EDA Case Study

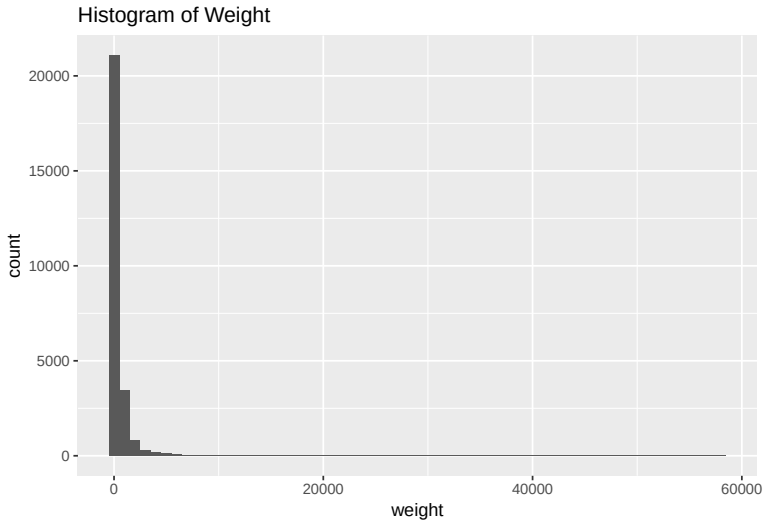
Nick Sun

May 19, 2019

Airbags and car accidents case study

Investigating weight

After making a histogram of the weights in this dataset, we notice that this distribution is heavily right skewed.

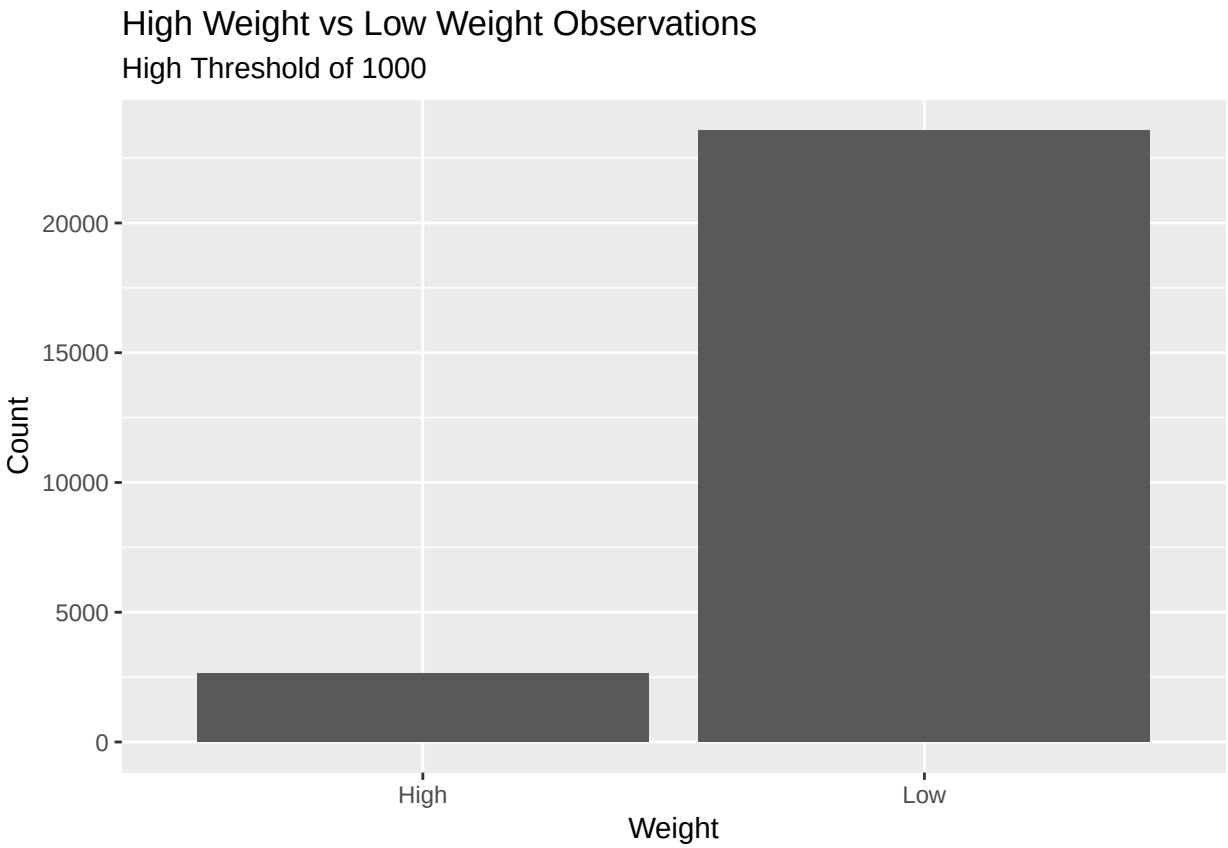


In fact, if we create a table summary we see that 50% of all of the observations have a weight less than 87 and 75% are less than 364. The mean of the weights meanwhile is 462.81, indicating that there are significant number of extreme points. This is corroborated by the fact that the maximum value in this dataset has a weight of 57871.59.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	32.47	86.99	462.8	364.7	57872

With this wide variability, I would say there is certainly some suspcion that we have uncertain accuracy with these weights. From the NASS GES website, these weights represent the inverse estimate of selection probability. The higher the weight, the less it is likely that the particular observation will be sampled in a NHTSA or sponsored study/simulation. It is interesting then that the estimates are not precise given that they are derived from some agreed-upon methodology, although the details on the NASS website are sparse on where these weights actually come from.

I would use some type of threshold to identify which cases have high weights and low weights. An example would be something like splitting the dataset into two parts and then using `geom_bar` on the counts.

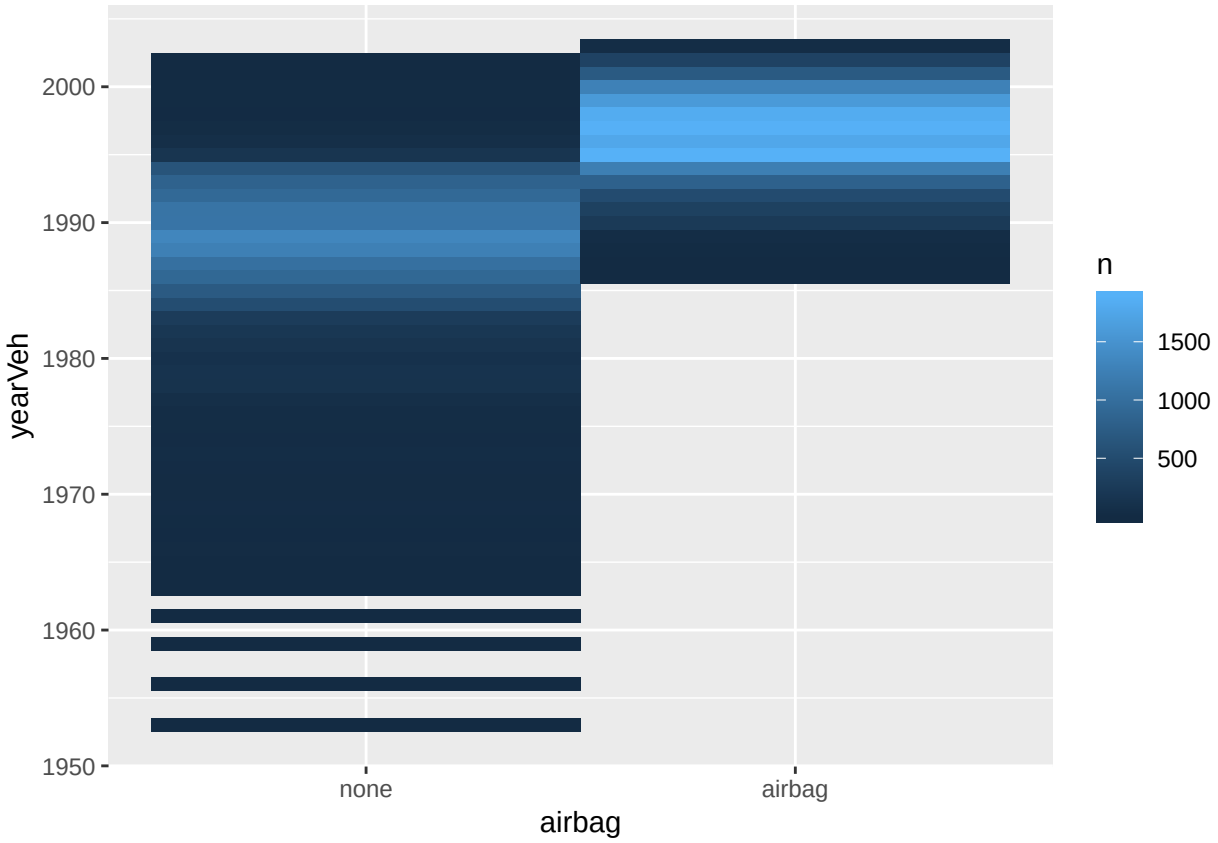


We can see that there are *far* more low weight cases than high weight cases, which is by design according to the NASS.

Availability of airbags and age of the vehicle

Here we can use `geom_tile` to investigate the relationship of car age and number of airbags.

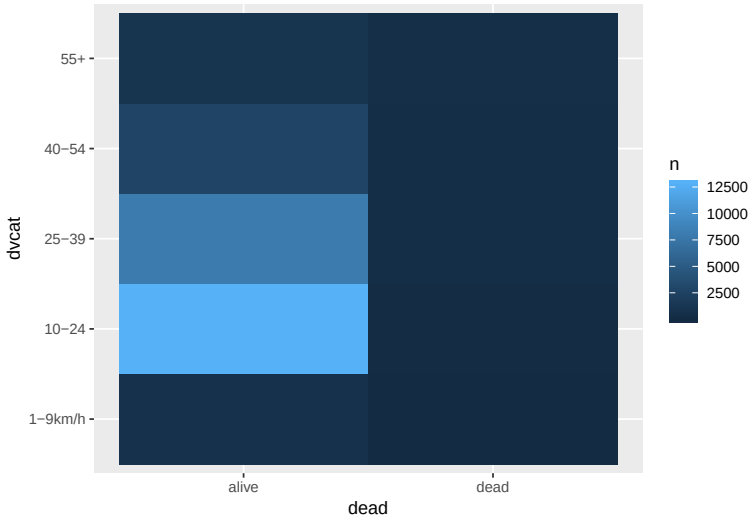
```
## Warning: Removed 1 rows containing missing values (geom_tile).
```



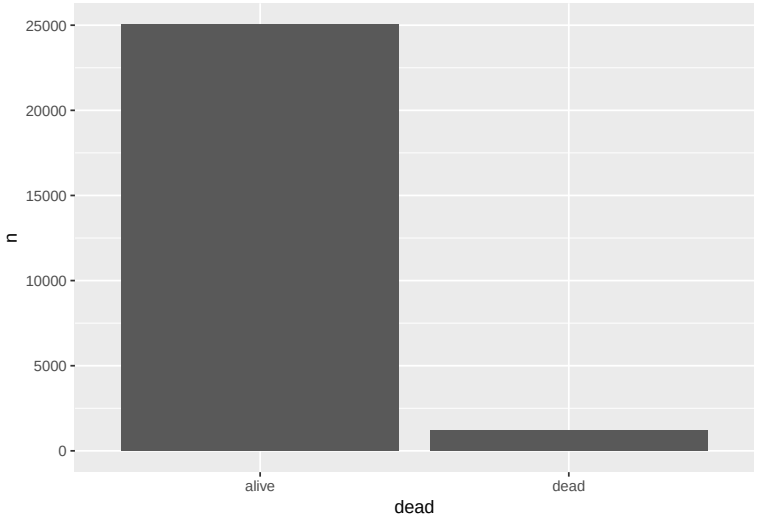
This visualization allows us to see that most of the cars that have airbags were made after 1985, which makes sense since the Federal Motor Vehicle Safety standard which mandated that seatbelts need to be put in cars was passed in July 11, 1984. Meanwhile, the cars that do not have airbags are mostly produced before 1984 and then drops to 0 after 1985.

Death rate and Speed

Since both variables `dead` and `dvcat` (which measures impact speed) are actually categorical variables, we can use `geom_tile` again.

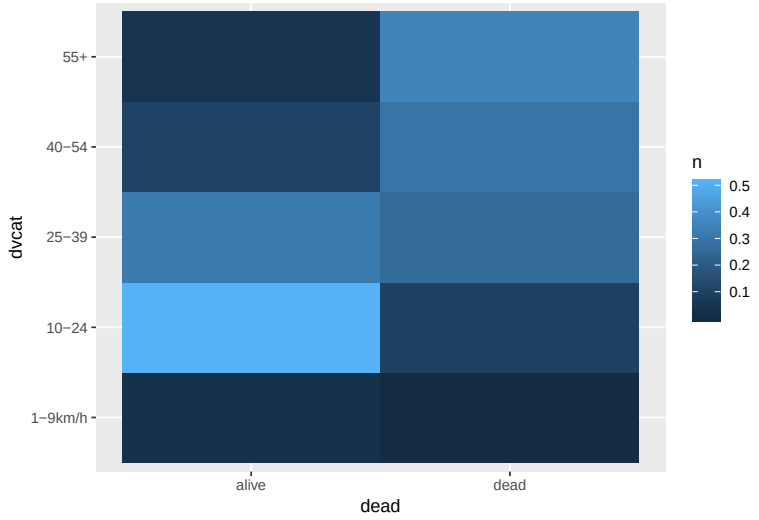


It looks like the impact speed of a car is related to whether the accident is fatal. The accidents with no deaths amass around 10-24km/h impact speed. However, we should also investigate if the proportions between accidents labeled 'alive' and 'dead' are approximately equal. This is particularly important because currently, the 'dead' part of this `geom_tile` gives us no new information.



We can see that this is not the case in this bar plot. One of the reasons are `geom_tile` plot looks the same in all categories of `dead` is that there aren't that many accidents which result in death relative to those which did not.

We would probably be better served looking at proportions within each category of `dead`.

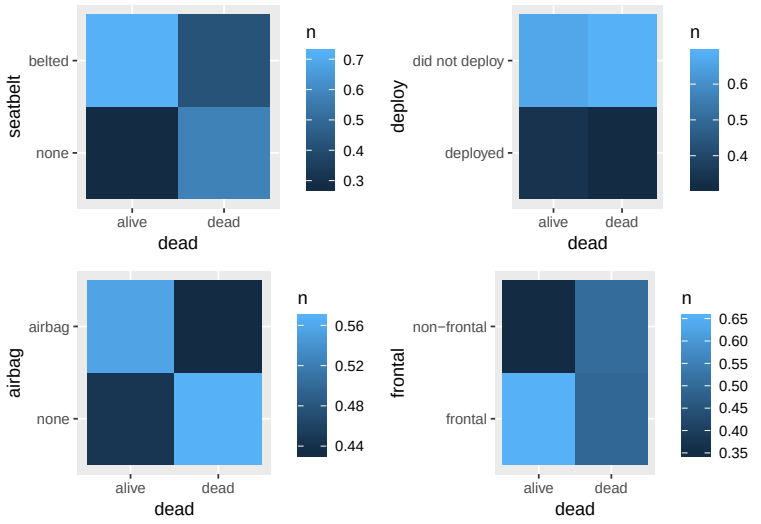


Now we can see that most of the accidents which resulted in death were at high rates of speed with the biggest proportion being at 55+ kmph. Conversely, most of the crashes which resulted in low rates of speed were at relatively low impact speeds.

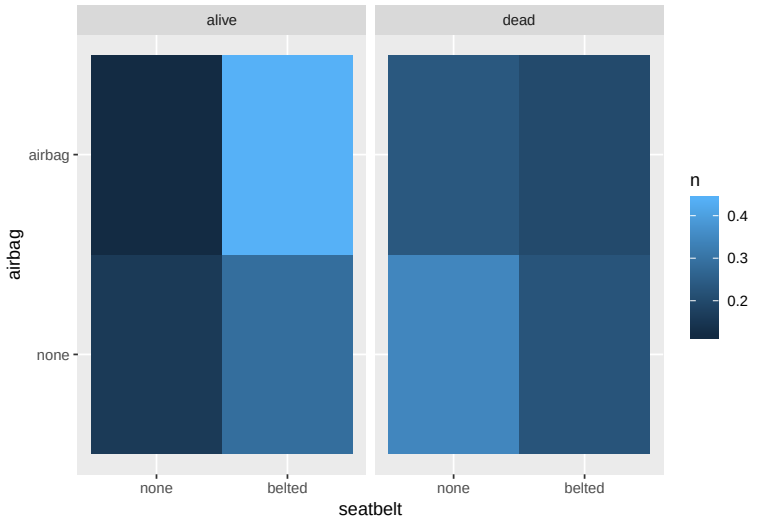
Death rate vs seatbelt, airbag, deploy, and frontal

Again, all of these variables are categorical. Using the previous strategies employed in the parts of this study, we can investigate each variable individually.

Here we have created `geom_tile` plots with `dead` plotted against `airbag`, `seatbelt`, `deploy` (as a factor), and `frontal` (as a factor).



We immediately notice that `airbag` and `seatbelt` have a very distinct pattern between the fatal and non-fatal accidents. We can split the data by these two categorical variables using `facet_wrap` and for further investigation.



This `geom_tile` is looking at proportion within each group `alive` and `dead`. We can see here that in the `alive` cases, the participants were usually belted in and had airbags. Wearing a seatbelt appears to be particularly important in minimizing fatalities.

In the `dead` group, we see that the biggest proportion had neither airbags nor were buckled in! Conversely the lowest proportion were both buckled in and had airbags.

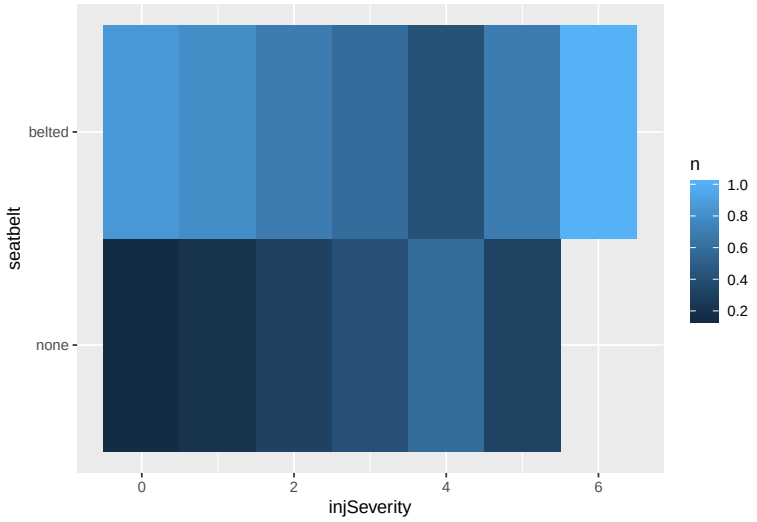
Other interesting explorations

I came up with several exploratory hypotheses with this data:

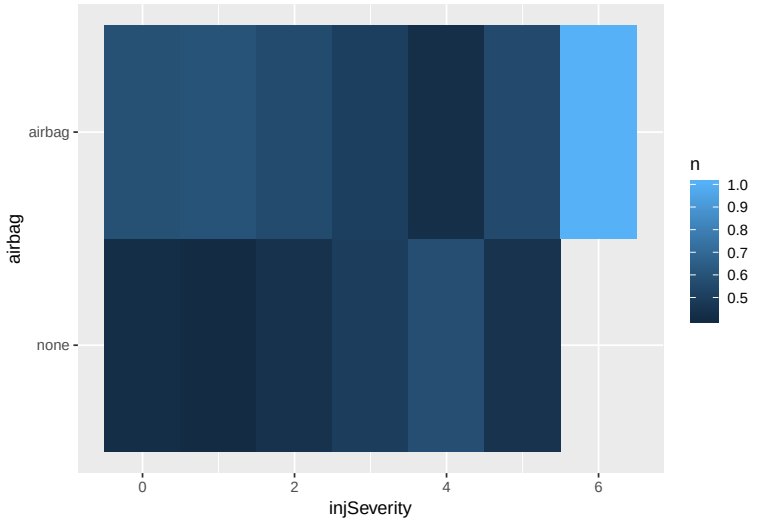
- Injury severity is related to seatbelts and airbags
- Mostly old people get into accidents
- Young people are more likely to drive faster
- Young people are more likely to not wear seatbelts than older drivers

I started by exploring injury severity.

```
## Warning: Removed 2 rows containing missing values (geom_tile).
```

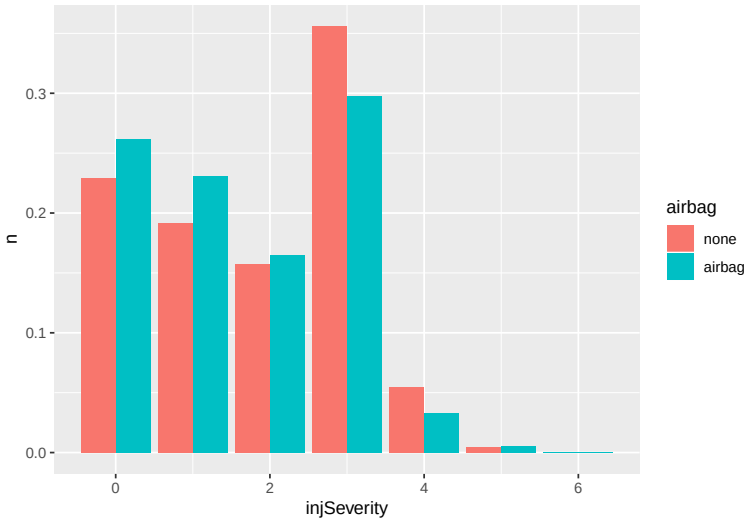



```
## Warning: Removed 2 rows containing missing values (geom_tile).
```



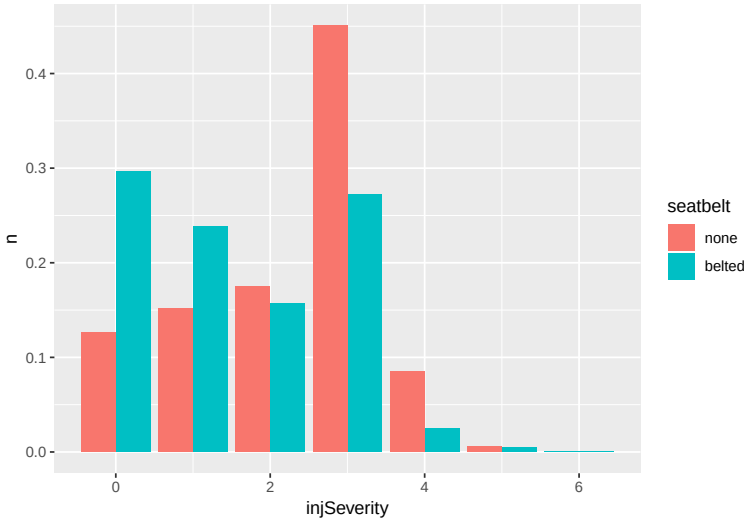
These plots are tough to interpret however since most drivers wear seatbelts and have airbags. This leads these plots to be useless.

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```



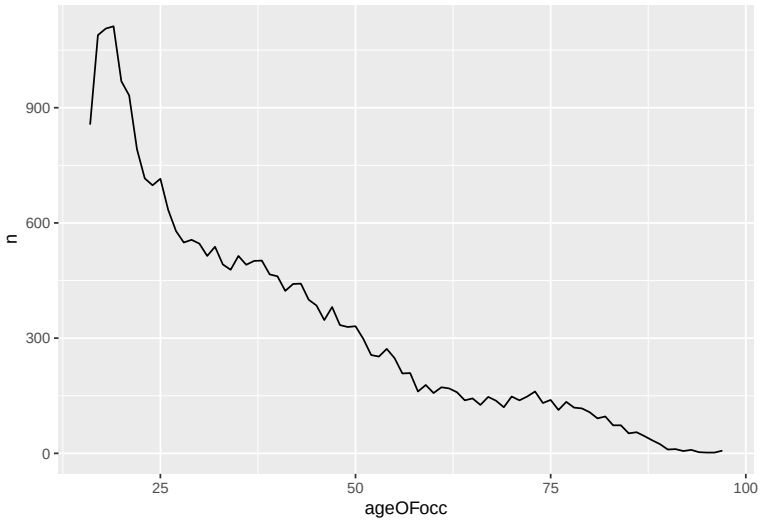
This plot comparing injury severity against airbag is better. We can see that most non-serious injuries (0-2) have an airbag involved while most of the serious injuries (3+) do not have an airbag involved.

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```



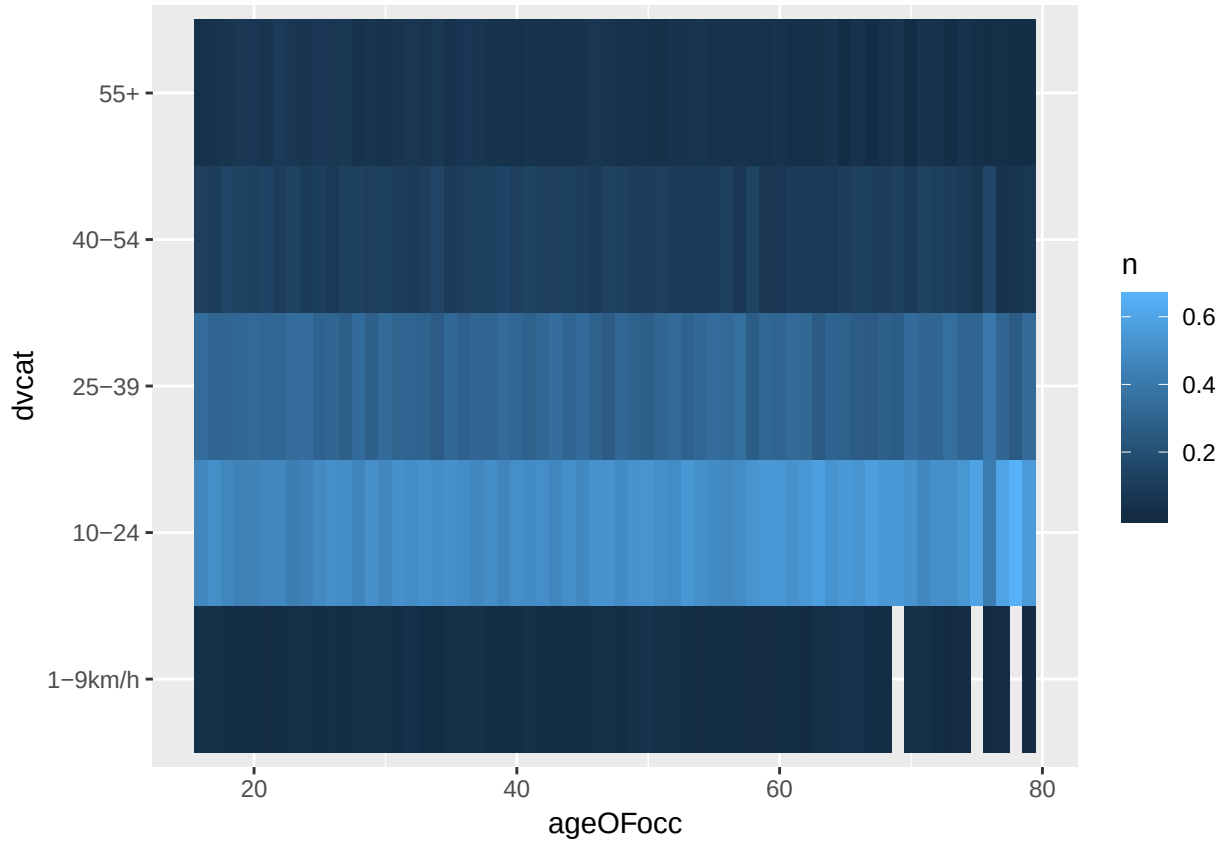
This pattern is even more apparent with `seatbelt`. Good lesson - wearing a seatbelt minimizes the risk of serious injury. Not really surprising, but it's nice to see the numbers hold up.

I wanted to begin investigating age by seeing the rough age distribution of this data. Here I chose to use `geom_line` instead of `geom_histogram` or `geom_density`, but this is essentially a density curve.



This isn't extraordinarily insightful though since old people are just less likely to drive. We can't really answer this question accurately with the data we have.

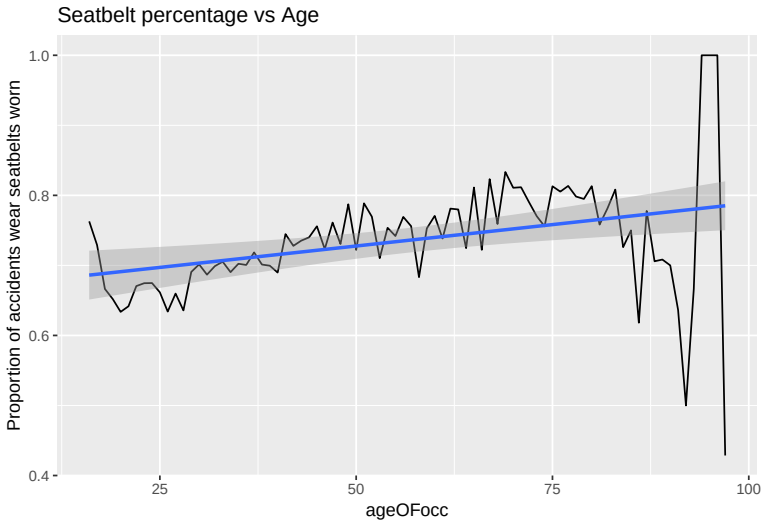
What we can do instead is investigate the age distribution of people who get into accidents split by impact speed. Perhaps younger people are more likely to drive faster so their accidents are at higher speeds?



Surprisingly, the distribution of accident impact speeds is pretty constant regardless of how old the driver is. Pretty unexpected!

Lastly I wanted to investigate if young people are less likely to wear seatbelts. I produced the following plot

with `geom_line` and `geom_smooth`.



We can see that there is a slight upward slope to the line, indicating that the older someone is, the more likely they are to wear a seatbelt. However, this isn't statistically rigorous since the method being used here is `lm` and what we should be using is a logistic model since the response is binary. It also does not to be a significantly large difference over age, at least not enough to be practically significant.