Analysis of Factors that Impact Alcohol-related Car Accidents

Tong Wu

5/01/2021

1. Introduction

1.1 Background

Alcohol-related car accidents are a major problem in the United States. In 2016, 10,497 people died of alcohol-related car accidents and accounted for 28% of all traffic-related deaths. Extensive studies have investigated how the odds of alcohol-related vehicular crashes are associated with factors such as race, age, car types, and socioeconomic status. Younger drivers are associated with an increase in risk of alcohol-related accidents in both sexes [1]. Alcohol-related accidents are also more common among male drivers than female drivers. [2] Latino drivers have consistently had a higher prevalence of being involved in drunk driving crashes. [3] RAM 2500, Chevy S-10, and BMW 4-series were found to have the most DUIs according to data from Insurance company Insurify. [4] Furthermore, lower socioeconomic status was associated with increased risks of drunk driving. [5] However, there hasn't been a definitive study that, to our knowledge, accounted for all of these factors, along with other factors such as weather and road conditions, specifically in alcohol-impaired motor crashes.

The aim of our study is to determine significant factors that are associated with an increase in the probability of alcohol-related car accidents in California through constructing a statistical model. We are interested in the following questions:

1. Are certain car characteristics, such as the model year, car type, nationality of the car brand, whether the car is a premium brand or classified as a sports car, associated with increased odds of alcohol-related accidents?
2. Are demographic factors such as race, sex, and age significantly associated with the probability of alcohol-related accidents?
3. Is insurance status associated with lower risks of drunk-driving accidents, and how does this association vary by race, sex, age, and whether a driver has a premium car or a sports car?

We want to examine these factors while also controlling for weather conditions, crash types, and county population. We hope that our analysis will provide helpful insights for insurance companies and policymakers to develop solutions to combat alcohol-related car accidents.

1.2 Data Information

The dataset we used was from the California Traffic Collision Data from Statewide Integrated Traffic Records System (SWITRS) [6]. We only focused on automobile accidents that took place in 2020, which totaled 2,741,357 events. The dataset contains various information about each accident, including characteristics of the driver, insurance status, car involved, whether alcohol was involved, county in which the crash took place, as well as the weather and road conditions during the crash. We decided to focus only on crash events where the driver was at fault, because we were interested in understanding what factors about the driver are associated with increased rates of alcohol-related car accidents. Since all the observations only occurred in

the state of California, the insights we gain from our analysis will not be generalizable to areas of the US with very different traffic laws, weather conditions, and economy/weath.

1.3 Exploratory Data Analysis

Our response, or dependent variable, is a binary variable of whether the accident was alcohol-related or not.

To get a better understanding of vehicle characteristics associated with drunk driving crashes for the first question, we examined the vehicle's model year, type, nationality of the vehicle brand, whether the vehicle was premium, and whether the vehicle was classified as a sports car. The nationality of the vehicle brand was grouped into American, Asian, and European. Cars classified as premium included brands such as Mercedes, Lexus, and Audi. Cars classified as sports cars included brands such as Ferrari, Lamborghini, and Aston Martin. Vehicle type was simplified from the original dataset by classifying vehicles as only passenger cars, trucks, two-wheeled vehicles such as motorcycles, buses, and other vehicle types.

To examine demographic characteristics about the driver for the second question, we examined the age, race, and sex of the drivers at fault in the crash and determined whether certain characteristics are significantly associated with higher rates of alcohol-related accidents.

For the third question, we examined whether the driver is insured or not, and determined whether it is significantly associated with drunk crashes. In addition, we are interested in determining if significant two-interactions exist between insurance status and race, sex, age, and premium car status because insurance coverage rates could vary by these factors.

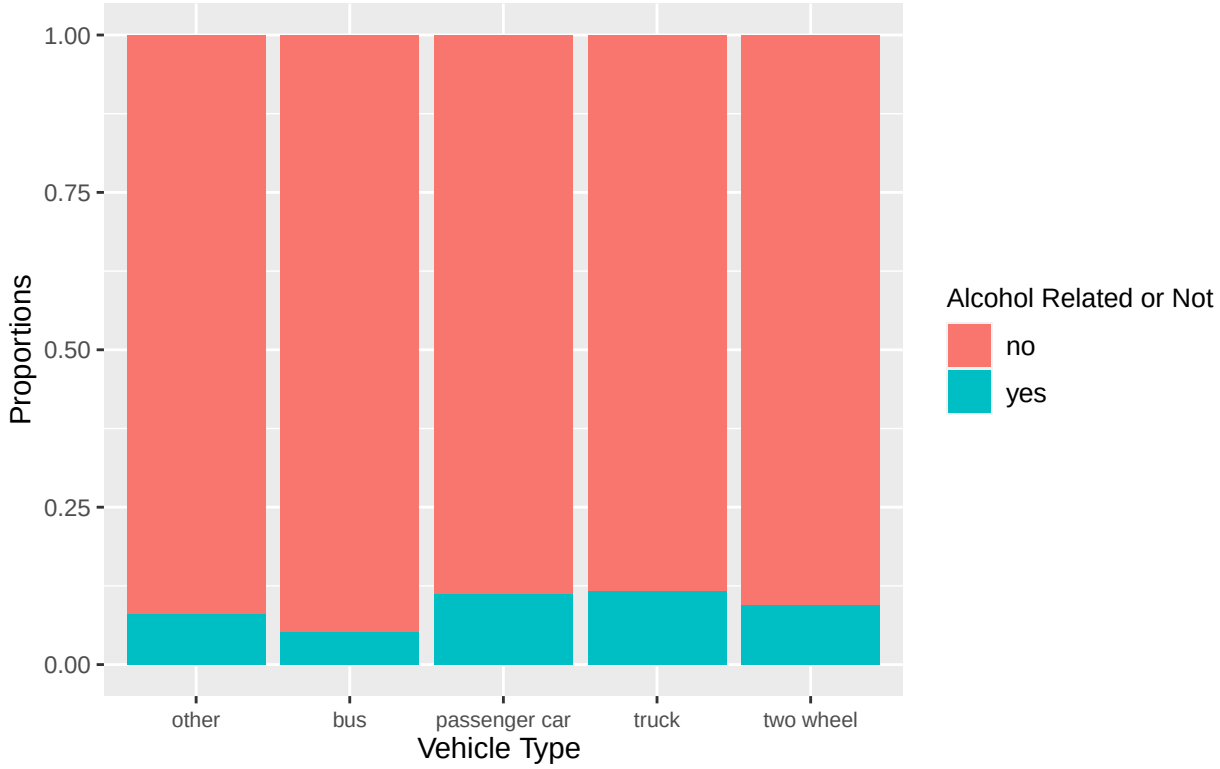
Finally, we also controlled for environmental factors, such as weather and road surface conditions, that can affect general car crash rates in our model. In addition, we also controlled for county population because it is reasonable to assume that higher density counties will have more accidents and thus more alcohol-related vehicle crashes.

1.3 Exploratory Data Analysis

To visualize how the rate of alcohol related accidents varies across the variables of interest, we generated stacked bar plots. Below are plots:

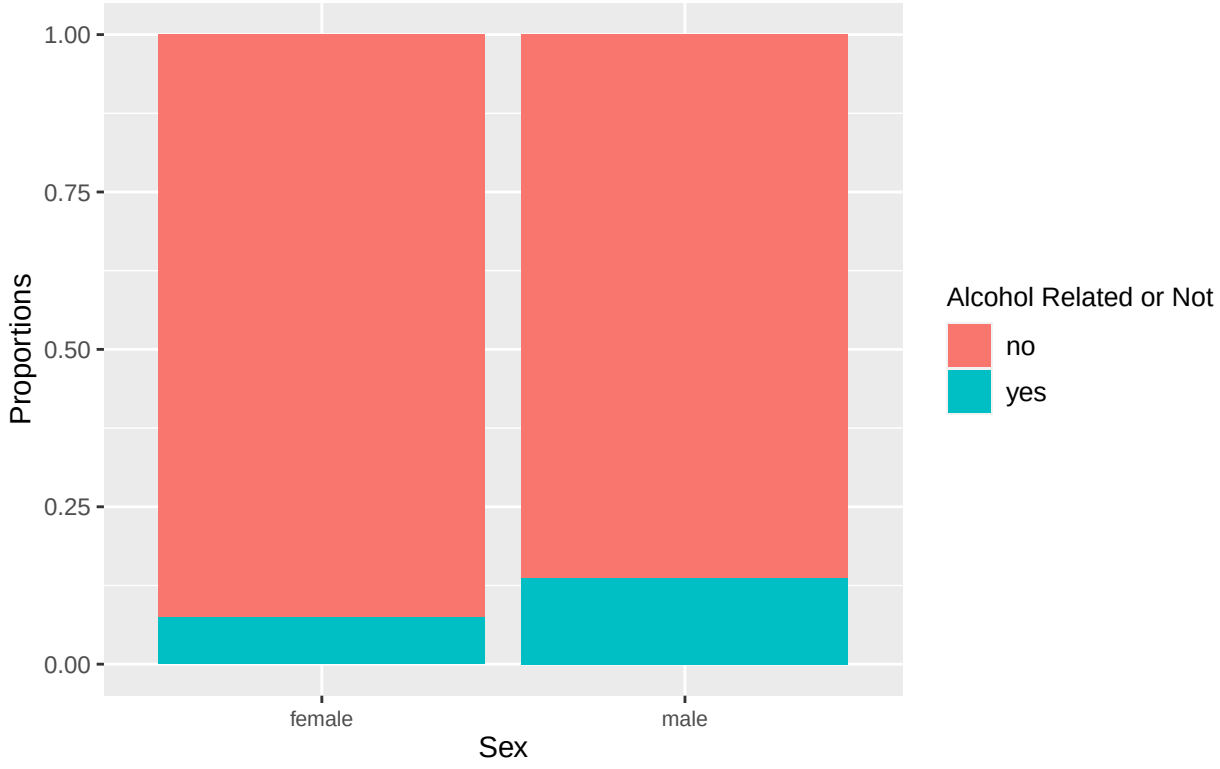
Vehicle Type vs Alcohol−related accident

Passenger cars, trucks, and Two−wheel vehicles have higher rates of Drunk−driving crashes



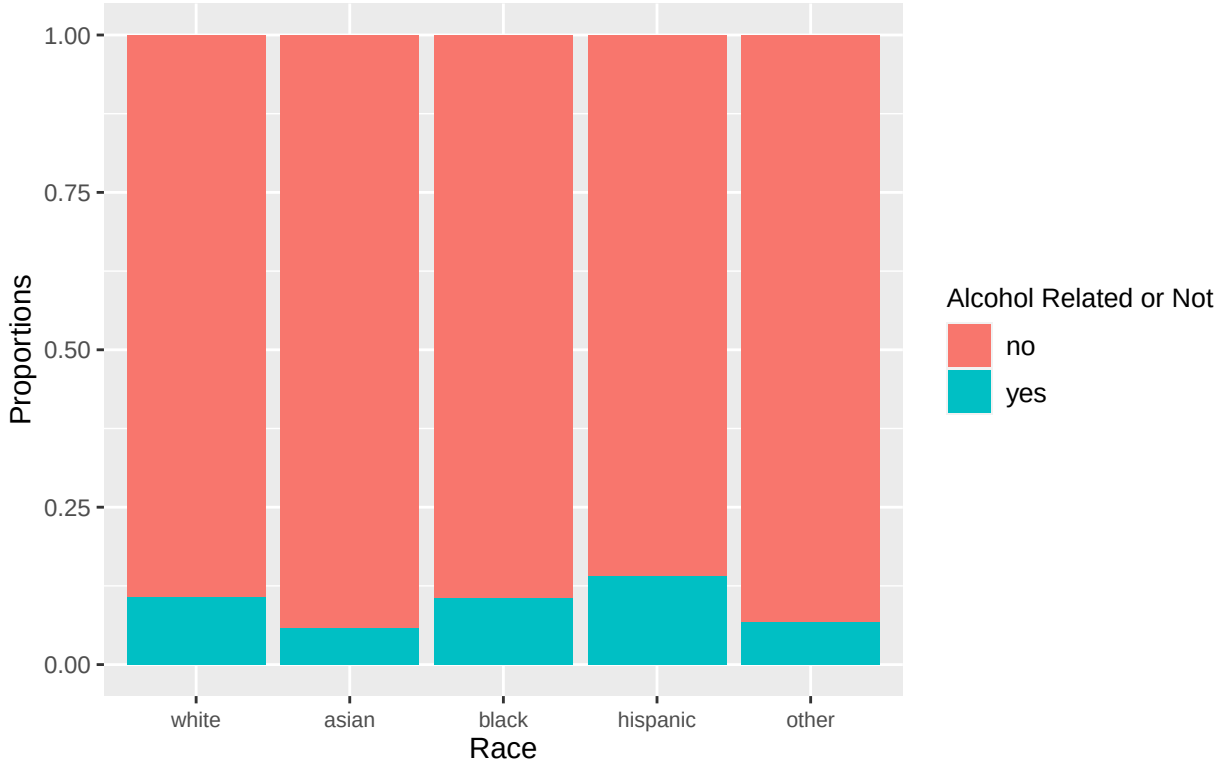
Sex vs Alcohol−related accident

Male drivers have higher rates of alcohol related accidents.



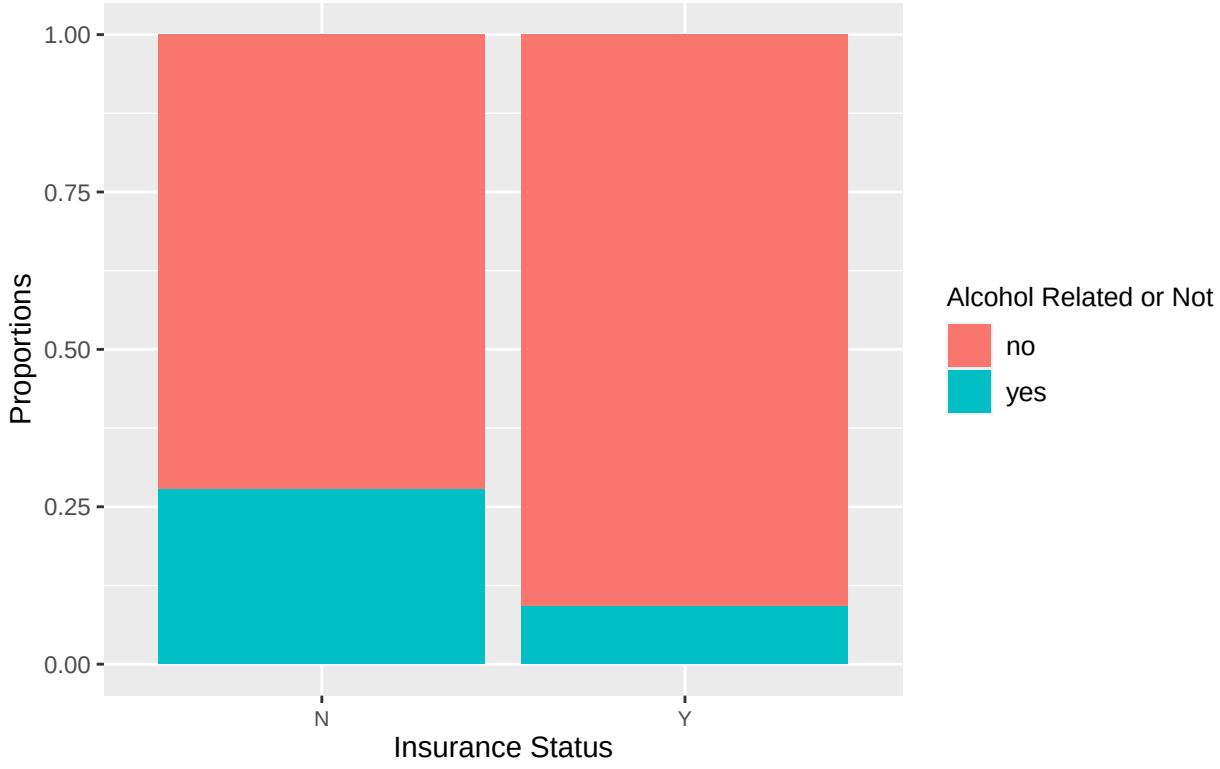
Race vs Alcohol−related accident

Hispanics have the highest rate of alcohol−related accidents compared to that of other races.



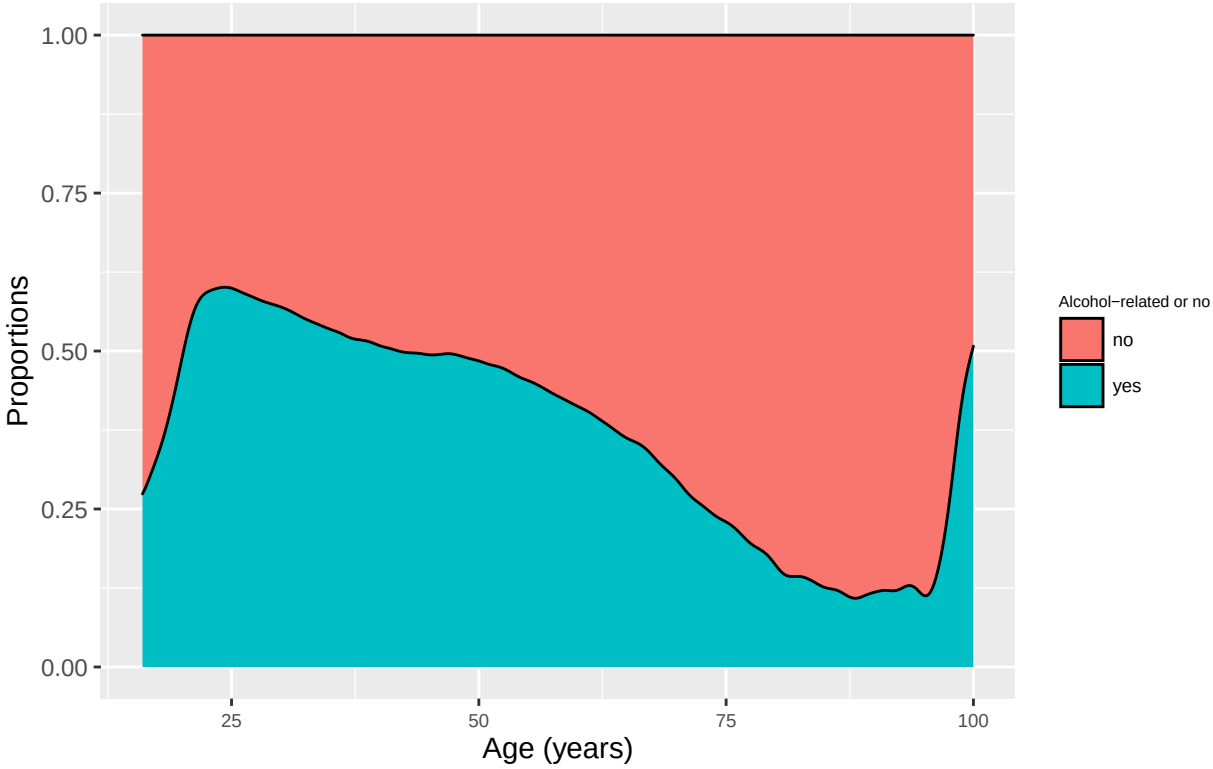
Insurance Status vs Alcohol−related accident

Insured drivers had lower rates of drunk driving accidents.



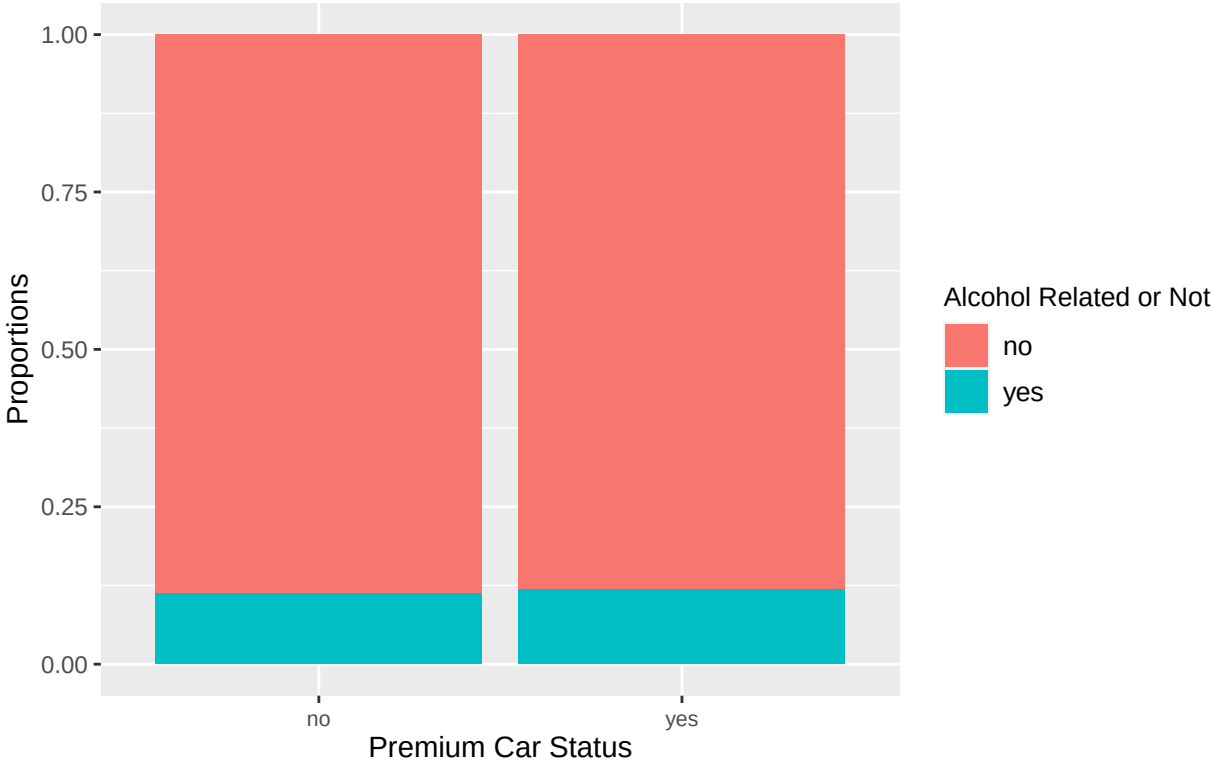
Age vs Alcohol−related Accidents

Drunk Crash rates generally decreases with age until around 90



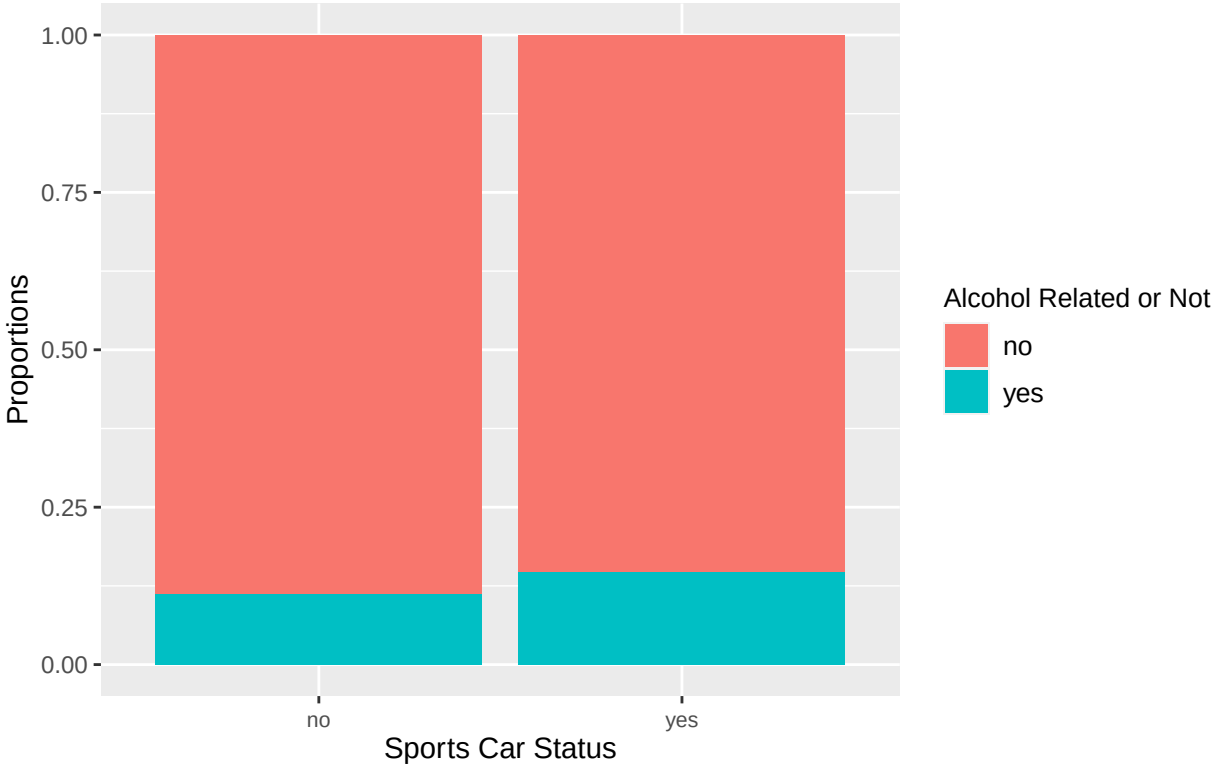
Premium Car Status vs Alcohol−related accident

Drunk crash rates are similar between premium and non−premium cars.



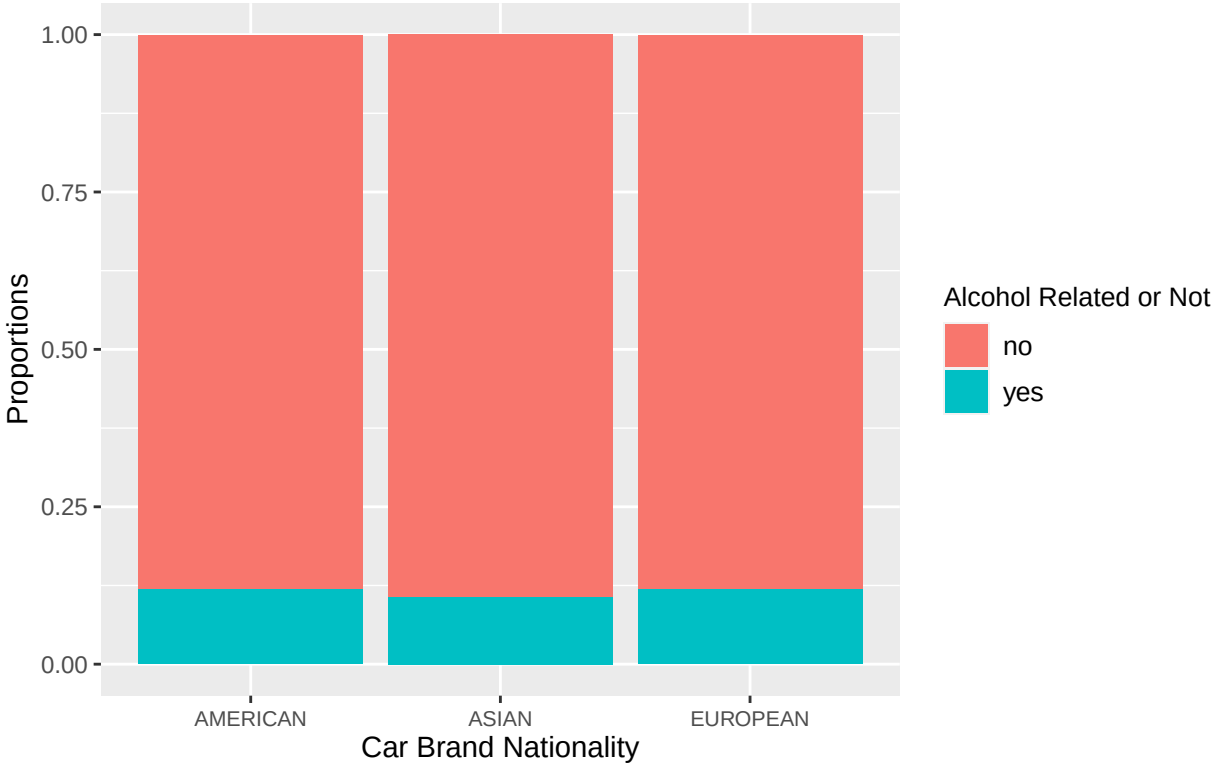
Sports Car Status vs Alcohol−related accident

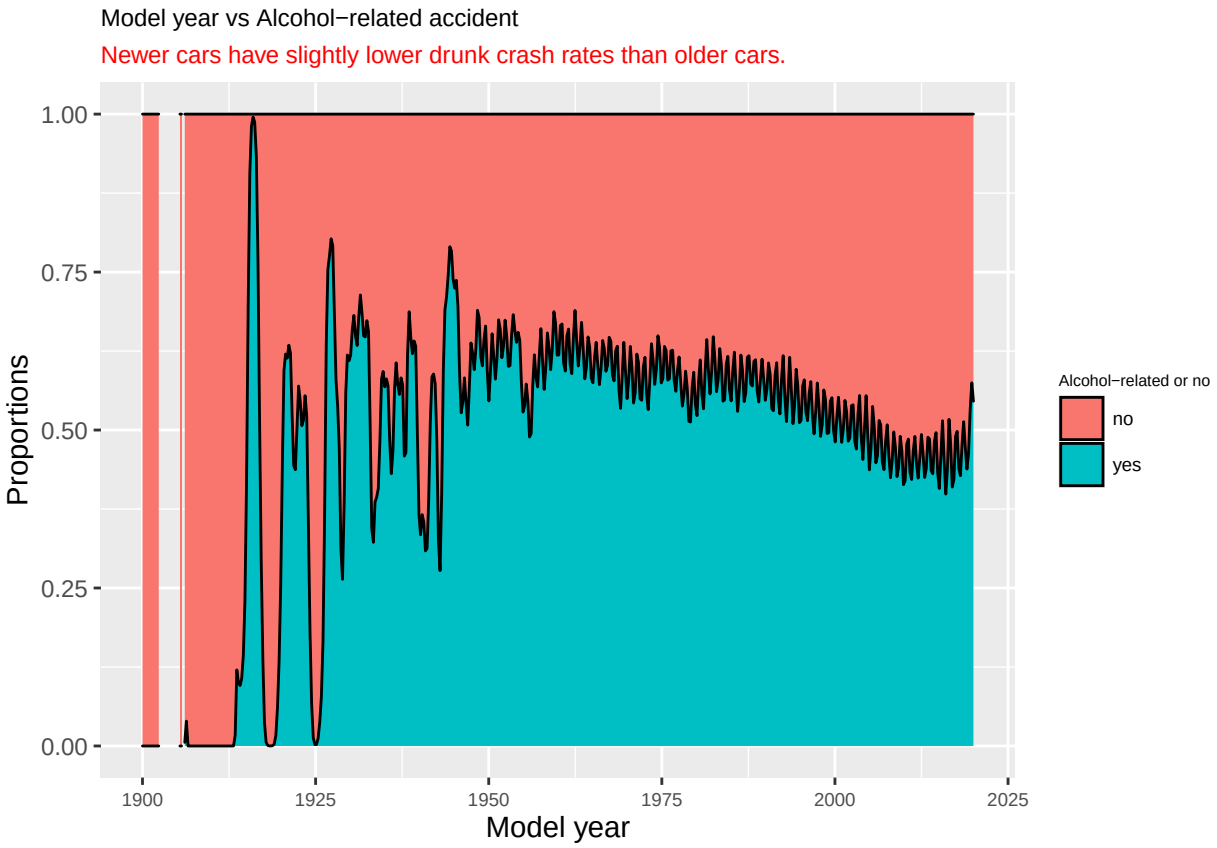
Drunk crash rates are slightly higher in sports cars than non−sports cars.



Car brand nationality vs Alcohol−related accident

Drunk crash rates are similar across car brand nationalities.





2. Methodology

2.1 Model Selection

Our goal was to develop an interpretable model that incorporated all the demographic, vehicle, and weather factors that we were interested in to determine which factors are significantly associated with higher/lower probability of alcohol-related car accidents in the context of all the predictors. Given these considerations, we decided that logistic regression was the best choice for our model. It provides interpretable coefficients that can help us determine which factors are significantly associated with an increase or decrease in probability of drunk automobile crashes. Furthermore, since our response variable is a binary variable indicating whether a car crash was alcohol-related or not, the logit-link from the logistic model is the most appropriate. A linear regression model, while also interpretable, would not allow our response variable to be bound between 0 and 1, which is inappropriate for analyzing our probability of interest.

2.2 Model Formulation

We propose the following generalized linear model (GLM) below:

$$
\begin{aligned}
\text{logit}(\mu_i) =& \beta_0 + \beta_1 DriverAge_i + \beta_2 I(VehicleType_i) + \beta_3 I(Race_i) \\
&+ \beta_4 I(Sex_i = Male) + \beta_5 I(PremiumCar_i = Yes) \\
&+ \beta_6 I(Insurance_i = Yes) + \beta_7 I(WeatherCondition_i) \\
&+ \beta_8 I(CollisionType_i) + \beta_9 (CountyPopulation_j) + \beta_{10} I(SportsCar_i = Yes) \\
&+ \beta_{11} I(Insurance_i = Yes) * I(Race_i) \\
&+ \beta_{12} I(Insurance_i = Yes) * I(Sex_i = Male) \\
&+ \beta_{13} I(Insurance_i = Yes) * DriverAge_i \\
&+ \beta_{14} I(Insurance_i = Yes) * I(PremiumCar_i = Yes) \\
&+ \beta_{15} I(Insurance_i = Yes) * I(SportsCar_i = Yes) \\
&+ \beta_{16} * (VehicleModelYear_i) \\
&+ \epsilon_i \\
\mu_i \sim Ber(\mu), & \\
\epsilon_{ij} \sim N(0, \sigma^2) &
\end{aligned}
$$

From our model shown above, i represents each crash event, j represents the index for the county where the crash event took place, and μ_i is the probability of the crash event being alcohol-related. The levels for vehicle type are: passenger car, truck, two-wheel vehicles, bus, and other vehicle types (baseline). The levels for race are: White (baseline), Black, Asian, Hispanic, and Other. The levels for Car Nationality are: Asian, European, and American (baseline). The levels for Weather Conditions are: Clear (baseline), cloudy, foggy, other, raining, snowing, and wind. The levels for Collision Type are: Broadside (Baseline), Head-on, hit object, other, overturned, pedestrian, rear end, and sideswipe.

2.3 Diagnostics and Assumptions Check for Model

The independence of observations assumption was met because every observation comes from one crash event. We checked for constant variance and lack of pattern in model residuals vs. fitted values by graphing a binned residual plot. To check for lack of multicollinearity, we calculated the variable inflation factor (VIF) values for all predictors in our initial model and made sure the VIF values are all less than 10. We also used drop-in deviance tests to assess whether interaction effects in our model were significant. Regardless of whether these predictors and interactions were significant or not, they were incorporated in our model for the reasons mentioned in the Variable of Interest section. Model diagnostic results can be found in the Appendix Section.

2.4 Sensitivity Analysis

We initially wanted to account for systematic variations that might exist at the county level in our model. Our rationale was that without controlling for county-level effects, the independence assumption might be violated because certain county-level characteristics that we do not have data on could contribute to variations in the rate of alcohol-related crashes. This variation could lead to different alcohol-related crashes based on location rather than the other factors we are interested in. To account for this, we added a county-level random effect to generate a mixed effects model. However, after generating the model, we observed the intraclass correlation coefficient (ICC) to be only 0.03. This value was very low and indicated that only 3% of the total variance in the probability of alcohol-related crashes was accounted for by county-level clustering. This means that the systematic variations at the county level were not very prominent in our model. Thus, we decided not to incorporate it and proceeded with regular logistic regression. Results for the mixed model can be found in the appendix.

3. Results

Term	Coefficient Estimate	p-value
Intercept	-2.632	<0.001
Sex: Male	0.578	<0.001
Insurance: Yes	-0.704	<0.001
Driver Age	0.005	<0.001
Race: Asian	-0.286	<0.001
Race: Black	-0.388	<0.001
Race: Hispanic	0.160	<0.01
Race: Other	-0.297	<0.001
Vehicle Model Year	-0.000	0.841
Vehicle Type: Bus	-0.150	0.156
Vehicle Type: Passenger Car	0.593	<0.001
Vehicle Type: Truck	0.463	<0.001
Vehicle Type: Two Wheel	-0.253	0.008
Premium Car: Yes	-0.020	0.129
Car Nationality: Asian	0.007	0.121
Car Nationality: European	0.094	<0.001
Sports Car: Yes	0.383	0.002
County Population	-0.014	<0.001
Weather Condition: Cloudy	-0.294	<0.001
Weather Condition: Foggy	0.076	0.008
Weather Condition: Other	-0.233	0.003
Weather Condition: Raining	-0.288	<0.001
Weather Condition: Snowing	-1.260	<0.001
Weather Condition: Windy	0.087	0.438
Collision Type: Head-On	1.062	<0.001
Collision Type: Hit Object	1.544	<0.001
Collision Type: Other	0.500	<0.001
Collision Type: Overturned	1.222	<0.001
Collision Type: Pedestrian	0.2705	<0.001
Collision Type: Rear-End	0.253	<0.001
Collision Type: Side-swipe	0.765	<0.001
Insurance:Yes*Sex: Male	0.007	0.522
Insurance:Yes*Driver Age	-0.013	<0.001
Insurance:Yes*Race: Asian	-0.346	<0.001
Insurance:Yes*Race: Black	0.261	<0.001
Insurance:Yes*Race: Hispanic	0.003	0.787
Insurance:Yes*Race: Other	-0.317	<0.001
Insurance:Yes*Premium Car: Yes	0.135	<0.001
Insurance:Yes*Sports Car: Yes	-0.209	0.118

4. Discussion

4.1 Interpretation

Question 1 - Vehicle Characteristics

After accounting for road/weather conditions and crash types, we found that vehicle type, vehicle brand nationality, and sports car status all had significant associations with the odds of a crash event being alcohol-related. A vehicle's model year and premium car status did not have significant associations with the response.

With regards to vehicle types, passenger car, truck, and two wheeled vehicles have significant positive associations with odds of drunk driving crashes relative to other vehicle types while buses did not have a significant association. The expected odds of a crash event being alcohol-related multiply by 1.81, 1.58, and 0.78 for passenger cars, trucks, and two-wheeled vehicles, respectively, relative to those of other vehicle types and holding all else constant. In addition, holding all else constant, the expected odds of a crash event being alcohol-related multiply by 0.86 for buses relative to those of other vehicle types, although this effect was not significant. Overall, these results suggest that passenger cars and trucks have the most positive association with alcohol-related accident rate, while two-wheeled vehicles like motorcycles, and other vehicle types have smaller magnitude associations with the response. Insurance companies should carefully consider drivers with passenger cars and trucks when evaluating their risk for alcohol-related crashes and potentially charge them at higher premiums. Policymakers should examine what factors could be making passenger and truck drivers more prone to drunk driving.

With regards to car brand nationality, relative to American cars, European cars had a significant positive association with odds of drunk driving crashes while Asian cars did not have a significant association. Holding all else constant, the expected odds of a crash event being alcohol-related multiply by 1.09 for European cars relative to those of American cars. In addition, holding all else constant, the expected odds of a crash event being alcohol-related multiply by 1.01 for Asian cars relative to those of American cars, although this effect was not significant. Overall these results suggest that European cars are associated with a higher probability of alcohol-related crashes relative to those of American and Asian cars, which were similar. Insurance companies should consider this factor when evaluating a drivers' risk for alcohol-related crashes and potentially charge European car owners higher premiums. Policymakers should examine what characteristics of European cars and its owners could potentially increase the probability of drunk driving.

Sports car status was found to have a significant positive association with odds of alcohol-related crashes, while premium car status did not have significant association with the response. Holding all else constant, the expected odds of a crash event being alcohol-related multiplies by 1.47 for sports cars relative to those of non-sports cars. In addition, holding all else constant, the expected odds of a crash event being alcohol-related multiplies by 0.98 for premium cars relative to those of non-premium cars, although this effect was not significant. Overall, these results suggest that insurance companies should charge higher premiums for drivers with sports cars and policymakers should research what actions or characteristics of sports car drivers could be increasing their association with probability of drunk crashes.

Question 2 - Driver Demographic Characteristics

After accounting for all other variables, race, age, and sex were all found to have significant associations with odds of drunk crashes. Holding all else constant, the expected odds of a crash event being alcohol-related multiplies by 1.78 for male drivers relative to those of female drivers. This suggests that male drivers are associated with higher probability of drunk driving crashes than female drivers. With regards to age, for every one year increase in age, the expected odds of alcohol-related crash events multiply by 1.01, holding everything else constant.

Regarding race, the expected odds of alcohol-related crash events multiplies by 0.75, 0.67, 1.16, and 0.74, for Asians, Blacks, Hispanics, and Other ethnicity drivers, respectively, relative to those of white drivers and holding all else constant. These results indicate that White and Hispanic drivers are associated with higher probability of alcohol-related crashes relative to Asian, Black, and other ethnicity drivers. Policymakers should investigate why racial disparities exist in alcohol-related car crashes and come up with ways to eliminate this disparity.

Question 3 - Financial Responsibility and Interactions

After accounting for all other predictors in the model, insurance status was found to have a significant negative association with alcohol-related crash rate, and the effect was one of the largest in magnitude. For an insured, white individual who doesn't drive a premium car, the expected odds of alcohol-related accidents multiplies

by 0.48 relative to those of their uninsured counterparts, holding all else constant. This result indicates that uninsured drivers are associated with higher probability of alcohol-related crash events. However, the model output also indicates that an individual's race, age, and premium car status significantly impacts the magnitude of the effects that insurance status has on the probability of alcohol-related car crashes.

We found that race significantly impacts the effects of insurancec status on the response. Holding all else constant, the expected odds of alcohol-related car accidents multiplies by 0.35, 0.64, 0.50, and 0.36 for insured Asian, Black, Hispanic, and Other ethnicity drivers, respectively, relative to those of their uninsured counterparts. The interaction between insurance status and hispanic was not found to be significant. When comparing the magnitude of insurance status' effect for these races to that of white drivers (0.48), we see that insurance status is associated with larger decreases in the probability of alcohol-related accidents for asian and other races, smaller decrease for blacks, and similar decreases for hispanics relative to whites. These results suggest that there is a racial disparity in the magnitude of negative association between insurance status and the response. Policymakers and insurance companies should investigate what confounding factors might explain why insured asian and other ethnicity drivers seem to fare better than insured black, white, and hispanic drivers with regards to the probability of alcohol-related accidents.

In addition, a significant negative interaction was found between insurance status and age. Holding all else constant, for every one year increase in the driver's age, the odds of alcohol-related crashes is expected to multiply by 0.49 for an insured individual. The direction of this association is opposite that of age's effect when the individual is uninsured (1.01). This result suggests that age is positively associated with probability of alcohol-related crash when the driver is uninsured and negatively associated when the driver is insured. Thus, policymakers should encourage older drivers to be insured and improve their chances of avoiding alcohol-related accidents.

Furthermore, we also observed that insurance status' effects on alcohol-related accident rates vary significantly by premium car status as well. The expected odds of alcohol-related crashes multiply by 0.57 for insured premium car drivers compared to those of their uninsured counterparts. Interestingly, the main effect for premium car was found to be not significant, however this interaction was significant. This result suggests that premium car drivers who are insured are associated with a lower probability of alcohol-related accidents relative to those who are premium car drivers but are uninsured. However, when comparing the magnitude of this association with that of an insured driver who doesn't drive a premium car (0.48), we see that being insured is associated with a smaller decrease in alcohol-related crash odds for premium drivers. Therefore, policymakers should encourage premium car drivers to be more insured to and insurance companies should consider charging higher premiums for premium car drivers when evaluating their risk of drunk driving crashes.

Finally, interactions between insurance status and sex, sports car status was also examined, however, these were not significant.

4.2 Critique on Methodology, Alternate Model Specification, Limitations and Future Directions

An alternative model specification that could potentially improve our current model would be to account for spatial dependence. While our sensitivity analysis demonstrated that there was no prominent systematic variation at the county level in terms of alcohol-related crash event rate, this does not mean that county-level spatial correlation does not exist. It is very possible that certain clusters of counties have higher or lower rates of alcohol-related accidents than average due to confounding factors specific to those counties, which are not present in the dataset. Thus, it would have been valuable to examine whether spatial effects do exist at the county level. Since our data does not include spatial data for the counties, this highlights a limitation in our dataset and analysis. A future direction would be to collect spatial data for the California counties, run the Moran I test to assess spatial dependency, and potentially fit a spatial error or spatial lag model on the same data.

Another potential limitation is the lack of data about county-level average alcohol consumption and in-

come/average wealth. It is reasonable to surmise that counties with higher alcohol consumption might have higher rates of alcohol-related car accidents. Furthermore, it has been shown that lower income is associated with higher rates of drunk driving and lower rates of insurance coverage [5]. Therefore, it is also reasonable to surmise that counties with lower average income could be associated with higher rates of alcohol-related accidents as well. Both of these county-level information could be confounding factors that we did not account for in our analysis. Thus, a future direction would be to gather county-level data regarding alcohol consumption and average income and incorporate these predictors in the model.

4.3 Conclusion

Despite the limitations, our analysis improved upon previous work by incorporating vehicle characteristics, driver's demographic factors, and insurance status all in one model, as well as controlling for crash type, county population, and weather conditions. Our analysis demonstrated that numerous demographic variables, vehicle characteristics and insurance status are significantly associated with the probability of drunk driving crashes. Furthermore, we also found interesting interactions which indicate that insurance status' effects on the response varies by race, age, and premium car status. We hope that these insights will help insurance companies and policymakers better assess the risks of alcohol-related accidents for certain individuals and develop solutions to combat the drunk car accident crisis.

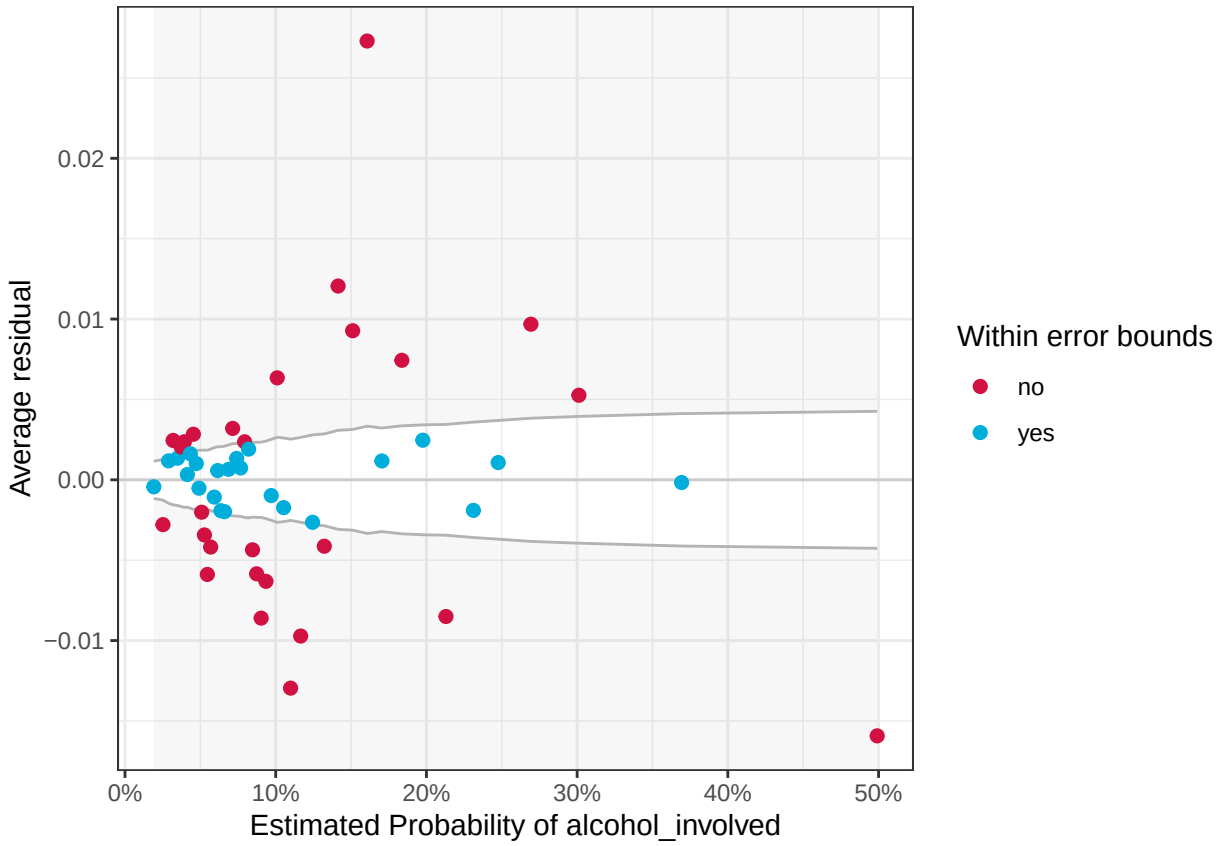
5. Appendix

5.1 Model Diagnostics

5.1.1 Binned Residual Plot for Logistic Regression Model

Binned residuals vs fitted values plot shows a lack of pattern.

```
## Warning: Probably bad model fit. Only about 46% of the residuals are inside the error bounds.
```



5.1.2 Multicolinearity Check

VIF values are displayed below. All VIF values are under 10, indicating there was no issues with multicolinearity.

	GVIF	Df	GVIF^(1/(2*Df))
party_sex	1.059539	1	1.029339
party_age	1.071112	1	1.034945
financial_responsibility	1.075635	1	1.037128
party_race	1.136301	4	1.016101
vehicle_year	1.062791	1	1.030918
vehicle_type_sort	1.236333	4	1.026873
premiumcar	1.571001	1	1.253396
carnation	1.759030	2	1.151644
sportcar	1.022064	1	1.010972
population	1.059326	1	1.029235
weather_1	1.026725	6	1.002200
type_of_collision	1.120828	7	1.008181

5.1.3 Drop-In Deviance Test

Drop-in-Deviance Test Results

Interactions	Chi-sq p-value
Insurance Status:Sex	0.522
Insurance Status:Race	<0.001
Insurance Status:Age	<0.001
Insurance Status:Premium Car	<0.001
Insurance Status:Sports Car	0.120

We see from the Drop-In Deviance Test Results that Insurance Status has significant interactions with race, age, and premium car status.

5.2 Sensitivity Analysis (Generalized Linear Mixed Effects Model)

Termin	Coefficient Estimate	p-value
Intercept	-3.077	<0.001
Sex: Male	0.578	<0.001
Insurance: Yes	-0.693	<0.001
Driver Age	0.005	<0.001
Race: Asian	-0.220	<0.001
Race: Black	-0.313	<0.001
Race: Hispanic	0.220	<0.01
Race: Other	-0.260	<0.001
Vehicle Model Year	0.000	0.606
Vehicle Type: Bus	-0.221	0.037
Vehicle Type: Passenger Car	0.585	<0.001
Vehicle Type: Truck	0.457	<0.001
Vehicle Type: Two Wheel	-0.295	0.002
Premium Car: Yes	-0.017	0.186
Car Nationality: Asian	0.009	0.050
Car Nationality: European	0.085	<0.001
Sports Car: Yes	0.397	0.002
County Population	-0.015	0.051
Weather Condition: Cloudy	-0.308	<0.001
Weather Condition: Foggy	0.055	0.055
Weather Condition: Other	-0.270	<0.001
Weather Condition: Raining	-0.309	<0.001
Weather Condition: Snowing	-1.286	<0.001
Weather Condition: Windy	0.073	0.517
Collision Type: Head-On	1.057	<0.001
Collision Type: Hit Object	1.564	<0.001
Collision Type: Other	0.519	<0.001
Collision Type: Overturned	1.267	<0.001
Collision Type: Pedestrian	0.209	<0.001
Collision Type: Rear-End	0.280	<0.001
Collision Type: Side-swipe	0.789	<0.001
Insurance:Yes*Sex: Male	0.008	0.492
Insurance:Yes*Driver Age	-0.013	<0.001
Insurance:Yes*Race: Asian	-0.338	<0.001
Insurance:Yes*Race: Black	0.253	<0.001
Insurance:Yes*Race: Hispanic	0.003	0.809
Insurance:Yes*Race: Other	-0.323	<0.001
Insurance:Yes*Premium Car: Yes	0.130	<0.001

Termin	Coefficient Estimate	p-value
Insurance:Yes*Sports Car: Yes	-0.240	0.075

ICC: 0.03

6. References

1. Zador PL, Krawchuk SA, Voas RB. Alcohol-related relative risk of driver fatalities and driver involvement in fatal crashes in relation to driver age and gender: an update using 1996 data. Journal of Studies on Alcohol 2000; 61:387-395.
2. https://pubs.niaaa.nih.gov/publications/arh27-1/63-78.htm#:~:text=Forty%E2%80%93six%20percent%20of%20male,pe
3. Cerdá, M., Diez-Roux, A. V., Tchetgen, E. T., Gordon-Larsen, P., & Kiefe, C. (2010). The relationship between neighborhood poverty and alcohol use: estimation by marginal structural models. Epidemiology (Cambridge, Mass.), 21(4), 482–489. https://doi.org/10.1097/EDE.0b013e3181e13539
4. https://insurify.com/insights/car-models-most-duis-2020/
5. Antti Impinen, Pia Mäkelä, Karoliina Karjalainen, Jari Haukka, Tomi Lintonen, Pirjo Lillsunde, Ossi Rahkonen, Aini Ostamo, The Association between Social Determinants and Drunken Driving: A 15-Year Register-based Study of 81,125 Suspects, Alcohol and Alcoholism, Volume 46, Issue 6, November-December 2011, Pages 721–728, https://doi.org/10.1093/alcalc/agr075
6. https://www.kaggle.com/alexgude/california-traffic-collision-data-from-switrs