Crashes in Delaware County in 2016

Graphics with ggplot2

Name: Rajnish Yadav

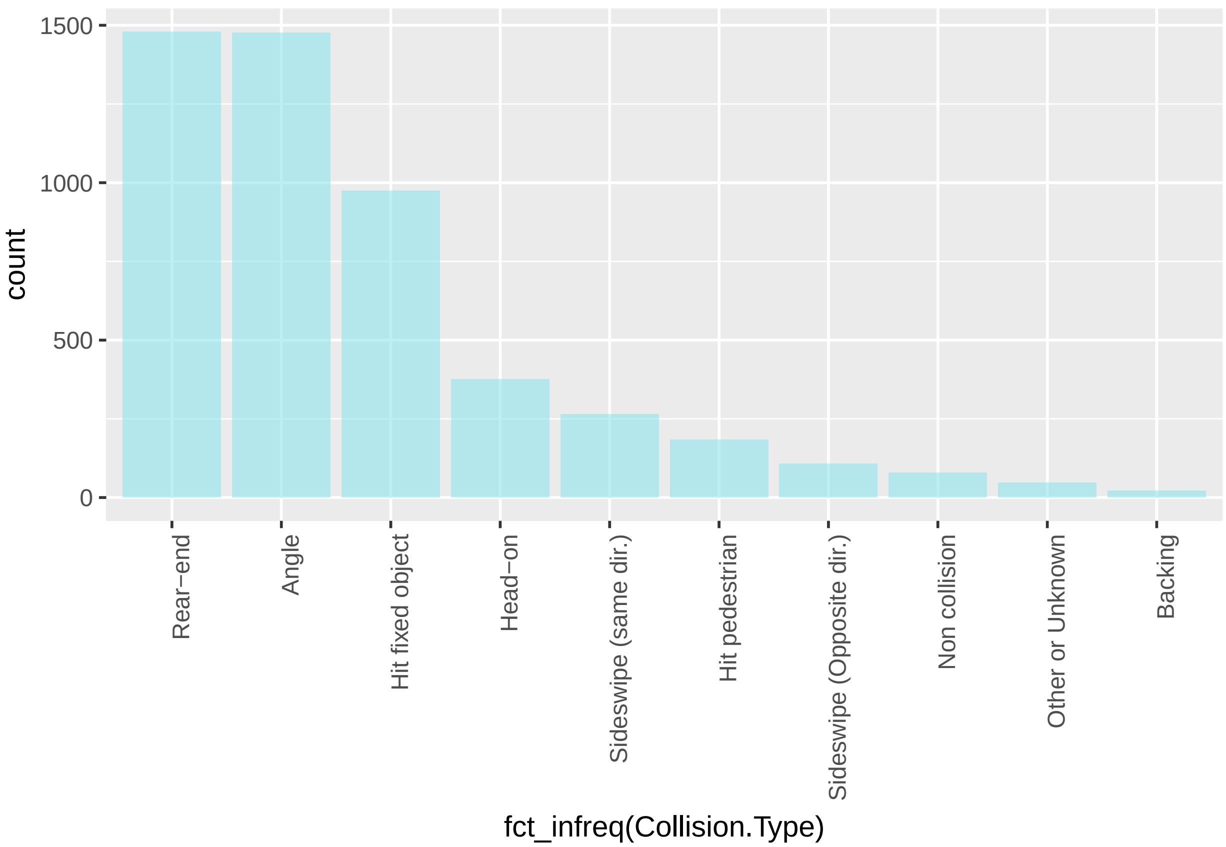
```r
#Put all necessary libraries here
library(ggplot2)
library(readr)
library(dplyr)
library(forcats)

#Load the dataset
crash_del <- read_csv("~/DataScience/Statistics/Data/crash_del.csv")
```

We will use data on all reported crashes in Delaware County for 2016. These data are provided by the Pennsylvania Department of Transportation. Information the various variables can be found at this link.

First of all, let's try to understand the frequency of the different types of collisions (`Collision.Type`).

For ease of reading, we would like to order the bars from most common to least common. In the `forcats` package, there is a function called `fct_infreq`, which will sort the categorical variable by frequency.

```r
ggplot(data = crash_del, aes(x = fct_infreq(Collision.Type))) +
  geom_bar(fill = "cadetblue2", alpha = 0.6) +
  theme(axis.text.x=element_text(angle=90, hjust=1))
```



Two most common collision types are 'Rear-end' and 'Angle'. 'Side swipe', 'Non collision', and 'Backing' are the least common collision types. From the bar plot, we can see that 'Rear-end', 'Angle' and 'Hit fixed object' account for ~80% of all collisions.

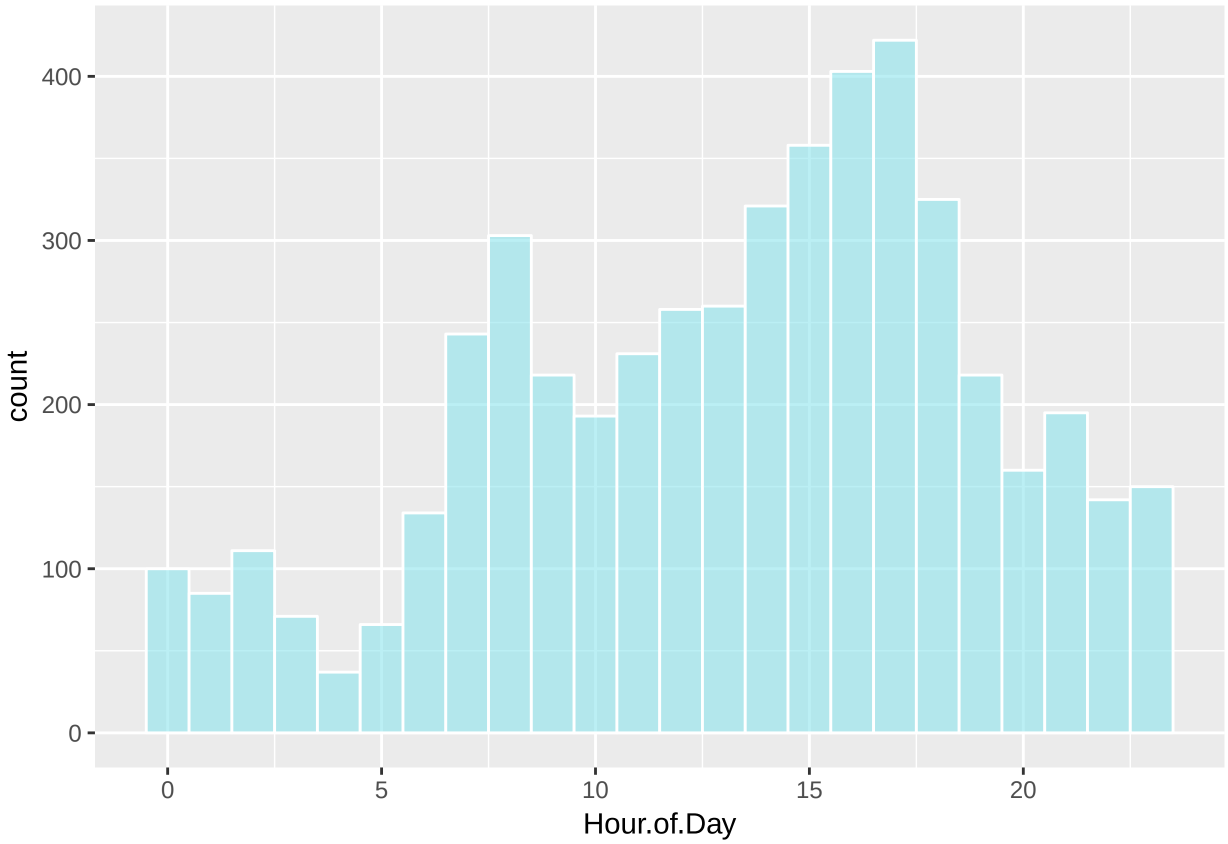
Now we want to plot the distribution of accidents over the hours of the day.

```
crash_del_2 <- filter(crash_del, Hour.of.Day != "99")
```

This code creates a new data frame where any observations where 'Hour.of.Day' = 99 is removed/filtered and this new dataset is assigned to crash_del_2. The size of the new dataset decreased by 3 to 5004 because there were exactly 3 values of 'Hour.of.Day' which were equal to 99.

Visually, some people prefer to see the separation between bars. This can be done by setting the color to white.

```
crash_del_2$Hour.of.Day <- as.integer(crash_del_2$Hour.of.Day)
ggplot(crash_del_2, aes(x = Hour.of.Day)) +
  geom_histogram(binwidth = 1, color = "white", fill = "cadetblue2", alpha = 0.6)
```

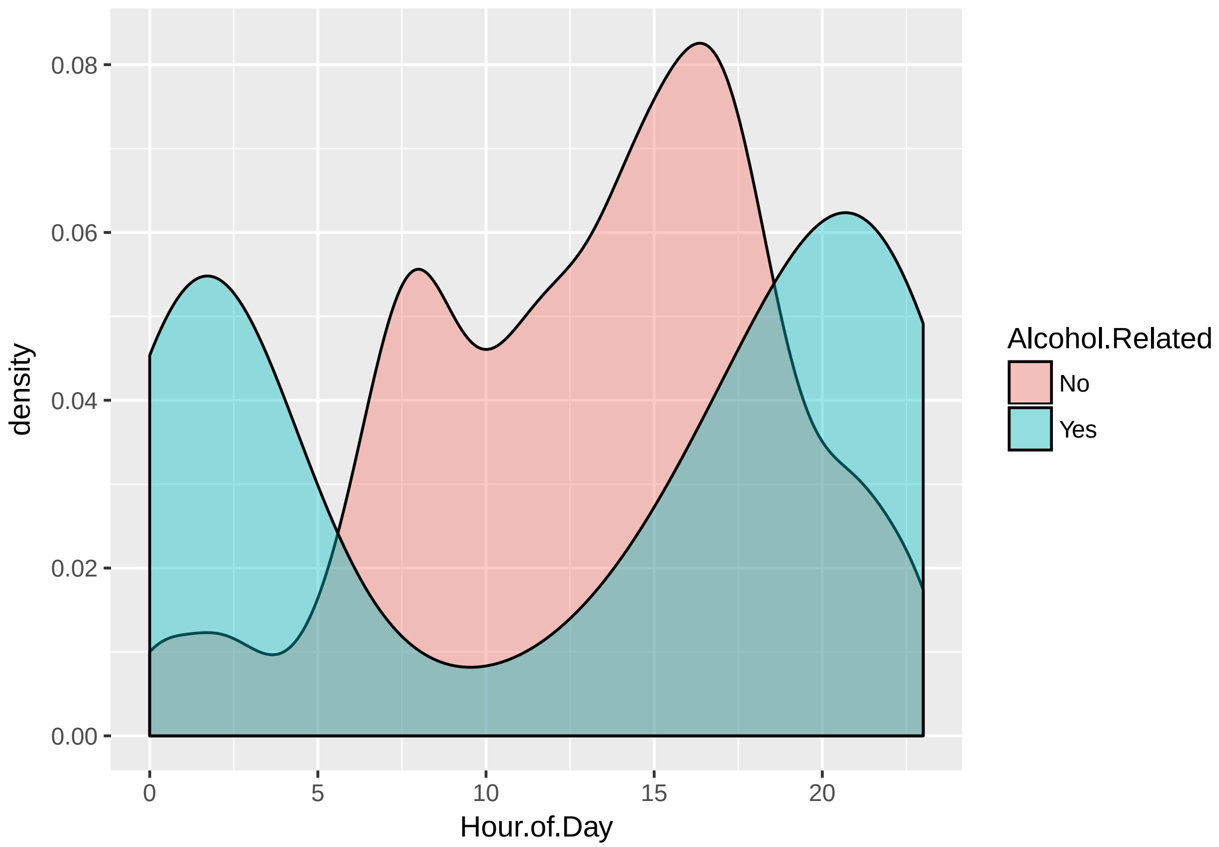


More accidents tend to happen during the morning rush hour (7 am to 8 am) and the afternoon rush hour (4 pm to 5 pm). However, the afternoon rush hour has more number of accidents compared to the morning rush hour. Not many people after midnight, so naturally less number of accidents occur from midnight to 5 am.

Note that we cannot conclude that it is safest to drive late at night because we do not know about the proportions of accidents occruing at night i.e there might be less cars on the road at night resulting into less number of accidents at any given hour. Moreover, it is also possible that more people might be driving under the influence of alcohol resulting into more number of accidents per 100 cars on the road. Thus, it is not a good idea to assume that it is safest to drive late at night.

Now, let's explore whether the distribution of accidents over the course of a day vary if the accident is alcohol related. We will use `crash_del_2`.

```
ggplot(crash_del_2, aes(x = Hour.of.Day, fill = Alcohol.Related)) +
  geom_density(alpha = 0.4)
```



Acidents related to alcohol occur more from midnight to about 2 am at night. One reason might be that people are driving back home from bars and parties at night under the influence. The number of alcohol related accidents start increasing from afternoon approx. 5 pm until night approx. 9 pm. Less alcohol related accidents occur during the day. More non-alcohol accidents occur during the day.

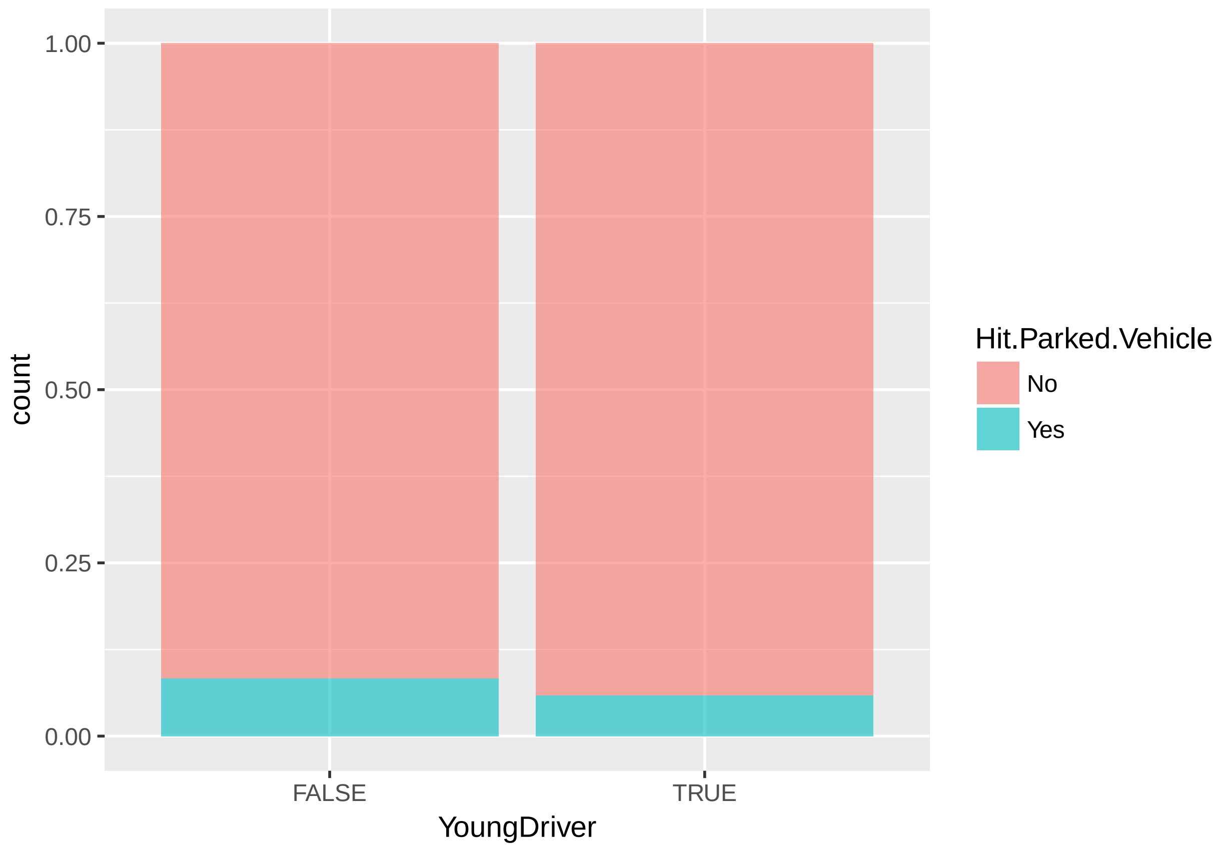
Young drivers and old drivers get a bad rap. Let's explore some accident trends associated with these types of drivers.

```
crash_del <- mutate(crash_del, YoungDriver = (Driver.16.Years.Old == "Yes" | Driver.17.Years.Old=="Yes"
```

The above code is adding two extra columns "YoungDriver" and "OlderDriver" to the crash_del dataset. 'YoungDriver' refer to the younger drivers who are between ages 16 and 18 inclusive. 'OlderDriver' refers to the older drivers who are older than or equal to 65 years. This adds the number of variables by 2 but the number of observations remain unchanged.

The plot below answers the following question: Given an accident occurred, are younger drivers more likely to hit a parked vehicle than non-young drivers?

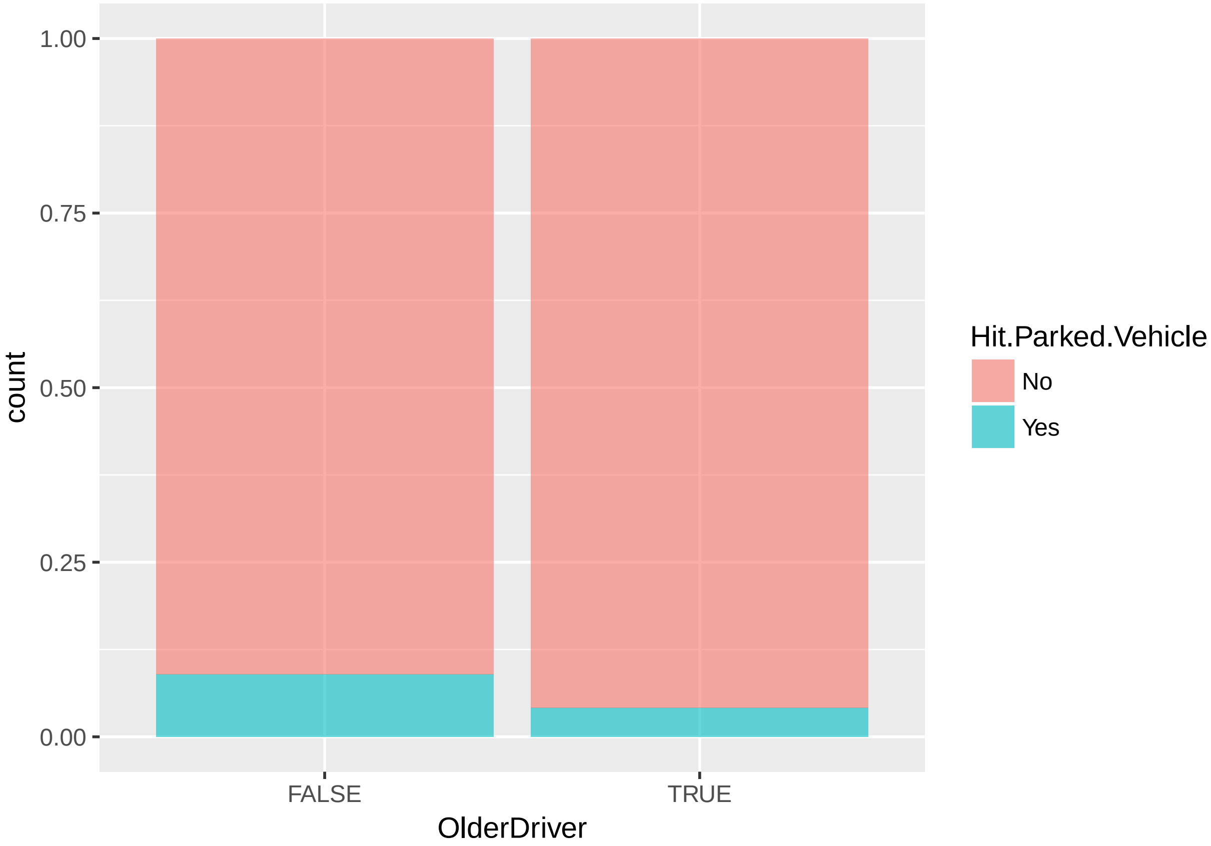
```
ggplot(data = crash_del, aes(x = YoungDriver, fill = Hit.Parked.Vehicle)) +
  geom_bar(position = "fill", alpha = 0.6)
```



Based on the bar plot above, we can say that given an accident occured, younger drivers between the ages of 16 and 18 inclusive are not more likely to hit a parked vehicle than non-young drivers.

The plot below answers the question posed above but this time for older drivers.

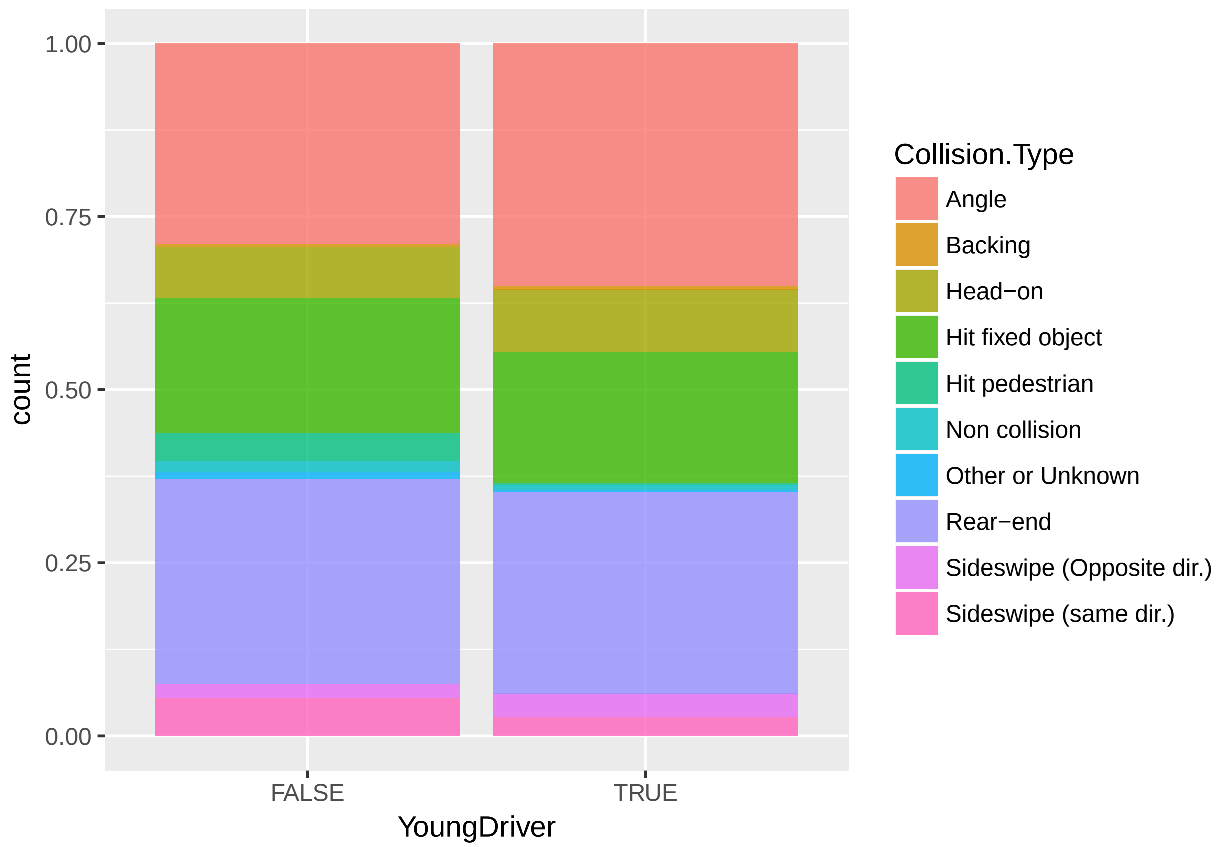
```
ggplot(data = crash_del, aes(x = OlderDriver, fill = Hit.Parked.Vehicle)) +
  geom_bar(position = "fill", alpha = 0.6)
```



Based on the bar plot above, we can say that given an accident occured, older drivers above age 65 (inclusive) are not more likely to hit a parked vehicle than drivers below age 65.

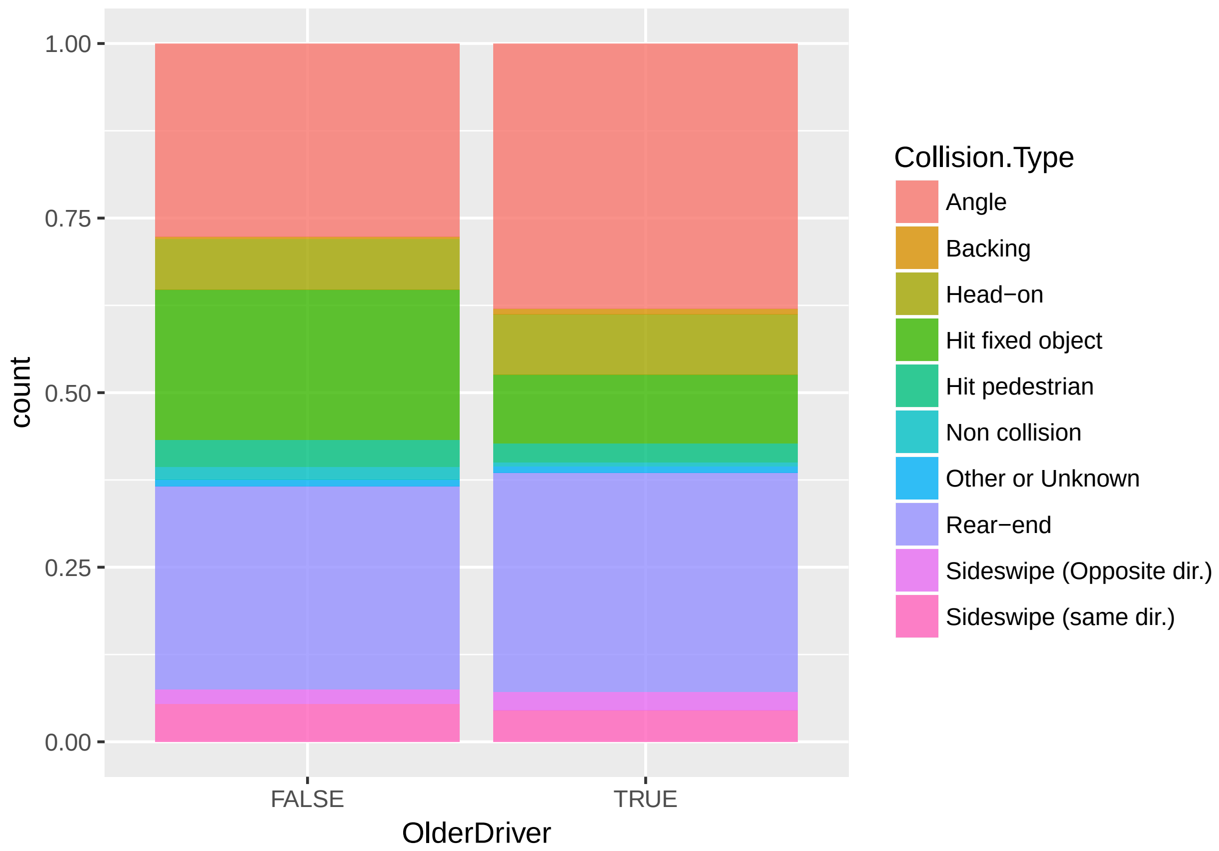
Let's create a barplot that displays the conditional proportions of the collision types for young and non-young drivers.

```
ggplot(crash_del, aes(x = YoungDriver, fill = Collision.Type)) +
  geom_bar(position = "fill", alpha = 0.8)
```



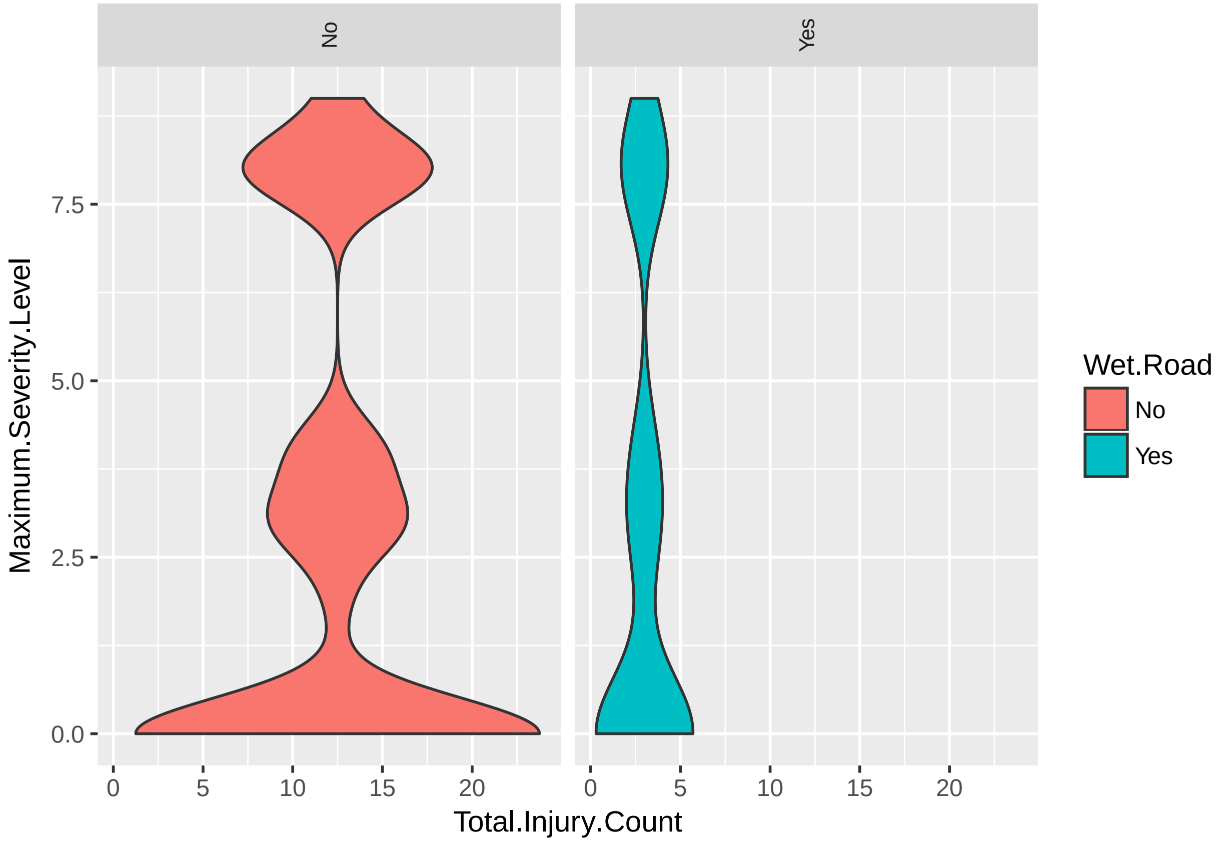
Let's create a barplot that displays the conditional proportions of the collision types for older and non-older drivers.

```
ggplot(crash_del, aes(x = OlderDriver, fill = Collision.Type)) +
  geom_bar(position = "fill", alpha = 0.8)
```



Based on your plots, 'Angle' collisions, 'Head-on' and 'Side swipe(Opposite dir.)' collisions collisions are more common when an younger driver is involved compared to non-young drivers. 'Angle' collisions and 'Rear-end' collisions are more common when older drivers are involved compared to non-older drivers.

```
ggplot(crash_del, aes(x = Total.Injury.Count, y = Maximum.Severity.Level, fill = Wet.Road)) +
  geom_violin() +
  facet_grid(. ~ Wet.Road) +
  theme(strip.text.x = element_text(size = 8, angle = 90))
```



Three variables are Total.Injury.Count (number of people getting injured in accidents), Maximun.Severity.Level (injury severity level; larger number implies more severe injury) and Wet.Road (whether the roads are wet or not). I would expect people to drive slowly when the road is wet. Hence, if an accident occurs less number of people are injured and their injury severity level is from small to severe; however, most people are not severely injured. When the road is not wet, people drive faster and might speed at times, resulting into more number of people getting more severely injured.
