore the areas to have future crimes occur. Therefore, the police force should be focused on these areas. Furthermore, future residents should avoid these areas. The data drew inconclusive results between the relationship of crime and pharmaceutical locations. A limitation this project faced had to do with the environment being used as the majority of the data to be used could not be included in the data set. An improvement to this experiment would be to use an environment capable of handling the full crime dataset, and perhaps this experiment could be redone with another parameter such as restaurant locations.

**Appendix**

|  |  |
| --- | --- |
| **File name** | **#lines** |
| Crime\_pharmacy.pyspark | 127 |

|  |  |
| --- | --- |
| **File name** | **#Spark Operations** |
| Crime\_pharmacy.pyspark | 92 |