Auto Insurance Analysis

Radha Avudaiappan, Abby Bilger, Dustin Fife

Purpose of the Project

- Analyze the factors that affect auto insurance rates including crash statistics, population traits, and socioeconomic factors
- Research and model data from a multiple data sources including the Census data
- Implement machine learning and statistical analysis tools such as pandas (Python) and Power BI

Exploratory Questions and Data Sources

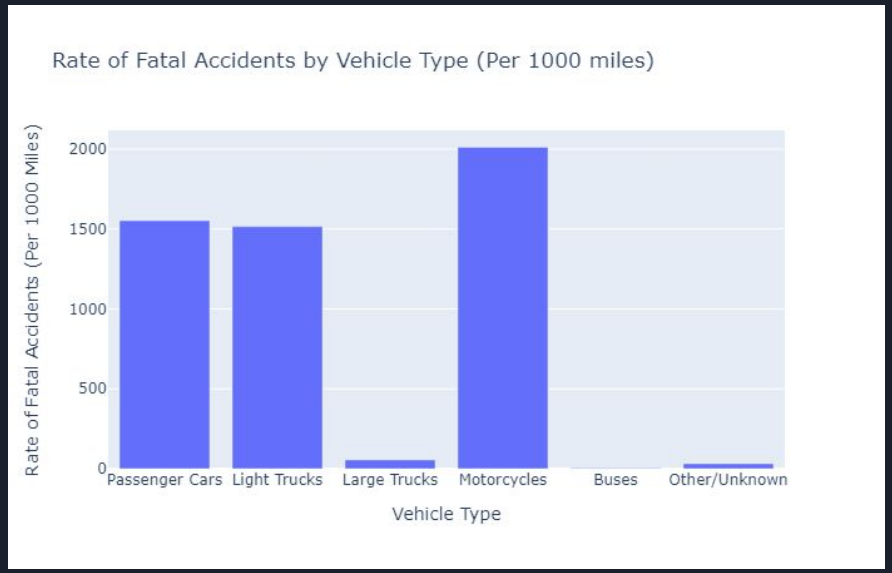
Questions

- What are the causes of car crashes?
- Which states have the most accidents?
- Which states have the highest/lowest insurance premiums?
- Can we accurately predict the price of a car insurance claim?
- Can we accurately predict whether a claim will be rejected?

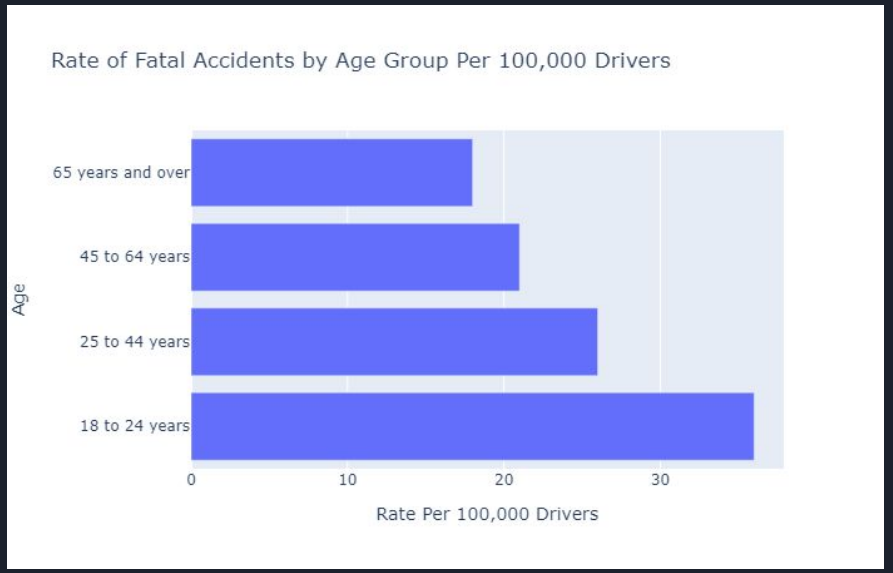
Data Sources

- Census Data Portal
- Insurance Information Institute
- National Highway Traffic Safety Administration
- Emcien Patterns Knowledge database

Types of Vehicles and Age Breakdown -Crash Analysis

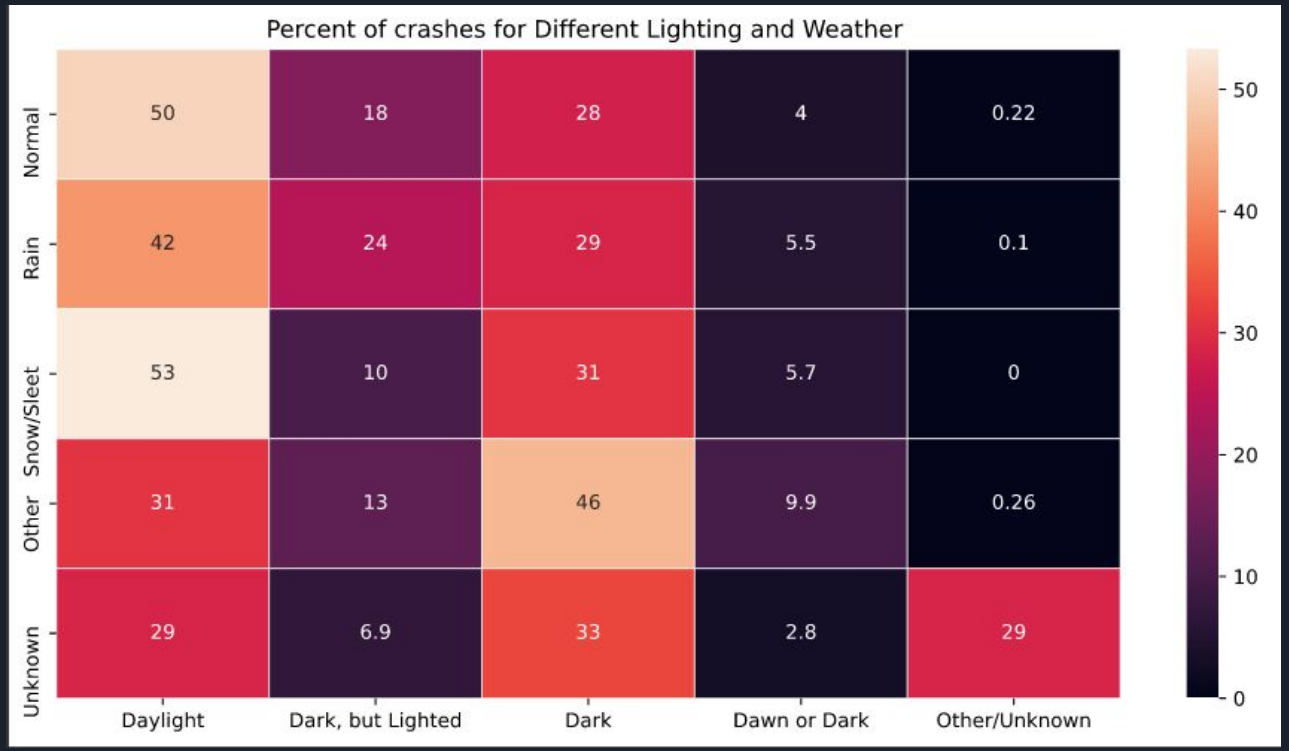


Motorcycles are most likely to be involved in a fatal accident



Younger drivers tend to be involved in more accidents

Weather vs. Daylight-Crash Analysis

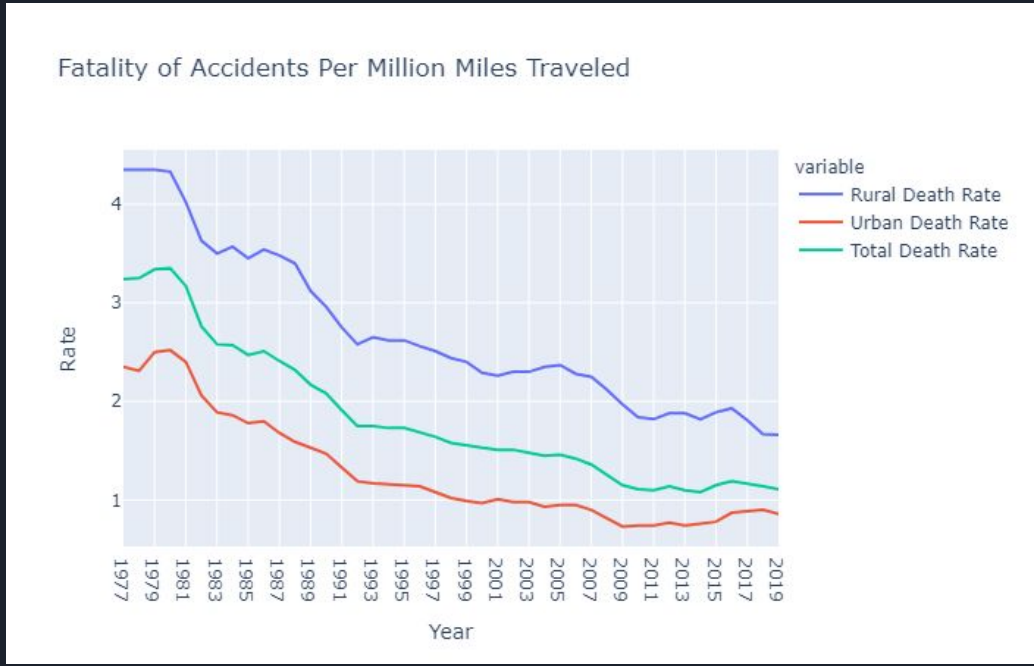


Urban vs. Rural-Crash Analysis

Possible causes:

- Rural roads have higher speed limits
- Roads tend to be in poorer condition
- Less likely to be plowed during winter time

Note: Only fatal accidents were analyzed.



Rural roads consistently have higher fatal collision rates

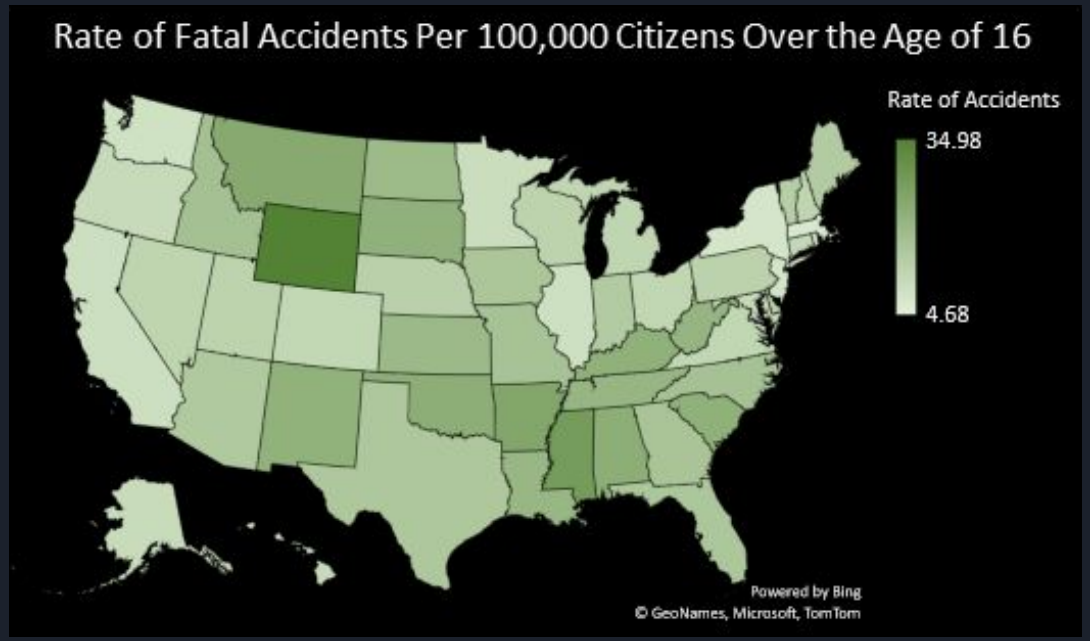
Breakdown of States-Crash Analysis

3 States with Highest Fatal Collision Rate:

1. Wyoming
2. Mississippi
3. Arkansas

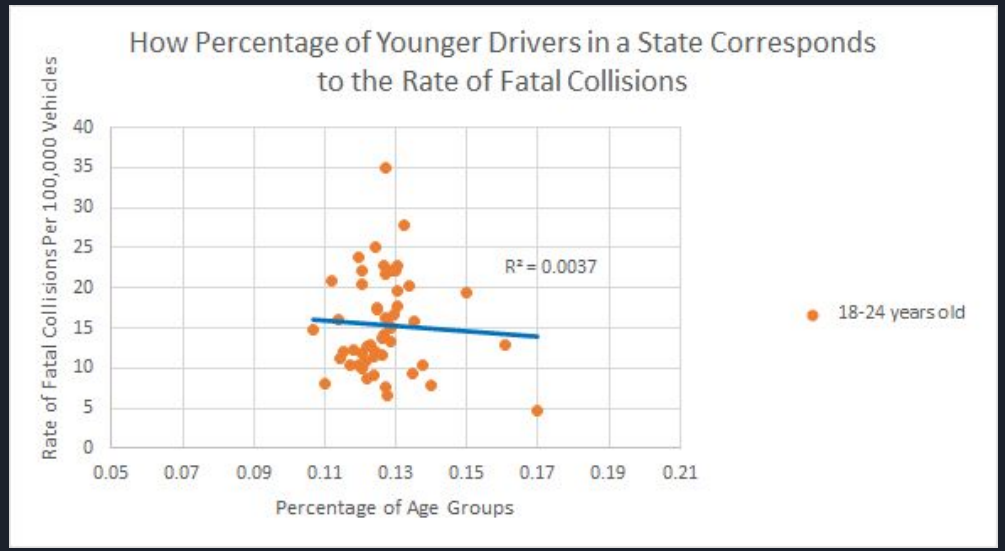
3 States with Lowest Fatal Collision Rate:

1. Washington D.C
2. Massachusetts
3. New York



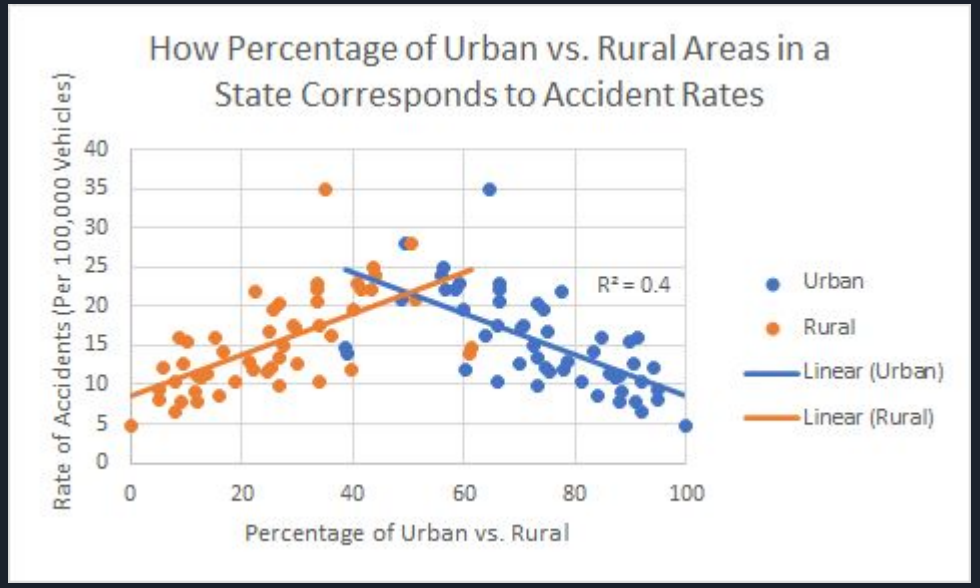
Rural states tend to have higher collision rates

How Big of A Factor is Age in Predicting a Fatal Accident?



Although younger drivers are more likely to get into a fatal accident, areas with more younger people do not necessarily have higher accident rates

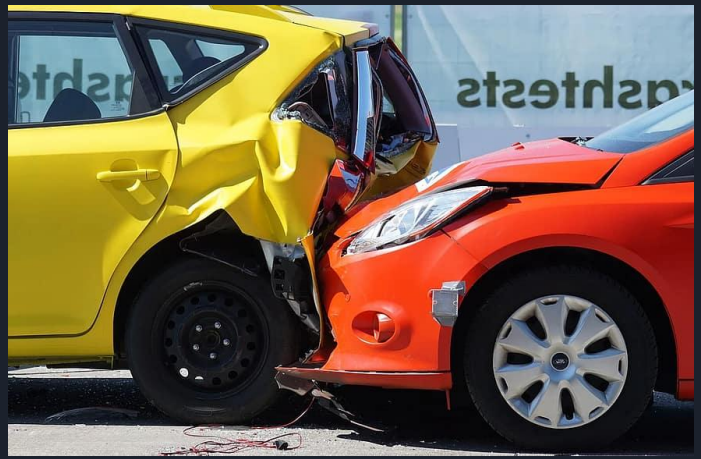
How Big of a Factor is Urban/Rural Distribution in Predicting Fatal Crashes?



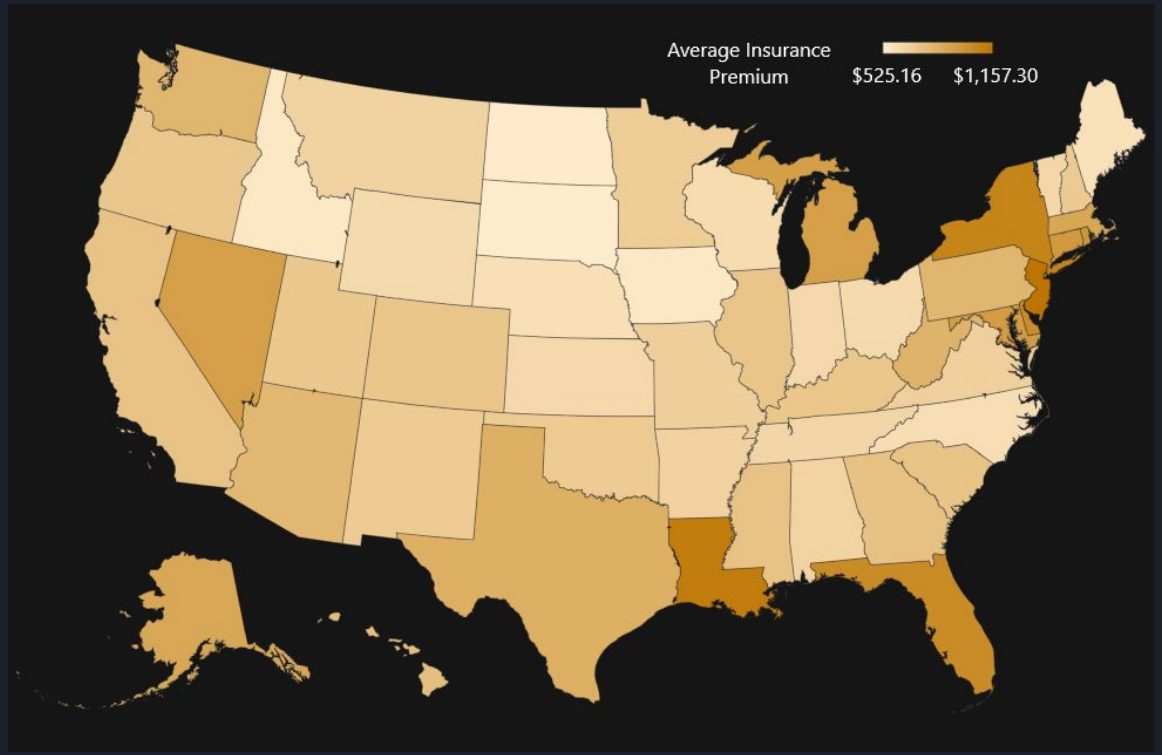
With a moderate correlation of 0.4, rural or urban location should be considered in predicting fatal crashes in an area.

Other Factors Worth Considering: Crash Analysis

- Make and model of vehicles
- Analyzing all crashes not just fatal crashes
- Deeper analysis on different factors that affect different climates
- Quality of Drivers' Education in each state
- State/City Ordinances (i.e. speed limits, laws that limit driving due to alcohol and curfews)



Average Insurance Premium By State (2010)



Highest Premiums:

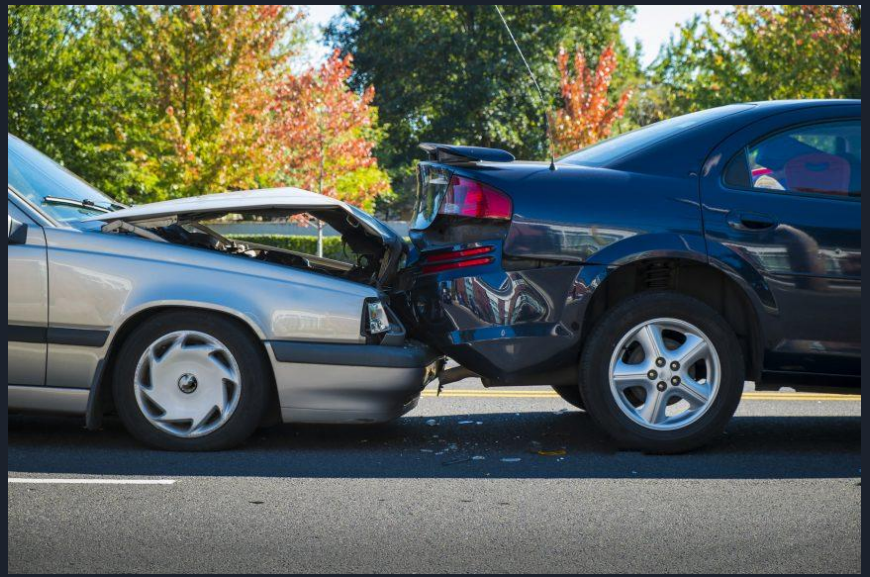
- New Jersey: $1,157.30
- Washington DC: $1,133.87
- Louisiana: $1,121.46

Lowest Premiums:

- South Dakota: $525.16
- North Dakota: $528.81
- Iowa: $546.59

Factors that Affect Insurance Premiums

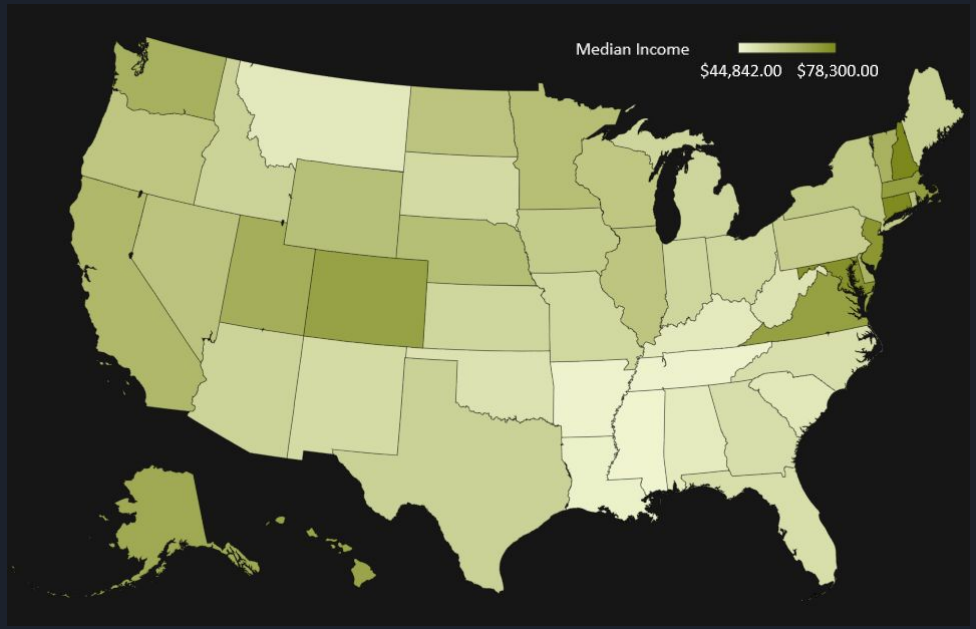
- Driving habits (Claim history, driving record, how much you drive, etc.)
- Vehicle Type
- Location
- Age and Gender
- Income and Credit History



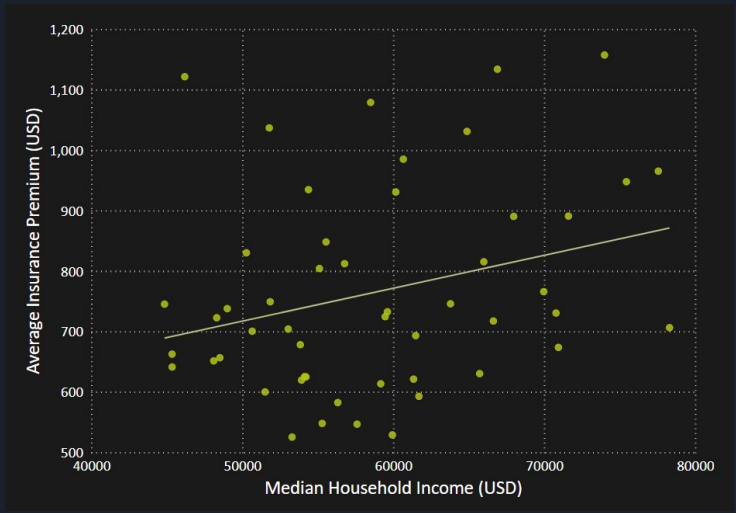
Information on Insurance Premium Factors taken from Allstate and Insure

Median Household Income

Median Household Income by State (2010)



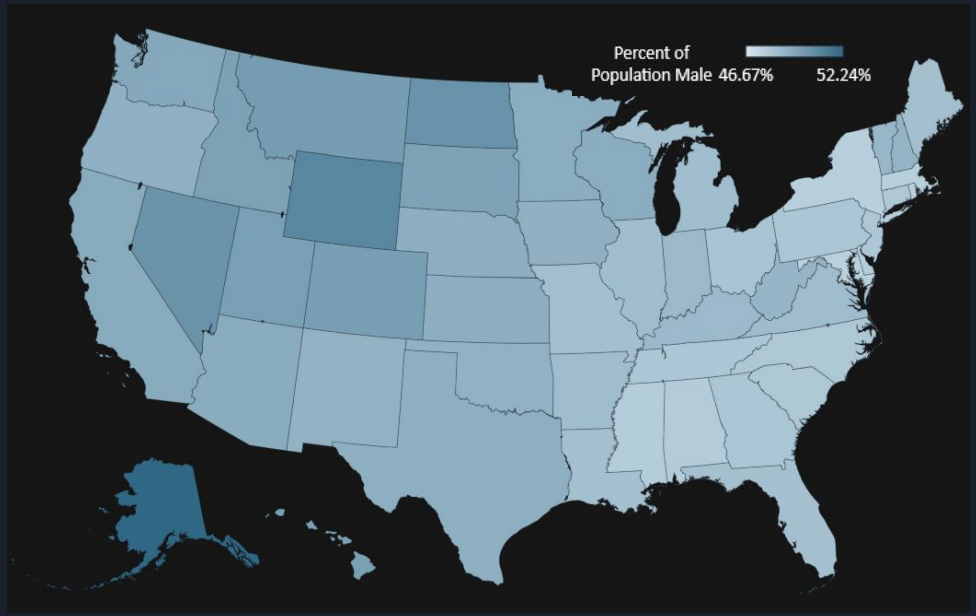
Insurance Premiums by Median Income



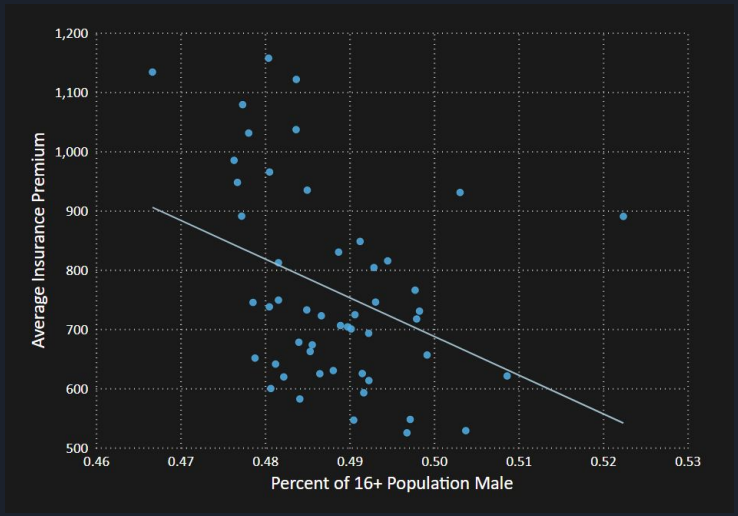
The correlation between median income and insurance premiums at the state level was 0.268.

Gender Distribution

Percentage of 16+ Population Male (2010)



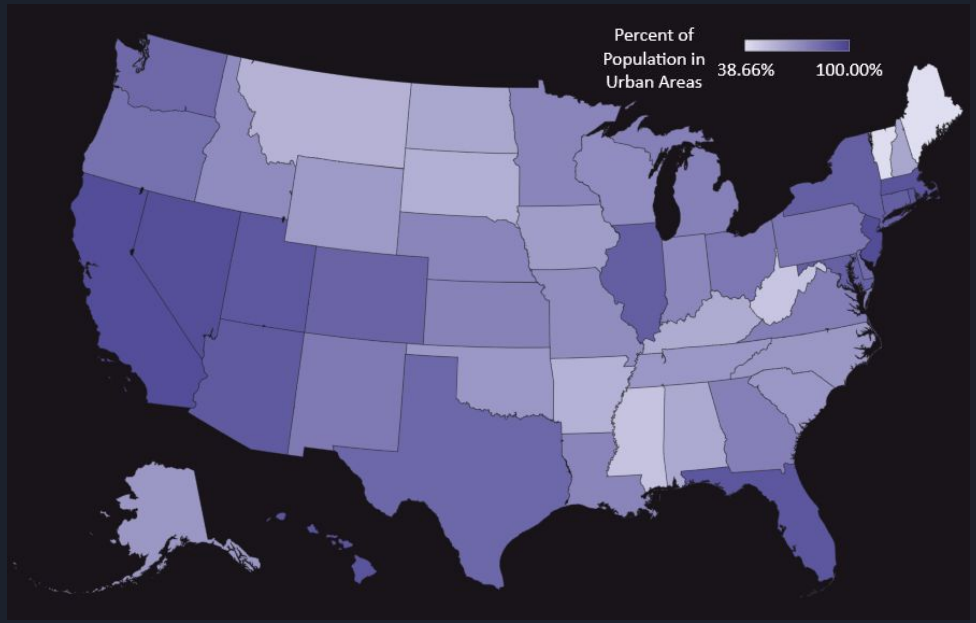
Insurance Premiums by Percent Male



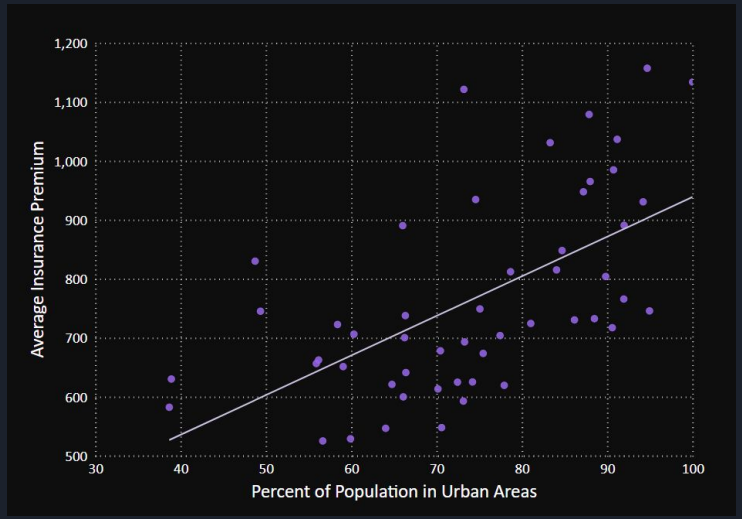
The correlation between median income and percentage male at the state level was -0.372.

Urban Percentage

Urban Population Percentage by State (2010)



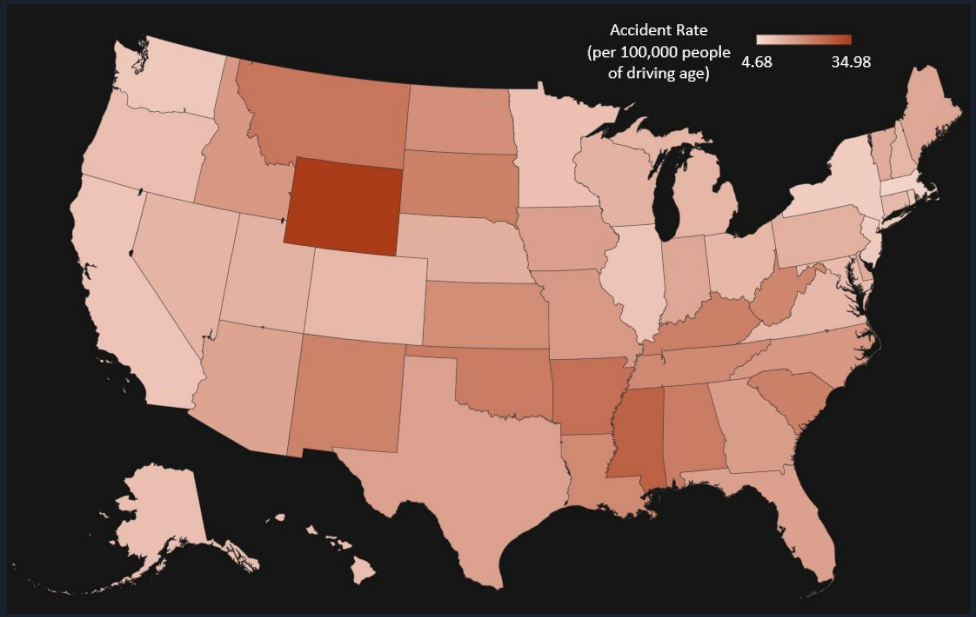
Insurance Premiums by Median Income



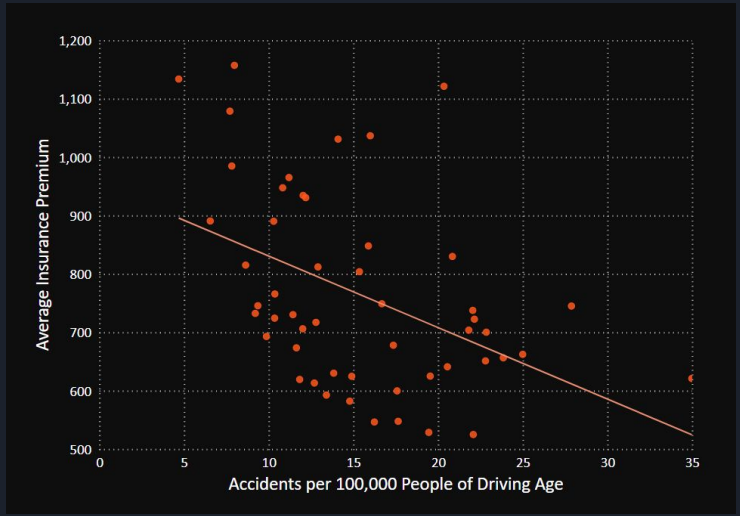
The correlation between urban percentage and insurance premiums at the state level was 0.598.

Accident Rate

Accident Rate by State (2010)



Insurance Premiums by Median Income



The correlation between median income and insurance premiums at the state level was -0.446.

Statewide Data Analysis Summary

- Insurance premiums are highly personalized, Census data is not
- Some relationships were unexpected (Accident Rate)
- Some expected relationships were not found (Income)
- A lot of data on important factors was unavailable
- Need data on more factors
- May not be possible to effectively predict statewide

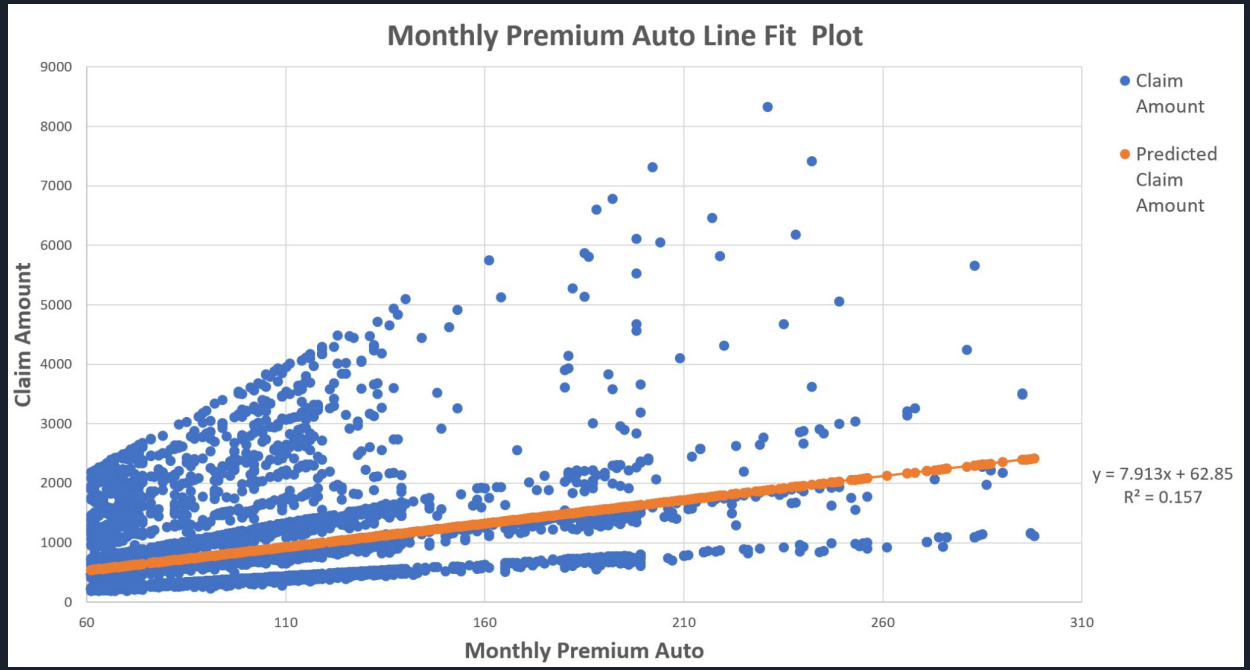
Machine Learning: Auto Insurance Claims

- Goals:
    1. Predict the cost of a claim
    2. Predict whether the claim is accepted or denied.
- Both Regression and Classification models used
- Cross Validation with training and testing datasets

Claims Correlation

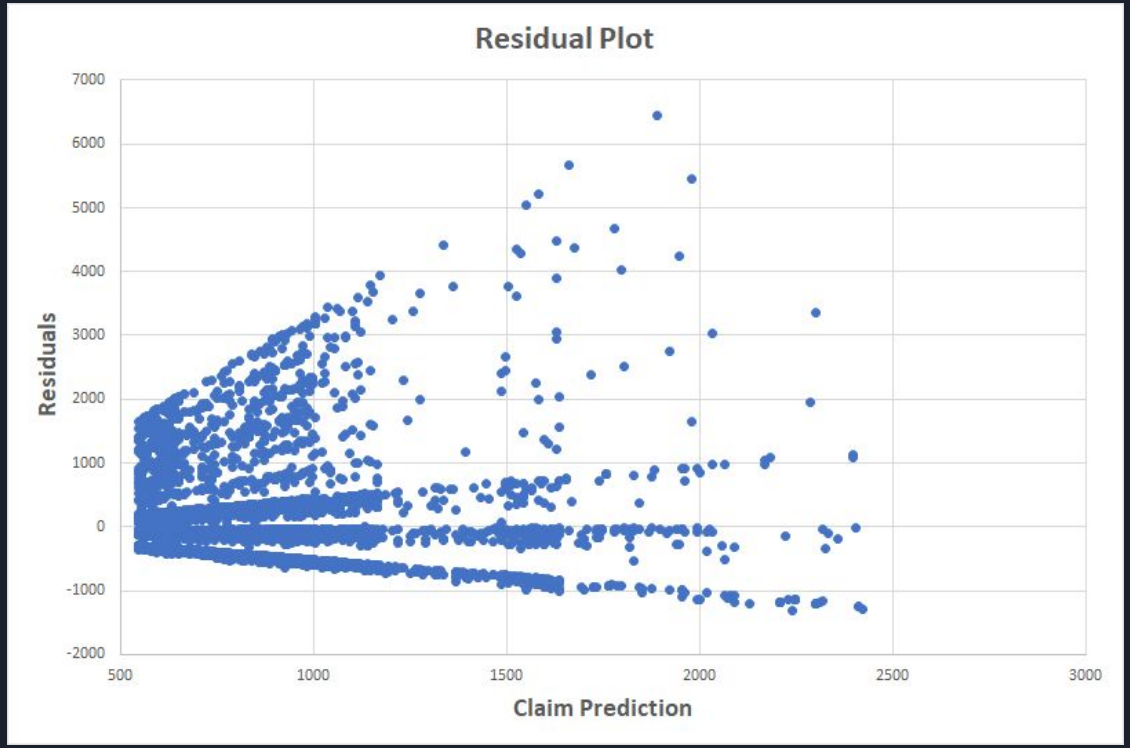
Linear Regression Model

- Multivariate Regression Accuracy: 0.28
- Monthly Premium is the most impactful predictor



Residuals

- Not very effective for predicting claim values.
- Multivariate regression only produced slightly better results
- Need more data to be able to predict this
  - Value of vehicle
  - Severity of damage

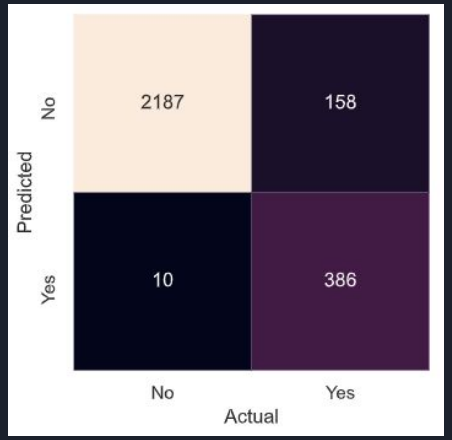


Classification Models: Confusion Matrices

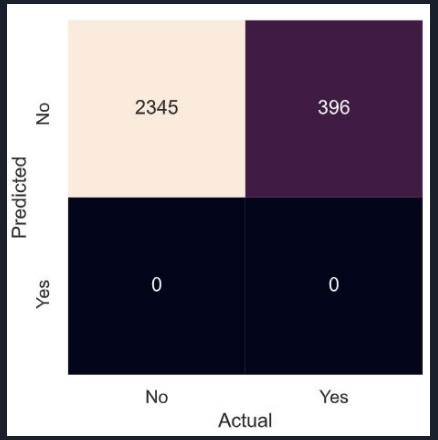
Performance of Algorithms:

- K-Nearest Neighbor: 0.94 Accuracy
- Naive Bayes: 0.86 Accuracy
- Random Forest: 0.98 Accuracy
- SVM: 0.83 Accuracy

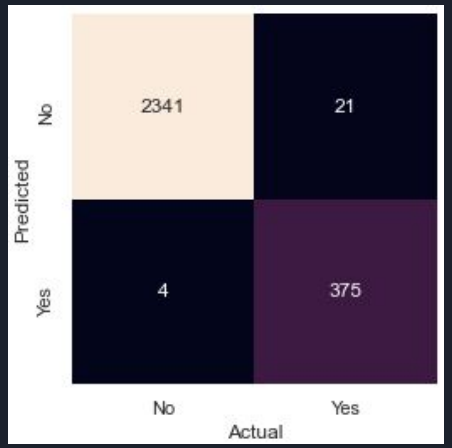
For our final classification algorithm, we used Random Forest.



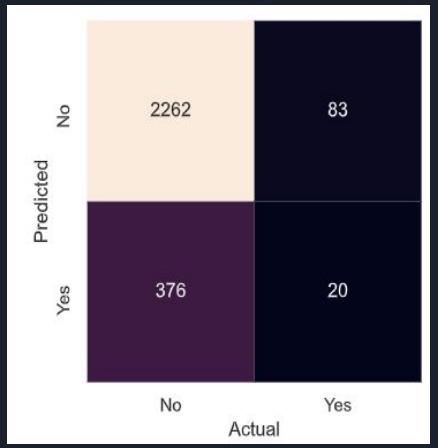
K-Nearest Neighbor



Naive Bayes



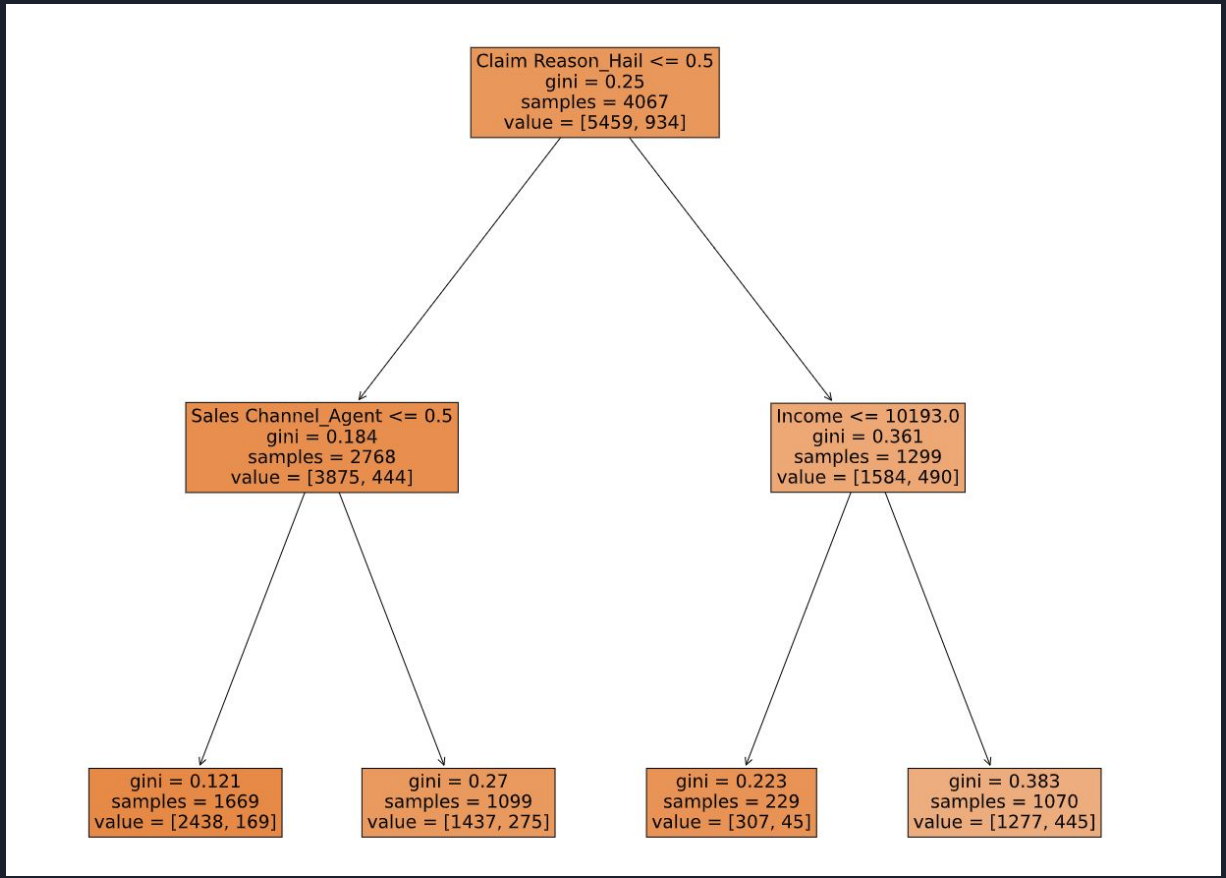
Random Forest



SVM

Classification: Random Forest

- Very effective for predicting if Claim will be accepted or denied.
- Most important factors
  - Claim Reason Hail
  - Sales Channel Agent
  - Income



Limitations

- Only had access to Fatal Crash Data and overarching data for important factors such as weather and types of vehicles
- Machine Learning data was oriented to Classification
- Census data is aggregated across a large population but auto insurance is personalized for each customer
- We did not have access to any data on many of the most important factors

Summary of Findings

- Factors that cause fatal crashes:
  - Younger Drivers, Snowy/Icy Weather Conditions, Motorcycles
- Rural roads cause more fatal accidents than urban roads and play as an important factor for predicting them.
- On a state level, there were no strong correlations between demographic data and insurance premiums. Only a few factors, such as Urban Population Percentage, showed meaningful correlation.
- Some of the trends on a state level contradicted expectations
  - Negative correlation between Insurance Premiums and Male Population Percentage
  - Negative correlation between Insurance Premiums and Accident Rate
- Our machine learning data was ineffective for predicting claim cost, however was more than adequate for predicting whether a claim would be accepted or denied.

Summary of Findings

- We were able to find a few factors in our data that were useful in predicting car accidents and insurance premiums on a state or national level
- Machine learning performed well for classification, but was missing important factors for regression.
- Future work should focus on identifying data for a wider range of factors and incorporating that data to make better predictions.

Thank you for listening! Questions?