ES218 Project: US Vehicular Accidents

Curtis Zhuang

5/4/2020

Table of Contents

[Introduction 2](#_Toc39667639)

[Methods 2](#_Toc39667640)

[Results 3](#_Toc39667641)

[Discussion 14](#_Toc39667642)

# Install packages  
library(stringr)  
library(tidyr)  
library(forcats)  
library(ggplot2)  
library(dplyr)  
library(tmap)   
#library(spData)  
#library(spDataLarge)  
library(sf)  
library(gridExtra)  
  
#install.packages('spDataLarge', repos='https://nowosad.github.io/drat/', type='source')

# Want to test if sex, car type,   
dat <- readRDS('../Data/farsp.RDS')  
  
# group the states into four main regions in U.S  
  
# Regroup and create region variable  
west <- c(2,4,6,8,15,16,32,35,41,49,53,56 )  
midwe <- c(17,18,10,20,26,27,30,31,38,39,46)  
south <- c(1,5,11,12,13,21,22,24,28,29,37,40,45,47,48,51,54)  
northea <- c(9,10,23,25,33,34,36,42,44,50,55 )  
oth <- c(43,52)  
  
# Regroup car type  
  
dat <- dat %>%   
 mutate('Region' = case\_when(state %in% west ~ 'West',  
 state %in% midwe ~ 'MidWest',  
 state %in% south ~ 'South',  
 state %in% northea ~ 'NorthEast',  
 state %in% oth ~ 'Other'))  
  
dat\_full <- dat %>%   
 select(-c(day, hour, minute, county)) %>%   
 # filter out severity  
 filter(inj\_sev %in% c(1,2,3,4,5,6))%>%   
 # filter out Virgin Islands and Puerto Rico  
 drop\_na(Region)

### Introduction

This study is conducted in light of the dataset recorded by the NHTSA for the years 1996 through 2016 on vehicular accidents. This is a large dataset as it contain many dimensions of an accident from the time the accident happens, the age of the driver caught in the accident, the location the accident, type of vehicle, etc. However, for this study, I only focused on the following variables:

1. Are there regional differences in traffic accidents?
2. Is the severity of the accidents affected by some of the variables that are included in this dataset?

In order to study the first question, I grouped the states into four main categories: West, MidWest, NorthEast and South (which is the most common way of grouping states). From there I created graphs to compare some of the selected variables between the four groups.

To study the second question, I used bivariate models to analyze if the severity of the accidents are related to the variables of my choice, namely: age, region, sex, collision type and people involved in the accident. From my study, I find that all the variables I choose are significant when it comes to severity.

The paper is structured as follows: the Methods section will touch briefly on what kind of graphs I made and what models I used; the Results section will focus on explaining the results shown by the graphs as well as interpreting the results from the models; the Discussion section will dig deeper into the findings of the results seciton.

### Methods

As discussed above, I followed the most common way of grouping states and came up with four groups. I excluded Puerto Rico and Virgin Island due to the scarcity of samples they have compared with other groups and their nature of being small islands which may cause issues. The original dataset has 1906184 entries.

I performed piping operations and removed the samples from Puertorico and Virgin Island after I finished grouping. Another group that I discarded in this study is the data entries reported as “Unknown”/“Not reported” which can be found for both the sex variable and the collision type variable. In total, they do not account for a large number and getting rid of these variables can help reduce the chance of having outliers. There are 1467635 samples left after these operations.

To study how the vehicular accidents can be affected by its location, I used univariate methods for some of the categorical variables and used bivariate methods for the quantative variables.

Using the graphs as guidance, I created two models. The first model has severity as the dependent variable, region and age as the independent variables. This models captures mainly how region can affect severity of the accident when age is controlled. The next model expands the scope of variables and studies more generally what variables can affect severity.

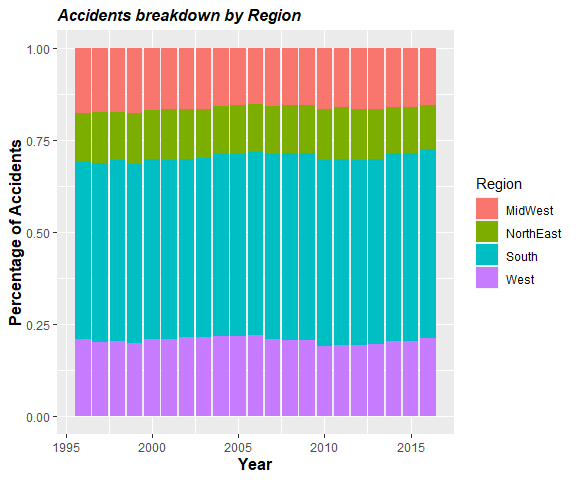
### Results

The results section looks at each variable following the order of: Region, Sex, Collision, Age, Number of ppl involved, and Severity.

#### Here I wan to Include a Map later on (Region)

#### Region

# Region by year  
dat\_reg <- dat\_full %>%  
 group\_by(year, Region) %>%   
 summarise(n = n()) %>%   
 ungroup() %>%   
 group\_by(year) %>%   
 mutate(percent = n/sum(n))%>%   
 drop\_na(Region)  
  
ggplot(dat\_reg, aes(year, percent, fill = Region)) +   
 geom\_bar(position = 'stack', stat = 'identity') +  
 xlab('Year') + ylab('Percentage of Accidents') +   
 ggtitle('Accidents breakdown by Region') +  
 theme(  
 plot.title = element\_text(color="black", size=12, face="bold.italic"),  
 axis.title.x = element\_text(color="black", size=12, face="bold"),  
 axis.title.y = element\_text(color="black", size=12, face="bold"))

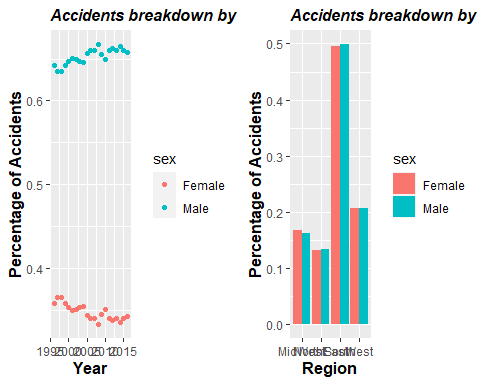


Based on the barchart showing distribution of accidents across the four regions, we can observe that the percentage of accidents for each region did not fluctuate a lot over the course of 22 years. Comparing the four regions, we see that the South has the most accidents which accounted for nearly 50%.

This makes intuitive sense as geographically, we can see that the South covers lots of lands and includes states like Texas, Maryland, Florida which are all populated states. The large portion of accidents happening in the south will affect a lot of the results we will get from the other graphs.

#### Sex

# Make a graph of how accidents happen related to sex  
# male vs female by year  
dat\_sex <- dat\_full %>%  
 group\_by(year, sex) %>%   
 summarise(n = n()) %>%   
 filter(sex %in% c(1,2)) %>%   
 mutate(sex = ifelse(sex == 1, "Male", "Female")) %>%   
 ungroup() %>%   
 group\_by(year) %>%   
 mutate(percent = n/sum(n))  
  
# plot the accident by sex  
ggplot(dat\_sex, aes(year, percent, col = sex)) + geom\_point() +  
 xlab('Year') + ylab('Percentage of Accidents') +   
 ggtitle('Accidents breakdown by sex') +  
 theme(  
 plot.title = element\_text(color="black", size=12, face="bold.italic"),  
 axis.title.x = element\_text(color="black", size=12, face="bold"),  
 axis.title.y = element\_text(color="black", size=12, face="bold")) ->p1  
  
# male vs female by region  
dat\_sex\_loc <- dat\_full %>%   
 group\_by(Region, sex) %>%   
 summarise(n = n()) %>%   
 filter(sex %in% c(1,2)) %>%   
 mutate(sex = ifelse(sex == 1, "Male", "Female")) %>%   
 ungroup() %>%   
 group\_by(sex) %>%   
 mutate(percent = n/sum(n))  
  
# univariate  
ggplot(dat\_sex\_loc, aes(Region, percent, fill = sex)) +   
 geom\_bar(position = 'dodge', stat = 'identity') +  
 xlab('Region') + ylab('Percentage of Accidents') +   
 ggtitle('Accidents breakdown by Region and Sex') +  
 theme(  
 plot.title = element\_text(color="black", size=12, face="bold.italic"),  
 axis.title.x = element\_text(color="black", size=12, face="bold"),  
 axis.title.y = element\_text(color="black", size=12, face="bold")) -> p2  
  
grid.arrange(p1, p2, nrow = 1)

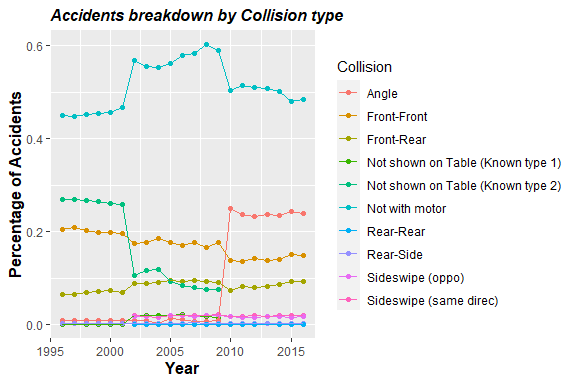


Here we consider how sex of the driver can contribute to the occurence of vehicular accident. From the graph on the left, we observe a very constant fluctuation over time. About 65% of the accidents are caused by male drivers and about 35% of the accidents are caused by female drivers.

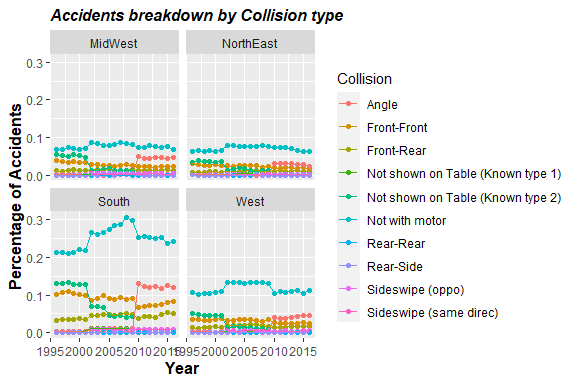
However, when we observe the graph on the right, we did divergence in sex when we neglect the time dimension and only studies region. So, I suppose that there is no divergence when it comes to the effect of sex of driver on the accident between the regions.

#### Collision

# collision by year  
  
dat\_col <- dat\_full %>%  
 group\_by(year, man\_coll) %>%   
 summarise(n = n()) %>%   
 mutate(Collision = case\_when(man\_coll == "0" ~ 'Not with motor',  
 man\_coll == "1" ~ 'Front-Rear',  
 man\_coll == "2" ~ 'Front-Front',  
 man\_coll == "3" ~ 'Not shown on Table (Known type 1)',  
 man\_coll == "4" ~ 'Not shown on Table (Known type 2)',  
 man\_coll == "6" ~ 'Angle',  
 man\_coll == "7" ~ 'Sideswipe (same direc)',  
 man\_coll == "8" ~ 'Sideswipe (oppo)',  
 man\_coll == "9" ~ 'Rear-Side',  
 man\_coll == "10" ~ 'Rear-Rear',  
 TRUE ~ 'Other')) %>%   
 ungroup() %>%   
 group\_by(year) %>%   
 filter(Collision != 'Other') %>%   
 mutate(percent = n/sum(n))  
  
ggplot(dat\_col, aes(year, percent, col = Collision)) + geom\_point() +  
 geom\_line() + xlab('Year') + ylab('Percentage of Accidents') +   
 ggtitle('Accidents breakdown by Collision type') +  
 theme(  
 plot.title = element\_text(color="black", size=12, face="bold.italic"),  
 axis.title.x = element\_text(color="black", size=12, face="bold"),  
 axis.title.y = element\_text(color="black", size=12, face="bold"))



# keep the Region variable  
dat\_col\_reg <- dat\_full %>%  
 group\_by(year, man\_coll, Region) %>%   
 summarise(n = n()) %>%   
 mutate(Collision = case\_when(man\_coll == "0" ~ 'Not with motor',  
 man\_coll == "1" ~ 'Front-Rear',  
 man\_coll == "2" ~ 'Front-Front',  
 man\_coll == "3" ~ 'Not shown on Table (Known type 1)',  
 man\_coll == "4" ~ 'Not shown on Table (Known type 2)',  
 man\_coll == "6" ~ 'Angle',  
 man\_coll == "7" ~ 'Sideswipe (same direc)',  
 man\_coll == "8" ~ 'Sideswipe (oppo)',  
 man\_coll == "9" ~ 'Rear-Side',  
 man\_coll == "10" ~ 'Rear-Rear',  
 TRUE ~ 'Other')) %>%   
 ungroup() %>%   
 group\_by(year) %>%   
 filter(Collision != 'Other') %>%   
 mutate(percent = n/sum(n))  
  
ggplot(dat\_col\_reg, aes(year, percent, col = Collision)) +   
 geom\_point() + geom\_line() + facet\_wrap( ~ Region) +  
 xlab('Year') + ylab('Percentage of Accidents') +   
 ggtitle('Accidents breakdown by Collision type') +  
 theme(  
 plot.title = element\_text(color="black", size=12, face="bold.italic"),  
 axis.title.x = element\_text(color="black", size=12, face="bold"),  
 axis.title.y = element\_text(color="black", size=12, face="bold"))

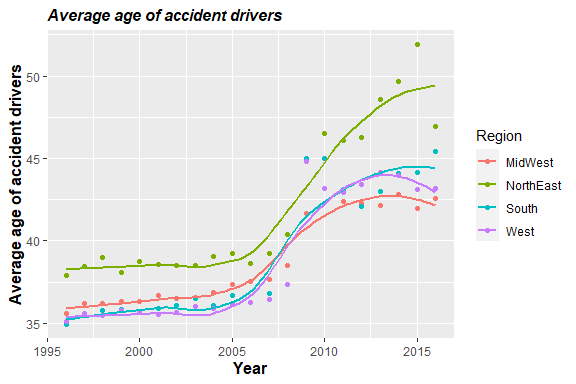
 The next variable that I considered in the study is the type of collision. The metadata for this variable is very incomplete as there is not explanation for type 3 and 4 which I think represents the collision between the front of one car and the side of the other and the collision between the front of one car and angle of anther. But without additional information, we cannot dig deeper into these two categories.

From the first graph, it is obvious that a majority of accidents happen between a vehicle and a non-vehicle (consistently over 40%). As time increases, we see that there are more and more accidents that happens at angle of the road (probably as a result of more roads).

When we break down the collision by state, we observe that the accidents in the South are nearly identical to the overall accident patterns. As we have talked about previously, a large proportion of the accidents happens in the South and this makes a lot of sense here. For the other three regions, though we see accidents with Non-motor also appears to be the most frequent type, it does not diverge that much from other types of collision.

#### Age

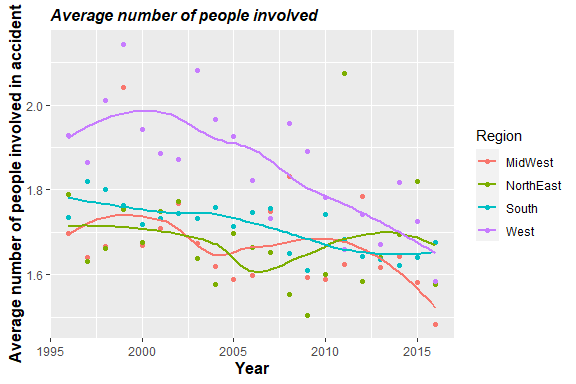
## AGE  
dat\_age <- dat\_full %>%  
 group\_by(year, Region) %>%   
 summarise(avg = mean(age))%>%   
 drop\_na(Region)  
  
ggplot(dat\_age, aes(year, avg, col = Region)) + geom\_point() +  
 geom\_smooth(se = FALSE, method = 'loess') +  
 xlab('Year') + ylab('Average age of accident drivers') +   
 ggtitle('Average age of accident drivers') +  
 theme(  
 plot.title = element\_text(color="black", size=12, face="bold.italic"),  
 axis.title.x = element\_text(color="black", size=12, face="bold"),  
 axis.title.y = element\_text(color="black", size=12, face="bold"))



We then shift our focus to numerical variables. The first of which is the average number of people involved.By using a loess graph, we find that, interestingly, as time goes on, the average age of accident victims also increases. With both increased about 16 years, we can suggest that there is a certain group of people that is very likely to get involved in traffic accidents. This trend is also observed for the four regions respectively, with a slight divergence after year 2015.

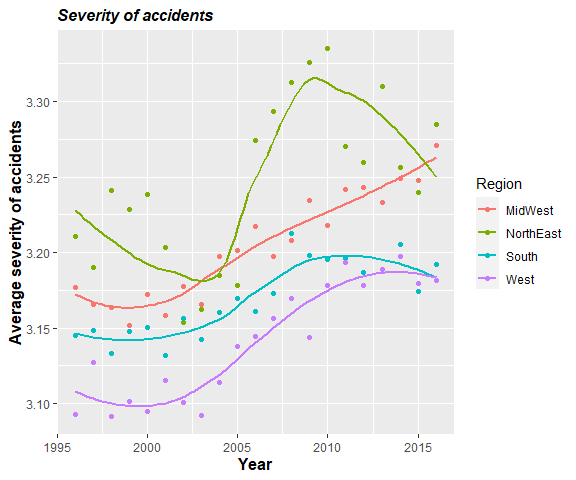
#### Number of person involved per accident

# Number of person vs year  
dat\_per <- dat\_full %>%  
 group\_by(year, Region) %>%   
 summarise(avg = mean(per\_no))  
  
ggplot(dat\_per, aes(year, avg, col = Region)) + geom\_point() +  
 geom\_smooth(se = FALSE, method = 'loess') +  
 xlab('Year') + ylab('Average number of people involved in accident') +   
 ggtitle('Average number of people involved') +  
 theme(  
 plot.title = element\_text(color="black", size=12, face="bold.italic"),  
 axis.title.x = element\_text(color="black", size=12, face="bold"),  
 axis.title.y = element\_text(color="black", size=12, face="bold"))

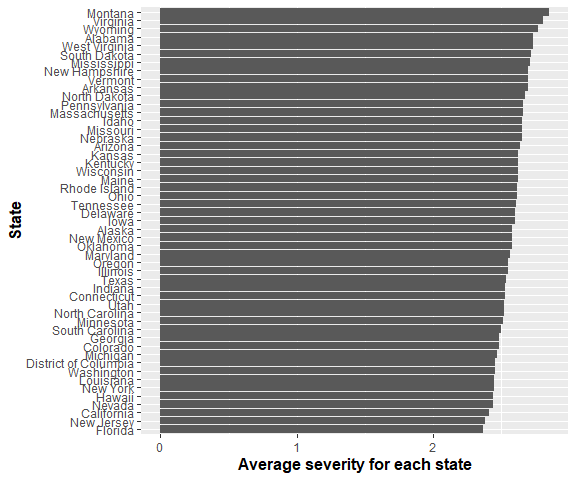
 Contrary to the average age which shows a positive trend, we see a downturn of average number of people involved in accidents which is an exciting finding. However, when we look at the data for 2016, three regions(West, South, NorthEast) average around 1.65 people compared to the Midwest that averages around 1.45 people per accident.. This is somewhat conterintuitive as we generally think that states in the North may have more people involved in a traffic accident due to more extreme weather. And the results here shown by the loess graph is completely opposite.

##### Severity

# Number of person vs year  
# severity by region  
dat\_sev <- dat\_full %>%   
 group\_by(year, Region) %>%   
 summarise(avg = mean(inj\_sev)) %>%   
 drop\_na(Region)  
  
ggplot(dat\_sev, aes(year, avg, col = Region)) + geom\_point() +  
 geom\_smooth(se = FALSE, method = 'loess') +  
 xlab('Year') + ylab('Average severity of accidents') +   
 ggtitle('Severity of accidents') +  
 theme(  
 plot.title = element\_text(color="black", size=12, face="bold.italic"),  
 axis.title.x = element\_text(color="black", size=12, face="bold"),  
 axis.title.y = element\_text(color="black", size=12, face="bold"))



# digging deeper into state  
  
state <- read.csv("../Data/state\_code.csv") %>%  
 rename(state = state\_code)  
  
# Create the names for states  
dat\_state <- left\_join(dat, state, by = 'state')  
  
dat\_states <- dat\_state %>%   
 group\_by(state\_name) %>%   
 summarise(avrg\_sev = mean(inj\_sev)) %>%   
 ungroup()  
  
# create plot  
ggplot(dat\_states, aes(fct\_reorder(state\_name, avrg\_sev, .fun = median), avrg\_sev)) +  
 geom\_bar(stat = 'identity') + coord\_flip() +   
 xlab("State") + ylab("Average severity for each state") +  
 theme(  
 plot.title = element\_text(color="black", size=12, face="bold.italic"),  
 axis.title.x = element\_text(color="black", size=12, face="bold"),  
 axis.title.y = element\_text(color="black", size=12, face="bold"))



When it comes to severity, I treat the original categorical variable, which has 6 levels (after dropping unknowns), as a numeric variable. The reason that I think it still functions is because in the dataset, the severity of accident increases with the figure assigned to it.

Thus, we can treat it as a quantitative variable without losing its original function. From our graph, we can observe a increase in severity for all Regions except NorthEast which shows a ‘S’ shape. But overall, the severity of NorthEast accidents also increase.The only draw back of categorical data is that when doing a spread location plot, medians for all four locations equals four and thus not variance can be found.

This is conflicting when we recall the previous result that number of people involved in the accidents decrease. Thus, to study the relationship between severity and the other variables, I will study it using linear models.

I also breakdown the severity into the state level, we can see that the average severity, between states does not differ a lot even when we compare Montana with Florida.

#### Model 1

lm\_sev\_age <- lm(inj\_sev ~ age \* Region, dat = dat\_full)  
summary(lm\_sev\_age)

##   
## Call:  
## lm(formula = inj\_sev ~ age \* Region, data = dat\_full)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.3313 -1.1193 0.7551 0.8333 2.9010   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.113e+00 3.185e-03 977.665 < 2e-16 \*\*\*  
## age 2.220e-03 6.073e-05 36.556 < 2e-16 \*\*\*  
## RegionNorthEast 5.889e-02 4.488e-03 13.120 < 2e-16 \*\*\*  
## RegionSouth 3.523e-04 3.574e-03 0.099 0.921482   
## RegionWest -5.121e-02 4.102e-03 -12.484 < 2e-16 \*\*\*  
## age:RegionNorthEast -6.066e-04 7.851e-05 -7.727 1.1e-14 \*\*\*  
## age:RegionSouth -8.358e-04 6.651e-05 -12.567 < 2e-16 \*\*\*  
## age:RegionWest -2.839e-04 7.568e-05 -3.751 0.000176 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.059 on 1467627 degrees of freedom  
## Multiple R-squared: 0.005501, Adjusted R-squared: 0.005496   
## F-statistic: 1160 on 7 and 1467627 DF, p-value: < 2.2e-16

In terms of the first model, I find that all the variables except South is significant. This fits my hypothesis that different areas have different impact on the accident and its variables which include severity. The South, just as we have been mentioning, accounts for nearly half of the sample of the overall dataset so we would expect it not to be significant.

Overall, we see that age has a positive retionship with severity, confirming our hypothesis.And control group, which is Midwest here, has lower severity than the other groups.

#### Model 2

lm\_all <- lm(inj\_sev ~ age \* Region + sex + man\_coll + per\_no, dat = dat\_full)  
summary(lm\_all)

##   
## Call:  
## lm(formula = inj\_sev ~ age \* Region + sex + man\_coll + per\_no,   
## data = dat\_full)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.4263 -0.9568 0.6392 0.7590 7.7120   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.467e+00 3.789e-03 915.089 < 2e-16 \*\*\*  
## age 2.213e-03 5.934e-05 37.288 < 2e-16 \*\*\*  
## RegionNorthEast 3.687e-02 4.381e-03 8.416 < 2e-16 \*\*\*  
## RegionSouth -1.121e-03 3.488e-03 -0.322 0.74782   
## RegionWest -4.158e-02 4.004e-03 -10.385 < 2e-16 \*\*\*  
## sex -1.028e-01 1.612e-03 -63.754 < 2e-16 \*\*\*  
## man\_coll -2.723e-02 2.029e-04 -134.189 < 2e-16 \*\*\*  
## per\_no -9.202e-02 4.254e-04 -216.309 < 2e-16 \*\*\*  
## age:RegionNorthEast -2.334e-04 7.664e-05 -3.046 0.00232 \*\*   
## age:RegionSouth -8.727e-04 6.492e-05 -13.443 < 2e-16 \*\*\*  
## age:RegionWest -2.149e-04 7.385e-05 -2.910 0.00362 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.034 on 1467624 degrees of freedom  
## Multiple R-squared: 0.05293, Adjusted R-squared: 0.05293   
## F-statistic: 8203 on 10 and 1467624 DF, p-value: < 2.2e-16

When we continue and include other variables into the model, the results is very satisfying as all variables turn significant (again except from South). Judging from the sign of the variables, we see that woman is less likely to involve in severe accidents, and the collisions with smaller index are more dangerous (e.g. accidents happen at crossroad is very likely to be severe accident). The number of person involved in accident is inversely related to severity, as we have discussed previously.

### Discussion

Overall, both the study on regional effect on traffic accident and the study on how other variables affect severity shows significant results. From the graphs, we find that the variables for accidents do vary from region to region. The South, compared with other groups, shows a strong impact on the overall data, especially when studying the type of collisions. Yet, when it comes to age of drivers, average number of people involved, and severity, South group is not significantly differnt from other groups.

The findings do suggest that where the accident do affect the accident from both the graphs and from our models. This can be due to many factors including geographic factors, cultural reasons, etc. It is important to understand what factors leads to what outcome in order to prevent severe traffic accident, whether it is smarter regulation, stricter road tests or better road condition.

On the other hand, we also tried to identify what factors can lead to more detrimental accidents. Results show that an aged male driver who is passing a crossroad with only himself in the car driving in the NorthEast will be the most likely to get involved in a severe vehicular accident than a light one. This makes sense from the outside perspective as if there is one more person sitting in the car, or if he is younger and can react faster, the damage can be reduced.

So this study provides insight of what affects severity of traffic accidents and how location can affect the accident. This is important for future decision making and accident prevention as we know under what circumstances the accidents may happen and how severe it can be. Still, this is just a very preliminary study and in order to diminish the accidents, more factors have to be included and the states need to learn from each other to find out what factors are contributing the road safety and what are not.

REFERENCE:

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.