**Project Outline**

**Sketch It Out:**

• Selected topic

For topic of the final project, we selected Burglary and Robbery as crimes indicators in the state of Colorado to discover data patterns and with the use of Machine Learning Models to predict when crimes will take place based on the indicators.

• Reason the topic was selected

Crimes are intricate and interesting, and can be both predictable and unpredictable. Crimes have existed for a long time and will remain parts of society. Therefore, I am interested in predicting when crimes will take place.

• Select your question -- Questions the team hopes to answer with the data

1. Can we predict when crimes will take place based on key indicators: historical crime data, incident date, type of crimes, location of crimes and number of crimes?

2. Can we predict types of crimes with key indicators: historical crime data, incident date, type of crimes, location of crimes and number of crimes?

3. What crimes are increasing or decreasing, at what rate and over what timeframes with a given regional?

4. Can we confirm the findings against published statistics?

• Description of the source of data

The datasets were sourced from Colorado Crime Data Explorer site. While data for other crimes are available, in this project the scope was narrowed down to Burglary and Robbery crimes. Datasets were available for download in comma separated values format files from the year 2016 to 2019. Since the data was being sourced from the state of Colorado only, the dataset will be smaller. Both files have 54,930 rows and NIBRS\_incidents\_16\_19.csv consists of 14 columns while NIBRS\_Offense\_16\_19.csv has eight columns. NIBRS\_incidents\_16\_19 dataset holds information about the incidents while NIBRS\_Offense\_16\_19 holds information about the offenses associated with the incidents.

The following are the column names and their data types from NIBRS\_incident and NIBRS\_offense datasets:

NIBRS\_incidents\_16\_19 dataset:

DATA\_YEAR INT

AGENCY\_ID INT

INCIDENT\_ID INT,

NIBRS\_MONTH\_ID INT

CARGO\_THEFT\_FLAG VARCHAR

SUBMISSION\_DATE DATE

INCIDENT\_DATE INT

INCIDENT\_Month VARCHAR

INCIDENT\_DAY INT

REPORT\_DATE\_FLAG VARCHAR,

INCIDENT\_HOUR INT

DATA\_HOME VARCHAR

ORIG\_FORMAT VARCHAR

DID INT

NIBRS\_Offense\_16\_19 dataset:

DATA\_YEAR INT

OFFENSE\_ID INT

INCIDENT\_ID INT

OFFENSE\_TYPE\_ID VARCHAR

ATTEMPT\_COMPLETE\_FLAG VARCHAR

LOCATION\_ID VARCHAR

METHOD\_ENTRY\_CODE VARCHAR

NIBRS Link: https://crime-data-explorer.app.cloud.gov/downloads-and-docs

• Description of the data exploration phase of the project

In the initial data analysis, we explored the structured dataset to uncover initial patterns, characteristics, creating a broad picture of important trends and major points to study in greater detail which took place in Tableau and will be incorporated in the dashboard.

The variable we will be included in the predictive models are Nominal data types. Nominal data cannot be used in statistical computation like mean and standard deviation. As shown in the DataExploration\_phase.ipynb, the descriptive statistics performed to both datasets did not show any significance output. Cross tabulations used with Nominal variables then chi-square test can be performed on a cross-tabulation but beyond the scope of this project.

• Description of the analysis phase of the project

Which models did you choose and why?

Logistic Regression and Support Vector Machines (SVM) models are classification algorithms so they are best applied to nominal data to predict discrete outcomes. The algorithms will attempt to learn patterns from the data. If the models are successful, they will help predict when the crime will take place? The datasets in this project contain categorical features that will be included in the machine learning algorithms. The categorical features will be converted to numerical data to use in the machine learning process. Also, the features needed with the Logistic Regression are pulled from two datasets and joined into a data frame which were coded in the DataAnalysis\_phase jupyter notebook.

For the new data frame we will be pulling the following features and targeting INCIDENT\_Month variable:

NIBRS\_incident dataset:

DATA\_YEAR INT

INCIDENT\_Month VARCHAR

INCIDENT\_DAY INT

NIBRS\_Offense dataset:

OFFENSE\_TYPE\_ID VARCHAR

LOCATION\_ID VARCHAR

The features below were encoded using python libraries:

INCIDENT\_Month VARCHAR

OFFENSE\_TYPE\_ID VARCHAR

LOCATION\_ID VARCHAR

How are you training your model?

The models were trained using sklearn.model\_selection importing the train\_test\_split. Logistic Regression and SVM models analyzed when the offense will take place.

What is the model's accuracy?

The Logistic Regression accuracy score was 9.5% while the SVM was 10.0%.

• Result of analysis

Both Logistic Regression and SVM accuracy results were close to each other and very low. They did not predict when the crime would take place. Both models were chosen for the types of data and both were resulted in similar accuracy, the dataset had issues or limited in rows or features.

• Recommendation for future analysis

Include datasets from the other states to improve the outcomes of this project.

• Anything the team would have done differently

Longer time exploring the dataset and prototyping machine learning models.

• Technologies use

1. Use Python or Excel to clean, prepare, and explore the data as well as to complete the initial analysis.

2. Use database integration (Postgres, MongoDB, or SQLite) to store your cleaned data.

3. Use machine learning to enhance your topic.

4. Use Tableau can be used to create visuals to help tell your Dashboard and Google Pages to complete the data story.

5. Use Tableau or JavaScript to build a dashboard to present your results.