Examining the Impact of Marijuana Legalization on Fatal Car Crashes

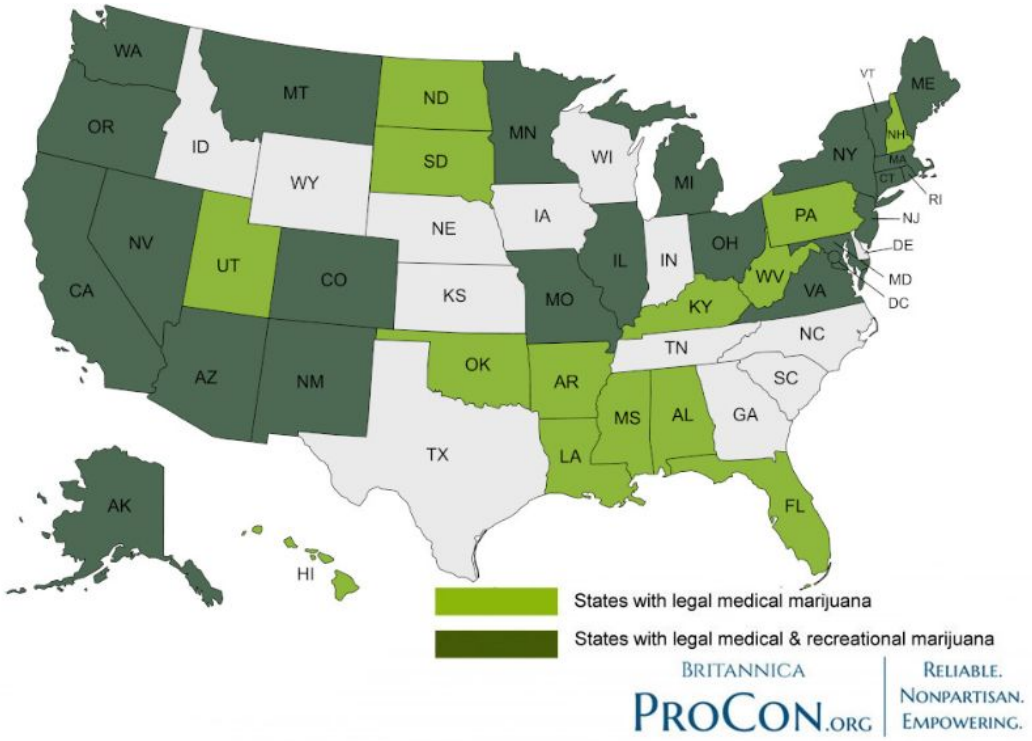
Alana Berson, Suleman Sadiq, Dennis Goldenberg, Solomon Gao

Agenda

1. Question
2. Dataset
   - Variables
   - Seasonality
3. Statistical Models
   - Data preprocessing
   - Hyper parameter tuning / variable selection
4. Models & Test Statistics
   - Paired sample t-test
   - Decision Trees & Random Forests
   - Linear/Log-Linear Regression
5. Best Models
6. Conclusion

Business Question:

Does Marijuana legalization have an impact on fatal car accidents?



Dataset

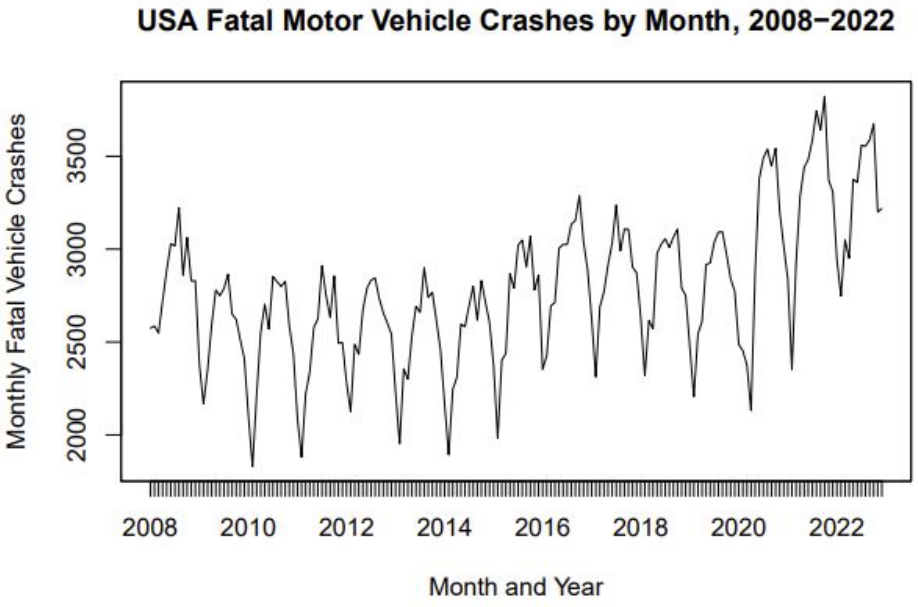
Response: Fatal vehicle crashes by state

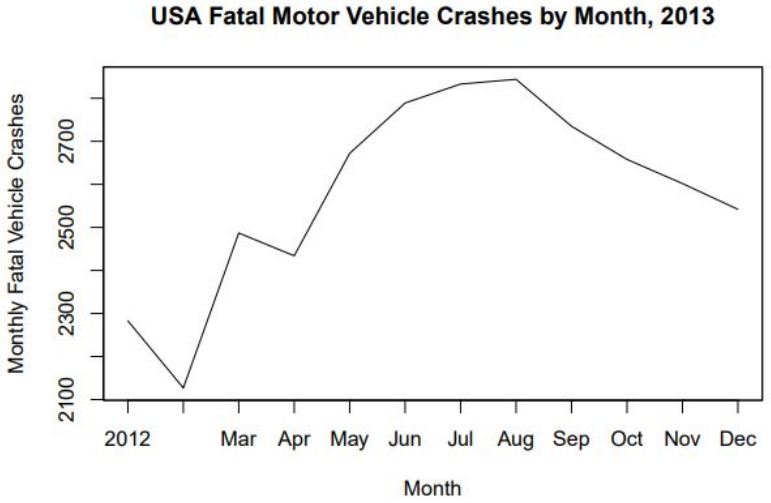
Predictors (at State Level):

- Highway-miles-driven in billions per 100,000 people
- % of population living in urban areas
- Urban highway speed limits
- % >70 years of age by state
- Adult binge drinking by state
- **Legalization of Marijuana (Pre 2022)***
- Damage per $100,000 by hazardous weather events, 2022

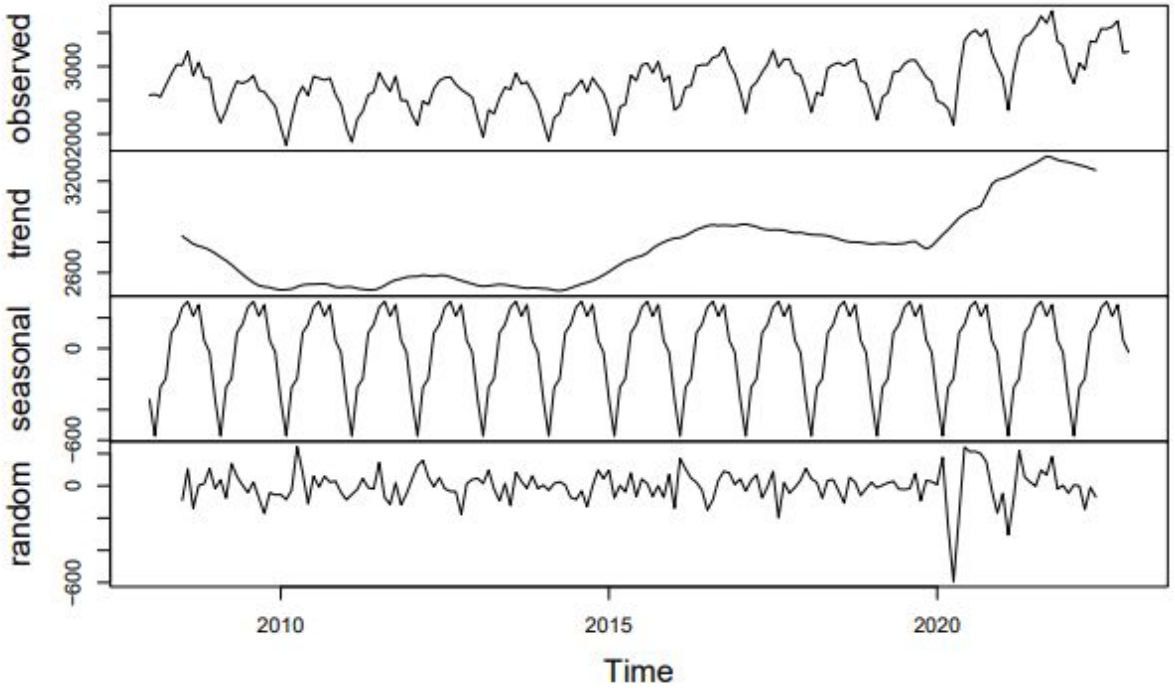
* Not including Rhode Island

Seasonality

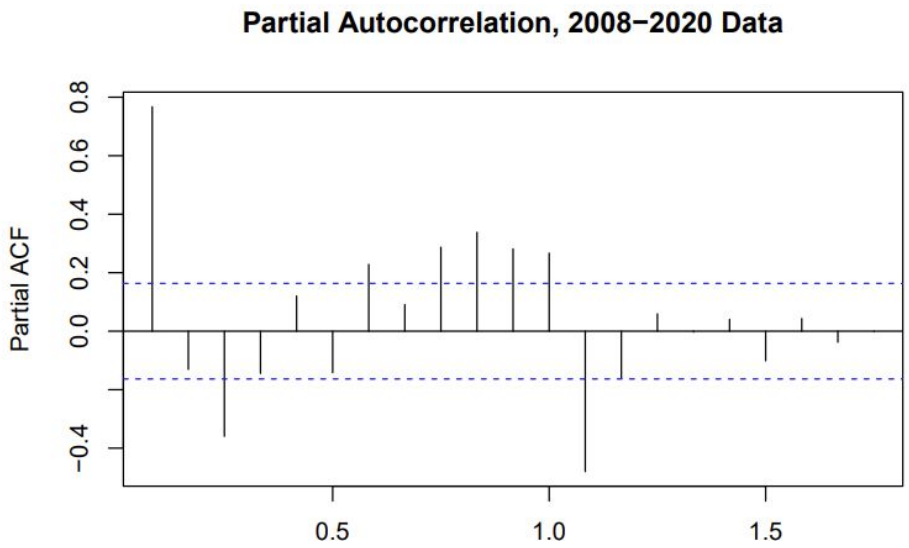


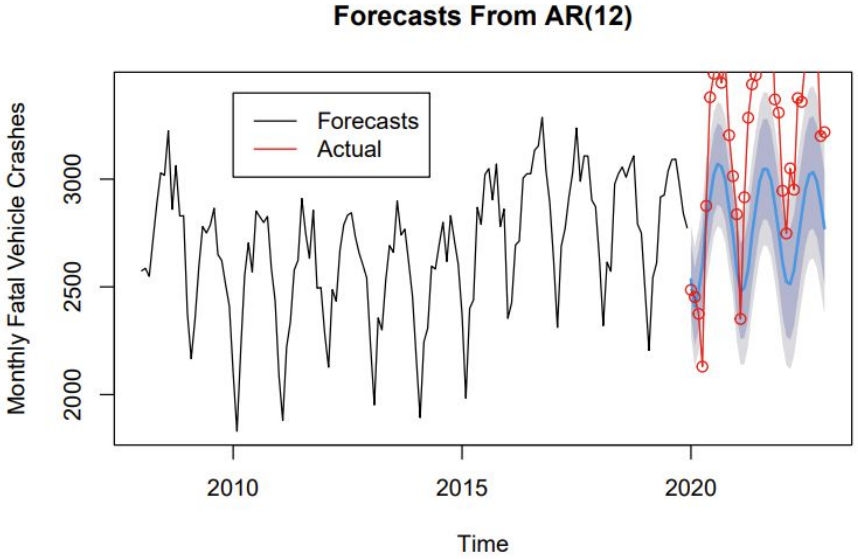


Seasonality: Decomposition of Additive Time Series



Running an Autoregression

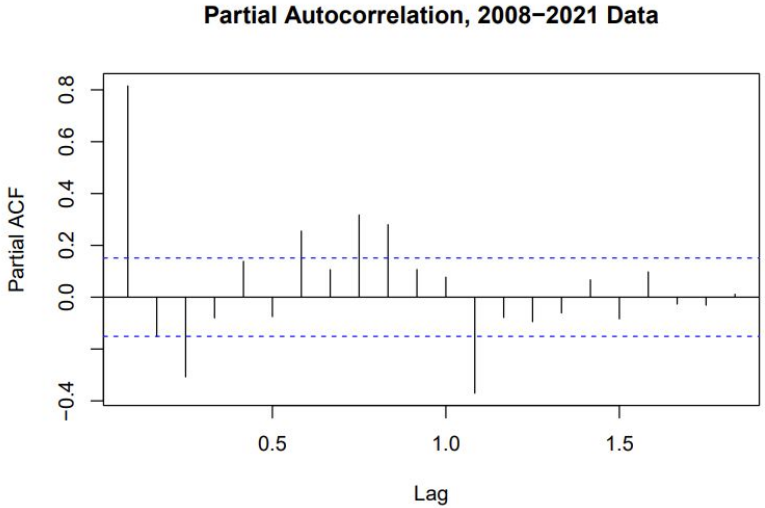


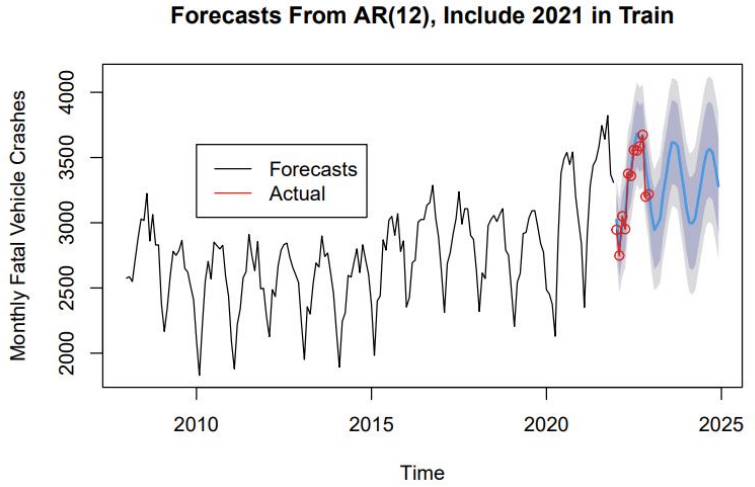



```
ar1_model <- ar(train_set, method = "mle")
```

- Use 2008-2020 as training data, 2021 and 2022 and test
- Partial autocorrelation shows that previous 12 values important
- Trend over 2021, 2022 not captured in training data

Running an Autoregression: Including 2021 in Training






```
ar1_model2 <- ar(train_set2, method = "mle")
```

- Use 2008-2021 as training data 2022 and test
- Partial autocorrelation very similar
- Trend over final years much more effectively captured

Paired Sample t-Test: Does Legalization Have an Immediate Impact?

t-test: Generating the Response

Examine the difference between average fatal car crash in state 12 months *after* legalization vs. 12 months *before* (per 100,000)

Steps to generate response variable:

1. Average number of fatal vehicle accidents in the 12 months prior to legalization
2. Linearly interpolate between population (in 100,000s) estimates of previous 2 years
3. Divide (1) by (2)
4. Repeat for 12 months post legalization

Ex. Oregon legalized on July 1, 2015

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

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

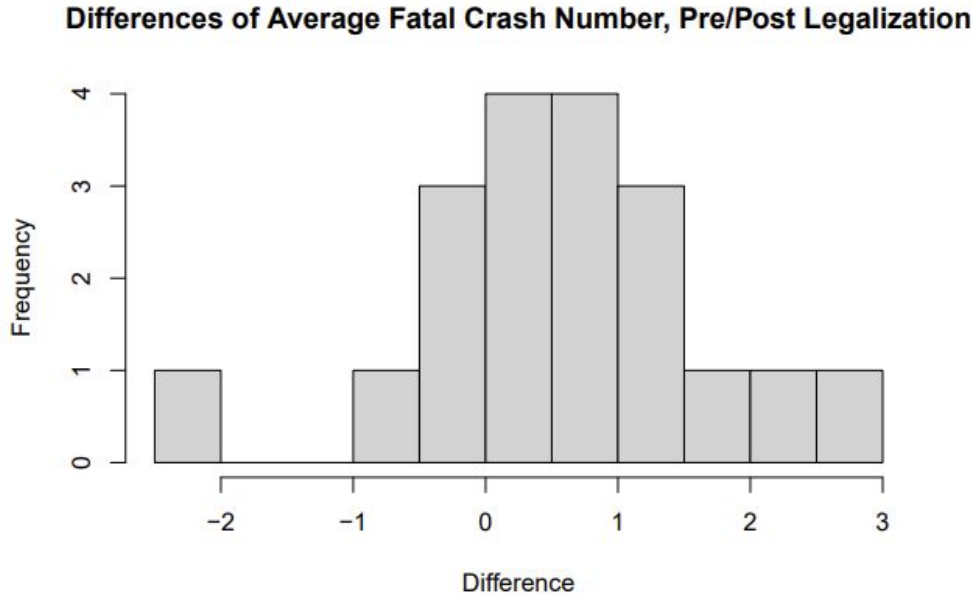
- **fatalities_i**: Number of fatal car accidents i months before/after July 2015
- Proportion of 2015 that passed before legalization: **1/2**

t-test: Examining the Assumptions

n= 19

Roughly normally distributed ✔

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$



T-test: Hypothesis Testing

```
testMeans <- t.test(pairedData$avg_post_Leg, pairedData$avg_pre_Leg,
        paired = TRUE, alternative = "greater")
```

$$H_0: \overline{X}_{post} \leq \overline{X}_{pre}$$

$$H_1: \overline{X}_{post} > \overline{X}_{pre}$$

Sample Mean Difference	0.60166
Test Statistic	2.41257
Degree of Freedom	18
p-value	0.01336
95% CI	0.1692114, ∞
Reject Null , α = 0.05	Yes

So, $\mathbb{P}(T_{18} > 2.41257) = 0.01336 < 0.05$; I reject H_0 at $\alpha = 0.05$.

Decision Trees, Random Forests & Regression: How Important is Marijunana Legalization?

Decision Tree: Setting up the Decision Tree

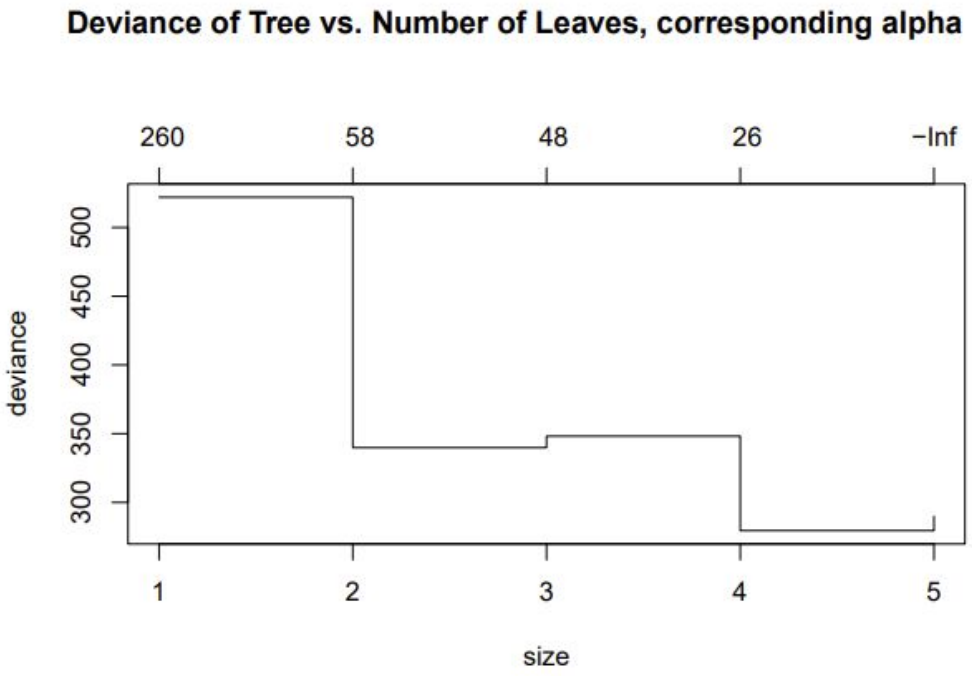
Train Test Split:

80% train

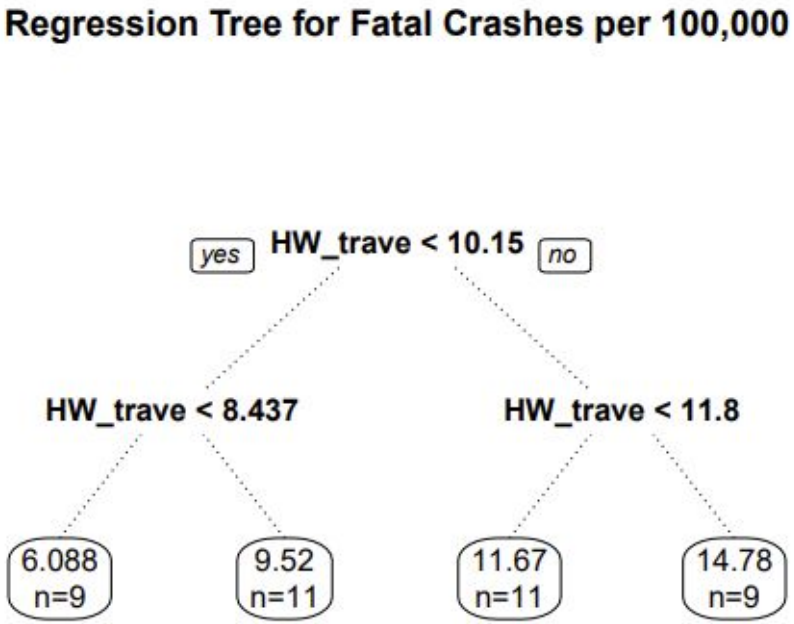
20% test

ii. 1 tree (for example)

```
set.seed(1)
train_index <- sample(1:nrow(finData),.8 * nrow(finData),replace = FALSE)
test_index <- setdiff(1:50, train_index)
rtree <- tree(resp ~ ., data = finData[train_index,-c(2)])
cv <- cv.tree(rtree, K = 5, FUN = prune.tree)
```



Decision Tree: Corresponding Graph



Generating the tree

```
rtree <- rpart(resp ~., data = finData[train_index,-c(2)], method = "anova",
               control = rpart.control(cp = 1/20))
prp(rtree, main = "Regression Tree for Fatal Crashes per 100,000",
    roundint = FALSE, extra = 1, digits = 4, branch.lty = 3)
```

Testing Tree

```
oneTree <- tree(resp ~ ., data = finData[train_index, -c(2)], method = "anova")
oneTree <- prune.tree(oneTree, k = 30)
predOneTree <- predict(oneTree, newdata = finData[test_index,3:9])
mseOneTree <- mean((predOneTree - finData$resp[test_index])^2)
pseudoROneTree <- 1 - (mseOneTree * 50)/(var(finData$resp) * 49)
```

Bagged Trees: An Explanation

Reduction in Variance: By averaging multiple trees, bagging reduces the variance of the prediction

Improved Accuracy: Bagging can yield improved accuracy over single decision trees

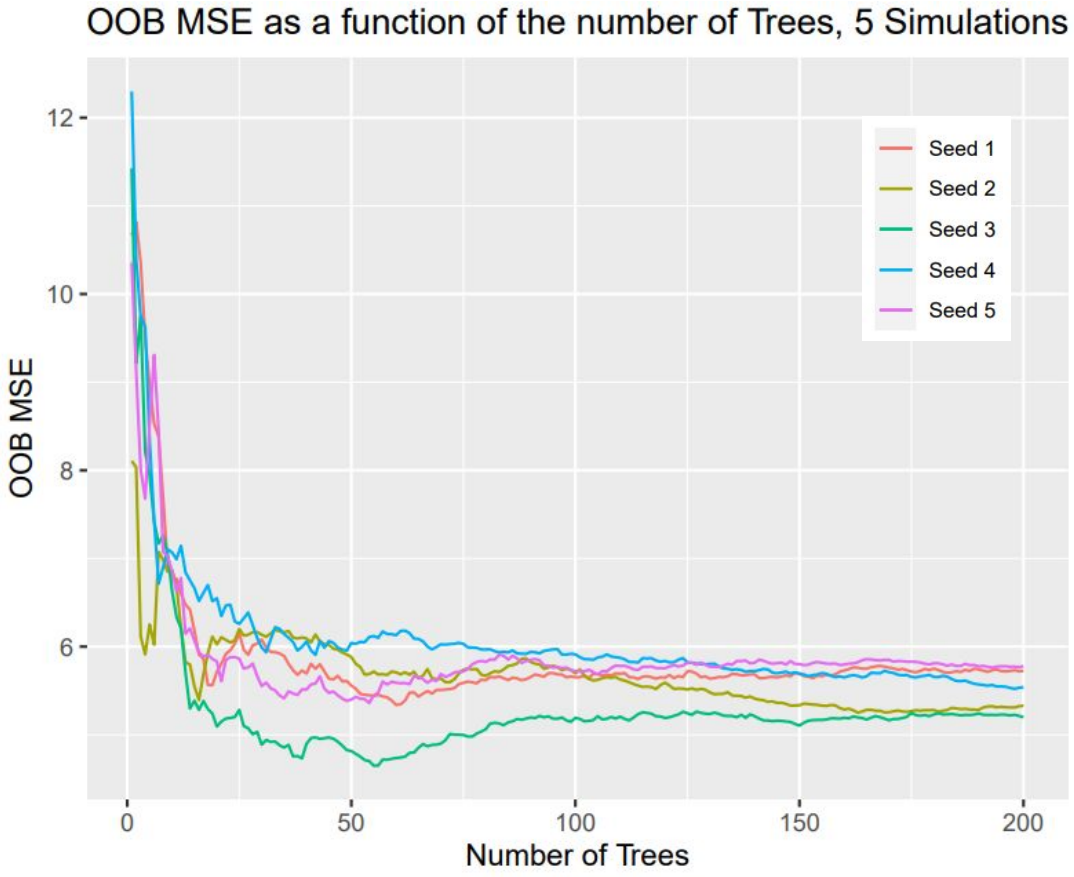
Handles Overfitting: Bagging helps mitigate overfitting by average multiple trees, each of which might overfit in different ways

Robustness to Outliers: No single outlier can influence all the trees in the ensemble

Bagged Trees: Hyperparameter Tuning

The optimal number of trees is 50

Adding more trees beyond 50 does not continue to decrease the MSE



Variable Importance

- Highway miles traveled is the most important variable
- Adult binge drinking is slightly important
- Legalization is not important when compared to other variables

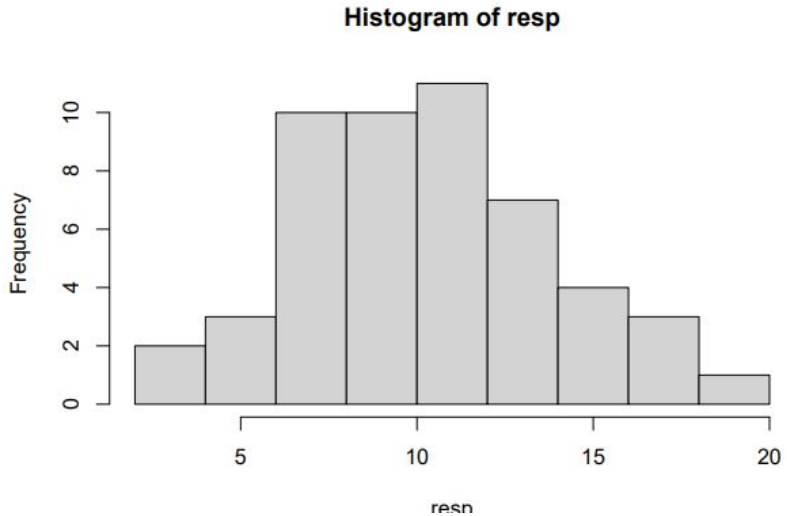
Variable	% Increase to MSE	Increase to Node Purity
Legalization	0.04685631	1.713326
Highway Miles Traveled	11.59499367	426.179536
% of Pop in Urban Areas	-0.19137532	24.914778
Urban Highway Speed Limit	-0.21115978	27.582473
% Pop > 70	-0.33457491	31.267251
Hazardous Weather	-0.01103727	28.871072
Adult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

Regression Analysis: Examining the Response Variable

i. Testing distribution of Response

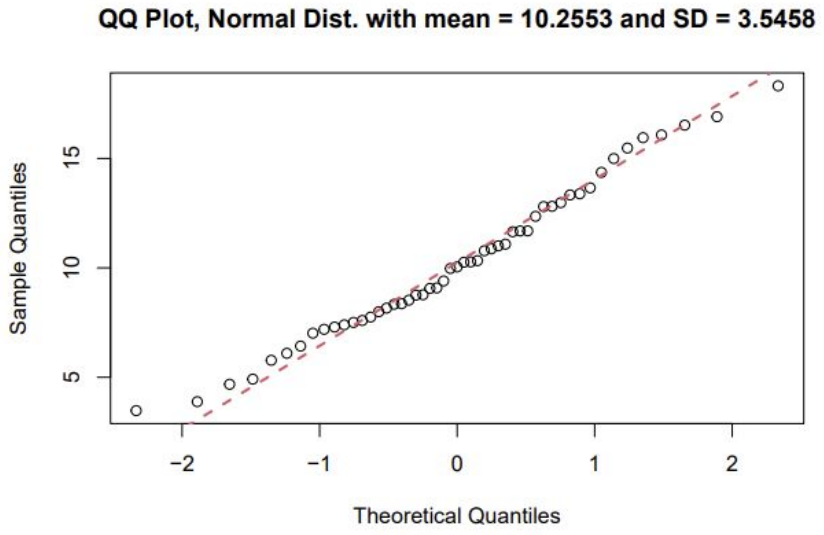
```
set.seed(1)
lambdaMLE <- mean(resp)
#hist(resp, breaks = 20)
#qqplot(resp, distribution = "poisson")
PoissonDist <- rpois(length(resp), lambdaMLE)
title <- sprintf("QQ Plot, Poisson Dist. with lambda = %.4f",lambdaMLE)
#qqplot(resp * 12, PoissonDist, main = title)
#abline(0,1,col = 'red')
hist(resp, breaks = 10)
```

Roughly Normal Distributed ✔

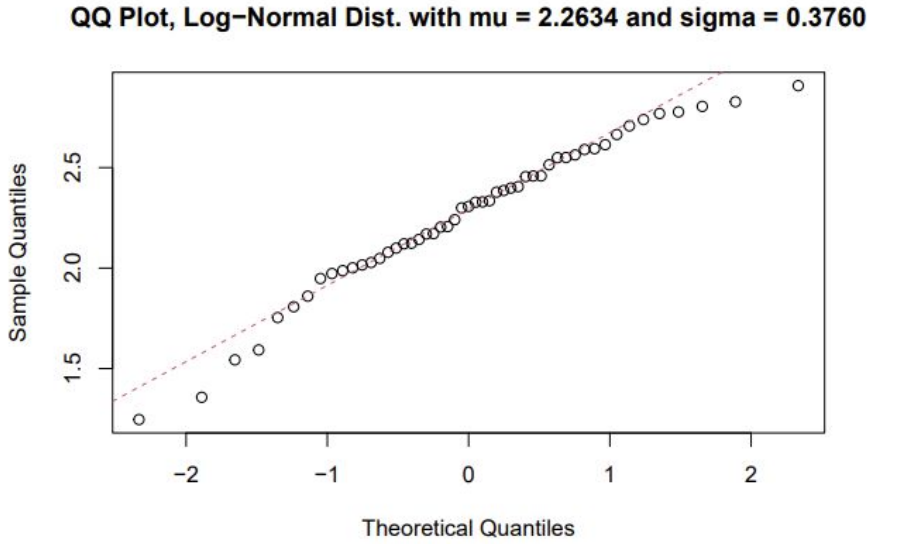


Regression Analysis: Verifying Normality Through qq-plots

```r
set.seed(1)
meanResp <- mean(resp)
sdResp <- sd(resp)
NormalDist <- rnorm(length(resp), meanResp, sdResp)
title <- sprintf("QQ Plot, Normal Dist. with mean = %.4f and SD = %.4f",
                 meanResp, sdResp)
qqnorm(resp, main = title)
qqline(resp, col = 2, lwd = 2, lty = 2)
```




```r
set.seed(1)
meanLogResp <- mean(log(resp))
sdLogResp <- sd(log(resp))
LogNormalDist <- rlnorm(length(resp), meanLogResp, sdLogResp)
title <- sprintf("QQ Plot, Log-Normal Dist. with mu = %.4f and sigma = %.4f", meanLogResp, sdLogResp)
qqnorm(log(resp), main = title)
qqline(log(resp), col = 2, lty = 2)
```



Regression Analysis: Feature Selection Using Best Subset

#	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

ii. Best Subset Selection

```
train_data <- finData[train_index,-c(2)]
test_data <- finData[test_index, -c(2)]
linmodel <- lm(resp ~., data = train_data)
best_sub <- ols_step_best_subset(linmodel)$metrics[,c("predictors", "rsquare",
                                                      "adjr", "cp", "aic", "sbic")]
best_sub
```

Based off CP, AIC, SBIC & R^2_{adj} the best subset is : Model 2

Regression Analysis: Running Best Linear Model

iii. Running best Model

```
linBestModel <- lm("resp ~ HW_travel + binge", data = train_data)
predlinBest <- predict(linBestModel, newdata = test_data)
linBestMSE <- mean((test_data$resp - predlinBest)^2)
summary(linBestModel)
```

```
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
```

Best Models

Results of our Models

Algorithm	MSE	Proportion of Variance Explained (R^2 / pseudo R^2)
Decision Tree	6.605976	0.4375485
Bagged Forest	5.609170	0.5224193
Multiple Linear Regression	3.652816	0.6817754

- Multiple Linear Regression is the best model
- MSE calculated on test set (20% of data / 10 data points)

Conclusion

Takeaways

DON'T SMOKE AND DRIVE

- Evidence of increase in monthly fatal vehicle accidents directly preceding legalization
- Marijuana legalization not statistically significant, however when evaluated in the context of other predictors
- Variables that were important in predicting fatal vehicle accidents
  - Billions of highway miles traveled, per 100,000
  - Adult binge drinking %

Considerations for Further Study

1. Does the impact of legalization remain constant or vary over time?

2. Does the legal possession limit have a significant impact of legalization overall impact on vehicle safety?

3. In terms of miles-driven, is there an increasing hazard rate for the distributions of time of first accident?

Sources of Data - Response

- Inspiration for Project: *The Insurance Institute for Highway Safety*
  - https://www.iihs.org/topics/fatality-statistics/detail/state-by-state#yearly-snapshot
- Source of Fatal Crash Data: *National Highway Traffic Safety Administration*
  - https://www.nhtsa.gov/crash-data-systems/fatality-analysis-reporting-system
  - *for Querying Data:*
    - https://cdan.dot.gov/query

Sources of our Data - Predictors

- Legalization of Marijuana: *U.S News*
  - https://www.usnews.com/news/best-states/articles/where-is-marijuana-legal-a-guide-to-marijuana-legalization
- Population Estimates Data: *St. Louis Fed*
  - https://fred.stlouisfed.org/release/tables?rid=118
- Highway-miles-driven: *Bureau of Transportation Statistics*
  - https://www.bts.gov/browse-statistical-products-and-data/state-transportation-statistics/state-highway-travel
- % Urban Data: *U.S Census Bureau*
  - https://data.census.gov/
- 70 and Older percentage by State: *U.S Census Bureau*
  - https://www.census.gov/data/tables/time-series/demo/popest/2020s-state-detail.html#v2022
- Total Damage per 100,000 by Weather : *National Oceanic and Atmospheric Administration*
  - https://www.weather.gov/media/hazstat/state22.pdf
- Adult Binge Drinking Percentage: *Centers for Disease Control*
  - https://www.statista.com/statistics/378966/us-binge-drinking-rate-adults-by-state

THANK YOU!
Questions?