CSCI 5408 PROJECT REPORT

By: Bhargav Dalal (B00785773) & Hardik Galiawala (B00777450)

AUGUST 5, 2018

DALHOUSIE UNIVERSITY

Table of Contents

1. Summary:

The objective of the project was to identify the relationship between the weather and the type of crime reported. The journey for identifying the relationship was motivated by the research paper posted by Mattew Rason [1]. In the paper, there was the discovery of the short-term relationship between the change in temperature and the rate of crime.

The dataset used in this project were downloaded from the Kaggle dataset library. The weather dataset consists of separate CSVs for each type of weather characteristic such as temperature, humidity, pressure, wind direction, wind description, etc. While the crime dataset consists of a varied type of information such as date of crime reported, type of offense for the crime, the medium through which crime was reported and many more. The meaningful data were extracted from both datasets.

The extracted data was then cleaned. and processed. The processing for both datasets was done separately. The processing steps include explicitly converting the date field into date time field, converting categorical string fields into numerical as well as combining all the separate CSVs into one file. Aggregation functions were computed on this file based on the date. The CSV file which consists of aggregated functions was then combined with the crime data to generate the final CSV file.

The final CSV file was used for data visualization as well as for obtaining the relationship between the two datasets. Thus, we perform Extraction from Dataset, Transformed, and finally Loaded the dataset in the visualization tool. After plotting analytical graphs, there was a slight indication of having a relationship of temperature or pressure with the type of crime.

Hence, we try to perform analysis based on the indications to identify the relationship. Given the pattern observed it is safe to say that there may be involvement of more than one factor other than weather.

2. Problem Statement:

The crime in the city is rising exponentially. Many kinds of research are in progress for predicting the crimes to decrease the crime rate. The problem statement of the project touches on the idea of using machine learning techniques to decrease the crime rate. But, the scope of the project is focused on identifying the relationship between the weather and the type of crime.

There is a lack of study and research in this specific field whether the crime is somehow related to weather conditions. It is not yet explored. These facts and the paper by Matthew Rason becomes motivation to focus on this specific problem. Thus, by the exploring the datasets, will be trying to identify the relationship or pattern with the crimes and weather dataset to decrease the crime rate.

3. Value Proposition:

The report can be helpful to government units as well as police officers to access the condition and take necessary action to prevent the crime and thereby decrease the crime rate.

4. Data Sources:

- **Crime Dataset:**

  "New York City Crimes 2014-2015," [Online]. Available: https://www.kaggle.com/adamschroeder/crimes-new-york-city

- **Historic Weather Dataset:**

  "New York City Crimes 2014-2015," [Online]. Available: https://www.kaggle.com/adamschroeder/crimes-new-york-city.

5. Implementation Steps:

- The datasets were initially collected from the Kaggle.
- Explored the datasets
- Perform the cleaning of the datasets
- After cleaning, we started extracting the meaningful data from both datasets.
- Once the meaningful data is extracted, started working on processing and transforming the data. For instance, we transformed the required string data fields into the numerical columns.
- The data transformation took most of the time in our project, as the data was in bad format. We had to convert several columns data types.
- The transformed datasets for the weather was merged into single CSV file from several CSV files.
- The combined CSV file was later uploaded in Azure Data Lake to calculate statistical measures such as mean, minimum, maximum and standard deviation for each weather condition and that CSV was downloaded.
- The CSV obtained from Data Lake and the transformed dataset was than combined into final CSV file.
- This file was loaded in python to apply sklearn's feature selection method which used Random Forest model to obtain the Gini Index for each weather condition in predicting the crime type.
- Based on the Gini Index, Dataset was loaded in Microsoft PowerBI to plot the visualizations for data analysis. Thus, we extracted the data and transformed the data and then loaded the dataset in PowerBI to obtain the visualizations.

6. Algorithm Used:

The machine learning algorithm used for trying to find the relationship was a combination of a random forest model and feature selection method. The area map (figure 4.3) was giving some indication of having a relation between weather condition. Hence, we decided to use the feature selection method to obtain the relation between each weather condition with the crime type. The output from this method is as follows:

Feature	Feature importance (Gini-index)
Temperature	0.3245
Pressure	0.2092
Wind_direction	0.1952
Wind_speed	0.1448
Humidity	0.1197
Weather_description	0.0066

Figure 5.1 – Feature Selection Method Output

The Gini index in simple words can give you the probability of the feature being used to predict the correctly classify the categorical variable. As it can be seen, temperature with 30% is top in the list. Hence, our cohort suggests that may be temperature is related to the type of crime.

Random Forest was chosen as a model, as it will divide into tree hierarchical structure and on every node different algorithms will be applied to it and the best one is selected for the final output. Hence, the Random Forest algorithm was chosen.

7. Visualizations:

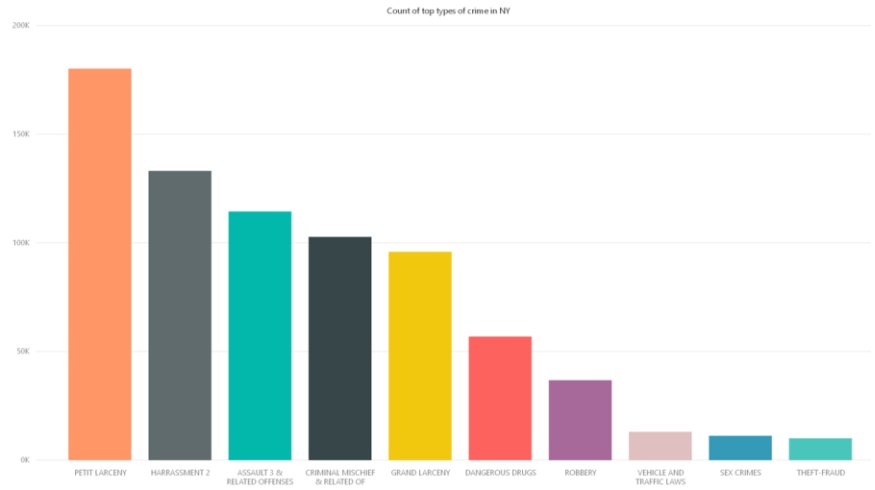
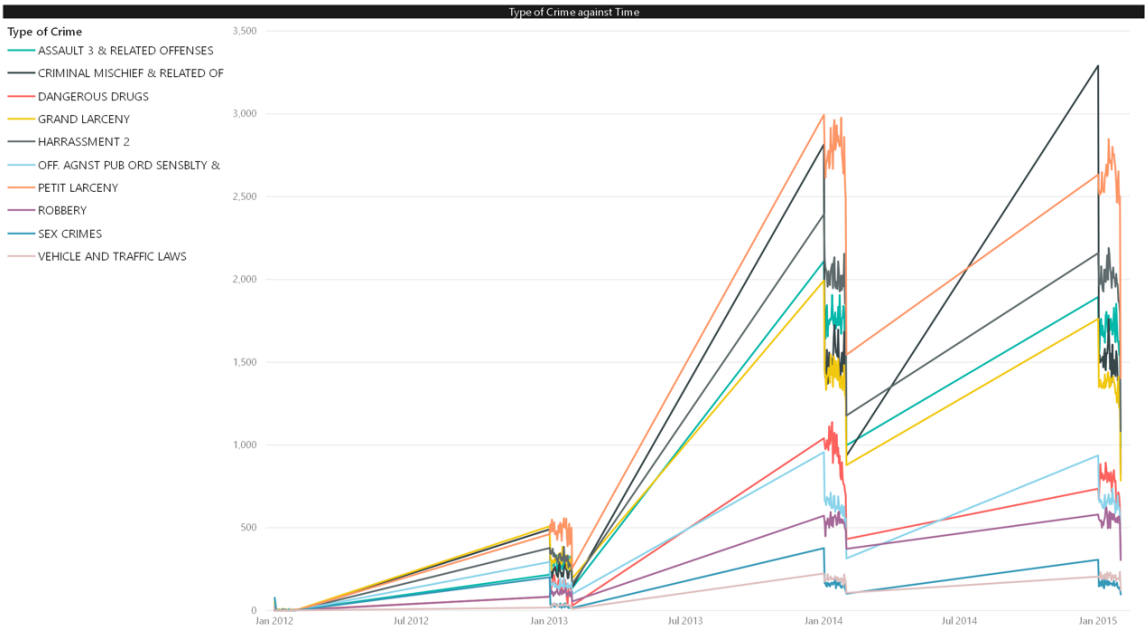


Fig 4.1 – Count plot for top 10 types of crime

The above graph is the count plot for the top 10 types of crimes in the dataset. As can be seen, petit larceny tops the list. Petit Larceny, in other words, can be explained as stealing small things (things which are less than $50) from the private properties. While harassment is second on the list. However, robbery and fraud seem to be lowest in this list.



8. Fig 4.2 – Type of crime against time

The above graph depicts the top 10 types of crimes which plotted against the time. It is evident from the graph that crime is increasing year by year. Also, we can see a constant increase in crime from July to January every year. However, there are some spikes and crime rate decrease during the starting month of the year to probably till March. But later again, it shows a constant increase in the crimes.

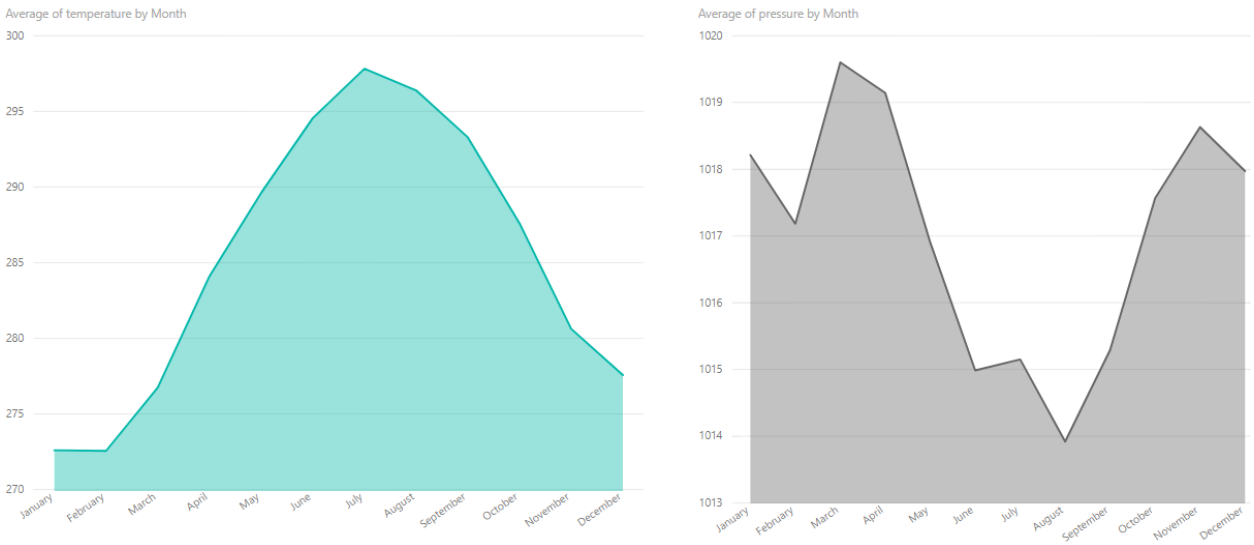


Fig 4.3 – Area Plot for average temperature & average pressure against month

As we identify the trend in figure 4.2, that near January to March there were some spikes in every type of crime and after that constant increase. So, this plot gives an indication as the crime may be related to the change in temperature or the pressure There are some spikes in the pressure in the area plot, but temperature seems to be fitting the trend from the previous figure.

Fig 4.4 – Bar chart with line of variance for temperature against each month

As we saw an indirect indication of temperature or pressure having relation with the crime rate. We decided to plot the bar chart to detect the whether the temperature is related to the spikes which were visible in figure 4.2. It seems that the temperature was related to crime type.

9. Data Analytics Tool & Services:

The programming language will be using for the data cleaning and processing is python (version 3.6). The other tools/packages/technologies which we intend to use is as follows:

- **Anaconda3:** To setup environment for python as well as using as package manager.
- **Jupyter Notebook:** To display the analysis along with the code if required.
- **Numpy, Pandas:** Packages Required for data cleaning and pre-processing process.
- **Microsoft Azure Data Lake:** To calculate the statistical measures such as average, minimum, maximum and standard deviation.
- **Microsoft Data Lake Storage:** The data lake storage is used as data service to store data onto the cloud.
- **Pyspark:** Package required to load the data and further pre-processing of data.
- **Microsoft PowerBI**: For Visualizing the data.
- **Sklearn:** To calculate the importance of ach weather factor with the help of feature selection method which uses random forest algorithm.

10. Work Breakdown:

Sprint #	Use case	Hardik Galiawala	Bhargav Dalal
1	Data cleaning and Preprocessing (1 week to complete the sprint)	Cleaning the csv file for weather dataset	Cleaning the CSV file of crime dataset
		Extracting meaningful data and perform basic pre-processing on weather dataset using Apache Spark	Extracting meaningful data and perform basic pre-processing on crime dataset using Apache Spark
2	Data Transformation (2 weeks to complete the sprint)	Transforming the string data into numeric for analysis.	The combined CSV was then uploaded in MS Azure Data Lake to obtain the statistical measure
		Combine different CSVs in a single file.	The CSV obtained from data lake was then combined with crime dataset and thus finally, processed CSV was generated
3.	Applying machine learning technique and data visualization (1 week to complete the sprint)	Applied feature selection method using Random Forest model to get the gini index of each weather condition.	Different graphs were plotted based on the gini index obtained from the feature selection method.

11. Future Work:

Our cohort concluded that the temperature from weather condition may have some relationship with the crime type. Hence, further detail analysis required to be carried out to get some strong facts that there may be a relation between temperature and the crime type. It seems to be a promising idea to perform a similar analysis with the datasets from different cities. This would help us generalize our analysis and study. It would also help us to get rid of the biased data analysis. We can also try to find out other features that may affect the crime. Once we find it, we should try to connect those features with the crime datasets and perform a similar analysis. Doing this would improve our model and give us a better idea about the relationship of these features with crime and type of crime.

12. Limitation:

The feature selection method was used to identify the relation between weather condition and the crime type. After obtaining the result and as it was indicated in the area plot (figure 4.3) also, temperature seems to be related to the type of crime. However, the temperature may be one of the factors from several factors which will lead to several other types of crime. The data which we used for predicting is biased and not generalized.

13. Critical Review:

The goal of the project was to identify whether climatic conditions affect the crime rate. The temperature out of given weather condition was showing some indication to have a relationship with the crime rate. This indication was obtained using a feature selection method using the Random Forest model. However, there may be involvement of other factors than the temperature which may influence the type of crime. Even in the paper by Matthew Ranson, a similar trend was observed. Matthew Ranson was able to predict that change from cold weather to hot weather has suddenly increased the crime rate. So, maybe temperature can be one of the major factors in predicting crime rate as well as crime type. But as compared to Matthew Ranson works, his work was generalized while our work is specific to New York City and not generalized. This means we had biased data set.

14. Role Based Distribution of Work:

Bhargav Dalal (B00785773)	• Performed cleaning and pre-processing of the crime dataset. • Performed ETL process on crime dataset. • Plotting the graphs in PowerBI. • Final Documentation of the Report.
Hardik Galiawala (B00777450)	• Performed cleaning and pre-processing of the historic weather dataset. • Performed ETL process on weather dataset. • Merged both datasets into one CSV. • Applied sklearns' feature selection method for identifying the relationship between weather and crime.

15. GitHub URL:

Project Repo: https://github.com/hardik0537/DWHProject

References

[1] M. Ranson, "Crime, Weather, and Climate Change," 10 November 2012. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2111377.

[2] "Historical Hourly Weather Data 2012-2017," [Online]. Available: https://www.kaggle.com/selfishgene/historical-hourly-weather-data/data.

[3] "New York City Crimes 2014-2015," [Online]. Available: https://www.kaggle.com/adamschroeder/crimes-new-york-city.