CMDA-3654 2019 Summer II

Final Project

Rafael Arbex-Murut

Introduction

Driving is something that many people partake in daily. People drive out of necessity, leisure, and convenience among many reasons. The US Census Bureau, the Federal Highway Administration, and the Department of Transportation estimate that there are around $227,754,100$ licensed drivers in the US in 2019.

Due to the ever-increasing number of drivers, driving safety is a big concern for many. Road safety encompasses many facets of driving such as public safety, property damage liability, social perceptions, insurance prices, and many other aspects.

Fortunately, there is plenty of public data available for entities and individuals too study and analyze. Much of this data is critical because it can lead to new legislation, changes in the automobile industry, and changes in car insurance rates. Additionally, much of this data can be also be used to simply confirm or disavow our conventional wisdoms and of driving safety.

In this report, I explore three different areas related to driving safety. I attempt to answer many questions that relate to these three topics by applying the statistical learning methods we learned in this course on a single data set.

Description of the Data

The data set for my project is a collection of information from all police-reports of car crashes in the United States between 1997-2002, in which there was a harmful event to a person or property, and which at least one vehicle was towed. The data has 15 variables, of different types that describe the factors of the car crash.

Here is a small sample of the data with all variables displayed:

Table 1: Data (continued below)

X	dvcat	weight	dead	airbag	seatbelt	frontal	sex	ageOFocc
1	25-39	25.07	alive	none	belted	1	f	26
2	10-24	25.07	alive	airbag	belted	1	f	72
3	10-24	32.38	alive	none	none	1	f	69
4	25-39	495.4	alive	airbag	belted	1	f	53
5	25-39	25.07	alive	none	belted	1	f	32
6	40-54	25.07	alive	none	belted	1	f	22

yearacc	yearVeh	abcat	occRole	deploy	injSeverity	caseid
1997	1990	unavail	driver	0	3	2:3:1
1997	1995	deploy	driver	1	1	2:3:2
1997	1988	unavail	driver	0	4	2:5:1
1997	1995	deploy	driver	1	1	2:10:1
1997	1988	unavail	driver	0	3	2:11:1
1997	1985	unavail	driver	0	3	2:11:2

- `dvcat` is the estimated speed upon impact in kilometers per hour. It is a categorical variable with the following speed ranges:

x
1-9km/h
10-24
25-39
40-54
55+

- `weight` is an estimated weight of the observation designed to account for varying sampling probabilities.
- `dead` is the status of the occupant. It is a binary variable that takes the values of "dead" or "alive".
- `airbag` is a binary variable indicating the availability of an airbag in the vehicle. It takes the values of "none" or "airbag".
- `seatbelt` indicates if the occupant had their seatbelt on. It takes the values of "belted" or "none.
- `frontal` is a dummy variable indicating if the accident was frontal. It takes the value of 1 for frontal collision, and 0 otherwise.
- `sex` indicates the sex of the occupant, with either "f" or "m".
- `ageOFocc` indicates the age of the occupant.
- `yearacc` indicates the year of the accident.
- `yearVeh` indicates the year of the vehicle model.
- `abcat` is a categorical variable indicating the role of the airbag. It assumes the values "deploy" for when the airbag was deployed, "nodeploy" for when it was present but did not deploy, and "unavail" for when the airbag was unavailable.
- `occRole` is a binary variable indicating the role of the person in the observation. It is only either `driver` or `pass` for passanger.
- `deploy` is a binary dummy variable indicating the functionality of the airbag. It is 1 for when the airbag deployed, and 0 otherwise.
- `injSeverity` is a categorical dummy variable. The numbers correspond as such: 0:none, 1:possible injury, 2:no incapacity, 3:incapacity, 4:killed; 5:unknown, 6:prior death
- `caseid` indicates the listed case number.

The data has 26,217 observations. This does not mean there were that many car crashes. It means that 26,217 people were involved in car accidents that met the aforementioned criteria. There are some pairs of observations that have the same `caseid`, this means that there was one passenger and one driver in the same accident. Their roles are indicated in `occRole`, as either `driver` or `pass`.

This data was collected from the National Automotive Sampling System / Crashworthiness Data System (NASS/CDS) and concatenated by researchers at Colorado State University. The NASS and CDS systems are kept and maintained by the National Highway Traffic Safety Administration (NHTSA) and can be found at https://www.nhtsa.gov/.

Method

Topic 1: Speed and fatality rates.

Arguably, the most important aspect of driving safety is speed. This is something people can largely be held accountable for, which is why we have legal speed limits on roads.

Intuitively, car accidents are more fatal at higher speeds. However, it is worth asking derivative questions such as "Is the growth exponential? Linear? Logarithmic?", or "What are the specific percentages of the fatality rates by speed range?".

This data set is more than adequate to answer such questions. To get a better understanding of the relationship between speed and fatality, I opted for a proportional stacked area graph. This way we can neatly see the proportion of fatal vs. non-fatal car accidents by the speed upon impact and how that proportion changes as speed goes up.

Topic 2: Fatality and its relation to other variables.

The end goal of driving safety is to reduce the number of deadly accidents on the road.

In this data set, fatality is described by the binary variable `dead` with the values of either `dead` or `alive`. To further understand how this variable related to the other variables, I chose to fit a logistic regression model with `dead` as the response variable, and `dvcat`, `deploy`, `occRole`, `ageOFocc`, `yearVeh`, `yearacc`, `seatbelt`, `frontal`, and `sex` as predictor variables.

I chose to exclude `weight` because of the uncertainty, `airbag` and `abcat` because the functionality of the airbags are better represented in `deploy`, `injSeverity` because it is better represented by `dead`, and `caseid` because it has no breaing on the accident.

I then used the logistic regression to perform backward model selection with `stepAIC()` in order to find the subset of variables that best indicate the outcome of fatality.

Topic 3: Driving habits by demographic.

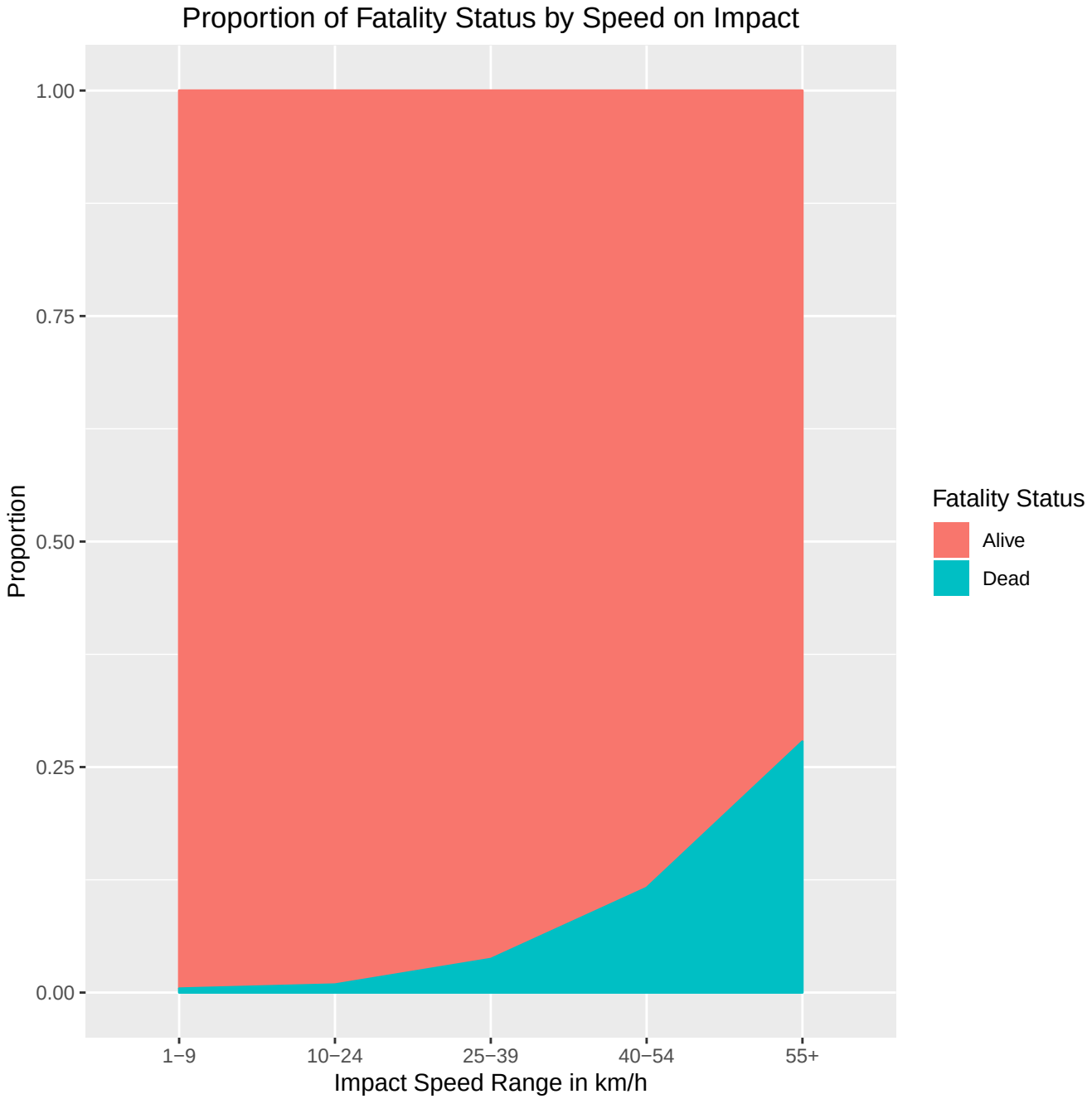
There are a lot of assumptions and discussions about certain demographics and driving safety. From a statistical perspective, driving demographics are often times very relevant because they justify laws, different insurance rates, and other things that factor into people's driving rights.

For example, many people are concerned with senior citizens driving, and claim that it is not safe to drive at more advanced ages. Many states have laws requiring senior citizens to retake their driving test. To explore this relationship, I create a plot with boxplots of age for each of the different speed ranges.

Another demographic that is thought to be correlated with certain driving habits is the sex of a person. There are many stereotypes that women have less spatial awareness than men when it comes to driving. On the other hand, insurance companies charge men higher rates for driving insurance because they have a higher propensity to drive dangerously. For the demographic of sex, I create a few proportion tables, with the `seatbelt` and `frontal` variables.

Results and Discussions

1. Speed and fatality rates.



As we can see from the data, fatality and speed upon impact are clearly proportional to each other. As one increases, so does the other.

Interestingly, this correlation is exponential. My initial assumption was that it would be logarithmic. My reasoning was that after a certain speed, fatality would not increase drastically. However, as we can see, this is not the case.

The fatality rates are almost negligable at lower speeds. It becomes noticeable at $25 - 39$ km/h. It reaches about 1/8 at $40 - 54$ km/h, and aroun 30% at 55+ km/h.

2. How does the fatality of a car crash relate to other variables?

Table 4: Logistic Regression Summary for dead

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	13.9	39.35	0.3532	0.7239
dvcat10-24	0.7306	0.5869	1.245	0.2132
dvcat25-39	2.168	0.5826	3.721	0.0001988
dvcat40-54	3.503	0.5831	6.009	1.872e-09
dvcat55+	4.734	0.584	8.106	5.219e-16
deploy	0.1147	0.08713	1.316	0.1881
occRolepass	0.2869	0.07818	3.669	0.0002435
ageOFocc	0.03375	0.001716	19.67	4.288e-86
yearVeh	-0.008813	0.007007	-1.258	0.2084
yearacc	-0.001608	0.02023	-0.07951	0.9366
seatbeltnone	1.044	0.0691	15.11	1.478e-51
frontal	-1.127	0.07013	-16.07	3.841e-58
sexm	0.1441	0.06918	2.083	0.03728

As we can see, the statistically significant variables are `dvcat`, `occRole`, `ageOFocc`, `seatbelt`, `frontal`, and `sex`.

As expected, the use of a seatbelt, and the speed of the vehicle are very significant. Their Pr values are way below 0.05. It is important to note that `dvcat` is treated as a dummy variable because it is categorical. Only the speeds of 25 km/h and above are significant, which is in accordance with the plot from Topic 1, that the fatality rate starts to become noticeable at that speed range.

The rest of the results are very intriguing. The occupant's role is significantly correlated as well. In this case, being a passenger is strongly correlated with fatality. The Pr value is 0.000243.

The age of the occupant is also correlated with fatality. The older an occupant, the higher the chances of fatality.

The frontal variable is very interesting. It is also significantly correlated. It appears that frontal collisions are less deadly than the non-frontal ones. I was not expecting this, because I assumed that frontal collisions had greater overall speeds because the oncoming object could have a speed of it's own. However, one must also consider the fact that there is more mass to shield a frontal impact than a side one, including the airbag.

Another interesting result is the value of `sex`. It is treated categorically, so when `sexm` is 1 (case where occupant is a male), the chances of fatality increase. The Pr value is slightly below 0.05, at 0.037279.

Start: AIC=7021.6 dead ~ dvcat + deploy + occRole + ageOFocc + yearVeh + yearacc + seatbelt + frontal + sex

```
       Df Deviance    AIC
```

- yearacc 1 6995.6 7019.6
- yearVeh 1 6997.2 7021.2
- deploy 1 6997.3 7021.3 6995.6 7021.6
- sex 1 7000.0 7024.0
- occRole 1 7008.7 7032.7
- seatbelt 1 7226.7 7250.7
- frontal 1 7256.2 7280.2
- ageOFocc 1 7375.6 7399.6
- dvcat 4 8716.5 8734.5

Step: AIC=7019.61 dead ~ dvcat + deploy + occRole + ageOFocc + yearVeh + seatbelt + frontal + sex

```
       Df Deviance    AIC
```

- yearVeh 1 6997.3 7019.3
- deploy 1 6997.3 7019.3 6995.6 7019.6
- sex 1 7000.0 7022.0
- occRole 1 7008.7 7030.7
- seatbelt 1 7226.7 7248.7
- frontal 1 7256.7 7278.7
- ageOFocc 1 7375.7 7397.7
- dvcat 4 8716.7 8732.7

Step: AIC=7019.34 dead ~ dvcat + deploy + occRole + ageOFocc + seatbelt + frontal + sex

`Df Deviance AIC`

- deploy 1 6997.8 7017.8 6997.3 7019.3
- sex 1 7002.0 7022.0
- occRole 1 7010.0 7030.0
- seatbelt 1 7238.9 7258.9
- frontal 1 7258.6 7278.6
- ageOFocc 1 7379.6 7399.6
- dvcat 4 8754.6 8768.6

Step: AIC=7017.84 dead ~ dvcat + occRole + ageOFocc + seatbelt + frontal + sex

`Df Deviance AIC`

6997.8 7017.8 - sex 1 7002.4 7020.4 - occRole 1 7010.2 7028.2 - seatbelt 1 7239.4 7257.4 - frontal 1 7262.3 7280.3 - ageOFocc 1 7380.1 7398.1 - dvcat 4 8759.3 8771.3

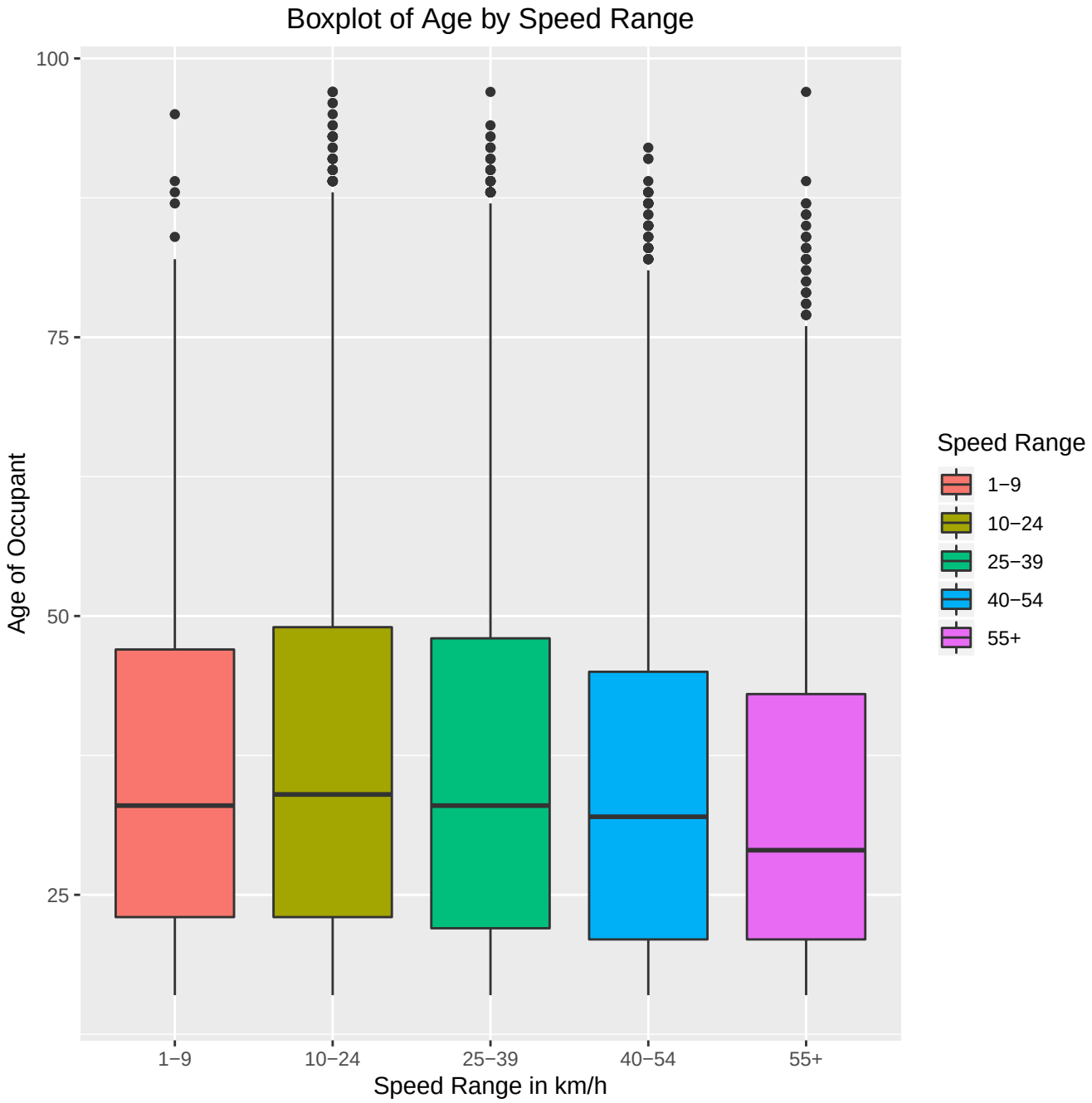
Table 5: Fitting generalized (binomial/logit) linear model: dead ~ dvcat + occRole + ageOFocc + seatbelt + frontal + sex

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-6.89	0.587	-11.74	8.193e-32
dvcat10-24	0.7535	0.5867	1.284	0.199
dvcat25-39	2.202	0.5822	3.781	0.0001559
dvcat40-54	3.542	0.5825	6.081	1.198e-09
dvcat55+	4.774	0.5834	8.183	2.761e-16
occRolepass	0.2775	0.07779	3.568	0.0003599
ageOFocc	0.03386	0.001716	19.73	1.155e-86
seatbeltnone	1.052	0.06818	15.43	9.751e-54
frontal	-1.105	0.06827	-16.18	6.518e-59
sexm	0.1471	0.06899	2.133	0.03296

The results of the backward model selection indicate that the ideal model for `dead` is:

$Y = -6.89 + 0.7535 * dvcat_{10-24} + 2.202 * dvcat_{25-39} + 3.542 * dvcat_{40-54} + 4.774 * dvcat_{55+} + 0.2775 * occRolepass + 0.03386 * ageOFocc + 1.052 * seatbeltnone - 1.105 * frontal + 0.1471 * sexm$

3. Driving habits by demographic.



Here is a boxplot of Age vs. Speed Range.

As we determined from topic 2, age is significantly directly correlated with fatality. However, this plot gives us more insight. It is clear that younger people drive faster. The average age, and interquartile ranges of age decrease as speed increases. As established in topic 1, higher speeds lead to higher fatality rates.

We can therefore conclude that the correlation between older ages and fatality is more likely not due to excessive speeds, but rather other factors such as reaction time, visibility, physical resilience to bodily harm, etc.

The only anomaly in this plot is the first speed range of 1-9 km/h. My best guess for this is that this is such a low speed for car accidents, that the circumstances in which these accidents occur are very special and in no way related to age.

Table 6: Proportion of Seatbelt Use by Sex The table above shows the proportion of male and female occupants who had their seatbelts on. About 3/4 of women drivers were belted, while only 2/3 of men were. This is about a 10% difference.

	belted	none
f	0.7665	0.2335
m	0.6575	0.3425

Table 7: Proportion of Frontal Accidents by sex (Drivers Only)

	0	1
f	0.3815	0.6185
m	0.3216	0.6784

This table shows the proportion of male and female drivers involved in frontal and non-frontal accidents. Note, this is not for all occupants, it is for drivers only, as this table was built to gauge the difference in peripheral and spatial awareness between the sexes.

In this case, about 38% of car collisions by women are non-frontal, while that number is only 32% for men.

Conclusions

Many of the findings from this data set can be explained by conventional wisdom. However, analysing how these variables relate to each other has provided me with additional insight.

Topic 1

It is not a breakthrough to know that driving faster is more dangerous. However, it was interesting to see by how much that danger increased. It is interesting to know that doubling my speed from 10 to 20 km/h might not be significantly more dangerous (to myself), but that increasing my speed to 55 from 45 km/h will about double the chances of a fatality occurring. The speed difference is the same, 10 km/h, but the fatality rate is much more drastic.

The big take away here is to exercise more precaution at higher speeds.

Topic 2

All factors of driving are correlated to fatality in some way. We should exercise safety and precaution for all of them. However, it was interesting to see which ones are strongly correlated with fatality that we did not expect. Things such as speed, age, and seatbelt use are obvious. The interesting results were non-frontal accidents and sex. Non-frontal accidents are much more likely to lead to a fatal accident than a frontal accident. This shows that peripheral awareness is very important.

Being male was also statistically significant in the prediction of fatality. Regardless of one's opinion on the morality, legality, and ethics of charging different insurance rates based on sex, it is very clear that from a mathematical standpoint, it makes absolute sense to charge men higher rates for car insurance.

Topic 3

It is conclusive that demographics can be an indication of driving habits - and consequently - driving safety.

Regarding the demographics of age, it is correct to assert that people of more advanced ages are more inclined to fatal car accidents. This was established in topic 2. However, the boxplot in this section explains why. It is clearly not because older people drive more recklessly and faster, since age and speed are inversely correlated. This danger might come from other factors, such as eyesight, reflexes, poor mechanical skills, among other reasons.

Regarding the demographics of sex, we can see that the stereotypes are true to a certain extent. Men do have certain dangerous driving habits. Men are 10 % less likely to wear a seatbelt. This contributes to the justifications that men are more dangerous drivers, mathematically.

It is also true however, that women have less peripheral awareness while driving. Women are 6% more likely to get into a non-frontal collisions.

Citations

The repository of the data set can be found here: https://vincentarelbundock.github.io/Rdatasets/datasets.html

Description of the variables can be found here: https://vincentarelbundock.github.io/Rdatasets/doc/DAAG/nassCDS.html

The data belongs to the National Highway Traffic Safety Administration: https://www.nhtsa.gov/

The original data was concatenated and cleaned by Dr. Meyer and Dr. Finney and used in:

Meyer, M.C. and Finney, T. (2005): Who wants airbags?

Farmer, C.H. 2006. Another look at Meyer and Finney's 'Who wants airbags?'

Meyer, M.C. 2006. Commentary on "Another look at Meyer and Finney's 'Who wants airbags?'

Appendix

Topic 1 Code

The following is the code generated for the stacked area graph in the first section of the project.

Firstly, I create a data frame that is suitable use for the `ggplot()` function with `geom_area()`.

For the first column, I list the x-axis variables (speed range) twice, one for alive and one for dead for each speed range. For the second column, I simply repeat "Alive" and "Dead" 5 times, one for each speed range. For the third column, I get the instances of fatal accidents and non-fatal accidents for each speed range, and place. Lastly, I bind the ccolumns together and calculate the proportions of dead and alive for each speed range.

```
#Creating the first column
col1 <- c("1-9", "1-9", "10-24", "10-24", "25-39", "25-39", "40-54", "40-54", "55+", "55+")
#Creating the second column
col2 <- rep(c("Alive", "Dead"), 5)
#Creating the third column
col3 <- c(
length(ds[which(ds$dvcat=="1-9km/h" & ds$dead=="alive"),][,1])
,length(ds[which(ds$dvcat=="1-9km/h" & ds$dead=="dead"),][,1])
,length(ds[which(ds$dvcat=="10-24" & ds$dead=="alive"),][,1])
,length(ds[which(ds$dvcat=="10-24" & ds$dead=="dead"),][,1])
,length(ds[which(ds$dvcat=="25-39" & ds$dead=="alive"),][,1])
,length(ds[which(ds$dvcat=="25-39" & ds$dead=="dead"),][,1])
,length(ds[which(ds$dvcat=="40-54" & ds$dead=="alive"),][,1])
,length(ds[which(ds$dvcat=="40-54" & ds$dead=="dead"),][,1])
,length(ds[which(ds$dvcat=="55+" & ds$dead=="alive"),][,1])
,length(ds[which(ds$dvcat=="55+" & ds$dead=="dead"),][,1]))

#Binding the columns into a data frame
newdf <- data.frame(col1, col2, col3)

#Calculating the proportions of fatality for each speed range, which is why i increases by 2.
i <- 1
```

```
while (i <= 9) {
  newdf[i, 4] <- newdf[i,3]/(newdf[i,3]+newdf[i+1,3])
  newdf[i+1, 4] <- newdf[i+1,3]/(newdf[i,3]+newdf[i+1,3])
  i <- i + 2
}

#Naming the columns for convenience
names(newdf) <- c("SpeedRange", "FatalityStatus", "Observations", "Prop")
```

Here is the data frame that I use to generate the area graph:

```
kable(newdf)
```

SpeedRange	FatalityStatus	Observations	Prop
1-9	Alive	673	0.9955621
1-9	Dead	3	0.0044379
10-24	Alive	12652	0.9910700
10-24	Dead	114	0.0089300
25-39	Alive	7861	0.9627679
25-39	Dead	304	0.0372321
40-54	Alive	2621	0.8839798
40-54	Dead	344	0.1160202
55+	Alive	1076	0.7216633
55+	Dead	415	0.2783367

Lastly, this is the `ggplot()` code used to generate the area graph.

```
#Using ggplot to build stacked area graph.
p <- ggplot(newdf, aes(x=newdf[,1], y=newdf[,4], fill=newdf[,2], colour = newdf[,2], group = newdf[,2])) + geo
p <- p + labs(title = "Proportion of Fatality Status by Speed on Impact", x = "Impact Speed Range in km/h", y =
p
```

Topic 2 Code

This is the code for the second topic.

Firstly, I fit the logistical regression model with the appropriate variables. I then displayed the results with `pander()`.

```
#Fit the model
logistic_model <- glm(dead ~ dvcat + deploy + occRole + ageOFocc + yearVeh + yearacc + seatbelt + frontal + se
#Display the results
logistic_model %>%
  pander(caption = "Logistic Regression Summary for dead")
```

Then, I use the `stepAIC()` function to perform the backwards stepwise algorithm. Again, I display the results with `pander()`.

```
pander(stepAIC(logistic_model))
```

Topic 3 Code

Here is the code for the boxplot in topic 3. I use `ggplot()` to produce the graph. I have to change the labels so that they are all standardized, since it should be "1-9" instead of "1-9km/h".

```
p <- ggplot(ds, aes(x = dvcat, y = ageOFocc, fill = dvcat)) + geom_boxplot()
p <- p + labs(title = "Boxplot of Age by Speed Range", x = "Speed Range in km/h", y = "Age of Occupant", fill =
p <- p + scale_x_discrete(labels=c("1-9", "10-24", "25-39", "40-54", "55+")) + scale_fill_discrete(labels=c("1
p
```

This is the proportion table for sex and seatbelt usage. I use the `xtabs()` function to produce the frequency table, the I use `prop.table()` to make the proportion one.

The 1 argument is to get proportions by rows, rather than columns or the entire table. I do this because I want the proportion by sex.

```
sex.belt_table <- xtabs(~ sex + seatbelt, data = ds)
prop.table(sex.belt_table, 1)%>%
  pander(caption = "Proportion of Seatbelt Use by Sex")
```

This table is the same, the only differences are that I filter the dataset for drivers only, and that it's frontal collisions instead of seatbelt usage.

```
ds.onlydrivers <- subset(ds, occRole == "driver")
sex.frontal_table <- xtabs(~ sex + frontal, data = ds.onlydrivers)
prop.table(sex.frontal_table, 1)%>%
  pander(caption = "Proportion of Frontal Accidents by sex (Drivers Only")
```
