COMP6235-Group Coursework Report

Hello UK
School of Electronics and Computer Science
University of Southampton
Southampton,United Kingdom, SO17 1BJ

***Abstract*—With the improvement of living quality, people want to pursue more comfortable and safer trips. Chicago, the second largest commercial center in the United States, attracts an increasing number of people to travel. However, crime often happens there. Personal safety has attracted people's attention, so we want to create an application to help travellers to choose a safer place to live in, as we also often think about it. There is no proper application focusing on safety factors when choosing an accommodation in Chicago at present. This is the reason why we want to build the application. We combine crime data and Airbnb data to show the crime number in the area where the Airbnb accommodations locate. Travellers can use the application to compare the accommodations they choose and know the details of specific areas. Moreover, we also do the prediction for travellers, so they can know the probability of every crime type during their journey or they can choose a time when there is a low criminal rate to travel. Therefore, people can have much safer journeys and enjoy their trips in Chicago.**

***Index Terms*—Airbnb crime analysis prediction**

I. Introduction

The aim of this project is to develop an application that help travellers find a safer area to stay in Chicago. To achieve this, crime history and Airbnb data in Chicago are the main sources used in this application. The application is divided into two parts which are analysis and prediction. For analysis part, we use Tableau to generate visualisations to show the overview of the crime situation and insights in Chicago together with Airbnb data. In prediction part, classification model of machine learning technique is applied to predict the crime probability that will happen in a specific location. This predicted result is used as a safety indicator to compare which accommodation is safer to live in. As a result, not only this application can give users more understanding in crime occurrences, but also allow users to find a safer Airbnb accommodation to stay.

II. Data Collection

Three raw datasets (Crime, Airbnb and zip code) were collected from the Open Data website. The first two are from data portals and the third is from a Web API. The description will be explained as follows.

A. Data Portal

1) Crime dataset: The dataset is in an Excel file. It contains the incidents of crime in Chicago with the location and date of the crime from 2012 to 2016. This data was obtained from Kaggle [1] where the raw data was published by the Chicago Police Department with the sample as shown in Figure 1.

Columns	Data type	Example data
Incident date	Quantitative , Discrete data	01/01/2012 12:01:00 AM
Primary crime type	Qualitative, nominal	THEFT
Community area	Qualitative, nominal	46.0
Latitude	Quantitative, Continuous data	41.831494381
Longitude	Quantitative, Continuous data	-87.667977499

Fig. 1. Sample of a record in crime data with data types

Measurement, biases and data quality:
We would like to measure crimes per area per day of the year. The preliminary crime could be changed later based on additional investigation. For that reason, the Chicago Police Department (CPD) does not guarantee the accuracy of the initial crime of record, correct sequencing of the information, completeness and timeliness. It could lead to data bias in the initial crime data and primary crime type.

2) Airbnb dataset: Data is in CSV format representing a summary information and detailed listing of Airbnb data in Chicago state. The information was collected on 10 May 2017 from Insider Airbnb [2] which provides a non-commercial set of Airbnb data in cities around the world. The data type of the primary records is structured as Figure 2:

Columns	Data type	Example data
Airbnb name	Qualitative, nominal	Lincoln Square Guest Suite 2 Bd 2Ba
Room type	Qualitative, nominal	Private room
Minimum nights	Quantitative , Discrete data	1
Number of reviews	Quantitative, Discrete data	25
Reviews per month	Quantitative , Continuous data	2.57
Community area	Qualitative, nominal	Lincoln Square
Latitude	Quantitative, Continuous data	41.97654639
Longitude	Quantitative, Continuous data	-87.68493431

Fig. 2. Sample of a record in Airbnb data with data types

Measurement, biases and data quality:
We would like to measure the Airbnb house per location.The provided data has been verified, cleansed, analysed and aggregated therefore the location information in the lists will be approximately 150 meters from the actual address.

3) Crime category group dataset: The dataset is in a PDF file. It divides the crime category based on the severity of the offensive behaviour into 8 categories. The crime category has been published since 2010 in a document named "USPC Rules and Procedures manual" by The United States Department of Justice [3]. The data is in text format and

shows each category with the description as follows:
Category 1 OFFENSES INVOLVING THE PERSON

Subchapter A - Homicide Offenses
Subchapter B - Assault Offenses
Subchapter C - Kidnaping and Related Offenses

It requires manual mapping each subchapter with the primary crime type in the crime dataset.

Measurements, biases and data quality:

The source is not up-to-date (2010). This could lead to incorrect mapping of the crime category that occurred after 2010.As there are 8 categories, some categories are not relevant to our crime dataset. In this case, only 6 primary crime types that can be mapped are selected. It could be biased in term of sample data and incorrect category.

B. Web API

The zip code data is collected from the Python Web API, searched by latitude and longitude coordinate from Crime and Airbnb dataset. The obtained zip code column has been updated in the Crime and Airbnb data as linking column. This Web API is developed from Python and published as an open source from [4]. It provides the US Zip code data with JSON format that is aggregated from 3 sources, Federal government zip code data from [5], US cities information such as population and wage from [6] and Geographical data from [7]

Measurement, biases and data quality:

The source of zip code data is not up-to-date (2012). This may cause incorrect mapping of zip code in some area that has been updated after 2012. As some zip codes are in overlapping areas, sometime a coordinate returns more than one zip code. In this case, only one zip code is selected. This could cause bias result in those areas. The data is in good quality and organized in JSON format easy to manage and parse with existing tools.

III. Data Cleansing and Preparation

The merged dataset is not ready immediately for analysis because there are three datasets that are inconsistent, incomplete and improperly formatted. In our group, we processed data using Microsoft SQL Server Management Studio 2017 (SSMS) which is a relational database management system tool. SSMS is an integrated environment to handle any SQL infrastructure from a SQL database [8]. Follows are the reason behind this choice.

1) Scalability: It supports large volumes of data of up to one terabyte in size while maintaining a high-performance [9]. The crime dataset containing more than a million records/rows was not fully accessible in Microsoft Excel.

2) Database structure: It is used to handle structured information. All datasets are stored in tables with primary and foreign keys to establish a relationship between tables. The schema always stays the same.

3) Reliability: The database transaction guarantees consistency and recoverability in case of system failure [9]. The entire transaction could be rolled back at any time in addition.

Extract Transform and Load process :

ETL process is applied in this preparation process as follows.

1) Extraction: After downloading the data sources, it is necessary to validate and convert each data value into its appropriate type before uploading into the SQL Server. It can be done in Microsoft Excel for Airbnb , zip code and crime category dataset. However, for the crime dataset, the data is over its limit and it cannot open in Microsoft Excel. The solution is to separate the dataset into 3 files before importing into SQL server and merge them before converting to correct data type.

2) Transformation: It involves the following tasks.

• Applying new calculation of new measures for the date, example day of week, day of month, and season.

• Designing tables and its column following the project requirement. For completeness of data, it requires joining tables and adding new columns or produced new table. For example, join table crime with table zip code and crime category for getting crime category in each zip code.

3) Cleaning:

• Checking duplicate data. There should be no duplicate data.

• Fixing null and empty data by replacing them with the numerical value 0.

• To map between crime and Airbnb data, community area is used as the foreign key. However, the values used between both dataset is not compatible. Therefore, this requires manual updates on the data to make the data consistent.

• Delete unused columns.

4) Loading: Exporting data from SQL Server and load data into Tableau, R Studio and Python for implementation.

IV. Data analysis and reporting

The aim of this analysis part is to show users an overview of the crime situation with the statistics behind the data in form of sophisticated visualisations. Tableau comes as our first choice to used as the main developing tool in this part. The reason behind this choice is that because tableau is a powerful tool which provides loads of sophisticated visualisations with user-friendly user interface. This enables user to analyse data easily and effectively. Moreover, Tableau provides Tableau Public services which allows dashboards to be embedded in an web application. This facilitates combining works from analysis and prediction parts.

A. Design

The visualisations are designed based on questions to lead users to follow the flow of the insight information. Firstly, the visualisation start from the question ”Where is the safest area?” to provide an overview of number of crimes that happened in each area. Users can have a glance on the map and will be able to answer that which zip code and community

area are the safest or the most dangerous ones. After users know where is the safe area, next visualisation is to answer the question of "When is the best time?". This is to help user find a proper period to travel. Following this, next visualisation "What is crime situation over years?" will allow users to see more details about crime types happening in each period of time. Finally, the last visualisation "Crime statistics" will show the very specific details about when and where each crime type happened in total.

Fig. 3. Design strategy to direct users to visualisations by questions

B. Insight from analysis

From the visualisations in analysis part, the travellers can see the details of crime which helps them choose a better time and safer place for travelling. As shown in Figure 4, the most dangerous day of the week is Friday, while the most dangerous month is distributed around mid year between May to August. This trend looks similar every year. Thus, we can imply that the way that crimes occur is *season-dependent*. We divide Chicago into different parts by zip codes and plot in the Chicago map. As the analysis result, we found that the most dangerous place is Austin with twice higher than the second rank. While the top crime that most happen is theft, with the total number of 321,569, which is about 55,000 more than the second crime.

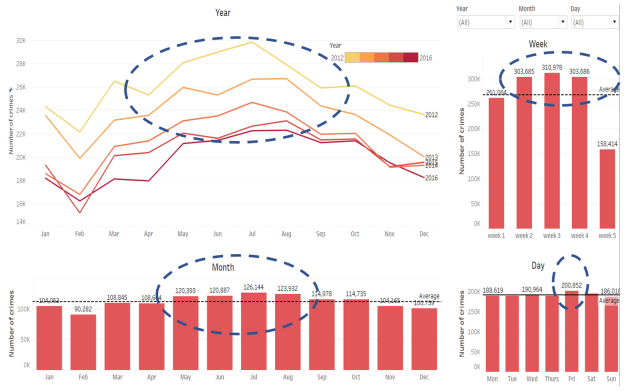

Fig. 4. Visualisations of question "When is the best time?"

In conclusion, the crime occurrence is season-dependent where the months from January to April and from September to December in any particular year are the safer time for travellers to stay in Chicago. In addition, Edison park is the most recommended place for visitors to stay because it has the lowest number of crimes as well as the Airbnb room price. On the contrary, Austin is the most dangerous area to live in the metropolitan area and its Airbnb prices are also a little more expensive.

V. Machine Learning

The machine learning model being used in this project is developed using scikit-learn. The main algorithms used are linear regression, time series prediction, and classification. Many examples have been tried to compose the most suitable model predictor for this crime dataset but the classification model produced the most satisfying results. It is deployed as the predictor function in the web application. Follows are the details for each algorithm used.

A. Linear regression

1) Processing: First, we used MATLAB to do a linear regression using an algorithm from neural networks toolbox. We found that in the training set, the accuracy of the linear regression is effective, however the resulting classifier is not sufficient with the training data provided. Next, we used the total data in year 2016 to do the linear regression in Python, but the fitting results are not ideal. We think the reason might be that the dataset is not big enough, so we expanded the training set by including 2012-2015 years' data for the linear regression in python, however, the accuracy is still very low.

2) Model Selection: We used the neural network toolbox to do the linear regression in Matlab, and, in python, we used a toolbox called MLPregressor.

3) Reason for Model Failure: The training result is good but it cannot be used in the test data for the accuracy of fitting is very low, so this model has "High Variance" and is not effective for our crime data.

B. Time series

1) Processing: We tried time series in R and found that the fitting result is good but it can only predict two weeks crime numbers after the last date in the dataset.

2) Model Selection: We first drew the time series plot in R and found that the crime numbers have a seasonal trend, so we chose Holt-Winter predict model to process the data.

3) Reason for Model Failure: The accuracy of prediction is effective but it can only do two weeks prediction which is not enough for our web application.

C. Classification

1) Processing: We divided all the crimes into 6 crime categories and tried classification by using the dataset of 2016 in Python and obtained the accuracy of each crime type. We found that the accuracy is efficient for our web application. We used a classifier model called MLPClassifier in python to implement the classification. The features used to train the model are location and date features as shown in Figure 5. The data in 2012-2015 is set to training set and the data of 2016 is used to test the model built by the training data. We then output the classification accuracy of each crime group and found that this model works effectively with our data. Finally, we obtained the possibility of each crime group that will happen on the dates that users choose for their travel.

Features	Sample
Latitude	41.880913
Longitude	-87.68152935
Day	1-31
Month	1-12
Day of week	1-7
Season	1-4

Fig. 5. Sample of location and date features that are used to train the classification model

2) Function for Users: The user can choose the date they want to travel and click on a certain point on the map to choose the place they want to travel to, then the classifier will show the probability of each crime type that will happen in that place around that date. Users can know the rate of crime of that date and place that they choose to travel therefore it helps users make a safer and better trip.

3) Principle of the Model: The MLPClassifier uses neural network to do the multi-classification which can return the probability of the six crime types that will happen in the future, it shows whether the place and date that the user choose are safe or not.

4) Evaluation result of Classification model: The model is evaluated by accuracy of the predicted results of each crime category from the test dataset. The accuracies shown in Figure 6 are calculated from the confusion matrices of each category by using this formula $Accuracy = \frac{TP+TN}{P+N}$.

Overall, the average of accuracy for all crime categories is around 86% which comes from 75% and 61% of person-related and property-related crimes [3] respectively. While the other categories has much higher rate of accuracy of over 90%. It is noticeable that person and property related crime have outstanding lower accuracy than the others. This is because there are lots of variation in the occurrence pattern of person and property related crime. This make them more ambiguous to be classified by the model. For example, theft, property-related crime, happens almost everyday and in various coordinates. Assuming that thefts often happen at two corners of an community area. In this case, the model will be misled that locations between this two points also have high chance of theft crime, while it is only the corners not in between.This can be improved by using distance measure from area that crime often happens such as distance from street as a new feature. For other categories, they happen less often than those prior categories. This means the model learned that it have low chance to occur and it will only happen in specific case which is easier to be predicted. For example, drug-related crime often happens in apartment on weekend. As a result, the accuracy is high because the model will only classify these crime categories on very specific features like above example, other than that will be classified as other crime categories. Figure 7 and 8 shows result of property and drug related crime. It can be seen in Actual axis on Has crime row, that property-related crime happen much more often than drug-related. This makes it has low chance to be correctly classified. In drug-related crime, mostly the model classify that the crime that happened is not this crime which is mostly correct. Only some specific cases like above example that the model will classify as drug-related crime.

Crime group	Person	Property	Drug	Weapon	Sexual Exploit	Other offense	Avg
Accuracy	0.7556	0.6102	0.9552	0.9856	0.9883	0.9004	**0.8659**

Fig. 6. Accuracy of classification model in each crime category

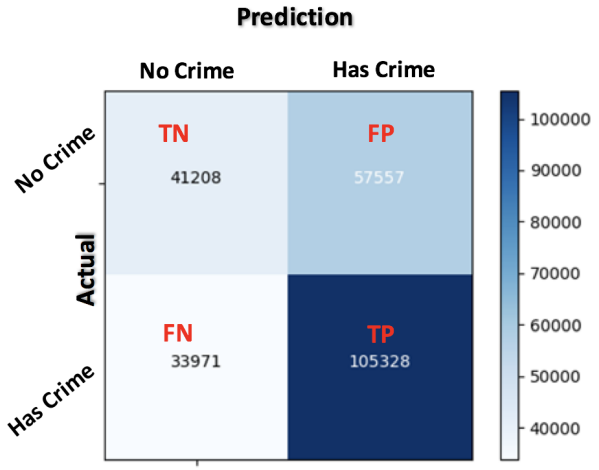

Fig. 7. The confusion matrix of Property-related crime

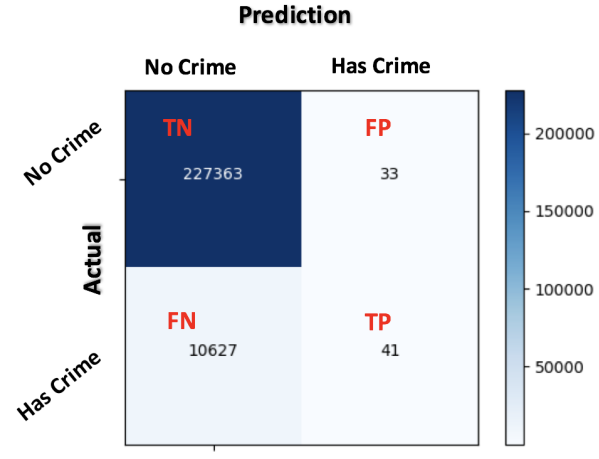

Fig. 8. The confusion matrix of Drug-related crime

In conclusion, Linear regression and time series are not suitable for our application, and classification by using neural

network can work well for the demand of the web application. Therefore, we finally choose this model to do the machine learning part for our project.

VI. Application

The application part is developed using the Python Django web framework. This Python framework is chosen as it is suitable for developing websites, also because it shares Python with the machine learning component scikit-learn which is widely supported by many additional libraries. Moreover, this web technology has great compatibility which can be viewed on all platforms by standard web clients and it is easy for online access. The application is separated into two parts, analysis and prediction. The analysis part is developed in Tableau, published to Tableau Public and embedded in the web application. The prediction component is developed using scikit-learn with a variety of standard machine learning algorithms.

A. Analysis

1) Goal of Analysis: The analysis component is important for users because it allows users to see the crime details of an area, which can help them make an informed decision. In the results given, users can see the area where they will stay and the crime rate of the neighbourhood. We show graphs and details of crimes in the area in order for users to determine whether the area they want to stay in is safe or not. Users can use filters to see particular information about the area with the details they consider the most important. With the inclusion of the date of the crime, users can determine if there is a particular time of year that is safer as well. This can reduce the degree of danger to a lower level in areas that perhaps have high crime rates throughout most of the year. With the Airbnb price data, users can choose to stay in cheaper and safer areas even if these are only available briefly.

2) Features: For features in this section, we can create a story with tableau, visualize the crime rate and Airbnb pricing information for users. In addition, users can choose different sections in one combined story. For example, users can choose among different time periods, crime types or other features in order to make their choices.

B. Prediction

1) Goal of Prediction: The prediction component of predictive analysis allows users to quantify the probability of crime that will occur in a specific location during the travelling period selected. The final goal of this part is to provide the users with the probability of crime occurrence as a safety indicator so that they can know where a safer place is in any particular location.

2) Features and Usage: Overall, there are two main modules in this analysis page, a map and chart. The map is designed to receive input location from users and transform to location features. The latitude and longitude can then be fed into to the prediction model together with the input dates. While the line chart is used to display the results returned from the model, it represents 6 crime categories with an extra line for average result of all crime groups. Follows are the features showing in accordance with the numbers in Figure 8

Fig. 9. Features in the Prediction Part

1. Interactive map with Airbnb locations – allows users to select Airbnb locations from the Google map API and shows the room details in a pop-up box. Users can use this information together with the prediction graph to decide where to stay.
2. Room Type filter – allows users to use Room Type as a filter to show/hide the Airbnb location on the map.
3. Community Area filter – allows users to use Community area as a filter to show/hide the Airbnb location on the map.
4. Date range input – allows users to select their travelling period as an input of prediction model.
5. Highlight selected period – the prediction chart is highlighted as green to show the travelling period. This is to make it easy for users to compare the crimes' probability with the surrounded dates.
6. Recommended date – the prediction chart will recommend the safest date in +/- 10 days from the selected date range. The recommended date is selected from the date with lowest crime probability in average.
7. Recommended date box – shows recommended date as a notification box.
8. Show/hide threshold lines – the threshold lines are included to allow users to check whether the crime probability in each group is higher or lower than the thresholds. The thresholds are pre-calculated from the average of the predicted results from all Airbnb locations and every day in a year (365 days from 01/01 to 31/12). The thresholds of each crime group are as follows:

a. Person Related: 16.85
b. Property Related: 66.05
c. Drug Related:5.97
d. Weapon Related:0.48
e. Sexual Exploitation:0.72
f. Other Offense:9.6

g. Average Threshold from all groups: 16.61

9. Show/hide prediction result of each crime group – allows users to show and hide the lines to make it clearer when comparing the result with the threshold line.
10. Compare locations – allows users to explicitly compare the prediction result between 2 locations. When the comparison mode is enabled, the map will allow users to select 2 locations on the map and compare the information from those two locations from the pop-up boxes as shown in Figure 9. The chart will show the predicted result of each location in 2 different lines to allow users to compare and decide where is the safer place to stay.

Note that all works in this project can be found in HelloUk Github [10]

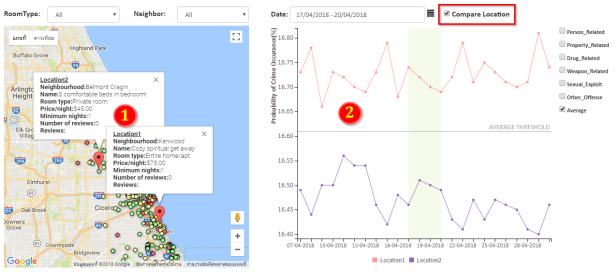

Fig. 10. Comparison mode that allows users to compare 2 locations

VII. Limitation and Improvement

First of all, the Airbnb data we use is incomplete because we cannot get the information about anonymous users. As for the data processing part, the zip code is used according to standard in 2012. The location could be bias if zip code is updated after that.In the data analysis part, since the size of the dataset is large, it always takes long time for Tableau to response. In the prediction part, we need more researches for factor why crime happen and add new features to make it more reasonable and improve the accuracy of classification. We consider about the distance from risky place, such as nightclub and crowded area, as new feature. Moreover, we can try to find the relationship between the special event date and peak crime number to determine whether we can use special event as a new feature for our classification. As for our web application, it should be allow users to compare each of the crime groups, not just the average of all the crimes.

VIII. Conclusions

To sum up, we solve the problem of how to make a travel safer in two parts. The first part is data analysis, users can know the criminal information of a specific area and safer month or day in previous period to help themselves choose a better time and place to travel. The second part is prediction, which allows users to choose the specific area and date they want to travel and give them a visual prediction result of the probability for each crime groups that will happen in the future. In conclusion, compared with other web applications which just show the details of Airbnb information and or just the crime analysis, our web application can give users the details of Airbnb and crime numbers and also has recommendation ability to make itself more useful for the users.

References

[1] "Crimes in chicago — kaggle," https://www.kaggle.com/currie32/crimes-in-chicago.
[2] "Get the data - inside airbnb. adding data to the debate." http://insideairbnb.com/get-the-data.html.
[3] "Uspc rules and procedures manual," https://www.justice.gov/sites/default/files/uspc/legacy/2010/08/27/uspc-manual111507.pdf.
[4] "Welcome to uszipcode documentation — uszipcode 0.1.3 documentation," http://pythonhosted.org/uszipcode/index.html.
[5] "Free zipcode database — free zip code database," http://federalgovernmentzipcodes.us/.
[6] "Census 2010 zip code zcta demographic profile dataset," http://proximityone.com/cen2010_zcta_dp.htm.
[7] "Google maps," https://www.google.co.uk/maps/@50.8892019,-1.4085896,14z?hl=en.
[8] "Download sql server management studio (ssms) — microsoft docs," https://docs.microsoft.com/en-us/sql/ssms/download-sql-server-management-studio-ssms.
[9] "Integrating_sql_platform.pdf," http://www.sage.com/na/~/media/Category/CRE/Assets/Documents/CRE%20Whitepapers/Integrating_SQL_Platform.pdf.
[10] "kinetiz/hellouk-cw3," https://github.com/kinetiz/HelloUK-CW3, (Accessed on 01/11/2018).