INTRODUCTION TO DATA SCIENCE

(CS418)

75°F with a Chance of Snow

Final Project Report

Team Snowflakes:

Keshav Malpani

Yashika Goyal

Jaspreet Kaur

Santhosh Mani

University of Illinois at Chicago

Spring 2018

Problem Statement:

For this project, we address two major problems: Crimes and Restaurants.
The hypothesis is that there is a connection between crime & weather, crime and liquor licenses and sentiments and Yelp review in Chicago.
This project is divided into different phases of data science:

- In the data discovery and extraction phase, we collect the census, crime, business, demographics, yelp and weather data.
- Using all the data, we perform data integration and analytics. In this phase, we see how the Crime is affected by the businesses, demographics and census. Considering the past crime and weather data, we will predict the robberies in summer 2018. Perform sentiment analysis of yelp reviews for each restaurant. Predict the review rating using precision and recall.
- To ensure the quality, we perform different techniques of Validation and Testing.
- Then we visualize the results for the user to skip the technicalities, and have a better look at the results.

Data Discovery:

- Looked for various types of data that are available about the City of Chicago, its people, businesses, health, transportation, land-use, government, etc from various sources. Submitted a list of more than ten sources.
- This list excluded data.cityofchicago.org, datausa.io and census.gov. Showed how its attributes are relevant/similar to the attributes in the datasets mentioned in the Data Extraction section.
- Identified key fields and provide potential ways in which the dataset may be joined with datasets mentioned in the Data Extraction section.

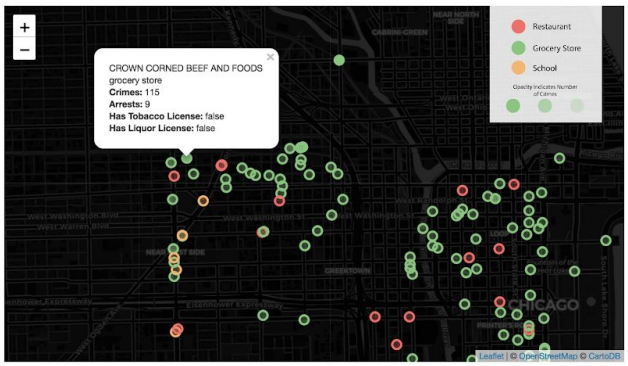
Data Extraction:

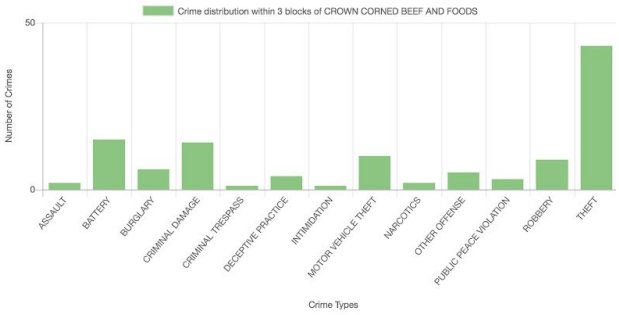
- For the following datasets, we are interested in data for the zip codes from 60601 to 60607 (in Chicago): Business Licenses, Food Inspections, Crime, Census, and Demographics from the City of Chicago's data portal and census.gov.
- Scrape and use external data for restaurants from Yelp for the zip code 60654.
- Used Wunderground for weather data.

Data Integration, Analytics and Visualization:

- **Query 1:** Crime type within 3 blocks of a Business type:
  This query combines Business and Crimes datasets on the basis of census tract which is obtained using latitude and longitude information. The businesses are classified into different types according to the license description. And then aggregations are performed on combined dataset to generate number of crimes, arrests, if business has tobacco or liquor license etc.

Using the geographic information, business types are plotted on map and color coded on the basis of business type. To represent more detailed information about each data point interactivity has been added such that on clicking a business details, crime distribution is shown in accompanying chart. (Visualization and resultant dataset is based on sampled data).

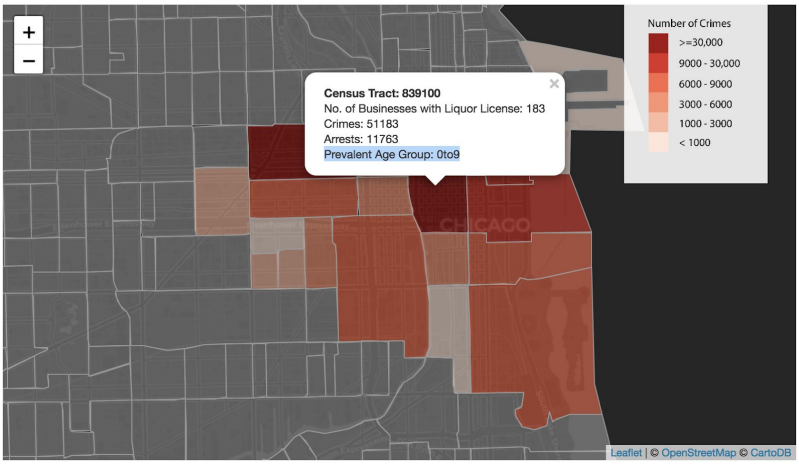




- **Query 3:** Crime and Age Group

  The Age Group data of 2010 gathered from American Factfinder website is first cleaned and aggregated to reduce it down to nine age groups. And then combined with the 2010 crime data. The resultant dataset is grouped by Census Tract (grouping is not done by census block because there are multiple census blocks of same code in different census tracts) and then the most prevalent age group and crime type is determined for a particular census tract.

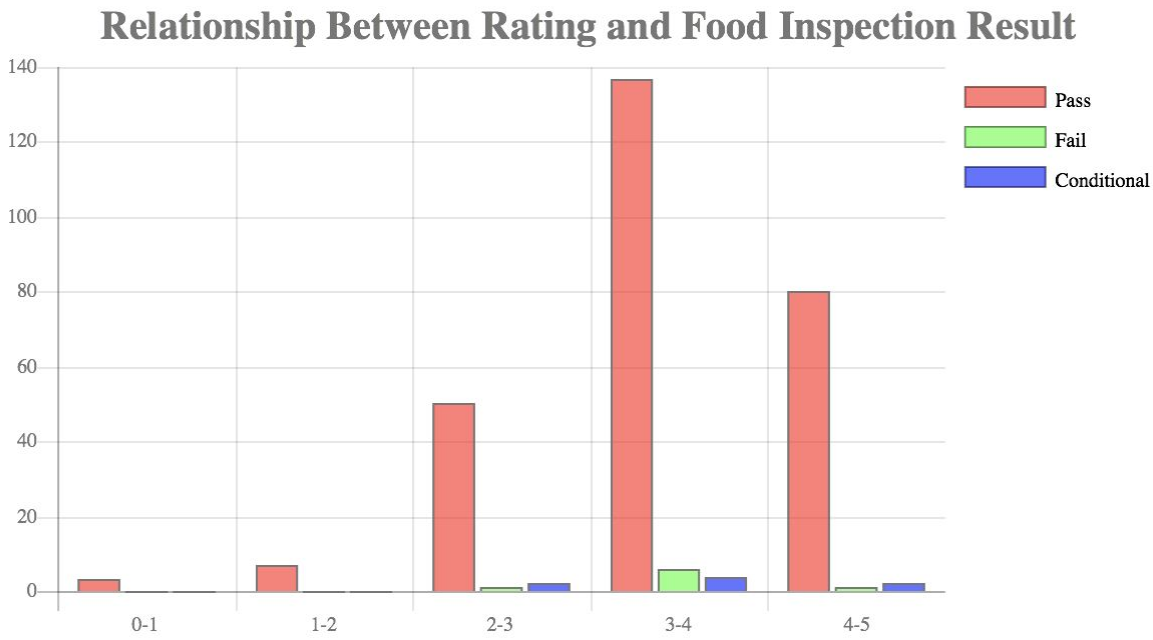
  Since the data is by census tract so for visualization this information is shown on clicking census tract in combination with data of other queries (details provided in specific queries).



- **Query 4:** Review Rating vs Inspection Result

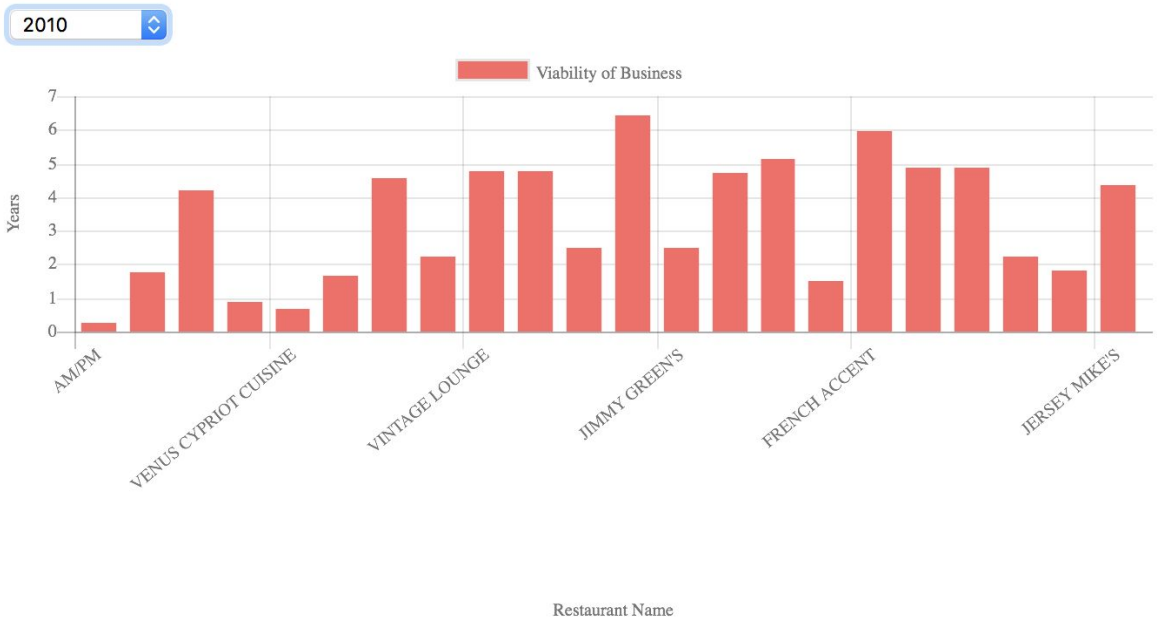
  This query reports the relationship between average review rating and food inspection result (pass/conditional pass/fail) for a restaurant. The datasets used in the query are yelp-reviews , yelp restaurant file provided and the food inspection data obtained from

City of Chicago. The average review rating file is obtained by merging restaurant file and yelp review file on restaurantId and later grouped by the restaurantId to obtain the average rating. The addresses of both the food inspection and average review rating files are normalized. The files are further merged on the names and addresses.
For each of the restaurants, the mode of most common food inspection result is calculated.
For visualization, the average rating is divided into 5 intervals of ratings. For each of the interval, the bar graphs show the count of restaurants against three food inspection result(pass, fail, conditional).



- **Query 8:** Failed Food Inspection and Restaurant Viability
  The problem here is to check the age of a business after a failed food inspection. The two datasets used for the data preparation are Business licenses and Food inspections both from City of Chicago Portal. Both the datasets are filtered for the required fields. Business Licenses has records of License Number, License Change Status and License Change Date , and Food Inspection has the License Number, Restaurant Name, Address, most recent failed inspection result, Inspection Date . The two datasets are merged together on the License Number.
  For each of the restaurants, then the difference between the Inspection date and License Status Change Date is calculated to get the age of a restaurant after the inspection.
  For visualization, Charts.js is used to plot the Bar Chart. In this chart , the restaurants age is plotted with their name on the x-axis and years on the y-axis. The restaurants are filtered on the basis of the year in which the inspection was conducted. Below is a

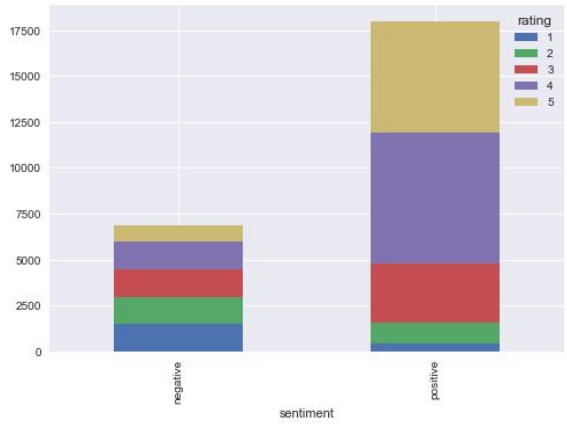
sample of visualization for all the restaurants with their age whose inspections were conducted in year 2010.

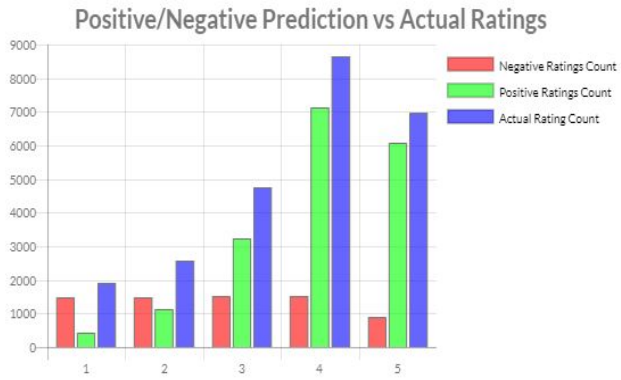


- **Query 9:** Does having liquor license influence crime incidents in neighborhood?
  For this query, 3 block radius is computed to determine the number of crimes around liquor businesses. The businesses are classified as liquor businesses on the basis of license description, considering the information provided on City of Chicago website about classes of liquor licenses.
  For visualization, census tracts are color coded by the number of crimes and clicking on census tract reveals further information like number of crimes and arrests and also prevalent age group data(Query 3). From the visualization, direct correlation can be seen between the number of business licenses in an area and the crime incidents. More crime incidents have occurred near areas where there are more number of business licenses and vice-versa.

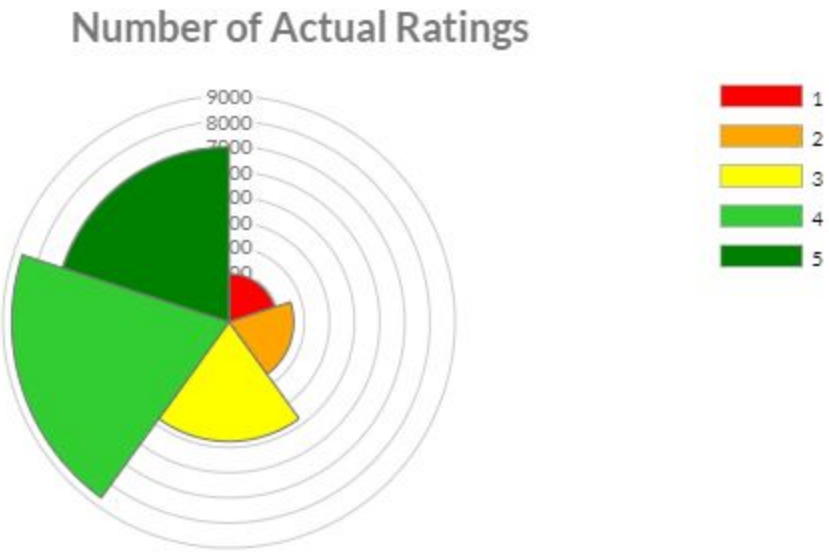
- **Query 5, 6:** Perform sentiment analysis and visualization:
  For this query we used TextBlob which takes in the sentence as input and gives out a polarity(-1 to 1) and subjectivity(0 to 1) values for that sentence. We used a threshold of 0.12 by trial and error to get the best performance for the polarity. It a sentence scored above 0.12, it was classified as positive else it was classified as negative. This gave us an accuracy of around 72%. We later(after the deadline) tried implementing an LSTM to do this binary classification. It gave us a much better score and took us upto 80% for accuracy.
  The first approach using TextBlob was more of a Parts of Speech (PoS) tags, semantics and grammar based approach and the second one was a deep learning approach. As expected the ML based deep learning approach out performed better than the non ML

approach. The performance of the deep learning model can be further improved by parameter tuning and can become a very good classifier for binary classification of text based sentences.



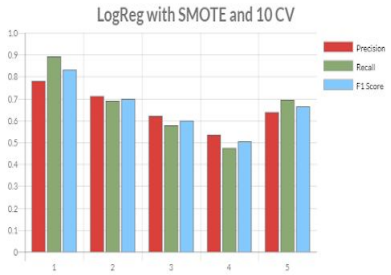


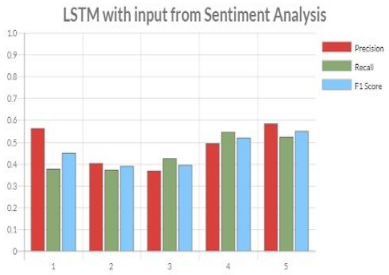
- **Query 7:** Predict Review Rating from Text:
  We used several methods for doing this. Since we have five classes and the data is imbalanced, we are using Precision, Recall and F1-Score as our performance metric. Firstly, we started off with a simple test train split of 80-20. It gave us okay performance with an average F1-Score of five classes as 48.6%. This improved when we tried using an LSTM with the input we got from the previous sentiment analysis step. Here our average F1 Score of 5 classes was 50.7% which was more than a 2% increase. This was mainly because we had 5 different classes and the data was not enough to properly classify it.



To overcome this problem, we used the technique SMOTE, which synthetically generated more data for the classes which had less data and basically made all the classes have same amount of data. This drastically improved the performance of our prediction and we got an average F1 Score of all classes as 65.8% which is a whopping 15% higher. One point to be noted is that since we do not use actual real data for

testing, we cannot be completely certain about this model's performance in the real world.





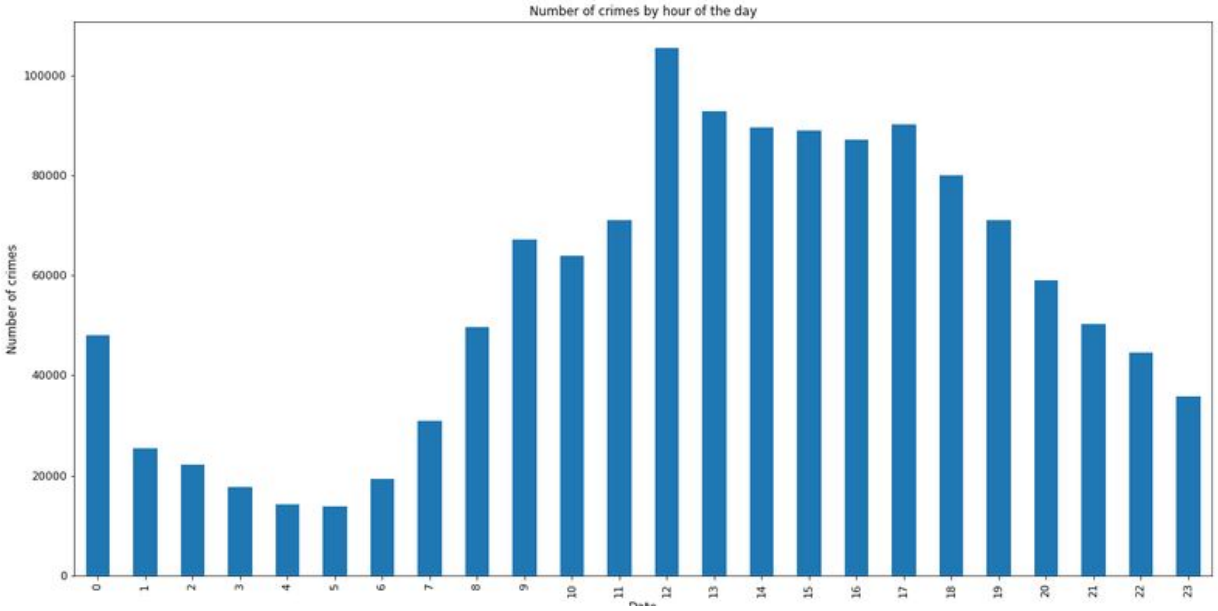
- **Query 2,10:** Prediction of type of crime and prediction of type of robbery for Summer 2018:

  Crime and Weather Data Preparation:

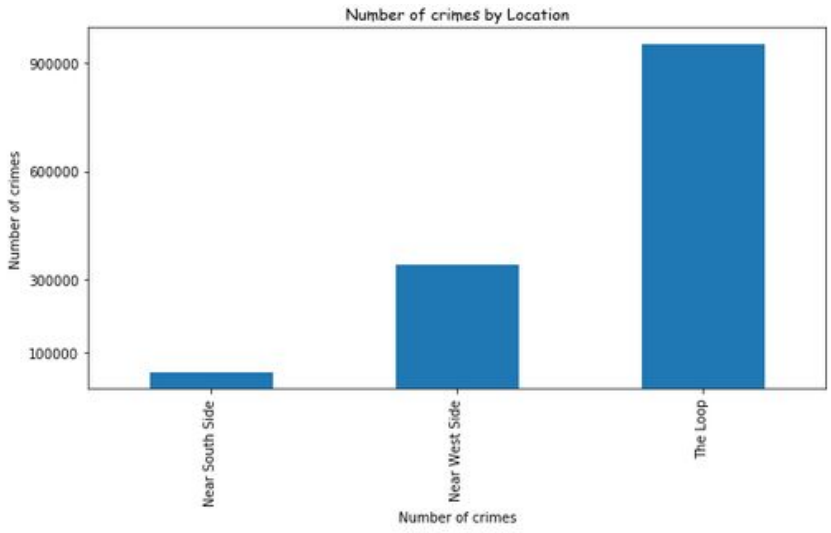
  Two main sets of data are used for this part, one is the weather data recorded at the Midway Airport, and the other is the crime data from City of Chicago Data Portal. The raw weather data contains over the period of 2001-01-01 to 2018-03-29, each recording contains the values of timestamp, temperature, dew point temperature, humidity,wind direction, wind,speed, precipitation, atmospheric pressure and other miscellaneous variables.Only mean temperature ws used for predictions. The raw crime data initially contains instances of crime in Chicago over the period of 2001-01-01 to 2018-03-29; each recording contains the value of timestamp, crime type, arrested, and other miscellaneous variables such as crime ID. To prepare the crime data for training and testing, all the irrelevant variables are first deleted.

  Dataset Analysis :

  Examining number of crimes by hour of day shows that frequency of crime is at its lowest at 5 AM in the morning. This is likely because in the early morning hours most people are sleeping or at home. Crime steadily increases from the low point at 5AM and reaches it maximum level at 12 PM. The spike in crime at 12 PM, which is largely driven by a high incident of theft during this period.



The dataset comes with location information (latitude and longitude) for each reported crime. Using this, we can sort the locations based on the number of crimes committed in a given Community area. For the crime data in the zip code 60601 to 60607, Loop communities has the highest crime count.



Modeling and Testing:

Random Forest was used for Crime Prediction because it is suited for multiclass problems, while SVM is intrinsically two-class. For multiclass problem we will need to reduce it into multiple binary classification problems.Also Random Forest works well with a mixture of numerical and categorical features and when features are on the various scales.

Using **Random Forest** we **predicted Robbery type for summer with an accuracy of 92%** and for a given address we **predicted crime type** (without considering weather data) with an accuracy of **94%**.We obtained this accuracy on train and test dataset ratio of 80:20. This could become more consistent if we run cross validation.

Other ML models we tried were:
Decision Tree - **92% accuracy without weather data** and **90% with weather data**.
KNN - **80% accuracy without weather data** and **78% with weather data**.

The modelling features were chosen using recursive feature elimination.For Robbery Type prediction features used are Latitude, Longitude, Weekend, hour, average temperature and non hispanic. For Crime Type prediction given a address, features used are Latitude, Longitude, Weekend, hour and season.

Conclusion:

- Understand the wide variety of concepts in Data Science by implementing different parts of the project: Data Discovery, Data Extraction, Data Analytics and Data Visualization.
- Gathered data about the City of Chicago, its people, health, businesses, transportation, government, census and demographics
- Extracted data about Business Licenses, Food Inspections, Crime, Census, Demographics from City Of Chicago and census.gov. Also scraped data from Yelp.
- Use the data from the Data Extraction phase and get the results for the 10 queries. Integrated and analyzed the data. The Python libraries used the previous steps are: Keras, TensorFlow, Scikit-Learn, Pandas, Numpy, NLTK, TextBlob, Pickle, CSV, Xgboost, etc.
- Built a website that displays the integrated data using Openstreetmap (OSM), Python (Matplotlib, Seaborn, Pyplot), JavaScript (Charts.js, Leaflet.js, JQuery), CSS and HTML. This site also includes charts for all the data analytics tasks.