Complete a case study

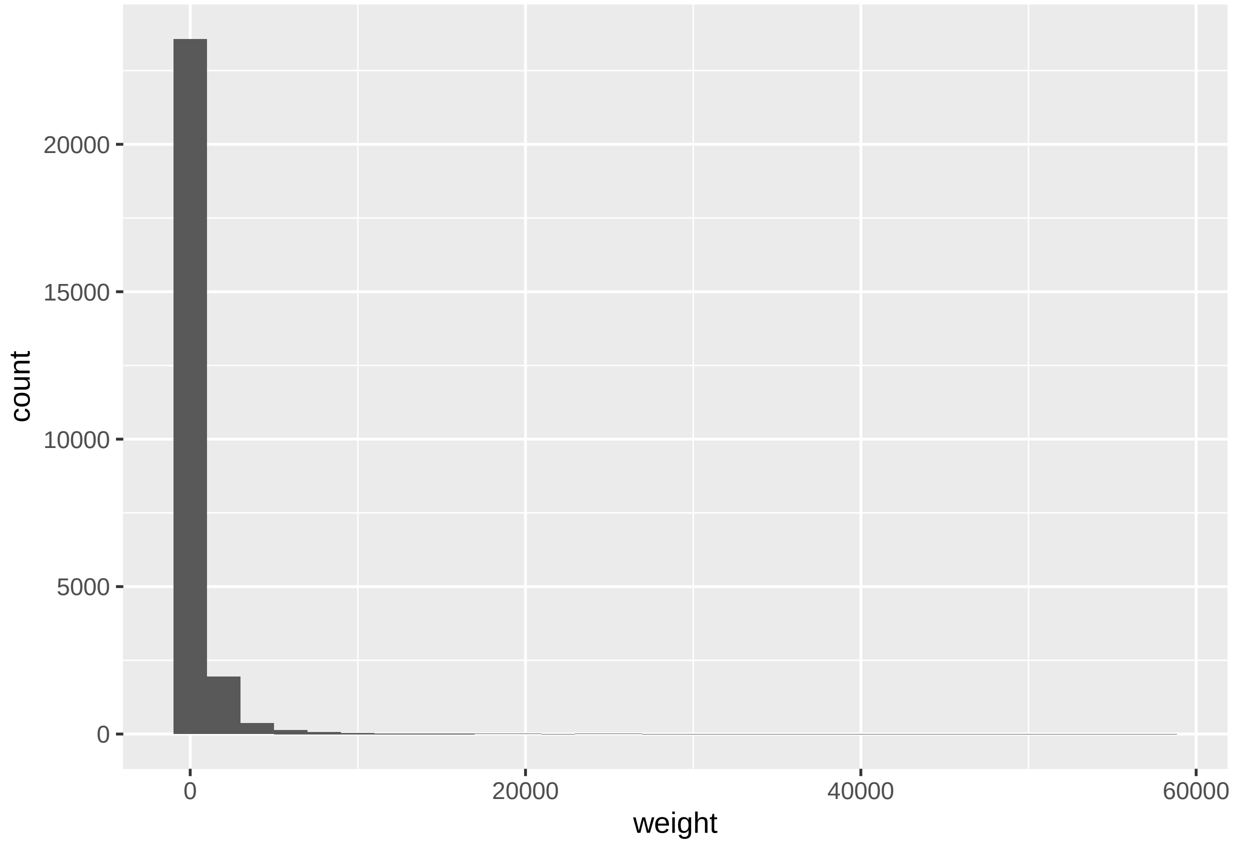
Nora Quick

```
#nassCDS
#?nassCDS
```

1. Draw a histogram of the variable weight. The Rhelp file for the dataset says that the observation weights are ”of uncertain accuracy”. Is there any evidence of this? What graphics would you draw to investigate which cases have high weights and which have very low weights?

```
ggplot(nassCDS, aes(x=weight)) +
  geom_histogram()
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```

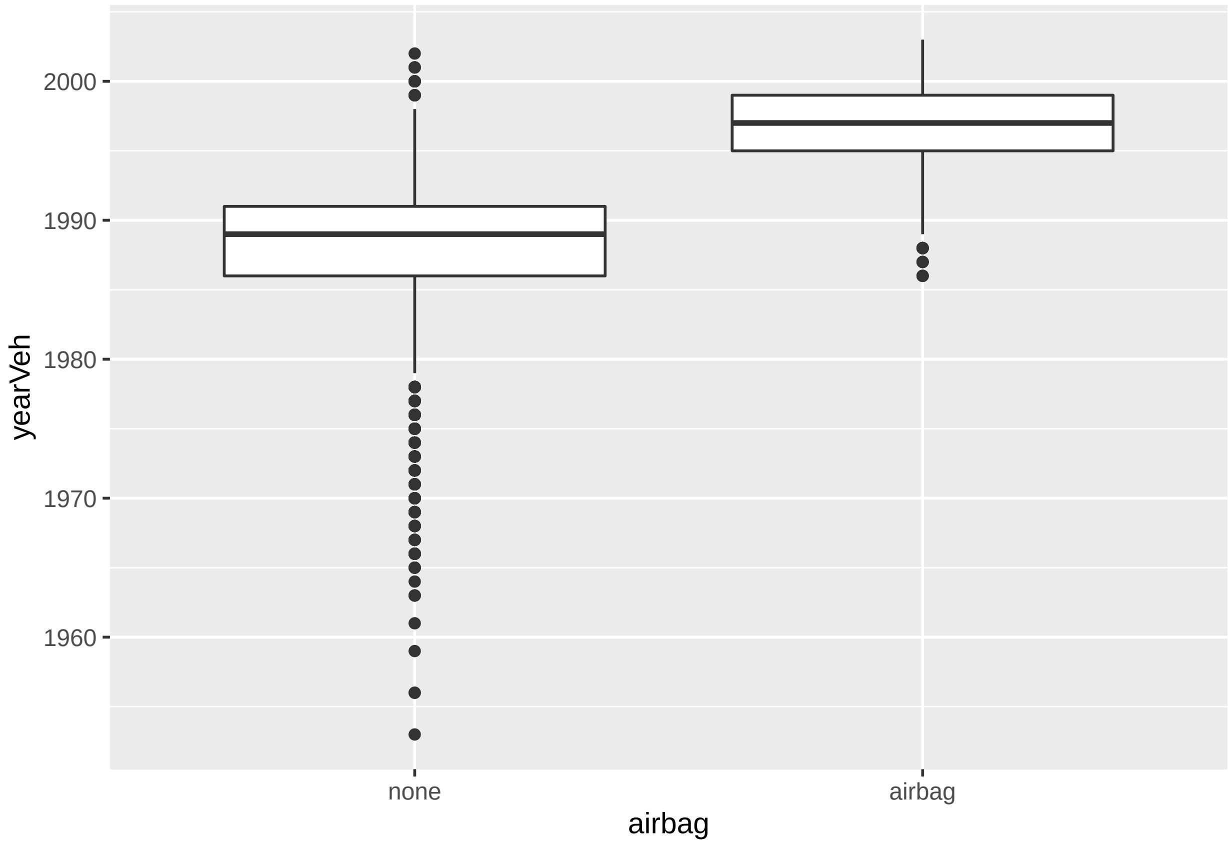


Yes,I believe we can see that there is some uncertainty in the accuracy of the data as there are too many zeros. Based on the first example in the reading I would choose to do a similar combination of a histogram but with cases as another parameter and a dot plot. I believe that a series of boxplots could also show the high and low weights.

2. How does the availability of airbags depend on the age of the vehicle?

```
ggplot(nassCDS, aes(x=airbag, y=yearVeh)) +
  geom_boxplot()
```

```
## Warning: Removed 1 rows containing non-finite values (stat_boxplot).
```

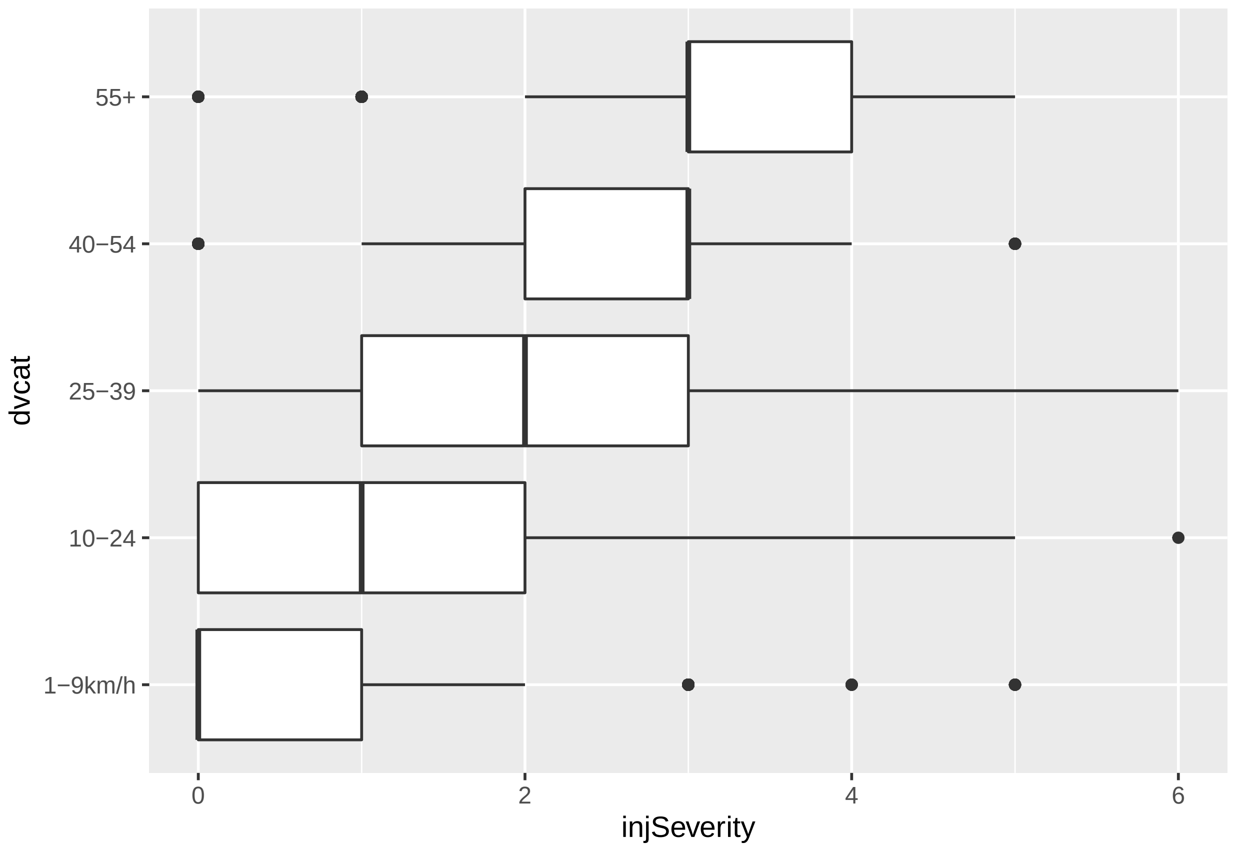


The newer the car the more likely it is to have the availability of airbags.

3. How does death rate depend on vehicle speed?

```
ggplot(nassCDS, aes(x=injSeverity, y=dvcat)) +
  geom_boxplot()
```

```
## Warning: Removed 153 rows containing non-finite values (stat_boxplot).
```

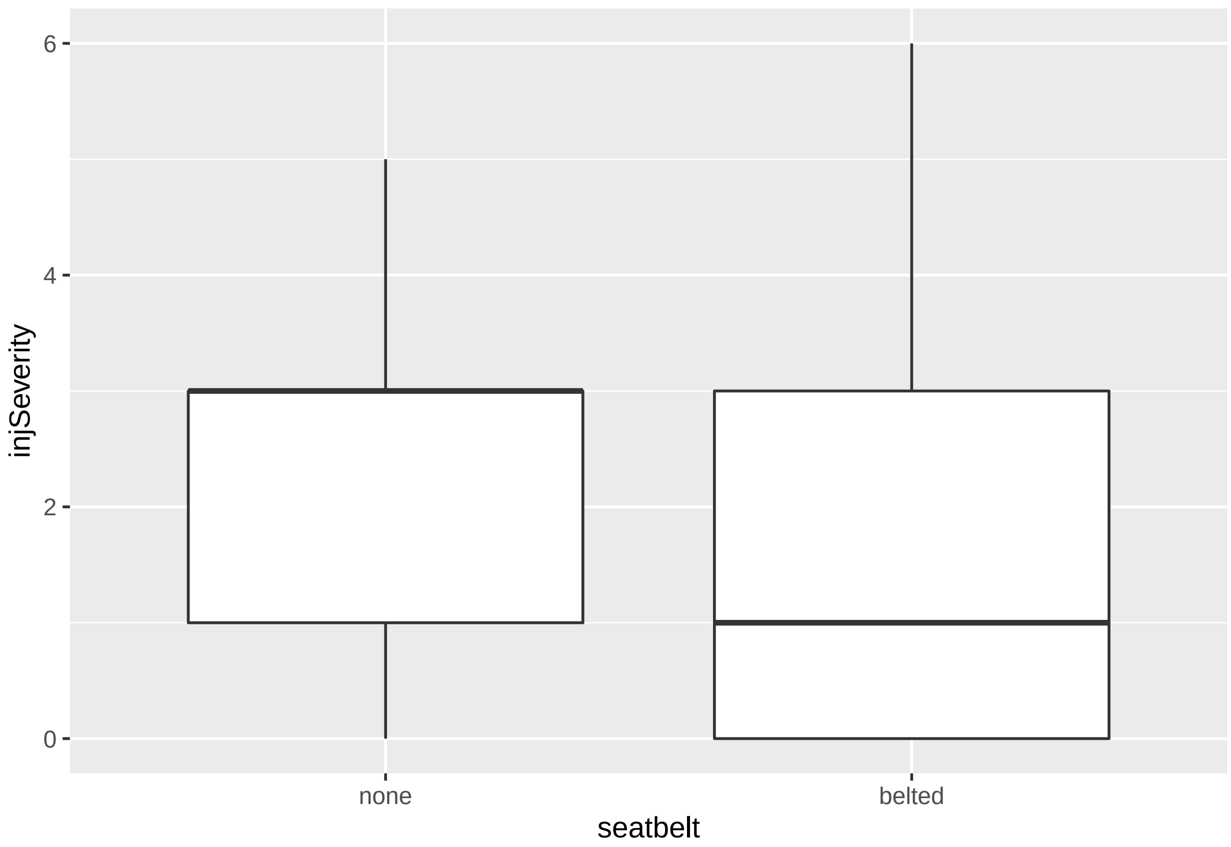


The death rate grows as the speed increases. In other words, the larger the speed is the likelyhood of injury and then death grow especially as 40+ MPH.

4. How does death rate vary with the variables seatbelt, airbag, deploy, and frontal? Which orderings of the variables and of the categories within the variables give the most convincing graphic?

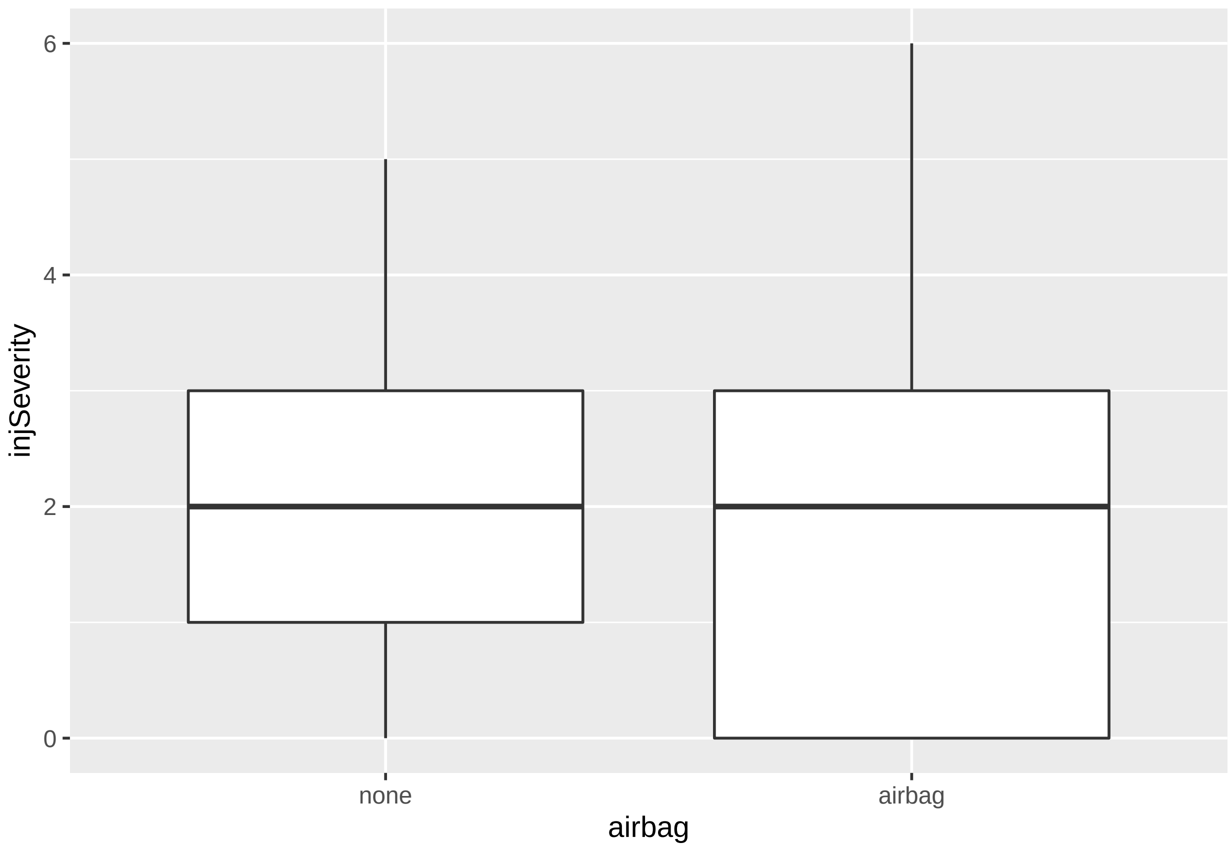
```r
ggplot(nassCDS, aes(x=seatbelt, y=injSeverity)) +
  geom_boxplot()
```

```
## Warning: Removed 153 rows containing non-finite values (stat_boxplot).
```



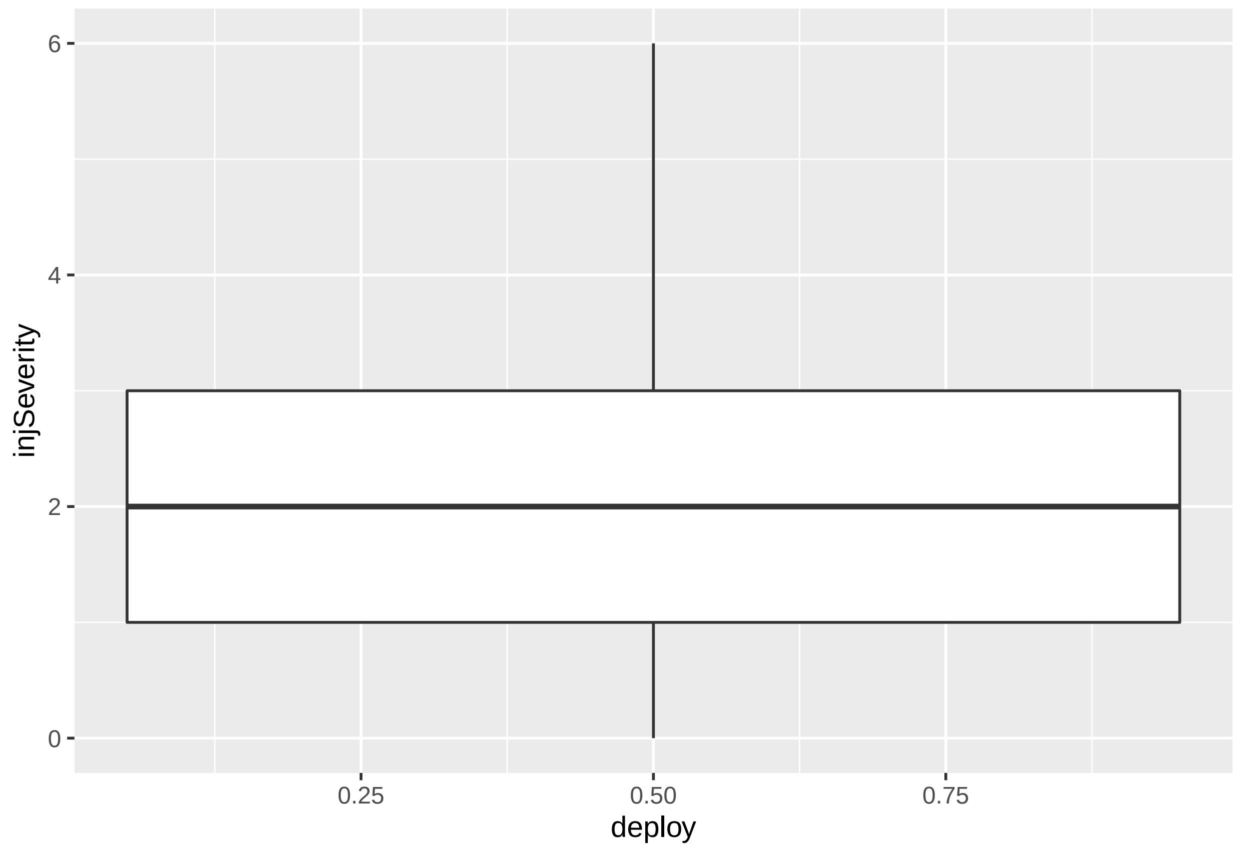
```r
ggplot(nassCDS, aes(x=airbag, y=injSeverity)) +
  geom_boxplot()
```

```
## Warning: Removed 153 rows containing non-finite values (stat_boxplot).
```



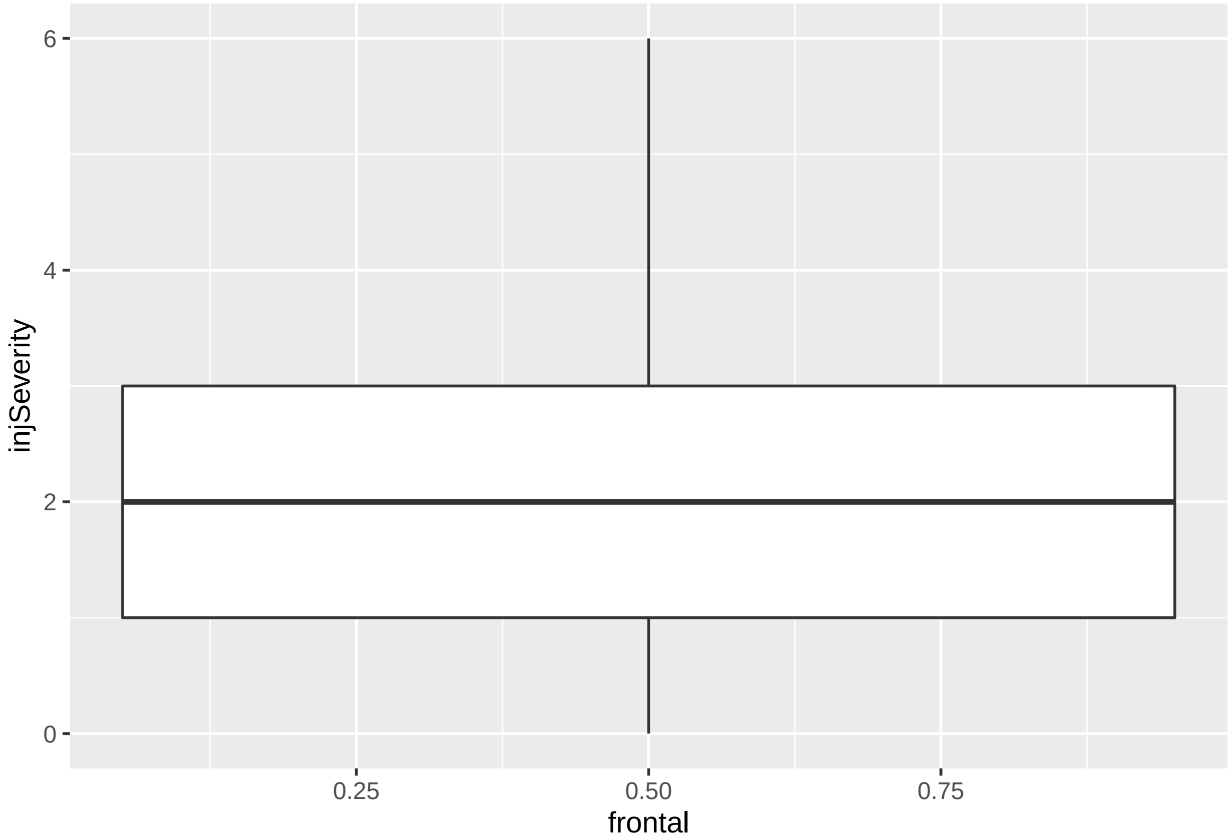
```
ggplot(nassCDS, aes(x=deploy, y=injSeverity)) +
  geom_boxplot()
```

```
## Warning: Continuous x aesthetic -- did you forget aes(group=...)?
## Removed 153 rows containing non-finite values (stat_boxplot).
```



```
ggplot(nassCDS, aes(x=frontal, y=injSeverity)) +
  geom_boxplot()
```

```
## Warning: Continuous x aesthetic -- did you forget aes(group=...)?
## Removed 153 rows containing non-finite values (stat_boxplot).
```



```
#grid.arrange(arrangeGrob(p1, p2, p3, p4, ncol=4),nrow=2, heights=c(1.25,1))
```

For the most part they were consistant accross the board. Most came out to be an injury of 2 which is "no incapasity". The only with more variation was the seatbelt which showed a higher injury rating of 3 with no seatbelt and a lower injury rating of 1 with a seatbelt.

I believe the order they're in is the best with the variation first and the consistant graphs next.

5. Are there other interesting patterns in the data worth presenting?

I think some interesting data to look at would be gender and age. For insurance purposes usually young men are charged the most. In addition to this there are a lot of distractions while driving and I believe mostly younger drivers do things such as text and drive so there could be a correlation there.

I would also be interested in looking at the weight data if it were more accurate.