IDS 702 [Optional] Team Project Final Report Draft

Robin Arun, Revanth Chowdary Ganga, Suim Park, Meixiang Du

TOTAL POINTS

0 / 0

QUESTION 1

1 Refer to class website for more information 0 / 0

✓ **+ 0 pts** *Click here to replace this description.*

1 more interesting to use a title that relates to the topic

2 no need for the sub-headings in this section (1.1, 1.2, etc). The introduction should be in paragraph structure and start with the motivation/background, then research questions, then give some information about the data. Be sure to cite the dataset appropriately (e.g., APA format) instead of just including the hyperlink.

3 this is all you need here; the table is not necessary as it doesn't provide any additional information. Simply state the percentage of observations that were missing/excluded. Also, not sure how it helps to classify observations as "unknown;" why not just exclude them?

4 Good info, but don't use actual variable names or columns. The reader will not see the dataset itself. Simply state that crimes were categorized into serious and non-serious, keep the examples of serious crimes like you already have, and give N (%) for each crime type. Including a bar chart is also not necessary as it doesn't provide additional information to N (%)

5 no need to justify each variable. Simply state that you used a priori selection based on domain knowledge and list the variables that are included in the model.

6 In the methods section, you should only state that multicollinearity is assessed with vif. The actual values and any modeling action taken in response to the values should be in the results section. Also, the sentence structure is a little confusing. Maybe something like this would be better: "VIF values for all variables except victim sex and victim descent did not exhibit evidence of multicollinearity." Is descent the same or similar to race/ethnicity? If so, I'm surprised it is collinear with sex; I'd keep both in the model since they still describe different demographic characteristics.

7 round - no need for that many decimals

8 round

9 "if the value approaches 1" not sure this makes sense. sounds like the 0.7 will approach 1 in a certain context, which is not actually what's happening

10 results

11 not sure what I'm supposed to get from this

12

focus on interpreting these in context. which is more important here, sensitivity or specificity? what does each one mean in context?

13 what is H?

14 missing interaction term?

15 do you have enough observations in each category to support so many? Your sample size is very large so maybe so

16 sounds good! If you run out of room, you can move the diagnostic plots/figures to an appendix

17 add some interpretation. don't need to do all variables but pick a couple that are interesting/relevant

18 be more specific about key takeaways. relate back to the background/motivation

19 is this the abstract? should have one abstract for the whole paper

20 this is good but should be in the results section

21 this is good but can be simplified/reduced, particularly if you run out of space

22 this should all be in the results section

23 good interpretation in context. can be in results or conclusion

24 Good start. It seems like the pieces are there for the first question, but it needs to be formatted into paragraphs to fit the report structure. Too many subheadings and bullets that take up space. The second question has appropriate formatting/structure but the content is a work in progress. Overall, a big area of improvement should be the separation of the methods and results sections. The methods section should just describe your process but not contain anything that you actually found from the data.

Final Report-Draft-Logistic Regression

Suim Park

1. Introduction

1.1 Background

The data used originates from the crime records of the Los Angeles Police Department(LAPD) from 2020 up to present day. The data is created by transcribing original crime reports that are typed on paper and is updated on a `weekly` basis.

The Dataset contains **820,599 observations**, encompassing **28 variables** as of the latest update on **18-Oct-2023**.

Each row of the Dataset represents a crime reported in Los Angeles, and contains the following categories of information(with selected examples):

1. `Location`: Latitude, Longitude, Area, Street, District
2. `Victim Demographic`: Age, Gender, Ethnicity
3. `Crime Description`: Type of Crime, Investigation Outcomes, Weapon Usage
4. `Date and Time`: Date Reported, Date Occurred, Time Occurred
5. `Identifier/Classifier`: Crime Record Identifier, Mocodes

1.2 Research Questions

We aim to answer the following two questions using the data:

1. What are the strongest indicating factors that influence the seriousness of crime committed(categorical outcome).
2. What are the factors which influence the number of crimes committed(continuous outcome).

1.3 Why these questions?

Based on the first research question, we aim to confirm the characteristics that are influential factors in crime commission, thereby helping to prevent potential criminals from engaging in such acts. Specifically, through the results of this research question, if we can identify the times and areas where crimes are frequently committed, it will enable people to exercise caution in these places. Measures such as installing additional lighting or assigning more police officers during these times can be considered. Moreover, targeted education can be provided to individuals who are more vulnerable to crime, based on factors like race, age and sex. This research question is crucial as it aids both the public and government in understanding, preventing, and addressing crime more effectively.

2. Data

2.1 Data Cleaning

The dataset has 2 variables which have blank values.

	Number of missing values	Type
Victim Sex	104,654	Blank, "-"
Victim Descent	104,663	Blank, "-"

In our dataset, which comprises over 8 million entries, we need to remove all rows that contain blank values or "-"s for two specific variables: victim sex and victim descent. Compensating for these missing values is not feasible, as they cannot be appropriately replaced with alternative values. Due to the dataset's large size, we replaced blank values in both variables with 'unknown' during the model configuration process. 3

2.2 Outcome Variables

1. A new variable `crime_type` is derived from the 'Part 1-2' column in the original dataset which classifies crime committed into two categories; `serious` for crimes such as felony offenses like criminal homicide, forcible rape, etc; and `non-serious` for less severe crimes.

   The bar plot below represents the count of the seriousness of crime by category in the dataset, about **60%** of the crimes committed are serious crimes. 4

3. Model

3.1 Priori Selection 5

Considering our focus on understanding the influence of time, place, weapon used and demographic information on crime type, we examined variables that appeared more relevant or were likely to confound the relationship between the independent variables and the dependent variable.

1. `Research Question`: according to the research question, these variables are necessary to be included in the model.
   - Outcome variable: crime type (serious/non-serious)

`Confounding`: we considered confounding while setting up the model because other variables can affect the occurrence of the crime type.

- Time hours occurrence: At specific times, such as night or early in the morning, serious crimes can occur frequently.
- Distance to precinct: There are specific areas where serious crime is committed more frequently.
- Weapon used: If a criminal uses a weapon, the probability of committing a serious crime is likely to be higher.
- Victim race: Depending on the race, the rate of serious crime occurrence can vary.
- Victim sex: Females might be more vulnerable to serious crime than males.
- Victim descent: Each descent has a different rate of exposure to serious crime.

3.2 Variables in Model

Based on priori selection through research question and confounding, these variables are on our model:

- `Independent variable`: time hours occurred, distance to precinct, weapon used, victim race, victim sex, victim descent
- `Outcome variable`: crime type

4. Assessment

4.1 Multicolinearity

Since the VIF (Variance Inflation Factor) values of all variables in the model, except for victim sex and victim descent, are around 1, this indicates a high correlation between these two variables because each values are around 17 or 18. It suggests that including them in the model may lead to instability.

4.1 Model Fit

4.1.1 Deviance

It is observed that the base model had the lowest deviance value, indicating that it provides a better explanation and a good fit for the data.

- Model (base model): 985184
- Alternative Model 1(excluded 'victim descent'): 989110.5
- Alternative Model 2(based on alternative model 1, excluded 'victim sex'): 1024930

4.1.2 Akaike Information Criterion

Among the AIC values, base model had the lowest value, indicating that this model is the best for preventing overfitting.

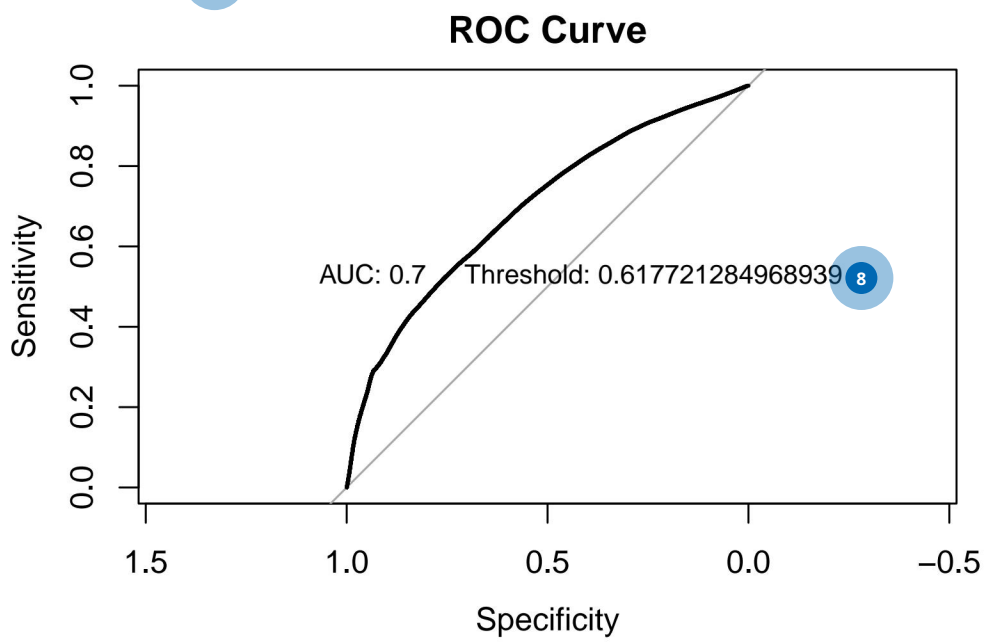
- Model (base model): 985236
- Alternative Model 1(excluded 'victim descent'): 989126.5
- Alternative Model 2(based on alternative model 1, excluded 'victim sex'): 1024940

4.1.3 McFadden's pseudo R^2

The McFadden's Pseudo R^2 value of 0.09032106 indicates that this model does not fit the data well, as it is relatively low compared to the typical range of 0.2 to 0.4. However, it's important to recognize that a higher McFadden's Pseudo R^2 value doesn't necessarily imply that the model is the best fit for the dataset. Therefore, further examination and analysis of the model are required.

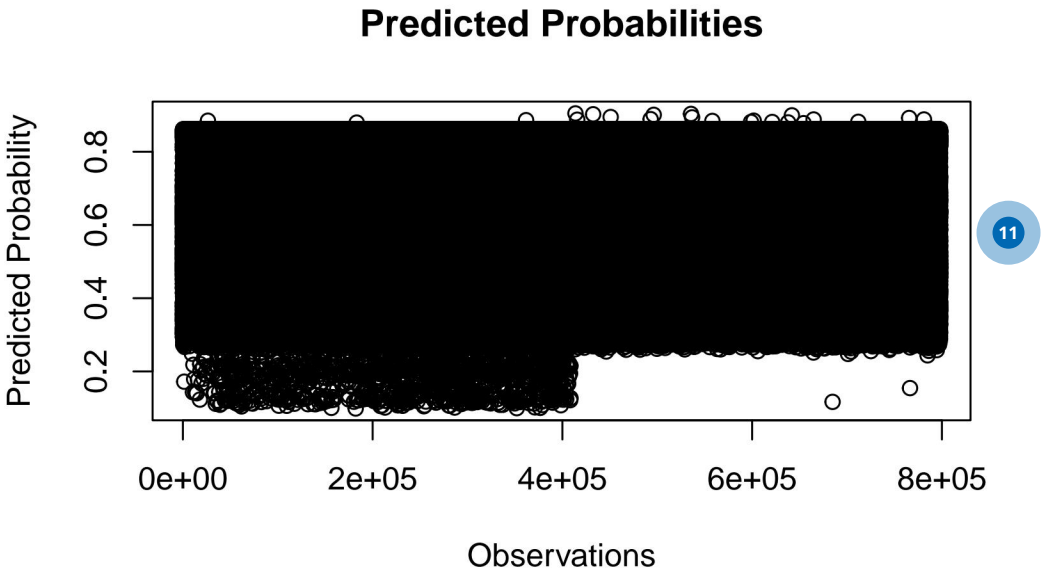
4.2 ROC Curve

Through the ROC curve, the best threshold for the model is found to be 0.618. This threshold allows us to comprehend the model's sensitivity and specificity. Furthermore, with an AUC value of 0.696(round value 0.7), it indicates that if the value approaches 1, the model's overall performance is considered excellent.



4.3Predicted Probabilities

Analyzing the probability plot, it's evident that if the predicted probability falls below the threshold value of 0.618, there is a higher likelihood of non-serious crimes occurring; conversely, probabilities above this threshold are indicative of serious crimes. However, the plot also reveals that the probabilities are not distinctly separated into clear 'serious crime' and 'non-serious crime' regions based on this threshold. This lack of clear demarcation suggests that the model may not be performing optimally in distinguishing between the two types of outcomes.



11

4.4 Confusion Matrix

The confusion matrix shows that the model has 497,513 true positive and true negative values (TP, TN), and 300,729 false positive and false negative values (FP, FN).

	Observed	
Prediction	**1**	**0**
1	241,207	211,267
0	89,462	256,306

With a Kappa value of 0.2634, the model's classification performance significantly deviates from random predictions, indicating a need for improvement.

	Value
Accuracy	0.6233
95% Confidence Intervals	(0.6222, 0.6243)
Kappa	0.2634
Sensitivity	0.5482
Specificity	0.7295

12

5. Results

5.1 Model Summary

The following table displays coefficient estimates, standard errors, p-values, and confidence intervals. The coefficient estimates and confidence levels are presented as log odds.

	Coefficient Estimates	Standard Errors	p-values	Confidence Interval (2.5%, 97.5%)
(Intercept)	1.1994074	1.795e-02	< 2e-16 ***	
Time hours occurred	1.0157575	3.695e-04	< 2e-16 ***	
Victim Age	0.9965660	1.511e-04	< 2e-16 ***	
Victim Sex: H	1.5608725	2.240e-01	0.04689 *	
Victim Sex: Male	1.8870123	5.262e-03	< 2e-16 ***	
Victim Sex: Unknown	3.4596957	2.501e-02	< 2e-16 ***	
Victim Descent: Black	0.7113850	1.690e-02	< 2e-16 ***	
Victim Descent: Chinese	1.6816080	4.519e-02	< 2e-16 ***	
Victim Descent: Cambodian	1.4936671	2.991e-01	0.17973	
Victim Descent: Filipino	1.1764790	4.062e-02	6.31e-05 ***	
Victim Descent: Guamanian	1.1197393	2.930e-01	0.69951	
Victim Descent: Hispanic/Latin/Mexican	0.6328782	1.632e-02	< 2e-16 ***	
Victim Descent: American Indian/Alaskan Native	1.0565499	7.896e-02	0.48600	
Victim Descent: Japanese	1.7841714	7.211e-02	9.89e-16 ***	
Victim Descent: Korean	1.1257615	3.635e-02	0.00112 **	
Victim Descent: Laotian	0.5476247	2.917e-01	0.03896 *	
Victim Descent: Other	0.7775481	1.779e-02	< 2e-16 ***	
Victim Descent: Pacific Islander	1.0373528	1.446e-01	0.79975	
Victim Descent: Samoan	0.6791198	3.072e-01	0.20785	
Victim Descent: Hawaiian	1.4673917	1.771e-01	0.03036 *	
Victim Descent: Vietnamese	1.9440009	8.485e-02	4.71e-15 ***	
Victim Descent: White	0.8580945	1.656e-02	< 2e-16 ***	
Victim Descent: Unknown	1.0444115	2.848e-02	0.12701	
Victim Descent: Asian Indian	1.8363606	1.215e-01	5.64e-07 ***	
Distance to precinct	0.9998391	6.238e-06	< 2e-16 ***	
Weapoon: Used	0.5976145	5.255e-03	< 2e-16 ***	

5.2 Results

We will present several plots here, including a map of California that depicts the counts of serious and non-serious crimes, a histogram showing the distribution of crimes based on the time of occurrence, a bar chart categorizing crimes by the weapon used, and pie charts illustrating the counts by race, sex, and descent.

6. Conclusion

1. **Strengths**

   - The model effectively explains how the independent variables within the model significantly impact the crime type from a statistical perspective.
   - By using the model, it is possible to identify factors that influence the crime type beyond time hours occurred, distance to precinct, weapon used and demographic information by controlling for other variables, obtain more accurate results.

2. **Limitations**

- Through various evaluation processes, it became evident that the model does not fit the data particularly well. In other words, it falls somewhat short in describing the data adequately.
- In order to find a model suitable for the research question, it may be necessary to add more variables or obtain new data by introducing new variables.

LAPD Crime Data Analysis - Poisson(Group 13)

Overview

In this section of the report, we present the second research question: **"What is the predicted number of crimes for a given area and time period?"** Addressing this question, we employ Poisson regression modeling, a choice driven by the count nature of our outcome variable and the need for predictive accuracy. Our model integrates various predictors, including the time and date of crime occurrences and geographic factors, to estimate the likelihood of crime counts in specific areas and time periods. We also explore potential interactions, such as between the day of the week and specific dates (e.g., holidays), to enhance the model's contextual relevance.

Given the dataset from the Los Angeles Police Department (LAPD), focusing on the period from 2020 to the present day, we look for predictive insights through the unfornuate individual crime incidents in Los Angeles. There are a wide array of variables such as location coordinates, victim demographics, crime descriptions, and date-time stamps, that provide a rich foundation for our exploratory and predictive analyses.

Our Poisson regression analysis on the Los Angeles Police Department's crime data reveals significant temporal and spatial variations in crime incidents. The model, with an R-squared value of 0.8246298, indicates a strong predictive performance, substantiated by the close alignment of the mean daily crime count estimate of 450.6 with the predicted values. Notably, areas like Devonshire and Foothill exhibit a lower incidence of crimes, as reflected by negative coefficients, while the Central area shows a slight increase. Weekday and month indicators suggest crime occurrences fluctuate with time, peaking in certain months and on specific days of the week, providing critical insights for targeted law enforcement deployment and community safety initiatives.

Introduction

Crime persists as a complex social issue with far-reaching implications, necessitating a rigorous analysis to inform public safety strategies. This report leverages an extensive dataset provided by the Los Angeles Police Department (LAPD), encompassing 820,599 crime reports. The granularity of the dataset allows for a detailed exploration of crime patterns across various dimensions, wherein each row of the datset contains location coordinates, area specifics, victim demographics such as age, gender, and ethnicity, crime specifics including type, investigative outcomes, and weapons involved, as well as temporal data like reporting and occurrence dates with corresponding times, all uniquely identified by record identifiers and mocodes.:

The focus of our statistical investigation addresses pivotal questions concerning the predictors of crime severity and the forecast of crime occurrences. Specifically, the second research question, **"What is the predicted number of crimes for a given area and time period?"** is crucial for allocating resources and enhancing preventative measures. By employing a Poisson regression model, we aim to dissect the frequency of crimes within the multifaceted urban setup of Los Angeles, identifying temporal and spatial hotspots of criminal activity.

Given the prediction nature, we've narrowed down our model to the following variables:

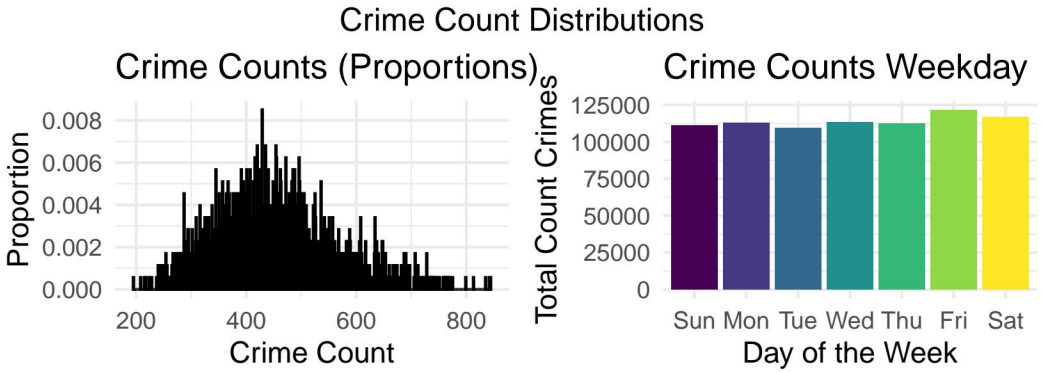
List of Predictor Variables

- Independent Variables
  - Time and Date: Time Occurred, Date Occurred
  - Geographic Factor: Area
- Outcome Variable
  - Count of Crimes (Generated by grouping the dataset by the selected predictor variables)

Understanding the dynamics of crime distribution is invaluable for law enforcement agencies, policymakers, and the broader community. It allows for a data-driven approach to crime prevention and the optimization of policing efforts. Moreover, deciphering the patterns in which crime manifests can lead to more effective community outreach programs and the potential to mitigate risk factors associated with crime. Our research contributes to this domain by offering empirical insights into the patterns of crime, which in turn could guide evidence-based resource allocation.

Crime Count Distributions by Weekday and Proportions 20

The composite visualization below merges two distinct perspectives on crime data: the aggregated crime counts by weekday and the proportional distribution of crime counts. On one side, the bar chart delineates the frequency of crimes for each day of the week, providing insights into the daily patterns of crime occurrence within the city of Los Angeles.



Methods

This section outlines the analytical procedures employed to address the research question 'what is the predicted number of crimes for a given area and time period?' posed by our study of the Los Angeles Police Department's crime data. Our methodology was designed to provide insights into crime patterns and to predict the number of crimes in a given area and time period.

Data Ingestion and Exploratory Data Analysis (EDA): 21

The crime dataset, sourced from the Los Angeles Police Department and publicly available via Dropbox, was then ingested into our R environment. This dataset underwent an initial cleaning process where we selected only the most relevant columns, including crime record numbers, dates of occurrence, area identifiers, and victim demographics.

The EDA was an integral preliminary phase where we immersed ourselves in the LAPD crime dataset to understand its intricacies and prepare it for in-depth analysis. Our EDA encompassed a multifaceted approach:

1. **Data Familiarization**: We began by acquainting ourselves with the dataset's structure, which comprised 820,599 observations and 28 variables, encapsulating a comprehensive range of crime-related information.
2. **Data Integrity Checks**: We conducted checks for missing values, particularly within victim demographics, recognizing that missing data could introduce bias or affect our models' robustness. Where missingness was systematic and substantial, we discussed potential strategies for imputation or cautious exclusion.

3. **Visual Exploration**: Our EDA involved the creation of bar plots and time series graphs to visualize the distribution of crime types, daily crime counts, and observe trends and outliers. This visual exploration allowed us to detect any peculiar patterns or anomalies, such as unexpected spikes in crime counts on specific days.

4. **Variable Relationships**: We assessed the relationships between various variables, including the time of crime occurrences (time of day, day of the week, and month), location (proximity to precincts), and victim demographics (age, gender, ethnicity). Through this, we observed minor variations in crime distribution, with more pronounced differences at certain times of the day.

5. **Outcome Variable Creation**: From the EDA, we derived crucial outcome variables. The `crime type` variable categorized crimes into serious and non-serious, providing a binary classification for further inferential statistics. Additionally, we aggregated the data to construct a daily crime count variable, offering a continuous measure for our predictive models.

6. **Data Cleaning and Preparation**: Variables not pertinent to our research, such as crime record identifiers and redundant information, were excluded. We also addressed categorical variables with numerous levels by consolidating them under broader categories to simplify our analyses and reduce model complexity.

Modeling and Preliminary Insights:

The decision to utilize Poisson regression was underpinned by the nature of our dependent variable, which counts the number of crimes—a typical example where Poisson regression is deemed appropriate due to its count-based distribution assumptions.

In constructing the model, we included area name, day of the week, and month as independent variables, hypothesizing that these factors significantly impact crime rates. A stepwise selection method, guided by the Akaike Information Criterion (AIC), was instrumental in refining our model. This iterative process of variable inclusion and exclusion allowed us to identify the most statistically significant predictors while maintaining model parsimony.

Preliminary examination of the model outputs suggests distinct spatial and temporal crime patterns. Certain areas, indicated by negative coefficients, such as Devonshire and Foothill, were associated with lower crime rates relative to the baseline, while the Central area showed a higher propensity for crime occurrences, as suggested by its positive coefficient. Moreover, the day of the week and month variables revealed variations in crime occurrences, with specific days and months exhibiting notable deviations from the average crime count. For example, coefficients for certain months, like September through December, pointed to a significant decrease in crime counts, potentially reflecting seasonal crime trends.

(AIC - 15358) - Work in Progress, TBD

(Overdisperson) - Work in Progress, TBD

Our preliminary evaluation of the model's predictive accuracy, gauged through the RMSE, yielded a value of 42.58718. This metric provides an estimate of the average difference between the predicted and observed crime counts, and is particularly useful for stakeholders to grasp the model's performance in concrete terms.

Additionally, the R-squared value of 0.8246298 reflected a strong linear relationship between the observed and predicted crime counts, explaining a substantial proportion of the variance in the crime data.

22

Results(WIP)

Our analysis using a Poisson regression model has yielded preliminary insights that are promising in addressing our research question: "What is the predicted number of crimes for a given area and time period?" The initial results point towards significant spatial and temporal effects on crime rates within the city of Los Angeles.

From the model outputs, two insights emerge at first glance. Firstly, the area named 'Central' has a higher relative crime rate compared to the baseline, which suggests that this area experiences a greater frequency of reported crimes. Secondly, the coefficients for the months indicate seasonal variations, with certain months like September showing a marked decrease in crime occurrences. These patterns are critical for understanding the ebb and flow of crime across the city and may provide valuable guidance for law enforcement resource allocation.

(NOTE: These results are placeholders and preliminary, they offer a starting point for further discussion with professor Andrea to refine the model and interpret the findings accurately.)

1 Refer to class website for more information 0 / 0

✓ **+ 0 pts** *Click here to replace this description.*

1 more interesting to use a title that relates to the topic

2 no need for the sub-headings in this section (1.1, 1.2, etc). The introduction should be in paragraph structure and start with the motivation/background, then research questions, then give some information about the data. Be sure to cite the dataset appropriately (e.g., APA format) instead of just including the hyperlink.

3 this is all you need here; the table is not necessary as it doesn't provide any additional information. Simply state the percentage of observations that were missing/excluded. Also, not sure how it helps to classify observations as "unknown;" why not just exclude them?

4 Good info, but don't use actual variable names or columns. The reader will not see the dataset itself. Simply state that crimes were categorized into serious and non-serious, keep the examples of serious crimes like you already have, and give N (%) for each crime type. Including a bar chart is also not necessary as it doesn't provide additional information to N (%)

5 no need to justify each variable. Simply state that you used a priori selection based on domain knowledge and list the variables that are included in the model.

6 In the methods section, you should only state that multicollinearity is assessed with vif. The actual values and any modeling action taken in response to the values should be in the results section. Also, the sentence structure is a little confusing. Maybe something like this would be better: "VIF values for all variables except victim sex and victim descent did not exhibit evidence of multicollinearity." Is descent the same or similar to race/ethnicity? If so, I'm surprised it is collinear with sex; I'd keep both in the model since they still describe different demographic characteristics.

7 round - no need for that many decimals

8 round

9 "if the value approaches 1" not sure this makes sense. sounds like the 0.7 will approach 1 in a certain context, which is not actually what's happening

10 results

11 not sure what I'm supposed to get from this

12 focus on interpreting these in context. which is more important here, sensitivity or specificity? what

does each one mean in context?

13 what is H?

14 missing interaction term?

15 do you have enough observations in each category to support so many? Your sample size is very large so maybe so

16 sounds good! If you run out of room, you can move the diagnostic plots/figures to an appendix

17 add some interpretation. don't need to do all variables but pick a couple that are interesting/relevant

18 be more specific about key takeaways. relate back to the background/motivation

19 is this the abstract? should have one abstract for the whole paper

20 this is good but should be in the results section

21 this is good but can be simplified/reduced, particularly if you run out of space

22 this should all be in the results section

23 good interpretation in context. can be in results or conclusion

24 Good start. It seems like the pieces are there for the first question, but it needs to be formatted into paragraphs to fit the report structure. Too many subheadings and bullets that take up space. The second question has appropriate formatting/structure but the content is a work in progress. Overall, a big area of improvement should be the separation of the methods and results sections. The methods section should just describe your process but not contain anything that you actually found from the data.