Heath Thompson
Taylor Townsend
Jiyin Zhang
Scott Martin
Dawson Hill

Final Report – LA Crime Data Analysis with KModes

1. Choose an investigation and identify pre-existing sources of data that can address a particular data science goal: (7%)
    A. Choose, and state, the goal and reasons why the data sets were chosen and how they were found and managed, Min. 3-4 sentences
        - The dataset was generated to find statistical data about crime in the greater Los Angeles area from 2020 to present date. Our collection was driven by the need to better understand crime rates, specifically for 2021, in LA and to answer our hypotheses questions. The dataset was downloaded via data.lacity.org, that publishes weekly updates of crime statistics made available for public access. The CSV file we obtained contained too much data that was unnecessary for our specific analysis. To cleanse our dataset, we removed all but nine columns (DR_NO, DATE.OCC, TIME.OCC, AREA.NAME, Vict.Age, Vict.Sex, Vict.Descent, LAT, and LON) and removed fields that didn't contain any data.
    B. Document and discuss the data formats and any metadata standards/ conventions in use, and the method(s) of discovery and access and how they helped or hindered the process, Min. 3-4 sentences
        - The available data format for the dataset includes CSV, RDF, JSON, XML, and RSS. Providing multiple formats allows users the choice to determine what file type works best for them. We chose to use CSV, as it worked best for our group and analysis methods. The metadata uses the PREMIS data model which consists of 4 interrelated entities: object, event, agent, and rights. There was not a file to download for metadata, so we had to pull details from the main webpage. This hindered the process as having relevant metadata elements is necessary to better understand the data, how it functions, and the reuse of data in the future.

2. Data Analysis (12%)
    A. Develop and state two particular questions/hypotheses related to the goal of the investigation and that can be answered using the datasets under consideration. Design an analysis study (preliminary, full and post) to answer these questions and document the analysis design, Min. 3-4 sentences (3%)
        - Which victim descent (race) is more likely to be victims of a crime in Los Angeles?
            - To answer this question, we developed a pie chart using the Vict.Descent column in our cleaned data set. The original data set was cleaned of any null or empty values in each of the Time, Vict.Age, Vict.Descent, Lat, and Lon columns. This was done by simply removing the whole row from the data. With the pie chart, we were not only able to determine which race was targeted most, but also which was targeted the least. Overall, we

were able to gain a full understanding of victim descent within the data and answer our first hypothesis question.

- Is there a correlation between the age of victims and their associated gender?
    - To answer this question, we used the count-plots produced after running our KModes machine leaning model. We produced count-plots for victim age, race, gender, and time of incidence. Analyzing the age and gender plots, we found that male victims tend to be roughly ten years older than female victims in LA.
- We developed an analysis study to determine the traits of a person who was most likely to be victimized in LA.
    - Our preliminary analysis was to first find and understand key elements in the data. This was done by determining areas of interest and developing questions about those areas. We then cleaned the data of any null or empty values and produced a few charts to visualize what the data was showing.
    - Our full analysis consisted of further data preprocessing for machine learning integration. This preprocessing eliminated columns of data that were not necessary for our research and understanding. We chose to use an unsupervised machine learning approach to determine patterns in the data not recognizable in the data's raw state.
    - Our post analysis is responsible for concluding any patterns or observations within the data after machine learning integration has been done. Recognized patterns were then validated with other data analysis tools and charts to ensure authenticity.

B. You need to use at least one machine learning method in your data analysis. Only using linear regression is not enough for the assignment. Provide a description of the choice of tools/ methods used or a description of any code or scripts written, and describe how your results were stored and managed, min. 3-4 sentences (4%)

- Because we are looking for unknown patterns in our data, we don't have labels to guide and train a machine learning model. We needed to use an unsupervised approach that learned something not yet known. We also found that the majority of our most useful data was categorical rather than continuous and numerical. We chose to use a clustering algorithm that could potentially cluster common groups of personal traits like age, gender, and race. With this understanding, KModes served as a great option as it performs similar to KMeans, but allows for categorical features in the data. In short KModes builds a specified number of clusters based on the observations/rows in the data. It calculates the most frequent trait in each feature of the data and then uses those common traits to build a cluster. We were able to use these clusters to identify people who are most likely to become victims in LA.
- Our results of preprocessing and data cleansing were stored as CSV files. We saved the results of our machine learning output into a CSV file as well. This format falls in-line with the format we used to download the raw data. Data with

cleansing, preprocessing, and machine learning was stored as Pandas Data Frames for easy manipulation before being finalized and saved as CSVs. CSVs could then be read into data frames and manipulated furthermore as necessary.

C. Perform the analysis in a form that can be validated and describe the steps and results you took to ensure this validation, min. 5-6 sentences. (5%)
    - As part of our post analysis within the analysis study, we conducted validation in several areas of our model's output, we wanted to ensure this output was in-line with what the data originally represented. Validation was done using pre-existing charts as well as a series of other visualizations. We compared these charts against each other and against the output of the model to ensure oddities and other extreme values were not present. With all representations providing the same or similar information, we believe our results are accurate and complete.

3. Presentation/ Visualization (6%)

A. Prepare presentations/ visualizations of both the data (and any metadata, information) and the results of the analysis and describe them, min. 2-3 sentences. (2%)
    - Male vs Female crime rates
        - In total there were 178,163 entries with four categories given for Gender. Victims could be F (Female), H, M (Male), or X (Unknown). To aid in analysis, we wanted to know if males or females were more likely to become victims of crime in LA. The results of the analysis concluded that males were about 15% more likely to be a victim of a crime committed in LA.
    - Race distribution
        - For distribution on the basis of race, there were 178,161 records with six categories given for victims. Victim descent could be B (Black), H (Hispanic/Latin/Mexican), O (Other), W (White), X (Unknown), or Others. The visualization indicates that H (Hispanic/Latin/Mexican) individuals are more likely to be a victim of a crime in LA. When added in age as an additional factor, we see that a White individual aged 43 is likely to be a victim.
    - Choropleth map for incidents
        - For each row of our retrieved data set, each incident has several attributes for spatial information, such as Lon, Lat, and Area.Name. To get an intuitive impression of the crime incidents' geographical distribution, we thought the choropleth map would be a proper tool capable of demonstrating the location and incident counts simultaneously. We chose "los-angeles-county.geojson" from the Github project (https://github.com/codeforgermany/click_that_hood). And to map the incident points into the cities in geojson file, we implemented a python library named "shapely.geometry" to calculate the incorporated relationships of the coordinates. Finally, the outcome figure of the

Choropleth map for incidents is generated with the help of another python library named "plotly".
    - The choropleth map shows that the "Downtown" region has the highest incidents. We guess that is partly because the Downtown region is also where the Los Angeles Police Department is located.
- Time Occurrences
    - To get the information for the time occurrences, 5 different time objects were created (12:00 A.M. - 5:59 A.M., 6:00 A.M. - 11:59 A.M., 12:00 P.M. - 3:59 P.M., 4:00 P.M. - 8:59 P.M., and 9:00 P.M. - 11:59 P.M.). These objects were filled with all the occurrences of a crime that took place during that time frame. We used a filter in R Studio to be able to get the correct data in the objects.
- Age and Race Correlation
    - To find the extract the information needed to find a correlation between age and race, we used R Studio and created a variable for the 5 most common races. We then filtered the data by age, took the mean of the age for each group, and assigned it to the respective variable.
- Count-plots for model output visualization
    - For each column Vict.Age, Vict.Descent, Sex, and Time, we have a count-plot that sums the number of occurrences for each trait in each cluster. These count-plots illustrate each cluster's most frequent traits. For example, Cluster 0 has a gender count-plot with majority males, a count-plot with majority Hispanic descent, a count-plot with majority 30-40 years age, and a count-plot with majority 16:00-21:00 time frames. Cluster 1 shows comparable results but with other most frequent traits.

B. Document the management of the presentation/ visualization products and any associated metadata, etc. min. 2-3 sentences (2%)
    - We currently have our dataset, visuals, cleansed code, and machine learning information located on One Drive through the University of Idaho. This is password protected and only accessed by invitation. Again, we found very limited metadata and will focus on adding an actual file with relevant information in the future. Once we finalize our data and presentation, we plan to house this information on Heath's GitHub.
C. Describe how your presentation/ visualization meets the goal of the investigation and highlight any value that was gained, min. 3-4 sentences (2%)
    - We had three overarching hypotheses that we wanted to answer based on our dataset of 2021 Crime Rates in LA. These included, "Which victim descent (race) is most likely to be victims of a crime in Los Angeles?", "Within the city of LA, at any particular time, is there an increase or decrease in crime rates, in comparison to the location where the crime took place?", and finally, "Is there a correlation between the age of victims and their associated sex?". Through machine learning, we wanted to take these three questions to learn about patterns within the data to predict if certain individuals are most susceptible to

being a victim of a crime in LA. The visualizations provided in our presentation answer each of our three hypotheses and give users an accurate representation of potential victims based on age, gender, race, time, and location. For instance, when viewing the pie chart depicting victims based on gender, it is clear that males are more likely to be a victim.

- The analysis of our machine learning indicated there were two clusters, or "Expected victims", with predictable patterns on race, gender, time of day, and age. Our model determined that Hispanic males between the ages of 30 and 40 years old and White females between the ages of 20 and 30 years old are most likely to become victims of crime in LA. It also determined that 4:00pm and 9:00pm is the time a male is most likely to be attacked while 6:00am to 12:00pm is the time a female is most likely to be attacked.

4. Describe your overall data management plan for the results of questions 1, 2 and 3 using the 9 categories of data management from assignment 1, min. 1-2 sentences for each category (5%)

- Logical Collections
    - Within the dataset, there is one file to be downloaded with 28 columns of data. Although there isn't a Data Dictionary within the CSV file, the webpage provides additional information to give descriptions of each column name and the type of data (plain text, numbers, data and time). There is not a particular naming convention used throughout all the column names, as we have columns such as DR_NO, Crm Cd 1, and Location.
- Physical Data Handling
    - While there were multiple data formats available for download, our dataset was displayed in CSV format containing 28 columns with thousands of rows of data. There was no metadata file available for download, metadata was only presented on the webpage in short supply.
- Interoperability Support
    - This data collection is readily available via data sharing on data.lacity.org and is available for download in multiple formats (CSV, RDF, JSON, XML, RSS). Data owners provide ample opportunity to download their data with users being able to pick their preferred format. There is also a 'Table Preview' option on the webpage to view 13 records at a time to sift through data if a user doesn't want to download a complete file.
- Security
    - Data access did not require login credentials or authorization to obtain access or download the records. The dataset is made available for public viewing, as access methods are not controlled. We are unsure if the dataset is audited, but there is a counter to show how many times the dataset has been viewed and downloaded.
- Data Ownership

    - The data is provided by the Los Angeles Police Department, with the dataset owner being LAPD OpenData. Provided anyone should need to contact the owners, there is a pop-up modal to send a message with any comments, suggestions, or questions.
- Metadata
    - While there is little metadata available for interpretation, there are four interrelated entities – object, event, agent, and rights. When viewing the data.lacity.org website, there is metadata regarding Data Owner, Category, Tags, and information regarding location coordinates. With the limited amount of metadata available, in the future we will provide a more detailed metadata file showing more Dublin Core elements for users.
- Persistence
    - The data owners update this information on a weekly basis and make it available online. Although we have made changes to the CSV file, we currently house our dataset in OneDrive with restricted permissions. At the conclusion of our analysis, we plan to house our data and visualizations on Heath's GitHub repo.
- Discovery
    - The dataset is accessible by searching via data.lacity.org and navigating to Public Safety to view datasets based on different time periods. Search for crime data using keywords or tags of "lapd", "crime", "crime data", "police", "safe city", and "crimes". As noted on the data.lacity.gov website, the data was first made available in 2020 and is created from original crime reports.
- Dissemination
    - With the dataset being made publicly available and updated on a weekly basis, the responsibility of disseminating the data falls on the data owners. When viewing the data.lacity.org website, it provides "Data Last Updated" dates, as well as when the Metadata was last updated to inform users of the newest available data.