Los Angeles Arrest Data (2010 – 2019)

Sidney Thomas
UCLA Extension - Intro to Data Science
Final Class Project

Introduction

Los Angeles (LA) has the second largest metropolitan population in the United States, with approximately 10 million inhabitants. It comes with no surprise that in such a large population, conflicts and crime will ensue, and local law enforcement will make arrests. This project sets out to explore arrest data for the Los Angeles metro area, and understand what demographic and geographic components are dominant in predicting arrests, are arrested the most frequently, as well as the location of most arrests.

Data Retrieval & Editing

The Los Angeles arrest data was obtained from the City of Los Angeles' open data website. The data set includes arrests logged between 2010 and 2019, with over 1.3 million records.

The following edits were made to the data:

- Excluded arrests for those under 13
- Excluded records where any fields have NA
- Truncated the dataset to 9 key fields pertinent to this analysis
    1. Arrest Date
    2. Arrest Time
    3. Time Category (categorized using the Arrest Time field)
        a. Morning: 5:00 am to 11:59 am
        b. Afternoon: 12:00 pm to 4:59 pm
        c. Evening: 5:00 pm to 9:59 pm
        d. Overnight: 10:00 pm to 4:59 am
    4. Area Name
    5. Age
    6. Age Range (categorized using the Age field)
        a. 18-24
        b. 25-34
        c. 35-44
        d. 45-54
        e. 55+
    7. Sex
    8. Charge Group
    9. Charge Group Description

After excluding records that didn't match the project definition, approximately 1.2 million records remained for data exploration.

Hypothesis

Upon the initiation of the project, a few assumptions were made about the outcomes and findings:

- Males are more likely to be arrested than females, regardless of age
- Arrestees in the 18 -24 age group category would make up the majority of arrests for both males and females
- Of the 21 distinct areas, Hollywood would log the most arrests, due to the large number of nightclubs, bars, and tourists
- Arrests would be disproportionately higher between the hours of 10 pm and 5 am, compared to other periods of the day. This assumption is also based on the hours most bars and nightclubs are visited

Exploratory Data Analysis

An iterative data exploration approach within R was taken to discover trends, summary level metrics, and potential issues within the dataset.

The following methods were implemented for data discovery: contingency tables, level function, summary function, plots, scatterplots, and barplots.

Using these exploratory data analysis methods, the following key trends were uncovered:

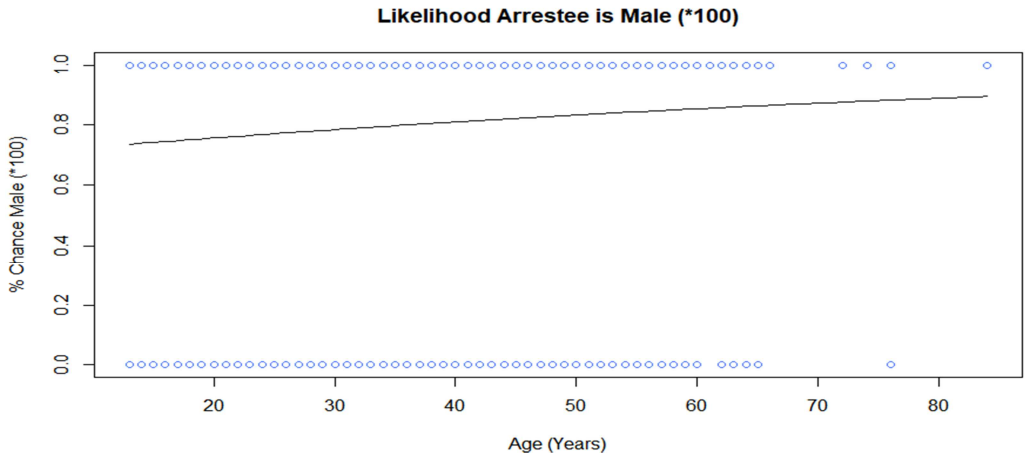
- Arrest Time: higher occurrence in the evening/overnight hours
- Area Name: 21 distinct areas, with Central (10%) and Hollywood (9.2%) logging the most arrests
- Age: range from 13 through 97
- Sex: approximately 80% of all arrestees are male
- Charge Group (Top 4 Categories): Narcotic Drug Laws (12.5%), DUI (9.2%), Drunkenness (8.8%), Aggravated Assault (6.6%)

Logistic Regression

In order to explore impact age has on the sex of the arrestee, classification using logistic regression was utilized. The step by step approach in the model is listed below.

- A new binary categorical variable named “ismale” was created
    - any observation where the sex = M set to true, and observations where sex is not M set to false
        - Note – there are only values of M & F within the sex variable
- The dataset was split into a trained data set of 787,779 observations, and a test dataset with 525,186 observations
- The r generalized linear model (glm) function was run on the training data set
- Due to the large amount of observations, a sample of 1,000 records was created

- The sample training set was plotted with a linear model sigmoid curve to show the likelihood of an arrestee being male (which increases with age)



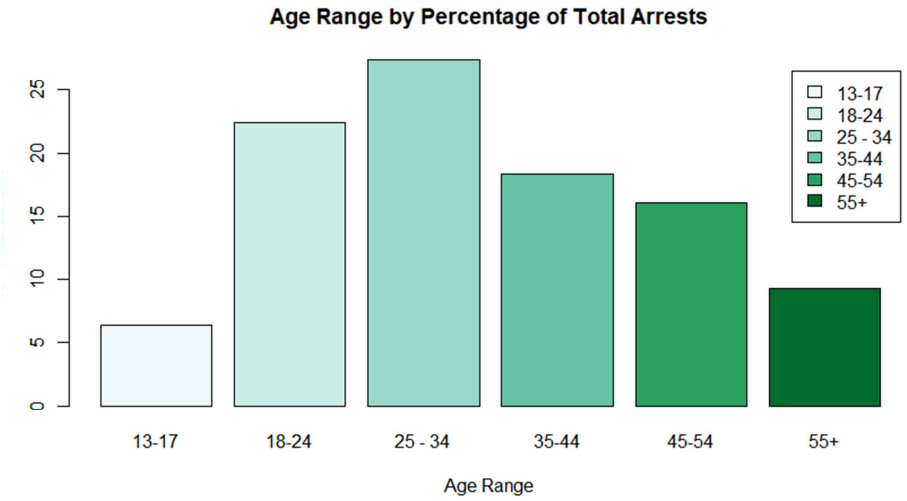
- The predict function was used on the trained model to make predictions about the testing dataset (probability an arrestee is male)
- It was discovered that when the model predicts more than a 76% chance the observation is male, it is correct 75% of the time
    - Through multiple attempts, 76% proved to be the most optimal probability predictor

Conclusion

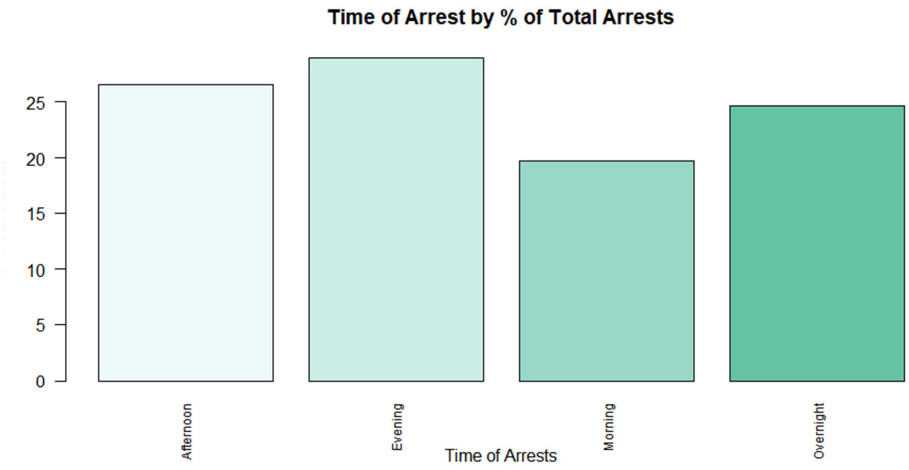
A vast amount of knowledge was obtained throughout the analysis, with some hypotheses proving true, while others were corrected through the data exploration and modeling.

As expected, males were much more likely to be arrested than females, regardless of other factors. Interestingly, as age goes up, the chance of the arrestee being male increases. At age 15, there's a 73% chance of the arrestee being male, and at 80, there's an 88% chance the arrestee is male.

Contrary to the hypothesis, the 25-34 age category actually has more arrests than the 18-24 age group.



Arrests were most prominent in the Central and Hollywood neighborhoods. And, a higher proportion of arrests occur in the evening (5 pm to 9:59 pm) than both the overnight and afternoon hours.



References

Arrest Data from 2010 to Present. Retrieved from https://data.lacity.org/A-Safe-City/Arrest-Data-from-2010-to-Present/yru6-6re4