Examining Marijuana Legalization and Its Impact on Fatal Car Accidents

By: Dennis Goldenberg, Alana Berson, Suleman Sadiq, and Solomon Gao

Introduction

Background Information

Recreational Marijuana use has been a contentious moral and legal issue in the United States. Under the Controlled Substances Act of 1970 (Drug Enforcement Administration, 2020), Marijuana is listed as a Schedule I drug, meaning its use is restricted with no federally accepted medical exceptions. However, legalization of Marijuana has steadily gained support from the public and Democratic and Republican politicians alike in the past 2 decades. Supporters argue that legalizing Marijuana would make it safer for users, it would free up law enforcement to focus on more dangerous crimes, and it provides a good source of tax revenue for state and local governments (McCarthy, 2019). As legal authority in the U.S is arranged in a two-tiered, federalist structure, individual states have begun to enact their own laws, legalizing recreational use of Marijuana, with 19 states having done so by 2022 (Hansen, Alan, & Davis, 2023).

There is opposition to legalization as well; some argue that it would harm society by encouraging more people to use Marijuana, lead to people using stronger and more addictive drugs, and it would increase the number of car accidents due to intoxicated drivers (McCarthy, 2019). The last of these reasons is of particular interest to actuaries, particularly those that work in pricing for auto insurers. Now that a handful of states have legalized marijuana and had it recreationally legal for up to a decade, the data exists to examine this claim.

Our Business Question

We looked to answer the following: ***Does Marijuana Legalization have a significant impact on the number of fatal car accidents in a state***? In order to answer this question, we examined the problem from 2 different lenses:

1. Is there a statistically significant immediate increase in the number of car accidents after a state legalizes marijauna?
2. Is the legalization status of marijuana in a given state an important determinant of the number of fatal car accidents in said state?

To answer these questions, we collected data from various sources, ran statistical tests and machine learning algorithms, and arrived at conclusions given the results.

Data Collection

Data Collected

Our models use fatal vehicle crashes as the response variable. The NHTSA (*National Highway Traffic Safety Administration, 2022*) contains data on the number of car accidents by state aggregated by a whole slew of factors Specifically, the Fatality Analysis Reporting System (FARS) contains data on fatal traffic crashes within the 50 states and the District of Columbia. To be included in FARS, a crash must result in the death of a vehicle occupant or nonoccupant within 30 days of the crash.

Our predictor variable of interest is marijuana legalization status. Each state has their own state Marijuana possession law for residents. On Election Day in 2012, Colorado and Washington became the first states in the U.S. to legalize the recreational use and sale of Marijuana. By the end of 2021, a total of 18 states plus the District of Columbia had legalized

recreational Marijuana (Hansen, Alan, & Davis, 2023). To aid in answering our first question, the legalization date (rounded to the nearest month) was used. For our second question, we encoded a simple binary variable with 0 representing the states that had not legalized by January 1st, 2022 and a 1 for those that did.

To evaluate the relative predictive power of marijuana legalization on our response, we sourced other potential predictors. From the Bureau of Transportation Statistics (2022), we obtained the total number of highway miles traveled in each state from the year 2022. To account for differences in population, we divided this value by the 2022 population estimate (in 100s of 1000s) for said state, provided by the Federal Reserve Bank of St. Louis (2023). From the Census Bureau (2023), we obtained the percent population of each state in urban areas to see whether urban traffic congestion caused a rise in accidents. To qualify as an urban area, the territory must encompass at least 2,500 people, at least 1,500 of which reside outside institutional group quarters. From the Insurance Institute of Highway Safety (2023, May) we obtained the speed limits on urban highways for each state to account for potential differences in driver recklessness.

To finalize our data set, we included 3 more variables: 70+ percentage, Adult binge drinking percentage, and total damage from hazardous weather events, such as hurricanes or floods. Binge drinking statistics were sourced from the CDC (Statista, 2022); this variable measures the percentage of adults in each state who answered in a survey that they had 4 or more (women) or 5 or more (men) alcoholic beverages on a single occasion in the past 30 days. Damage statistics from hazardous weather events was obtained from the National Weather Service (2022); both drunk driving and poor weather conditions have a clear connection to potential fatalities on the road. However, the percentage of a population 70 or older also has a

connection. According to Morris Bart LLC. (*What age group*, 2023), “when drivers age into their 70s, health conditions can begin to interfere with their driving abilities”.

Exploratory Data Analysis - Seasonality

To examine and demonstrate the seasonality of our data, we conducted several analyses and created various plots to illustrate the recurring patterns. In Figure 1, a graph depicting Monthly Fatal Vehicle Crashes against Month and Year from 2008 to 2022, a clear cyclical pattern is evident, wherein the number of vehicle crashes annually fluctuates within specific periods. Notably, after the year 2020, there was a discernible increase in the general number of vehicle crashes compared to previous years. Further analysis was performed by examining the data for the year 2013 specifically. Figure 2 shows that the number of vehicle crashes reaches a nadir in February before escalating to a peak in August, followed by a subsequent decline.

Additionally, we utilized a decomposition of the additive time series to better elucidate the seasonality in our data (Figure 3). This decomposition yielded four distinct plots: observed, trend, seasonal, and random components. The observed plot replicates the general trend of vehicle crashes over time, as displayed in the initial graph. The trend plot reveals the long-term progression, smoothing out short-term fluctuations. Here, the data suggests a stable trend from 2010 to 2015, with a notable increase from 2015 to 2020, and a significant surge post-2020. The seasonal plot explicitly demonstrates the annual repetitive pattern observed in the data, corroborating the cyclic nature of vehicle crashes as previously discussed. The random plot, on the other hand, indicates unaccounted variations in the data after removing the trend and seasonal

components. This plot highlights a substantial negative residual shortly after 2020, likely due to reduced vehicular activity during COVID-19 lockdowns.

To delve deeper into the seasonality and enhance our understanding of the data, our group conducted an autoregression model analysis. Employing the first 80% of the data as a training set and the remaining 20% as a testing set, we initially generated a partial autocorrelation function (PACF) plot. The PACF is instrumental in time series analysis, especially for autoregressive processes. It quantifies the correlation between a time series and its lags, controlling for correlations at shorter lags. The PACF plot (Figure 4) indicated that only the first 12 lags significantly influence the current observation, prompting us to implement an AR(12) model.

An AR(12) model, an autoregressive model of order 12, predicts future values based on the past 12 months of data. Although the initial predictions approximated the overall trend, they did not align precisely with the actual data (Figure 5), necessitating further refinement. By adjusting our approach to use data from 2008 to 2021 (Figure 6) as the training set and only 2022 data as the testing set, the revised model demonstrated improved accuracy (Figure 7), with predictions closely matching the actual data for 2022.

These analyses and visual representations underscore the presence of significant seasonality in the data, characterized by consistent annual patterns in vehicle crashes. This finding emphasizes the importance of considering seasonal effects when analyzing such data to avoid potential biases introduced by averaging over time periods with inherent seasonal variability.

Statistical Methods Used

Paired Sample t-test

In order to examine the first question, we needed a way to examine the immediate effect of legalization. A way to examine this is via a paired sample difference of means test, or taking the number of fatal car crashes both before and after different states legalized, and seeing if there was a statistically significant difference, on average. The dates that were used as the inflection point of comparison were the dates that legalization **went into effect** in that state (more precisely, the date used was the 1st of the month that was closest to the true effective date, as the FARS data was on a monthly, not daily basis). These dates were obtained from several sources, such as the *Marijuana Policy Project*, *ballotpedia*, and different state-specific websites.

To generate said means, there are two complications that need to be addressed: seasonality and population changes. To address seasonality, the average number of fatal crashes of the 12 months preceding legalization and average number of fatal crashes post legalization for the pre-legalization and post-legalization means were used to encapsulate a whole season, as 1 of each month was included. To address population change, these means were divided by population estimates of the state divided by the Census bureau. Since some states' legalization went into effect in the middle of a year, the specific denominator was a linear interpolation between two years. For example, legalization in Oregon went into effect on July 1st, 2015; therefore, the means were calculated as follows:

$$\bar{x}_{\text{pre}} = \frac{\frac{1}{12}\sum_{i=-12}^{-1} \text{fatalities}_i}{\frac{1}{2}(\text{Oregon Pop., 2015}) + (1-\frac{1}{2})(\text{Oregon Pop.,2014})}$$

$$\bar{x}_{\text{post}} = \frac{\frac{1}{12}\sum_{i=0}^{11} \text{fatalities}_i}{\frac{1}{2}(\text{Oregon Pop., 2015}) + (1-\frac{1}{2})(\text{Oregon Pop.,2016})}$$

Here, i is the month index related to the legalization month, and ½ is the proportion of the year that passed before legalization. Since legalization occurred directly halfway into the year, the population estimate for the state pre-legalization was half of the 2015 estimate added to (1 - ½), or ½ of the 2014 estimate. A similar argument followed for post legalization.

Since the data from FARS only went to the end of 2022, and 12 months were needed to calculate the means, we only used the states that legalized pre-2022 (so $n = 19$). A t-test of this nature typically needs roughly 20 points, so this is enough. Next, the distribution of this difference of means needed to be examined for normality. As evidenced by figure 8, the difference of means was roughly normally distributed. All assumptions for the paired sample t-test were met.

Decision Trees/Random Forests

To examine the second question, we modified the response variable to eliminate the time element. First, the most recent data for fatal car crashes was used for each state - that is, the average monthly total in the year 2022. As with the t-test, this monthly average over a year eliminated the worry of seasonality, as one of each month was included. Second, these averages were divided by the 2022 population estimates, but no linear interpolation was necessary, as each state was evaluated at the same time, and that time was the whole of 2022, beginning to end. Additionally, marijuana legalization was converted to a response variable, with 1 representing a state that legalized by January 1st, 2022 and a 0 for states that did not. It is noteworthy that, as Rhode Island legalized in the middle of 2022, it was excluded from analysis, as it would obfuscate the interpretation of our results. Thus (with the addition of Washington D.C), we had $n = 50$ data points.

Decision Trees and Forests were used due to the easy interpretability of relative feature importance in predicting the response variable. First, the data was split randomly 80%-20% into training and testing sets, respectively. A tree was trained on the training set, and then pruned using 5-fold cross validation. The deviance of trees with different numbers of leaves is shown in figure 9. The tree with 4 leaves minimized the deviance, so it was selected; that particular tree's splits and predictions are highlighted in figure 10.

However, individual trees suffer from extraordinarily high variance. For this reason, we implemented a bagged (bootstrap aggregated) tree model, where we repeatedly sampled with replacement from our data and fit trees to each resampling, with predictions being averages on each tree. This solved two problems: first, it reduced variance due to the law of large numbers, and second, it accounted for our relatively small dataset. However, as these trees are all sampling from the same dataset, their results tend to be highly correlated; thus, to determine the optimal number of trees for minimization of M.S.E, we needed to balance between falling variance and rising covariance.

The way we did this was running 5 different simulations for up to 200 trees, and then calculating the Out of Bag (OOB) M.S.E, or average M.S.E per tree. This OOB MSE results from the fact that, when you perform a sampling with replacement from a data set, roughly a 3rd of the data points are excluded from the resampling, and can be used as test points. The results of these simulations can be shown in figure 11. The plot highlights that average OOB MSE steadily fell until roughly 50 trees, at which point the addition of more trees was either insignificant or actually worsened predictions. Thus, $n = 50$ trees were chosen, and we ran the algorithm again with the hyperparameter set.

We analyzed the bagged trees "variable importance" attribute, whose output is shown in figure 12. The results were consistent with the individual tree, in that *highway miles traveled* was by far the most significant predictor, with *adult binge drinking percentage* second in terms of increasing node purity. By comparison, our *legalization* variable was negligible in the vast majority of trees produced.

Linear Regression

To examine specifically linear relationships between fatal car crashes and our predictors, linear models are useful. First, we had to determine the distribution of our response variable to determine the appropriate link function. Figure 13 shows a histogram of our response (the same response variable as in the above tree analysis) and its relative normality. To ensure that the distribution was normal, however, we compared the distribution of the response to normal and lognormal distributions with Maximum Likelihood Estimates as parameters (figure 14 and 15, respectively). As the quantiles form a relatively straight line throughout in the normal qq plot, there is significant evidence of the response being normally distributed; thus, we used the identity link and our model was multiple linear regression.

However, 7 features is a large amount for only 50 data points; thus, we performed feature selection using the best subset algorithm on the same training set that the tree was used to train on. Figure 16 illustrates the best model by R-squared for each number of predictors. The Adjusted R-squared, Mallow's Cp, AIC, and SBIC statistics were used to compare the models of different predictor amounts, and unanimously chose Model 2. This model had *highway miles traveled* and *adult binge drinking percentage* as its two predictors.

We ran the model on the training set, and the results are shown in figure 17. The *highway miles traveled* variable was very statistically significant, and more miles traveled per person was associated with a higher number of average fatal crashes, an intuitive result. Surprisingly, higher *binge drinking percentage* in a state was associated with lower numbers of average fatal crashes, though this was less statistically significant. We believe that there is a confounding variable; it is likely that the states with higher binge drinking rates have some other quality that makes driving safer, as increases in the proportion of the population that is drunk is not likely to reduce the danger on the road.

Results and Takeaways

Numerical Results

The results from the paired t-tests were that there was a sample mean difference of 0.60166, a test statistic of 2.41257, and a p-value of 0.01336. The value of 0.60166 can be described as the average month in the year following legalization having 0.60166 fatal car accidents per 100,000 people more than the average month in the year pre-legalization. The p-value of 0.01336 being less than 0.05 shows that this is significant and we reject the null hypothesis.

Comparing the decision tree, bagged forest, and multiple linear regression models, they had the following MSEs: 6.605976, 5.609170 & 3.652816 respectively. Note that the MSE on the individual tree and regression models was calculated on the same 10 test points, while the stand-in for the test MSE for the bagged forest was the OOB MSE. Furthermore, the proportion of variance explained (R2 / pseudo R2) was: 0.4375485, 0.5224193 & 0.6817754 respectively.

This led to the best model being multiple linear regression since it had the lowest MSE as well as the highest proportion of variance explained.

Conclusion and Takeaways

The main takeaway from this project was that there is evidence that the amount of monthly fatal vehicle accidents increased directly preceding legalization but marijuana legalization is not statistically significant when compared to the other predictors. After completing all of our tests and analysis the results did present that the two most important predictors were billions of highway miles traveled per 100,000 per state & the adult binge drinking percentage.

Furthermore, for future studies this project left us with three main ideas: firstly, it would be interesting to test whether the impact of legalization remains constant or varies over time. This would be interesting to see whether legalization has a short-term or long-term impact, as our t-test only examined the immediate impact of legalization. Secondly, different states have different legal possession limits; the differences in outcomes between states with these limits warrants further study as other states look for guidance on the specific amounts they want to allow in crafting legalization legislation. Finally, our tree and regression models seemed to suggest that the more an average person drove in a state, the more fatal car accidents there were. This suggests that there is an increasing hazard of accidents with each mile driven; therefore, the "time to 1st accident" distribution for individual drivers deserves analysis to see if the hazard rate is indeed increasing.

Appendix

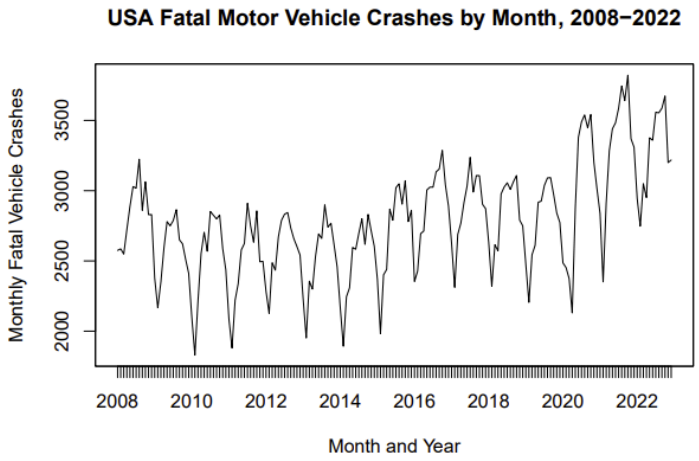

Figure 1. Vehicle Crashes vs. Time 2008-2022

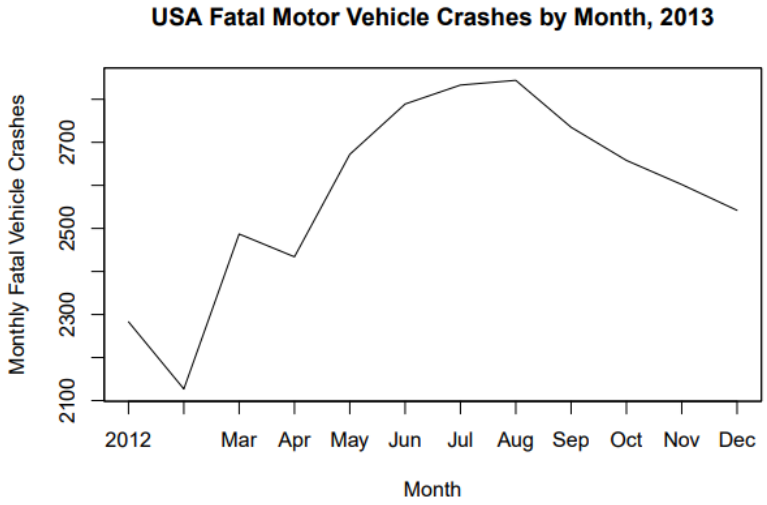

Figure 2. Vehicle Crashes vs. Time 2013

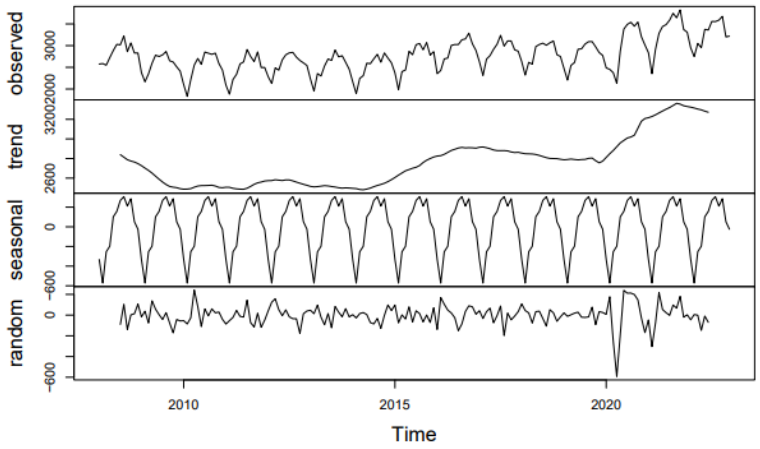

Figure 3. Decomposition of Additive Time Series

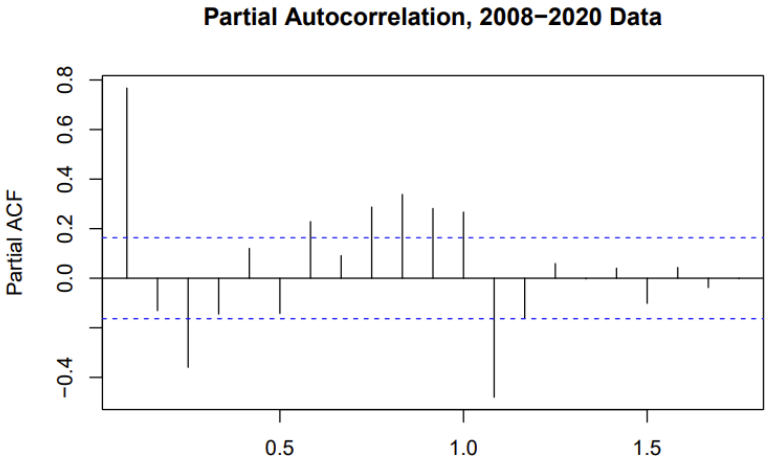

Figure 4. Partial Autocorrelation, 2008-2020 Data

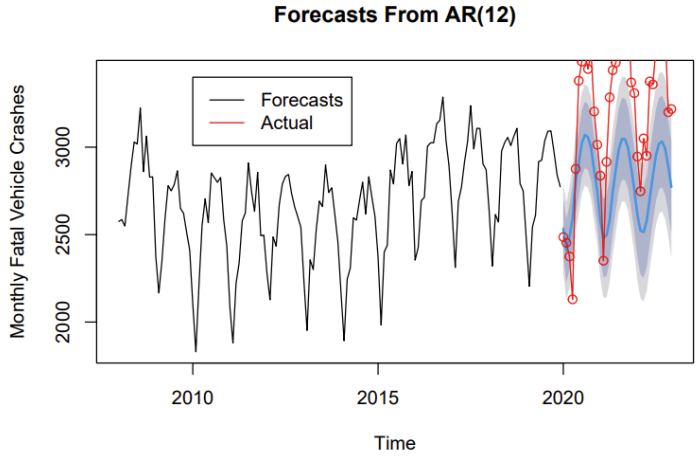

Figure 5. .Forecasts from AR(12), 2008-2020 Data

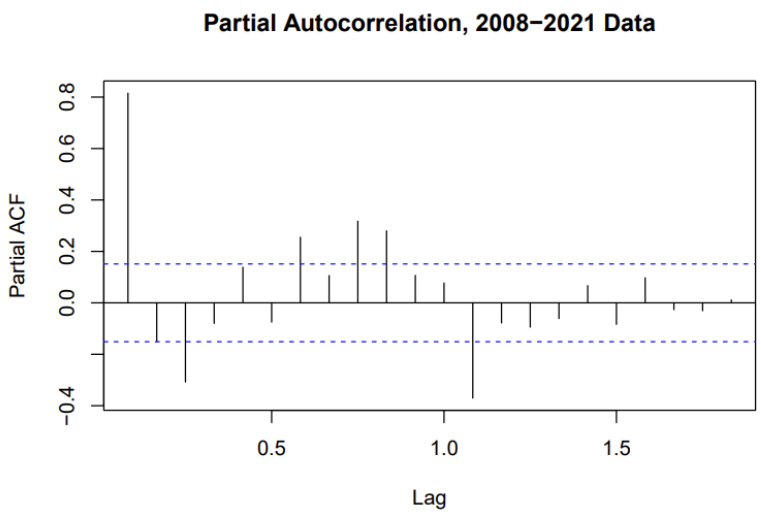

Figure 6. Partial Autocorrelation, 2008-2021 Data

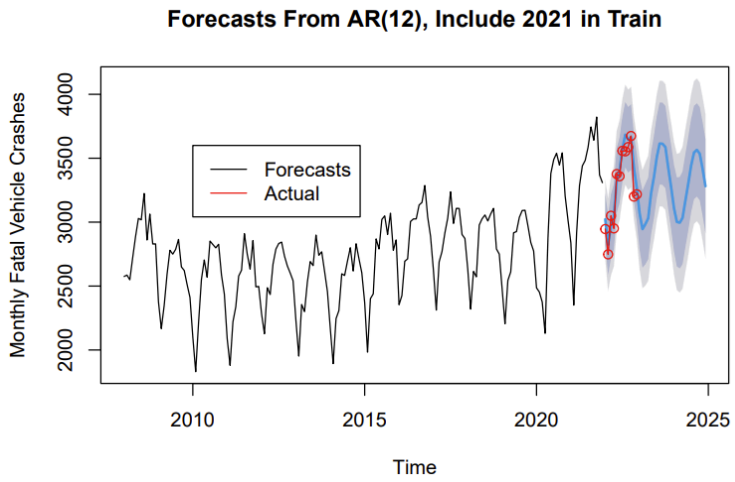

Figure 7. Forecasts from AR(12), 2008-2021 Data

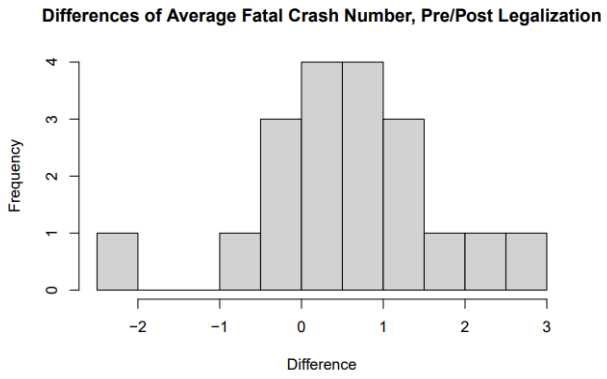

Figure 8. Difference of Average Fatal Crashes, Pre/Post Legalization

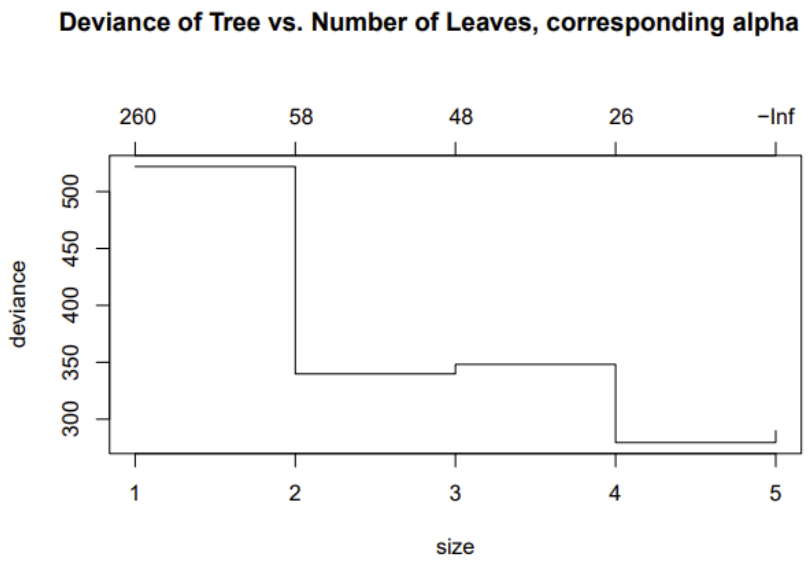

Figure 9. .Deviance of Tree vs. Number of Leaves

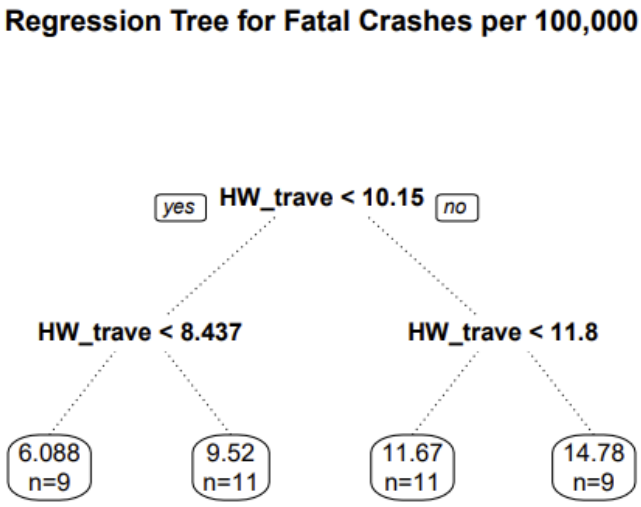

Figure 10. .Regression Tree for Fatal Crashes per 100,000

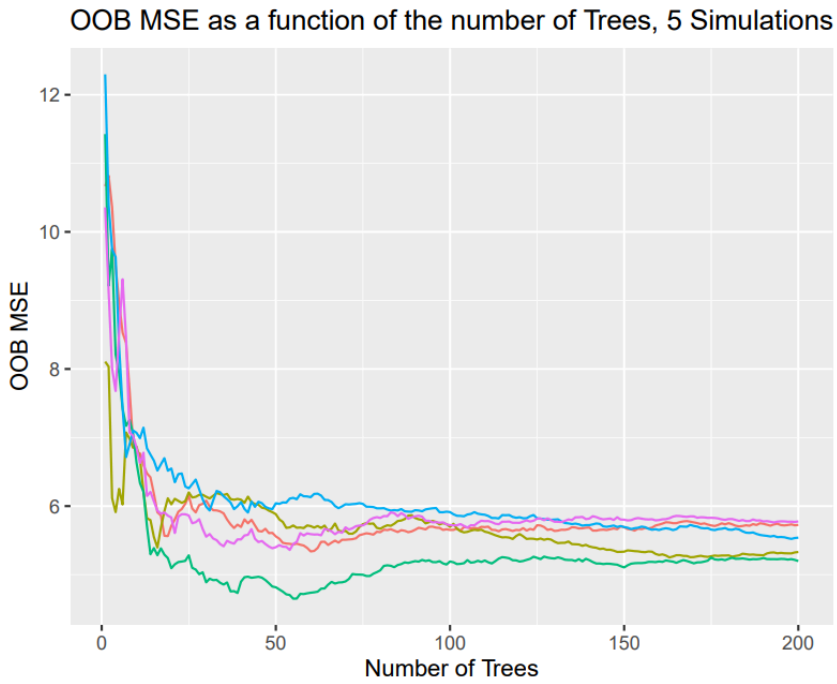

Figure 11. .OOB MSE as a function of the number of Trees

Variable	% Increase to MSE	Increase to Node Purity
Legalization	0.04685631	1.713326
Highway Miles Traveled	11.59499367	426.179536
% of Pop in Urban Areas	-0.19137532	24.914778
Urban Highway Speed Limit	-0.21115978	27.582473
% Pop > 70	-0.33457491	31.267251
Hazardous Weather	-0.01103727	28.871072
Adult Binge Drinking	0.34190161	34.615816

Figure 12. Variable Importance from Bagged Tree

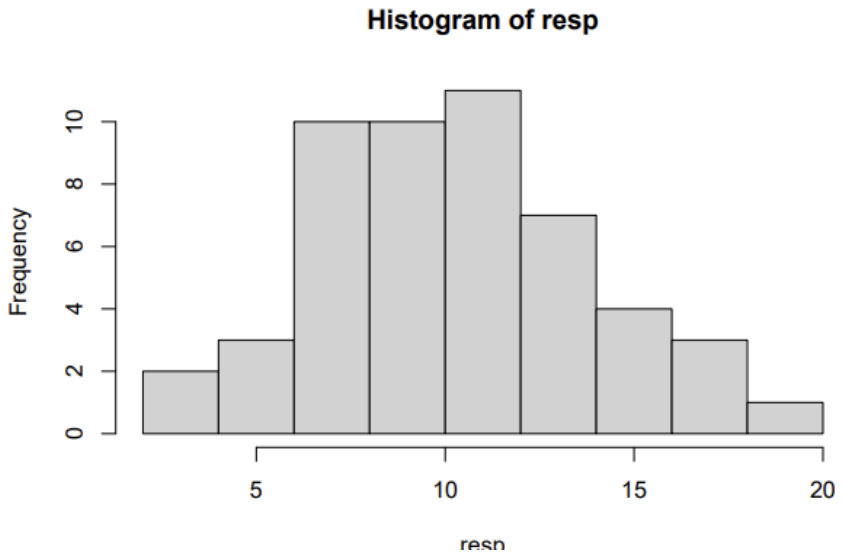

Figure 13. Testing Distribution of Response Variable

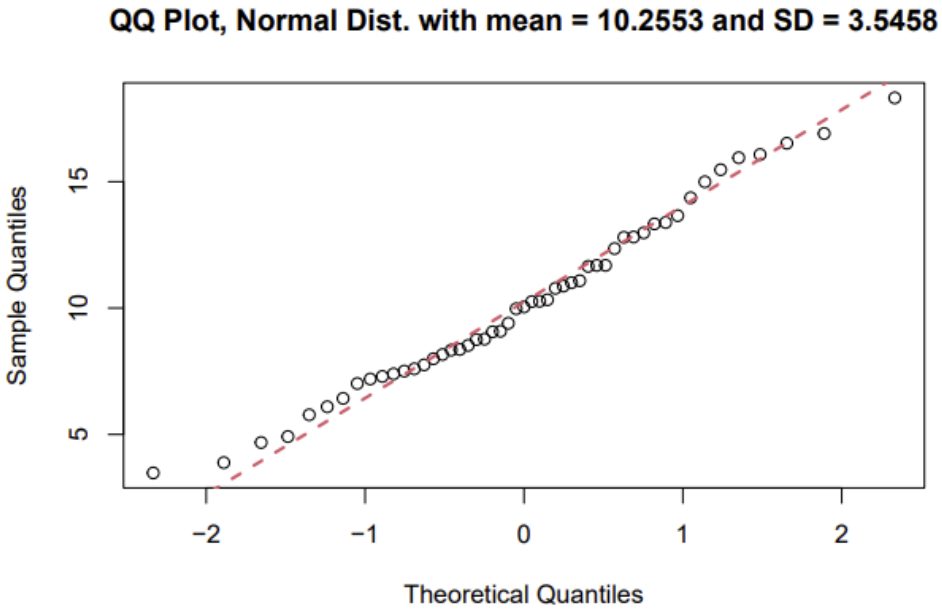

Figure 14.QQ Plot with Normal Distribution

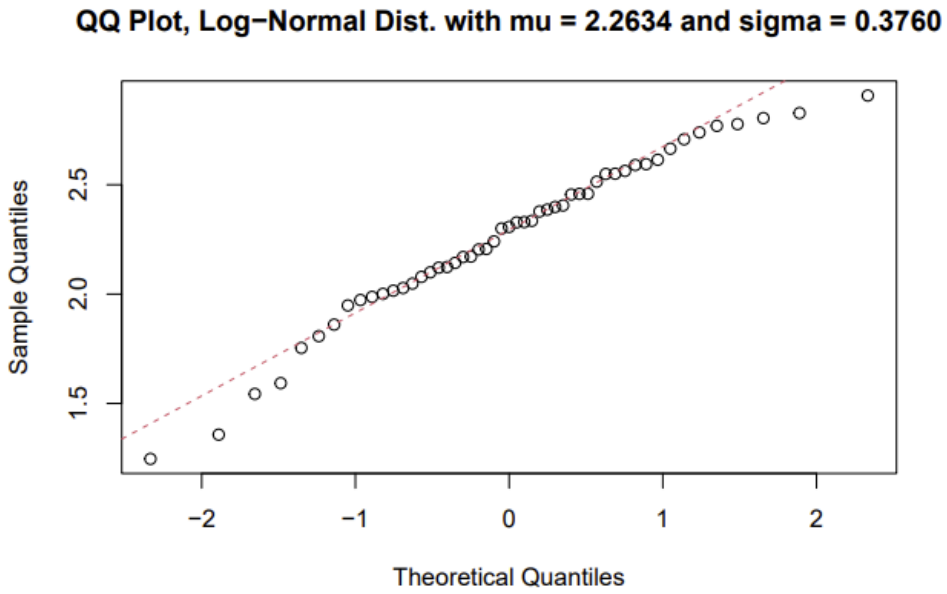

Figure 15. QQ Plot with Log of Normal Distribution

#	Predictors	R-Square	Adjusted R	CP	AIC	SBIC
1	HW_travel	0.6475	0.6382	-3.9562	179.2986	66.65335
2	HW_travel + binge	0.6818	0.6646	-5.0711	177.2081	65.69632
3	HW_travel + Speed_Lim + binge	0.6876	0.6193	4.3991	186.4687	67.88781
4	HW_travel + Speed_Lim + over_70 + binge	0.6906	0.6108	6.1259	188.0821	70.43232
5	HW_travel + Speed_Lim + over_70 + hazard + binge	0.6912	0.5986	8.0692	190.0013	73.22550
6	Legal + HW_travel + Speed_Lim + over_70 + hazard + binge	0.6920	0.5858	10.0000	191.9026	76.02018
7	Legal + HW_travel + Urban + Speed_Lim + over_70 + hazard + binge	0.6920	0.5710	12.0000	193.9026	78.87731

Figure 16.Best Subset Selection Result

```
Residuals:
    Min      1Q  Median      3Q     Max
-5.3293 -0.9234 -0.1180  1.5654  4.1128

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.7993     3.0577   0.588   0.5598
HW_travel     1.2838     0.1641   7.821 2.35e-09 ***
binge       -25.2905    12.6708  -1.996   0.0533 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 17. Best Linear Model Result

Works Cited

Bureau of Transportation Statistics. (2022). State Highway Travel. https://www.bts.gov/browse-statistical-products-and-data/state-transportation-statistics/state-highway-travel

Drug Enforcement Administration. (2020, April). Drug fact sheet: Marijuana/cannabis. https://www.dea.gov/sites/default/files/2020-06/Marijuana-Cannabis-2020_0.pdf

Federal Reserve Bank of St. Louis (FRED). (2023). Resident Population by State, Annual. https://fred.stlouisfed.org/release/tables?rid=118

Hansen, C., Alas, H., & Davis, E. (2023, November 8). *Where is marijuana legal? A guide to marijuana legalization*. U.S News. https://www.usnews.com/news/best-states/articles/where-is-marijuana-legal-a-guide-to-marijuana-legalization

Insurance Institute for Highway Safety (IIHS). (2023, May). Fatality Facts 2021 State by state. https://www.iihs.org/topics/fatality-statistics/detail/state-by-state#yearly-snapshot

McCarthy, N. (2019, June 14). *The arguments for and against marijuana legalization in the U.S. [infographic]*. Forbes. https://www.forbes.com/sites/niallmccarthy/2019/06/14/the-arguments-for-and-against-marijuana-legalization-in-the-u-s-infographic/?sh=550738eb678b

National Highway Traffic Safety Administration (NHTSA). (2022). https://www.nhtsa.gov/crash-data-systems/fatality-analysis-reporting-system

National Weather Service (NWS). (2022). Summary of Natural Hazard Statistics for 2022 in the United States. https://www.weather.gov/media/hazstat/state22.pdf

Statista. (2022). Binge drinking prevalence among adults in the United States as of 2022, by state. https://www.statista.com/statistics/378966/us-binge-drinking-rate-adults-by-state

United States Census Bureau. (2023). State Population by Characteristics: 2020 - 2023. https://data.census.gov/

What age group has the most accidents?. Morris Bart, LLC. (2023, September 13). https://www.morrisbart.com/faqs/what-age-group-has-the-most-accidents/#:~:text=According%20to%20the%20AAA%20Foundation,from%20any%20other%20age%20group