****

The Analysis and Aggregation of Felony Location and Time using Machine Learning: Project Specification

Table of Contents

[Introduction](#bdft4pv40nqv) 2

[Focus and Objectives](#ota166pymgha) 2

[Research and Functionality Specification](#scspqt7hy7iv) 3

[Potential risks/issues](#m3gii375s309) 8

[Project Timeline](#uuw8oyf1s40s) **9**

[References](#u8hch72gzgn) 1**0**

# 

# 

# 

# 

# Introduction

The objective of this document is to provide a detailed and specific overview of what is expected to be the base foundations that this project will be built upon. The document will be broken down into 5 sections with each section giving an in-depth description of its respective topic. The sections are as follows.

**Section 1**: *Focus and Objectives -* The first section spotlights the foundations of the research aspect of the project, describes the area’s chosen for the research and goes over the team’s goal when it comes to the project

**Section 2**: *Research and Functionality Specification-* This sections is focused on where the data is coming from, is it reliable, how the team plans to handle the data and what tools, algorithms and libraries will use to reach the goal.

**Section 3**: *Potential risks/issues -* The section gives an in-depth look at the risks and issues which the project faces that the team has to consider while moving forward with the project

**Section 4**: *Project Timeline -* Outlines the expected dates this project and its milestones are expected to be completed by

**Section 5**: *References -* The last section is a compilation of links to the data set used as a base for this project

# Focus and Objectives

The primary focus of the research of this project is to determine if there is any correspondents between a crime being committed, the hour of the day and the location that the crime took place at. In order to accomplish the goal the team will use data sets from 3 major US cities Austin (Texas) with the population of almost 951,715, Boston (Massachusetts) with the population of 685,084 and Chicago (Illinois) with the population of 2.716 million between the years 2015 and 2018 to compare and contrast the data collected. Furthermore, the team will use the data between the years of 2001 and 2014 for the city of Chicago and 2002 and 2014 for the city of Austin in order to find if there were any past hotspots of crime that reformed into areas with medium to low crime.

The data sets used by the team were compiled by the police departments and published by the government of the respective cities and so the information in this research is only applicable to the cities themselves, the metropolitan area’s are not covered by any of the data. Moreover, due to the fact that each data set was compiled and published by the local police department in their respective cities, the standard and regulation slightly varies from state to state and so the data sets have a slight variation in their columns and size as the data set from Austin has 2,173,230 unique rows, Boston has 282,517 and Chicago 6,577,865. However none of the differences are expected to provide any issues while compiling the data. All of the data sets used were provided by the local government of each respective city.

While the team don’t know the current police strategy, the overall aim of this entire project is to assist police hone in on the hours and locations that crime is the most prevalent at with the objective of cutting down crime rates while hopefully saving police resources as well human lives in the process. Moreover, with the added benefit of saving resources, the police could reinvest the saved funds into getting better training, equipment, new staff and or more patrols, which would inadvertently cut crime even further. The team plans to provide the police departments with the aggregated information from the aforementioned data sets which would allow them to adopt new tactics.

# Research and Functionality Specification

**Data**

There is a direct correlation between the integrity of the data science project and choosing an appropriate source to acquire data sets from. The team can not be fully confident in the findings of the results if there is no confidence in the source from which the data was inherited from. For these reasons datasets must not be skewed, contradictory or differ majorly from set to set.

Due to the size, population and number of crimes reported in Ireland over the past two decades, Irish crime reporting cannot be used as a viable option for this project. Instead the chosen subjects are three major cities from the United States of America with multiple data source options and years of documented crime reports. Although there are many sources that could have been used to gather the data sets from, the decision was influenced by consistency, uniform data structure and web source reputation. Data sets for Boston, Chicago and Austin were acquired from the following websites respectively; [data.boston.gov](http://data.boston.gov), [data.cityofchicago.org](http://data.cityofchicago.org) and [data.austintexas.gov](http://data.austintexas.gov).

The websites [knoema.com](http://knoema.com), [kaggle.com](http://kaggle.com) and [justice.gov](http://justice.gov) were all considered as appropriate data sources for this project and were useful in guiding the team toward the final decision.

Police departments have used Uniform Crime Reporting (URC) since the Federal Bureau of Investigation(FBI) published the standard in the early 20th century. The data sets used for this project follow the structure and guidelines of these reports with consistency.

**Preprocessing**

The size and type of data sets in question will vary, because of this it is necessary to use ‘pre-processing’. This will help eliminate anomalies like incorrect data types and missing data fields. It would be negligent and an insufficient use of time to use these data sets without considering this step. Data preprocessing is a data mining technique that involves transforming raw data. Real-world data is often incomplete, inconsistent, and is likely to contain errors. Data preprocessing prepares raw data for further processing. Some data techniques include preprocessing includes cleaning, instance selection, normalisation and transformation. The team are currently considering libraries like ‘Numpy’ for all things numerical, ‘Matplotlib’ to use with charts and ‘Pandas’ as an import and managing tool.

The raw data sets originally exist in .csv format. This introduces several problems [1]:

1. Commas in CSV are treated as information separators. A misplaced comma (which is plausible in large data sets) can lead to invalid data.
2. CSV files only support a maximum of 15 columns. The data sets may have to be rearranged or stripped off excess columns.
3. “Input form invalid error” - thrown when an empty row is encountered. These will have to be handled during processing.
4. There is a possibility of inconsistent date formats. It is likely that the data sets use American date format and will need to be formatted.

When parsing, if any of the aforementioned issues surface, the data will have to be cleaned. ‘CSV Kit’ is a useful Linux tool that can help with this. It can clean and validate the files off common syntax errors. Any unhandled problems will have to be dealt with separately. Also, since the data sets are too large to be handled at once, it will be more efficient to split the data and handle each section individually.

Inherently however, CSV files fall short in comparison to the more popular and powerful database solutions. For easier data handling and storage, the raw data can be transferred onto one of the following open source database solutions:

* postgreSQL
* MongoDB
* MariaDB
* SQLite

The team expects to use SQLite [2] as it is suitable for this project’s use case. SQLite is easy to use, is self-contained, serverless, fast and lightweight making it ideal for getting up and running. Additionally, Python has a useful library called “csv-to-sqlite” [3] - a script that processes the input CSV files and copies them into a SQLite database. This could eliminate some of the potential issues up-front. Since the data sets contain multiple columns, those that do not relate to crime, location, time etc. will have to be stripped off. This should hopefully produce a more relevant and clog free set of data to work with.

**Algorithms**

Possibly the most important decision the team will face in this project will be deciding what algorithms to use. The decision will be influenced by the type of data selected, constraints and the problem in question. Before making a decision on the algorithms it will be required to categorise the data. This is to ensure better error handling as labelled data may be a supervised learning problem compared to unlabelled data which may be an unsupervised problem. Output will have to be categorized before deciding on an appropriate algorithm. Data storage capacity, prediction time and learning time are all factors that need to be considered before any decision is made.

Some of the most popular algorithms used in data science are ‘Linear Regression’, ‘Logistic Regression’, ‘K-Means Clustering’, ‘Decision Trees’, ‘PCA(principal component analysis)’, ‘Naive Bayes’, ‘K-means’ and ‘Random Forest’. Each algorithm will be analyzed extensively and assessed for usability of this project. Some algorithms work well with certain kinds of data, for example ‘Naïve Bayes’ is said to work well with categorical input. Drawing comparisons between data and algorithms will help teasing out the most useful tools.

Once the data has been successfully transferred to SQLite, machine learning algorithms will have to be selected. The algorithms must align with the project’s goal (ie. to identify crime clusters at city locations and time periods). Below are useful algorithms that can help achieve this goal:

* K-means clustering [4]:

A scalable, general purpose algorithm for separating samples in n groups of equal variance; requires a number of clusters to be specified.

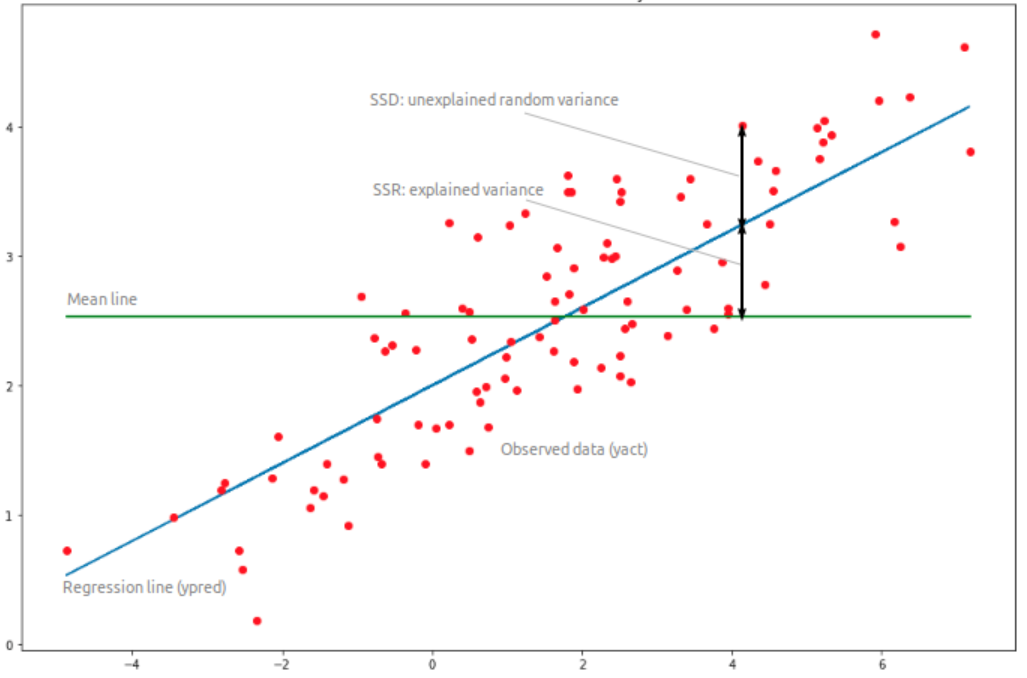
* Mini Batch K-Means:

A variant of K-Means which uses mini-batches to reduce computation time while still optimising for precision.

* Agglomerative Clustering:

This algorithm performs hierarchical clustering using a bottom-up approach - in the beginning, every point in the data set becomes a cluster. Two closest points are then subsequently found to form new clusters. Finally, all clusters are merged together to form one giant cluster.

* Linear Regression:



An example of linear regression

A popular go-to solution for predicting real world data such as number of calls or costs of

houses based on continuous variables. In the context of this project, it can be used to estimate the likelihood of certain crimes in particular areas and times of day.

All of the above clustering algorithms are *unsupervised* ie. self-organised learning that helps find previously unknown patterns in a data set without pre-existing labels. Linear regression on the other hand, is *supervised* ie. it models a relationship between a dependent variable and one or more independent variables [5].



An example of K-means clustering

The aforementioned algorithms are expected to be tested and evaluated on the basis of precision. The most effective ones will be used as the primary algorithms.

The IDE of choice is Jupyter Notebook due to its versatility, shareability and ability to perform data visualization in the same environment. It will host the backend code and algorithms will be imported to process the data one section at a time. Over subsequent iterations, the resulting ML model is expected to become more accurate and reliable. The resulting clusters can then be illustrated through plot diagrams using the MatPlotLib library. This will give a rough idea of the existing crime patterns and aid with the final conclusions.

For frontend, the team expects to use Python’s Flask library because of its simplicity and flexibility. A Rest API will be written with Flask. API calls will trigger the machine learning model(s) to draw new data predictions which will then be displayed/updated on the website. Intractability is expected to be added as a feature to allow visitors to play around with the data and note real-time changes in the visualization.

# Potential risks/issues

This project may involve handling sensitive data, in which case it ought to be treated and stored with caution. Technical problems may also arise during the development phase. Some of the technologies selected are novel in regards to the team’s experience. The overall potential issues are:

* IT Carlow reputation:

It is crucial that the information (public, internal, sensitive) causes no adverse effect on the operations, assets or reputation of the institute. Mishandling of such data could put the institute’s reputation at risk.

* Breach of General Data Protection (GDPR) [6]:

GDPR aims to strengthen personal data protection in Europe, the following rules must be followed:

1. No storing of personal data or unique identifiers that link to personal data in GA e.g. emails, customer ID’s; must use hashing.
2. Consent is required before saving tracking cookies or any personal data.
3. Any personal data must be removed from the database upon request.
4. A Privacy Officer is required if a company deals with sensitive personal data or has more than 250 employees.
5. Only store data that is directly relevant to one’s business activities.

* Breach of Data Protection Act 2018:

Similar to GDPR, it states that:

1. An individual has the right to have their data removed if it’s used for purposes of direct marketing.
2. Data protection rights apply to data stored digitally or paper files.
3. Individuals can complain to the Data Protection Commissioner who has powers to enforce the provisions of the Act.
4. If one suffers damage due to a breach of data protection rights, they have the right to sue the business held responsible.

Breaking any of the outlined rules can lead to legal matters.

* Security of the proposed website:

The website to be delivered during this project may be under the risk of being targeted by hackers. It will be a priority to make the website as secure as possible to protect sensitive data.

* Lack of algorithm precision:

While a selection of algorithms have been picked, there is no certainty that they will yield accurate results, this could undermine the legitimacy of the project findings.

* Lack of conclusive outcome:

There is no guarantee that the findings acquired through machine learning will fall in line with the team’s expectations. There may not be useful patterns in the data to allow for a crime prediction model - one of the research aims. Worst case scenario, there may not be any useful discoveries made.

* Poor compatibility of frontend and backend modules:

Novel development modules and libraries will be used throughout the development phase. The team could potentially face incompatibility problems when integrating the functional components together. There is a risk of losing valuable time if the current technology stack will have to be scrapped and rebuilt.

* Difficulty when processing data:

Since the data sets contain millions of records there is a plausible chance of encountering empty or unusable fields. These issues can cause crashes when loading data or they may interfere with the efficiency of the algorithms. As an alternative solution, the datasets may be transferred to a database for reliable handling or another existing platform.

* Time constraints:

The project must be submitted by a specific deadline. The aforementioned issues may delay progress in which case the team will have to fall back to the back-up plan. Under the back-up plan, secondary features will be stripped with only the core ones remaining.

* Data loss:

The data stored on the files can become corrupt or damaged. To prevent this the data sets and any other usable data will be backed-up on the Cloud and local hard drives.

# Project Timeline

The proposed dates are the expected deadline however they are subject to change.

**Project Proposal**: 11/10/2019 **Status:** Submitted

**Project Specification:** 18/10/2019 **Status:** Submitted

**Research/Application Design:** 08/11/2019 **Status:** Incomplete

**Up to Date Presentation:**  03-10/12/2019 **Status:** Incomplete

**Research Implementation:** 07/02/2020 **Status:** Incomplete

**Final Project Report:** 14/02/2020 **Status:** Incomplete

**Final Presentation:** 21-28/02/2020 **Status:** Incomplete

# References

* [1] - Onlinesurveys.ac.uk. (2019). *Common problems when importing or exporting respondents | Online surveys*. [online] Available at: https://www.onlinesurveys.ac.uk/help-support/common-csv-problems/ [Accessed 18 Oct. 2019].
* [2] - Docs.python.org. (2019). *11.13. sqlite3 — DB-API 2.0 interface for SQLite databases — Python 2.7.17rc1 documentation*. [online] Available at: https://docs.python.org/2/library/sqlite3.html [Accessed 18 Oct. 2019].
* [3] - Benc, L. (2019). *csv-to-sqlite*. [online] PyPI. Available at: https://pypi.org/project/csv-to-sqlite/ [Accessed 18 Oct. 2019].
* [4] - Scikit-learn.org. (2019). *Unsupervised learning: seeking representations of the data — scikit-learn 0.21.3 documentation*. [online] Available at: https://scikit-learn.org/stable/tutorial/statistical\_inference/unsupervised\_learning.html [Accessed 18 Oct. 2019].
* [5] - En.wikipedia.org. (2019). *Linear regression*. [online] Available at: https://en.wikipedia.org/wiki/Linear\_regression [Accessed 18 Oct. 2019].
* [6] - General Data Protection Regulation (GDPR). (2019). *General Data Protection Regulation (GDPR) – Official Legal Text*. [online] Available at: https://gdpr-info.eu/ [Accessed 18 Oct. 2019].
* Boston Police Department (2018). *Crime Reports*. Available at: <https://data.boston.gov/dataset/crime-incident-reports-august-2015-to-date-source-new-system/resource/12cb3883-56f5-47de-afa5-3b1cf61b257b>

[Accessed 17 Oct. 2019].

* Chicago Police Department (2019). *Crime Reports.* Available at: [https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2/](https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2/data)*[data](https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2/data)*

[Accessed 17 Oct. 2019].

* Austin Police Department (2019). *Crime Reports.* Available at: [https://data.austintexas.gov/Public-Safety/Crime-Reports/fdj4-gpfu/data#expand](https://data.austintexas.gov/Public-Safety/Crime-Reports/fdj4-gpfu/data%23expand)

[Accessed 17 Oct. 2019].