BDA Project Report

December 1, 2024

1 Introduction

1.1 Motivation The motivation behind this project is to apply Bayesian modeling techniques to a real-world dataset in order to understand the underlying structure and make predictions. In this report, we focus on crime data and explore the ways in which we can make use of Bayesian methods to do crime prediction. Additionally, the results can be useful for individuals on knowing perhaps under what criteria more violent crimes may happen more often.

1.2 The Problem The primary goal of this analysis is to analyze crime rates in different areas of Los Angeles, US, based on historical data. The dataset includes various features such as crime types, times, locations, and victim demographics, which can be used to build the models.

1.3 Main Modeling Idea In this project, we utilize a Bayesian non-hierarchical model and a hierarchical model using logistic regression to account for the inherent uncertainty in crime data. The idea is to estimate probability of a crime occurrence having a weapon involved in it based on various predictors such as area codes, time of day, and victim characteristics.

In the case of our model, a weapon is classified as some kind of bodily force (arms, fist, etc), a rock/thrown object, a knife, some kind of object, a pistol, gun, machete, and so on. The most common weapon to be used in a crime occurrence was some kind of strong-arm (which is defined to be hands, fist, feet, or bodily force). It is important to note that the police data sometimes classified serious verbal threats as a form of weapon.

The data set contained all weapons used in a crime as a description of the weapon and a weapon code number.

2 Data and Analysis Problem

2.1 Data Source The dataset used in this analysis was obtained from the public government records on crime statistics. We obtained the data from Kaggle (link). It contains data from 2020 to present.

2.2 Data Description Dataset is composed of 1000 crimes committed in the LA area. There are many features such as the time of occurence, data, what kind of crime, where, on what road, victim details, etc.

2.3 Previous Analyses We did not actively search for any previous analyses done on this data set. However the Kaggle page with this dataset has some machine learning projects that other users have done. Upon initial inspection none of the projects are doing a Bayesian Analysis.

2.4 Illustrative Figure of Data This subsection will illustrate the data points on a map of the Los Angeles area. This was done in R using the leaflet package (link) and the htmlwigets package in r (link).

Figure 1 shows the dispersion of the crime data in our data set. It is clear to see that a lot of the data is centered around one area with a bit spreading out above, to the left, and down. The center condensed area is the Downtown Area. A closer look is provided in Figure 2.

This zoomed in map shows the data points on the different streets in the Downtown area of LA. Each data point can be clicked and it displays the exact description of the crime. The final map, Figure 3, shows the differentiation in crime occurrences that involve a weapon and those that do not involve a weapon. A red data point indicates that the crime that occurred at the point was a crime involving a weapon. On the other hand, data points that are yellow did not involve a weapon. Orange points may be seen and those are due to the overlapping nature of yellow data points.

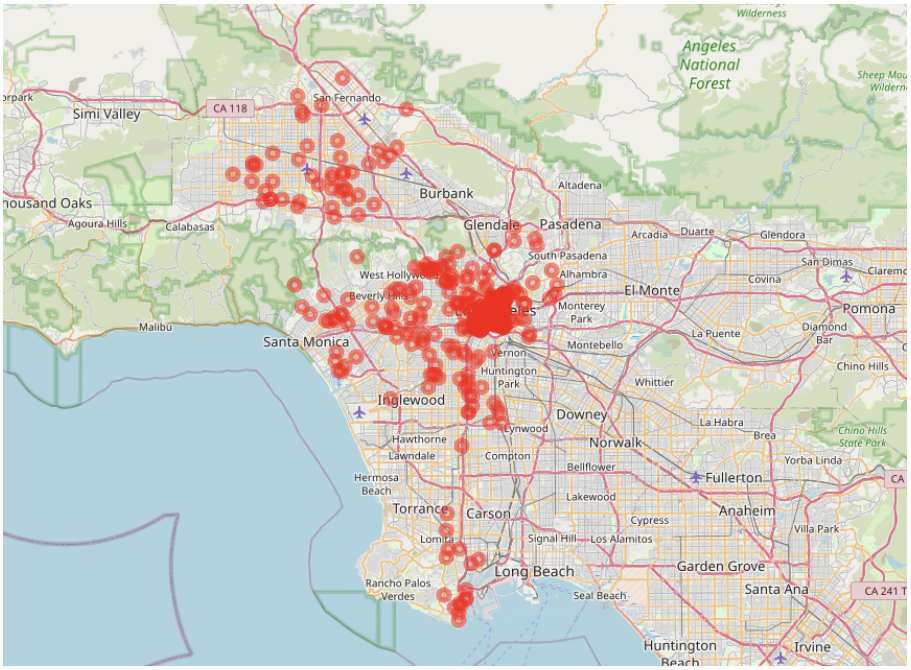


Figure 1: Map of Los Angeles Area with Crime Scenes Plotted

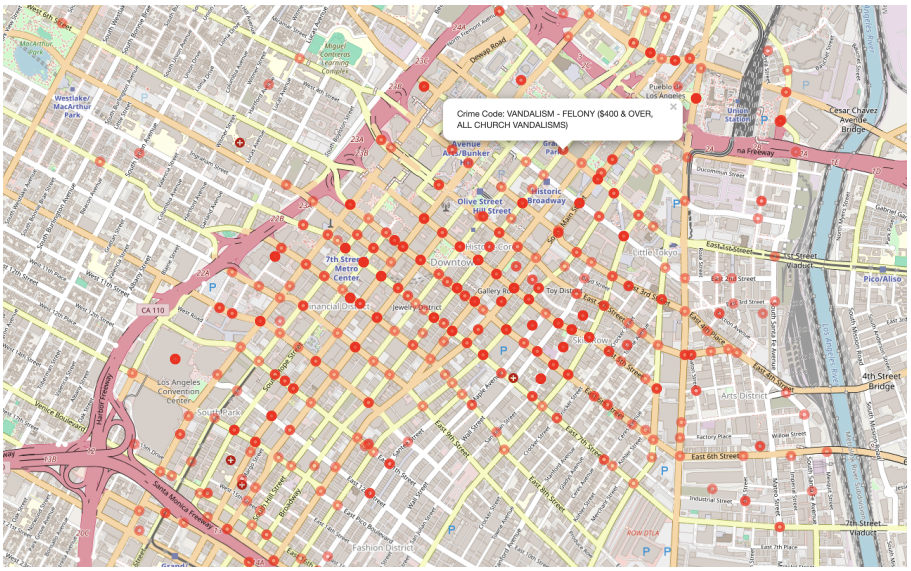


Figure 2: Zoomed in Picture of Downtown LA

3 Model Description

3.1 Model 1: Non-Hierarchical Bayesian Model The first model we apply is a basic logistic regression model. The data is preprocessed to create categorical variables such as time of day (night, afternoon, morning, or evening), day type (weekend vs weekday), and then area type (high crime area, low crime area, or other). Each of

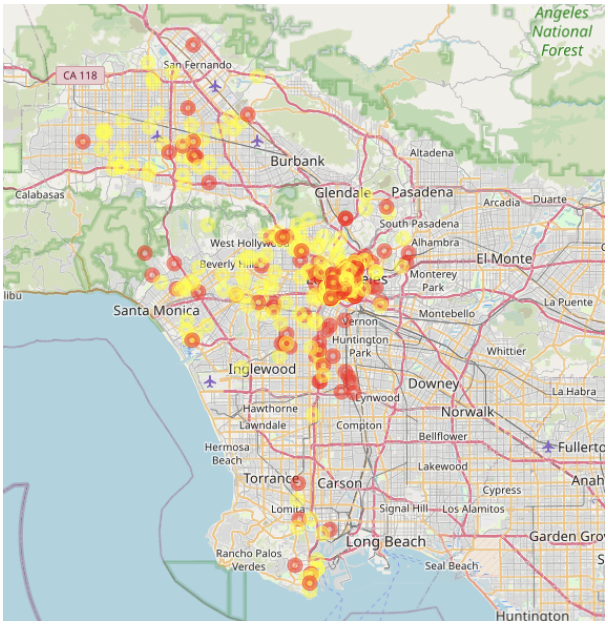


Figure 3: High Violence Crimes

these categorical features were created from the existing data. Date, area name, and time of crime occurrence were preexisting features in the dataset.

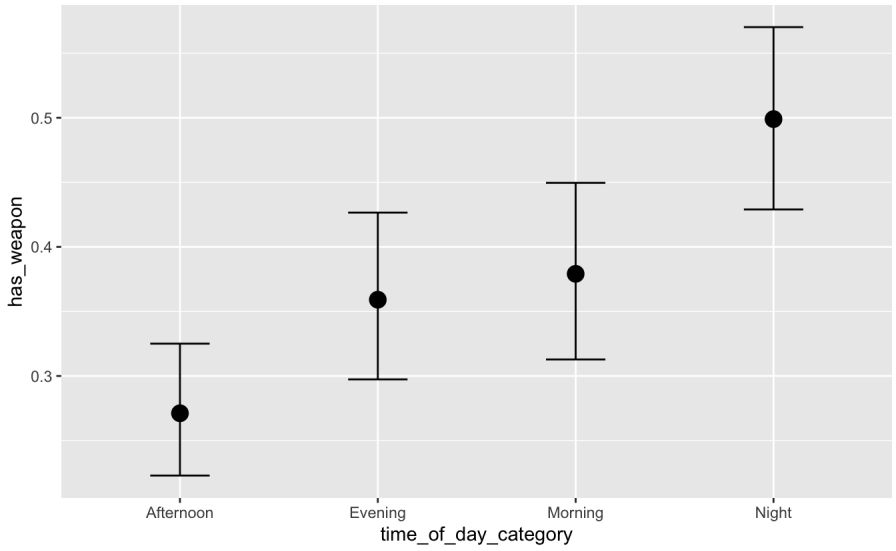


Figure 4: weapon is involved in crime by time of day

Figure (4) shows a conditional effects plot of the probability of a crime having a weapon given that the time of day is either afternoon, evening, morning, or night. It is clear to see that the probability of a crime having a weapon is much higher in the nighttime.

3.2 Model 2: Hierarchical Bayesian Model

While analyzing crime related statistics, grouping based on demographic factors such as age, ethnicity, and gender can unveil insightful patterns and differences in crime involvement and victimization.
Thus, we decided to create a model that extends a hierarchical Bayesian framework by incorporating area-based grouping, in addition to the gender, descent, and age group hierarchy. The formula predicts whether an individual is carrying a weapon (`has_weapon`, a binary outcome) based on variables like victim's age (`Vict.Age`), day type (`day_type`), time of day (`time_of_day_category`), location type (`Premis_Group`), sex (`Vict.Sex`), and descent (`Vict.Descent`). The random effects part of the model introduces two levels of grouping: random intercepts for combinations of sex, descent, and age group (`Vict.Sex:Vict.Descent:Age_Group`) as well as random intercepts based on geographical area (`AREA_NAME`). This allows for capturing the variability in weapon possession by both individual characteristics and spatial factors. The model uses a Bernoulli distribution with a logit link function to estimate the likelihood of weapon possession.

The Table obtained for this model and corresponding conclusions are at 2.

3.3 Models 3-5: Additional Models

In addition to the above models using categorical data, we created multiple models to test out the relationship between the time of day (as a continuous numerical variable) and the probability of a crime having a weapon. This was done by taking intervals of one hour as buckets and counting the number of occurrences in that interval where a crime had a weapon associated with it. Taking that number over the total amount of crimes in the time interval yields the probability of a crime having a weapon in 24 1-hour intervals.
To start, we decided to test a logistic regression model with the goal of predicting the fraction of weapon usage as a relationship of time, the logistic regression is shown in the top right of Figure 5. Clearly the logistic regression model does not fit the data well enough. The bottom left figure in Figure 5 shows a polynomial model fitted to the data using a quadratic formula. It is clear to see that it fits the data somewhat better than the logistic regression model but not as well as we would ideally want. To make the model even better we have a spline model shown on the bottom right of Figure 5. The spline model, it is clear to see, fits the data pretty well. It seems like there are some data points

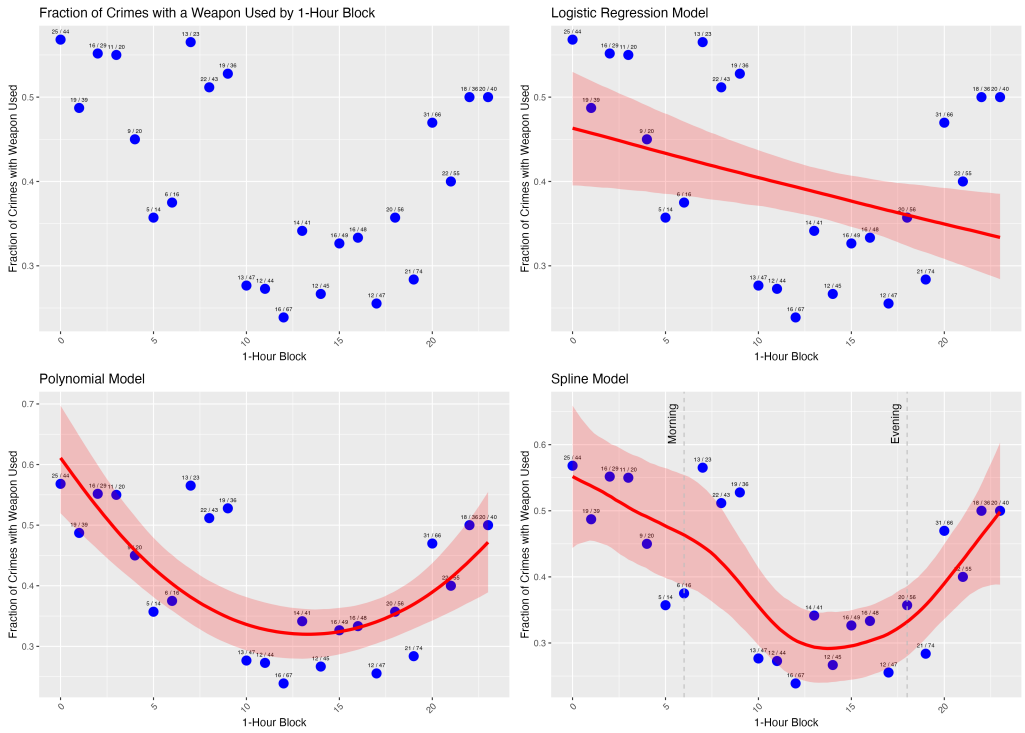


Figure 5: Base Data and Three Models

from 6am to 9am that are not fitted by the model very well.

The priors used for these models were all weakly informative models. For the spline model we used a normal distribution normal(0, 2) for the intercept class. Below we provide the brms code for the polynomial and spline models.

```
--- Polynomial Model -----
fit_poly <- brm(
  formula = weapon_used | trials(total) ~ poly(hour, 2),
  data = merged_data,
  family = binomial(),
  prior = c(
    prior(normal(0, 2), class = "Intercept"),
    prior(normal(0, 1), class = "b", coef = "polyhour21"),
    prior(normal(0, 1), class = "b", coef = "polyhour22")
  ),
  chains = 6,
  iter = 2000,
  warmup = 1000
```

```
)


--- Spline Model ----
fit_spline <- brm(
  formula = weapon_used | trials(total) ~ s(hour),
  data = merged_data,
  family = binomial(),
  prior = c(
    prior(normal(0, 2), class = "Intercept")
  ),
  chains = 4,
  iter = 2000,
  warmup = 1000
)
```

4 Priors and Justification

Weakly-Informative priors were selected since no strong assumptions could be made about the priors based on the data source. For the hierarchical model, the priors for the model parameters include a normal prior for fixed effects with wider variance, a student-t prior for the standard deviation of random effects, and appropriate priors for each coefficient.

5 Code and Inference Setup

5.1 Code Example Below is our code used to fit the non-hierarchical Bayesian model using the ‘brms‘ package in R:

```
model <- brm(
  formula = has_weapon ~ Vict.Age + day_type + time_of_day_category + Premis_Group +
  family = bernoulli(link = "logit"),
  data = crime_data,
  chains = 4,
  iter = 2000,
```

```
  warmup = 1000,
  thin = 1,
  prior = c(prior(normal(0, 5), class = "Intercept"),
            prior(normal(0, 2), class = "b", coef = "Vict.Age"),
            prior(normal(0, 1), class = "b"))
)
```

This code snippet shows the creation of the model. The formula for our predictive has_weapon contains the age of the victim, the type of day, and the time of day. These three features were found to have significant affects on the probability of a weapon being involved in a crime occurrence.
Below is our code used to fit the hierarchical Bayesian model using the 'brms' package in R:

```
model_hierarchical <- brm(
  formula = has_weapon ~ Vict.Age + day_type + time_of_day_category + Premis_Group +
            (1 | Vict.Sex:Vict.Descent:Age_Group:Premis_Group),
  family = bernoulli(link = "logit"),
  data = crime_data,
  chains = 4,
  iter = 2000,
  warmup = 1000,
  thin = 1,
  prior = c(
    prior(normal(0, 5), class = "Intercept"),
    prior(normal(0, 2), class = "b", coef = "Vict.Age"),
    prior(normal(0, 1), class = "b"),
    prior(student_t(3, 0, 2.5), class = "sd")
  )
```

5.2 Inference Setup and MCMC Options We used four chains with 2000 iterations each, a warm-up period of 1000 iterations, and weakly informative priors for the model parameters. The reason for choosing these options was to ensure that the

model had enough time to converge to the true posterior distribution, while also allowing for sufficient exploration of the parameter space.

6 Convergence Diagnostics

This section will cover the convergence diagnostics of both models, starting with the non-hierarchical model.

6.1 Diagnostic Results Results for the non-hierarchical model are in the following table 1. As we can see for the Rhat values, all variables converged correctly. The estimate for the intercept was -1.9 indicating that the baseline log odds of a crime having a weapon is fairly low if all other variables are at their reference levels. The victim age is somewhat significant because it has a positive estimate of 0.02 and a lower 95% CI of 0.01 and a upper 95% CI of 0.03. This means that there is a slight two percent increase in the chance of a crime containing a weapon for every increment of age. The type of day (e.g. weekend or weekday) does not appear to have any significant meaning as the lower 95% CI of -0.1 and the upper 95% CI of 0.48 contains 0. The time of day appears to be significant and will be elaborated on later. The location of the crime is fairly un-predictive of the crime having a weapon.

The Bulk ESS values for each variable is fairly big, more than 1000, indicating that there was a lot of independent draws that contributes to the bulk of the posterior. Similarly for the Tail ESS all the values were fairly big meaning that the draws contributed well to the tail of the distribution.

Results for the non-hierarchical model are in the table 1. As we can see for the Rhat values, all variables converged correctly.

Table 1: Regression Coefficients for Logistic Regression Model (Bernoulli Family, Logit Link)

Predictor	Estimate	Est. Error	l-95% CI	u-95% CI	Rhat	Bulk ESS	Tail ESS
Intercept	-1.90	0.25	-2.39	-1.42	1.00	3537	3372
Vict.Age	0.02	0.00	0.01	0.03	1.00	8864	2821
day_typeWeekend	0.20	0.15	-0.10	0.48	1.00	6198	3386
time_of_day_categoryEvening	0.40	0.19	0.02	0.76	1.00	3014	3205
time_of_day_categoryMorning	0.48	0.19	0.11	0.87	1.00	3176	2831
time_of_day_categoryNight	0.98	0.19	0.62	1.36	1.00	3236	2635
Premis_GroupGovernment	-0.08	0.42	-0.93	0.74	1.00	3731	2748
Premis_GroupHealthcare	0.42	0.77	-1.15	1.88	1.00	5012	2805
Premis_GroupOther	0.09	0.53	-0.95	1.10	1.00	4687	3263
Premis_GroupPublicSpaces	0.29	0.21	-0.12	0.72	1.00	2634	3221
Premis_GroupResidential	0.34	0.25	-0.14	0.82	1.00	3118	3316

Table 2: Regression Coefficients for Heirarchical model)

Predictor	Estimate	Est. Error	l-95% CI	u-95% CI	Rhat	Bulk ESS	Tail ESS
Intercept	-3.31	0.64	-4.57	-2.06	1.00	3940	3222
Vict.Age	0.01	0.01	-0.00	0.02	1.00	5722	3483
day_typeWeekend	0.14	0.16	-0.17	0.45	1.00	9201	2945
time_of_day_categoryEvening	0.40	0.20	0.00	0.79	1.00	7390	3319
time_of_day_categoryMorning	0.42	0.21	0.01	0.83	1.00	6299	3647
time_of_day_categoryNight	0.91	0.20	0.52	1.29	1.00	6224	3078
Premis_GroupGovernment	0.09	0.45	-0.81	0.99	1.00	6688	3065
Premis_GroupHealthcare	0.39	0.64	-0.84	1.67	1.00	7957	3348
Premis_GroupOther	-0.04	0.51	-1.04	0.97	1.00	6847	3185
Premis_GroupPublicSpaces	0.49	0.27	-0.04	1.02	1.00	4416	3224
Premis_GroupResidential	0.38	0.30	-0.19	0.98	1.00	5622	3800
Vict.SexF	1.40	0.55	0.32	2.47	1.00	4028	2938
Vict.SexM	0.68	0.56	-0.41	1.76	1.00	3935	2865
Vict.SexX	0.50	0.75	-0.99	1.96	1.00	5817	3144
Vict.DescentA	0.03	0.51	-1.00	0.98	1.00	3573	3454
Vict.DescentB	0.88	0.43	0.04	1.72	1.00	3237	3045
Vict.DescentC	-0.11	0.83	-1.78	1.53	1.00	8912	2617
Vict.DescentF	-0.74	0.81	-2.37	0.80	1.00	7695	3023
Vict.DescentH	1.28	0.42	0.45	2.12	1.00	3237	2929
Vict.DescentI	-0.22	0.96	-2.13	1.67	1.00	9671	2481
Vict.DescentK	-0.09	0.82	-1.73	1.48	1.00	8637	3008
Vict.DescentO	0.42	0.46	-0.47	1.33	1.00	3377	3046
Vict.DescentW	0.41	0.44	-0.43	1.28	1.00	3058	3105
Vict.DescentX	0.60	0.72	-0.81	1.99	1.00	7055	3378

```
Regression Coefficients:
           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept     -0.42      0.07    -0.55    -0.29 1.00     5312     4202
polyhour21    -0.75      0.32    -1.35    -0.12 1.00     5647     4292
polyhour22     1.30      0.30     0.71     1.88 1.00     6047     4663
```

Figure 6: Diagnostics and Convergence for Polynomial Fit

```
Smoothing Spline Hyperparameters:
              Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sds(shour_1)      1.92      0.93     0.66     4.26 1.00     1204     1528

Regression Coefficients:
          Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept    -0.39      0.07    -0.53    -0.25 1.00     4129     3051
shour_1       1.76      3.07    -4.19     8.17 1.00     1532     1313
```

Figure 7: Diagnostics and Convergence for Spline Fit

Figure 6 and Figure 7 show the convergence diagnostics for both the polynomial and spline models. Both models display a Rhat value of 1 which means that they converged well. The Bulk ESS for both models was all above 1000 indicating that they contributed well to the bulk of the distribution. Similarly all Tail ESS were also above 1000 meaning they contributed effectively to the tail ends of the distribution. At first the convergence was not that great for the polynomial model so the number of chains was increased from 4 to 6. For the other models the number of chains remained at 4.

Results for the regression coefficients are shown in Table 2. All variables have Rhat values of 1.00, indicating correct convergence. All variables have large Bulk ESS and Tail ESS values, suggesting that the model's sampling was effective with reliable estimates. The following offers more insight:

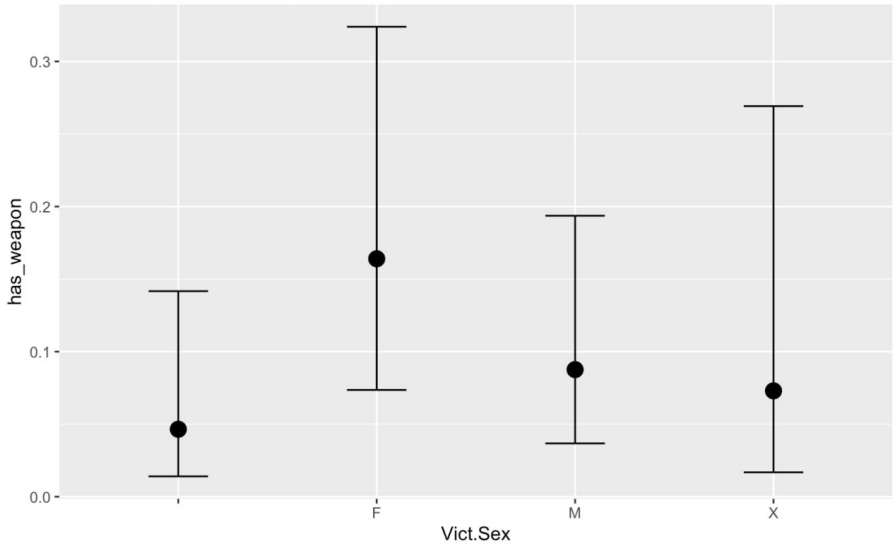


Figure 8: probability of a weapon used vs victim gender

Victim Sex The coefficients for *Vict.Sex* shows significant differences in probability of being a victim of a crime type involving a weapon

- **Female (F)**: 1.40 (95% CI: $[0.32, 2.47]$) shows a strong positive association, suggesting females are more likely to be victims of crime involving weapons.
- **Male (M)**:Similarly from Table 2, estimate 0.68 (95% CI: $[-0.41, 1.76]$) suggests a weaker association
- **Unknown (X)**: 0.50 (95% CI: $[-0.99, 1.96]$) shows high uncertainty.

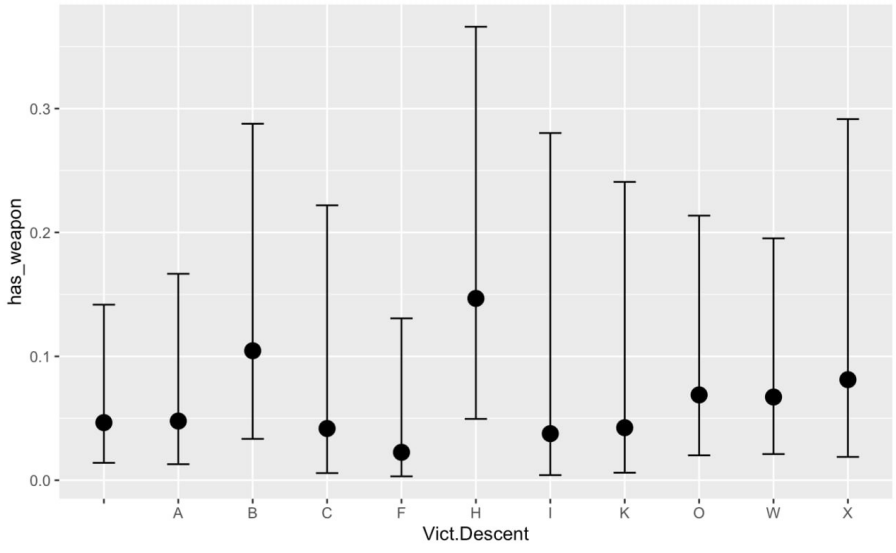


Figure 9: probability of a weapon used vs victim descent

Victim Descent The coefficients for *Vict.Descent* across demographic groups:

- "H" (Hispanic) shows the highest likelihood of weapon use (20–30%), with moderate uncertainty.
- "A", "B", and "W" fall in the mid-range (10–20%), with overlapping intervals showing no significant differences.

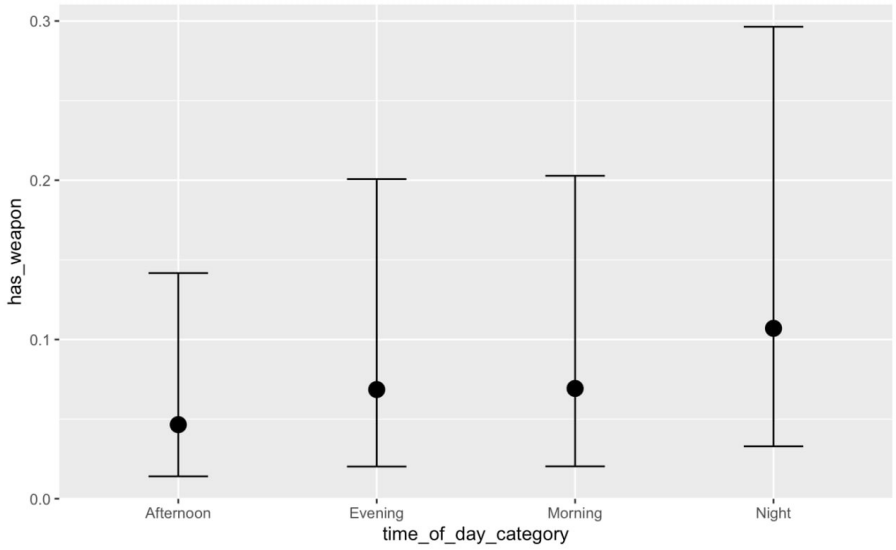


Figure 10: probability of a weapon used vs time of the day

Time of Day The coefficients for *time_of_day_category* shows significant differences for particular time of the day groups:

- **Evening**: 0.40 (95% CI: $[0.00, 0.79]$) shows a weak positive effect.
- **Morning**: 0.42 (95% CI: $[0.01, 0.83]$) also shows a weak positive effect.
- **Night**: 0.91 (95% CI: $[0.52, 1.29]$) indicates a strong positive association, suggesting crimes are significantly more likely to occur at night.

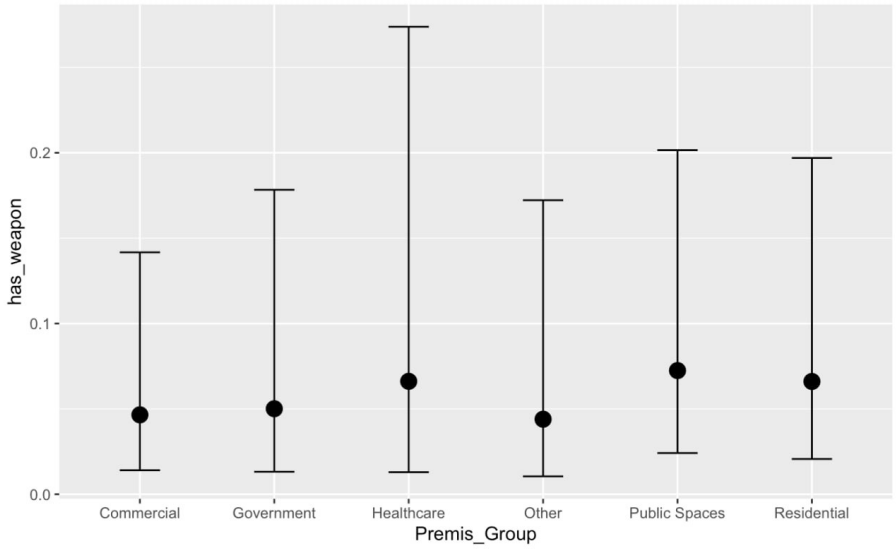


Figure 11: probability of a weapon used vs premises

Premises Group The coefficients for *Premis_Group* vary:

- **Government Premises**: Shows a lower association compared to the other groups. This shows that crimes may be more unlikely to take place at government premises.
- **Healthcare Premises**:Shows wide uncertainty.
- **Public Spaces**: Coefficients suggests a weak but positive effect.
- **Residential Premises**:Coefficients suggests a weak association.

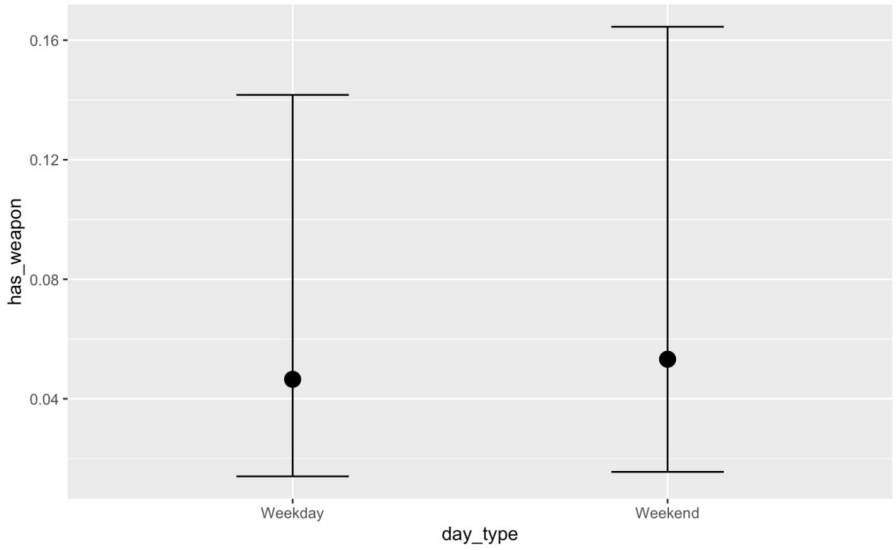


Figure 12: probability of a weapon used vs type of day

Day Type (Weekend) coefficients for *day_typeWeekend* indicates a weak and most likely insignificant effect of crimes being more likely o occur on weekends compared to weekdays.

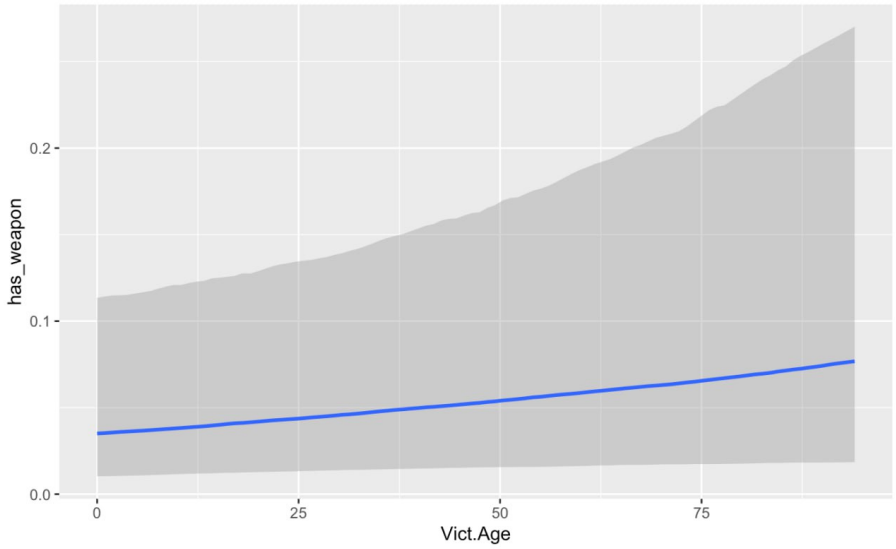


Figure 13: probability of a weapon used vs age of victim

Victim Age Analysis of victim age suggests a weak positive correlation of victim age with being a weapon involved in the crime against the victim. This correlation is more apparent if drawing analysis based on the upper bound.

6.2 Improvement Actions

Some suggestions which may lead to better model fit performance

Non-Heirarchical Model:

- Feature engineering: Add more predictors and interaction terms. Eg, time of day and area type

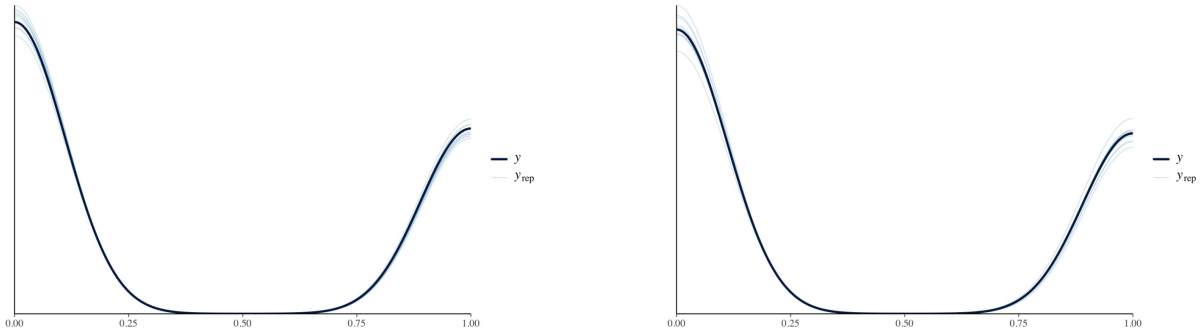
Heirarchical model:

- Add street-level or specific crime types as hierarchical levels.

Spline Model:

- Refinements: Use higher-order splines for more flexibility for the fit to the data.

7 Posterior Predictive Checks



(a) PP check fit for Non-Hierarchical Model (b) PP check fit for Hierarchical Model

Figure 14: Comparison of PP check fits for Non-Hierarchical and Hierarchical Models

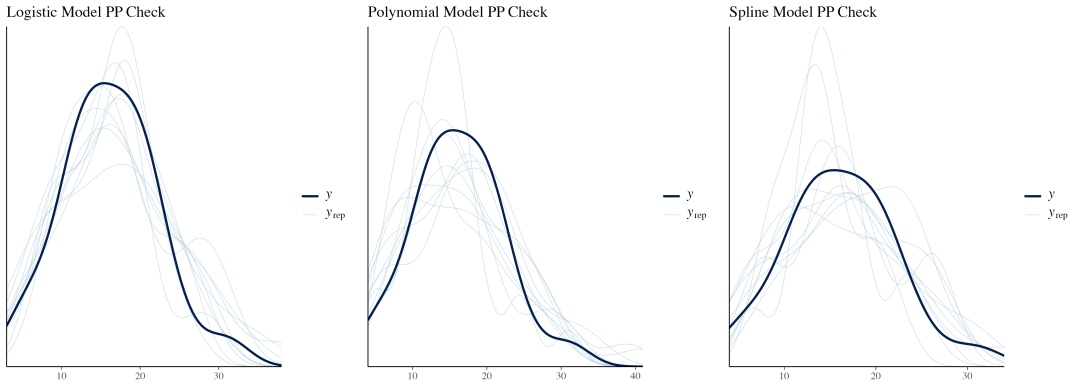


Figure 15: PP_Check Plot for Three Additional Models

Figure 15 shows the posterior predictive checks for all three additional models. It appears that all models are able to fairly well model the observed data except that there are some discrepancies in the polynomial and spline models. Overall the logistic model seems to do the best.

7.1 Results of Checks The posterior predictive checks show that both models fairly well simulate the observed data. The y_rep lines are very closely similar to the thick darker line. This shows that both models are fairly well simulating the observed date. However, the second model, the hierarchical model, has a bit more variation in the tails of the pp check plots.

8 Predictive Performance Assessment

Actual / Predicted	**0**	**1**
0	498	225
1	113	163

Table 3: Confusion Matrix for Non-Hierarchical Model

Accuracy for Non-Hierarchical Model: 0.6617

Actual / Predicted	**0**	**1**
0	508	195
1	103	193

Table 4: Confusion Matrix for Hierarchical Model

Accuracy for Hierarchical Model: 0.7017

Comparing the accuracies of the two models, it can be inferred that the Hierarchical model is slightly better at fitting to the data than the non hierarchical model and therefore its performance is slightly better.

Spline confusion matrix has limited entries as our data is processed into 24 one-hour intervals. The classification accuracy is calculated to be roughly 83.3%.

9 Model Comparison

Non-Hierarchical Model:

- Time of day shows strong positive association of crimes involving a weapon at nighttime.
- Victim age: Slight positive associations (estimate= 0.02)
- Generally weak effects; public spaces had a weak positive effect
- Day type is generally insignificant and no concrete associations can be made.
- All Rhat value $\simeq 1$ showing good convergence.

- Bulk and Tail ESS values are large, confirming sufficient sampling.
- Limited flexibility, cannot effectively account for variability between groups or areas or other factors.
- Posterior predictive checks indicate a good fit for the overall dataset.

Hierarchical Model:

- Victim Demographics: Significant effects for gender. Women are more likely to be involved in a weapon-related crime, estimate = 1.40.
- Victim Age: Slight positive effect.
- Premises: Generally weak effects; public spaces had a weak positive effect.
- Day Type: Insignificant, with wide confidence intervals (estimate = 0.20)
- Time of Day: Similar patterns as the non-hierarchical model. Nighttime remains highly significant.
- Similar Rhat $\simeq 1$ values and large ESS values, demonstrating convergence and reliable estimates.
- Better at capturing dependencies based on groupings according to age, sex, gender, Descent and other factors which cannot be accurately seperated in the Non-Heirarchical model.
- Highly flexible, allowing it to capture variability within and between different factor groups.
- Posterior predictive checks indicate a similar good fit for the overall dataset. This is also supported by Table 5

Spline model:

- Models non-linear relationships between time and the probability of crime involving a weapon better than other models.
- Slight underfitting for specific time ranges (e.g., 6 AM to 9 AM)

- Lower LOO (Leave-One-Out) error compared to logistic regression models, indicating better predictive performance.
- High flexibility in fitting non-linear patterns.

9.1 LOO-CV Model Comparison

The second `elpd_loo` estimate of -599.4 is 7.1 units higher than the first estimate of -606.5, but given the standard errors of 12.4 and 11.8, respectively, the difference is relatively small. This indicates that neither is better or worse compared to the other.

Table 5: Comparison of elpd_loo values for the Non-Hierarchical Logistic and Hierarchical Models

Model Type	**elpd_loo Estimate**	**Standard Error (SE)**
Non-Hierarchical Logistic Model	-606.5	11.8
Hierarchical Model	-599.4	12.4

Table 6: Comparison of elpd_loo values for the Logistic and Spline Models

Model Type	**elpd_loo Estimate**	**Standard Error (SE)**
Logistic Model	-74.1	5.3
Spline Model	-61.6	2.6

Table 6 shows the loo estimates for the logistic and spline model. It appears that the elpd_loo estimate for the spline model is lower and less negative which indicates that it is a bit better of a model.

10 Sensitivity Analysis

After we refit out models with informative priors, the estimate and the standard error of elpd loo for non hierarchical model becomes -607.3 and 12.8, while for the hierarchical model the corresponding values are -608 and 14.9. Thus the elpd loo values almost unchanged, signaling that the models are quite robust to the specifications of the prior. A sensitivity analysis was carried out for the three additional models (log, poly, and spline) as well, the priors were each changed somewhat to test for any changes in the

models. The results of this testing was that there were no significant changes and the plots remained very similar to Figure 5.

11 Conclusion

This report explores a non hierarchical logistic regression model, a hierarchical model and three additional models for predicting the probability that a weapon is used during the occurrence of a crime in the city of Los Angeles in California, USA. All of the models seem to converge well based on the regression coefficients. However, visual inspection of the predictions suggest that the dataset exhibits a rather non-linear relation that the additional models are better suited to capture better. The non-hierarchical and hierarchical models focus on categorical attributes of the data such as type of day, descent of victim, etc, which makes the modeling fairly calculated. The logistic, polynomial, and spine model on the fraction of crimes with weapon used in 1-hour blocks is more telling and provides a simplier and more effective model to fit the data.

12 Self-Reflection

We learned that pre-processing and coming up with meaning full features to focus on was the most difficult part. The coding portion can be fairly simple but the challenging part is finding some kind of relationship in the data and then focusing on that. In our case we had loads of different features and it was important to focus on only a few.