# Case Studies

Amy Ly

5/8/2022

## Airbags and car accidents

First, I need to do a lot of data wrangling:

```
accidents <- get("nassCDS")

age <- rep(NA, nrow(accidents))
age[which(accidents$ageOFocc>=16 & accidents$ageOFocc<30)]="16-29 years"
age[which(accidents$ageOFocc>=30 & accidents$ageOFocc<45)]="30-44 years"
age[which(accidents$ageOFocc>=45 & accidents$ageOFocc<60)]="45-59 years"
age[which(accidents$ageOFocc>=60 & accidents$ageOFocc<75)]="60-74 years"
age[which(accidents$ageOFocc>=75 & accidents$ageOFocc<100)]="75-100 years"
accidents$age<-as.factor(age)

decade_veh <- rep(NA, nrow(accidents))
decade_veh[which(accidents$yearVeh>=2000)]="2000s"
decade_veh[which(accidents$yearVeh>=1990 & accidents$yearVeh<2000)]="1990s"
decade_veh[which(accidents$yearVeh>=1990 & accidents$yearVeh<2000)]="1990s"
decade_veh[which(accidents$yearVeh>=1980 & accidents$yearVeh<1990)]="1980s"
decade_veh[which(accidents$yearVeh>=1970 & accidents$yearVeh<1980)]="1970s"
decade_veh[which(accidents$yearVeh>=1960 & accidents$yearVeh<1970)]="1960s"
decade_veh[which(accidents$yearVeh>=1950 & accidents$yearVeh<1960)]="1950s"
accidents$decade_veh<-as.factor(decade_veh)

accidents$frontal <- as.character(accidents$frontal)

accidents <- accidents %>%
  mutate(my_label = ifelse(weight >= quantile(weight, 0.99),
                           "Beyond 0.99 Quantile", NA),
         my_label2 = ifelse(weight <= quantile(weight, 0.01),
                           "Below 0.01 Quantile", NA),
         Quantile = coalesce(my_label, my_label2)) %>%
  mutate(Impact = recode(frontal,
                 "0" = "Non-frontal",
                 "1" = "Frontal Impact"),
         Deploy = recode(deploy,
                 "0" = "Deployed",
                 "1" = "Didn't Deploy"),
          Airbag = recode(airbag,
                 "none" = "No Airbag",
                 "airbag" = "Airbag"),
```

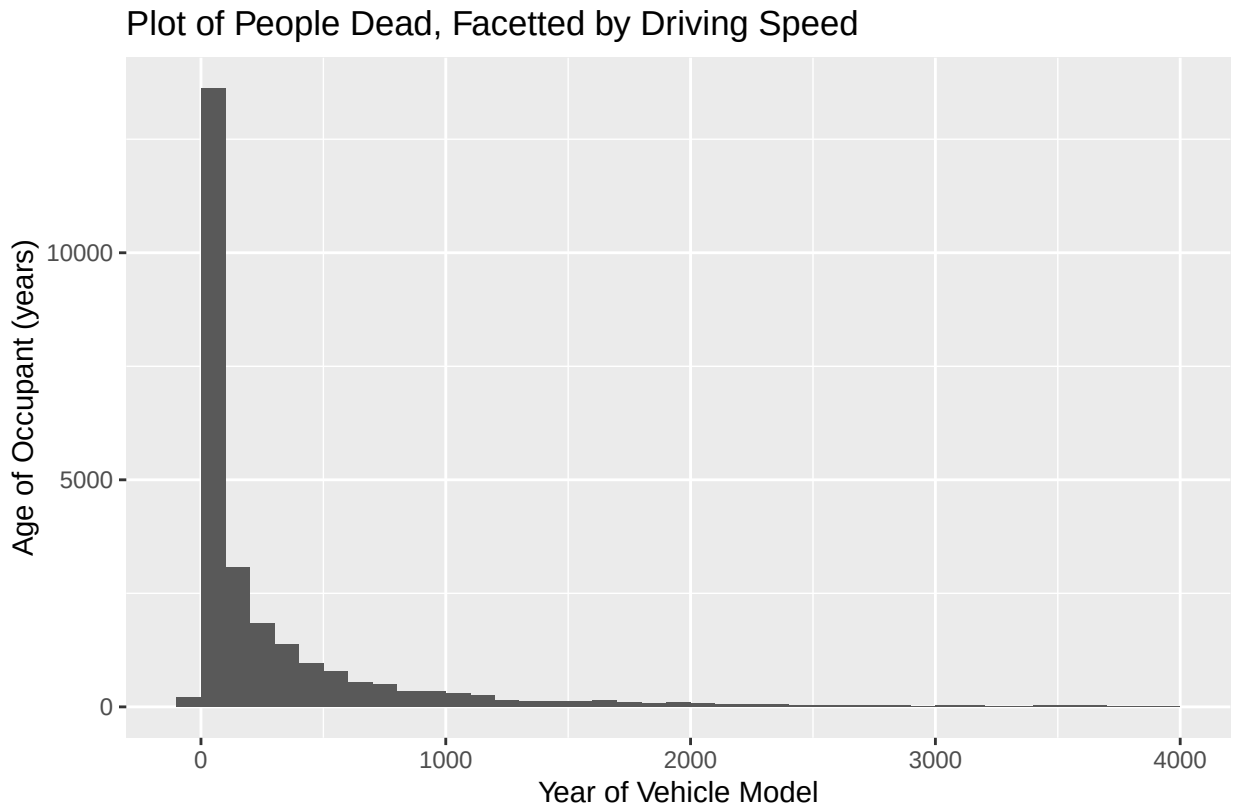
```
         Seatbelt = recode(seatbelt,
                "none" = "No Belt",
                "belted" = "Belted"),
         Severity = recode(as.factor(injSeverity),
                           "0" = "None",
                           "1" = "Possible Injury",
                           "2" = "No Incapacity",
                           "3" = "Incapacity",
                           "4" = "Killed",
                           "5" = "Unknown",
                           "6" = "Dead Prior to Crash"),
         Death = recode(dead,
                        "alive" = as.numeric("0"),
                        "dead" = as.numeric("1")))


accidents$Quantile[is.na(accidents$Quantile)] <- "Within (0.01, 0.99) Quantile"

accidents$frontal <- as.numeric(accidents$frontal)

#note that prior death refers to non-motor vehicle fatalities that are involved in a motor vehicle crash
```

1. Draw a histogram of the variable weight. The R help 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?

```
summary(accidents$weight)
```

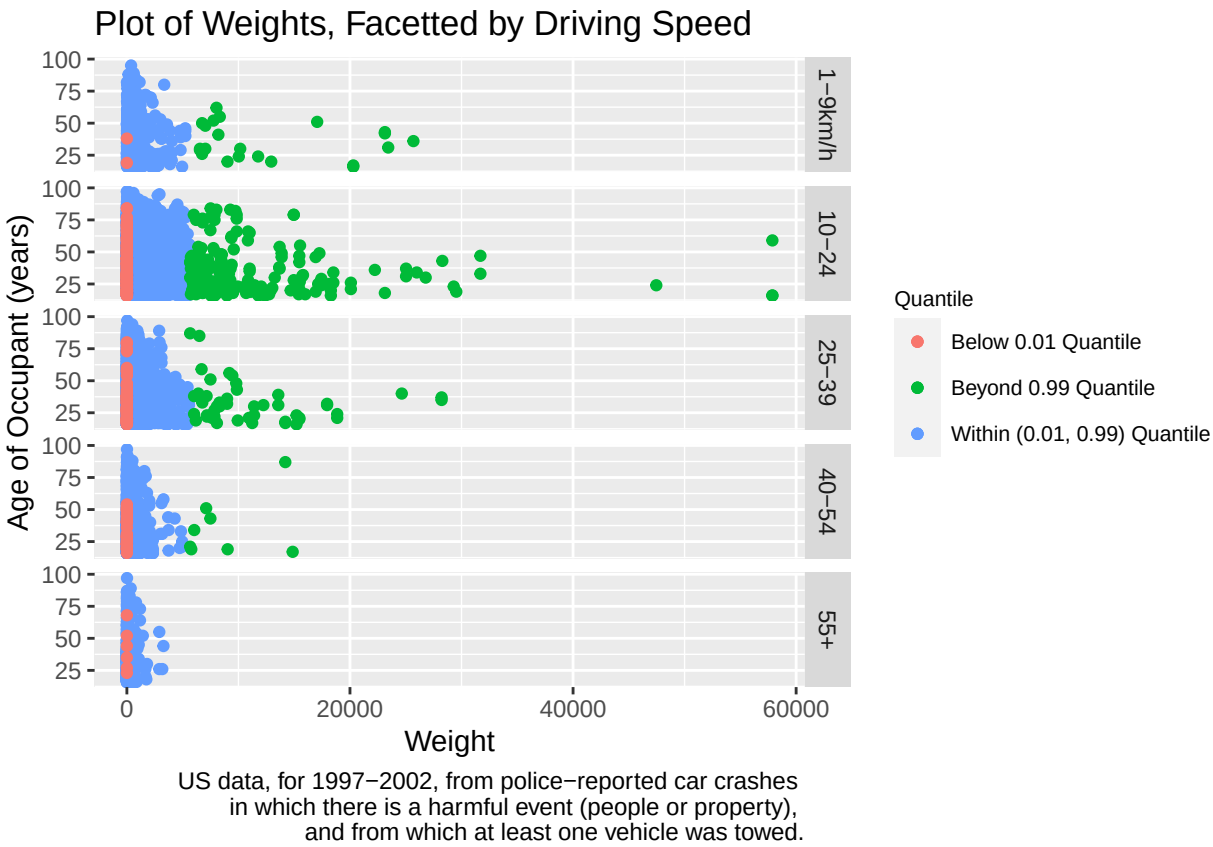
```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##     0.00    32.47    86.99   462.81   364.72 57871.59
```

```
ggplot(accidents, aes(x=weight)) +
  geom_histogram(breaks=seq(-100, 4000, by = 100))+
    labs(title = "Plot of People Dead, Facetted by Driving Speed",
       x = "Year of Vehicle Model",
       y = "Age of Occupant (years)",
       caption = "Heavily weighted cases are labeled in black and least weighted cases are labeled in bl
```




Heavily weighted cases are labeled in black and least weighted cases are labeled in blue.

```
ggplot(accidents, aes(y=ageOFocc,
                      x = weight,
                      color = Quantile))+
  facet_grid(vars(dvcat))+
  geom_point(data = subset(accidents, Quantile %in% c("Within (0.01, 0.99) Quantile"))) +
    geom_point(data = subset(accidents, Quantile %in% c("Beyond 0.99 Quantile")))+
  geom_point(data = subset(accidents, Quantile %in% c("Below 0.01 Quantile")))+
  labs(title = "Plot of Weights, Facetted by Driving Speed",
       x = "Weight",
       y = "Age of Occupant (years)",
       caption = "US data, for 1997-2002, from police-reported car crashes \n in which there is a harmfu
  theme(legend.title = element_text(size = 8),
                legend.text = element_text(size = 8))
```

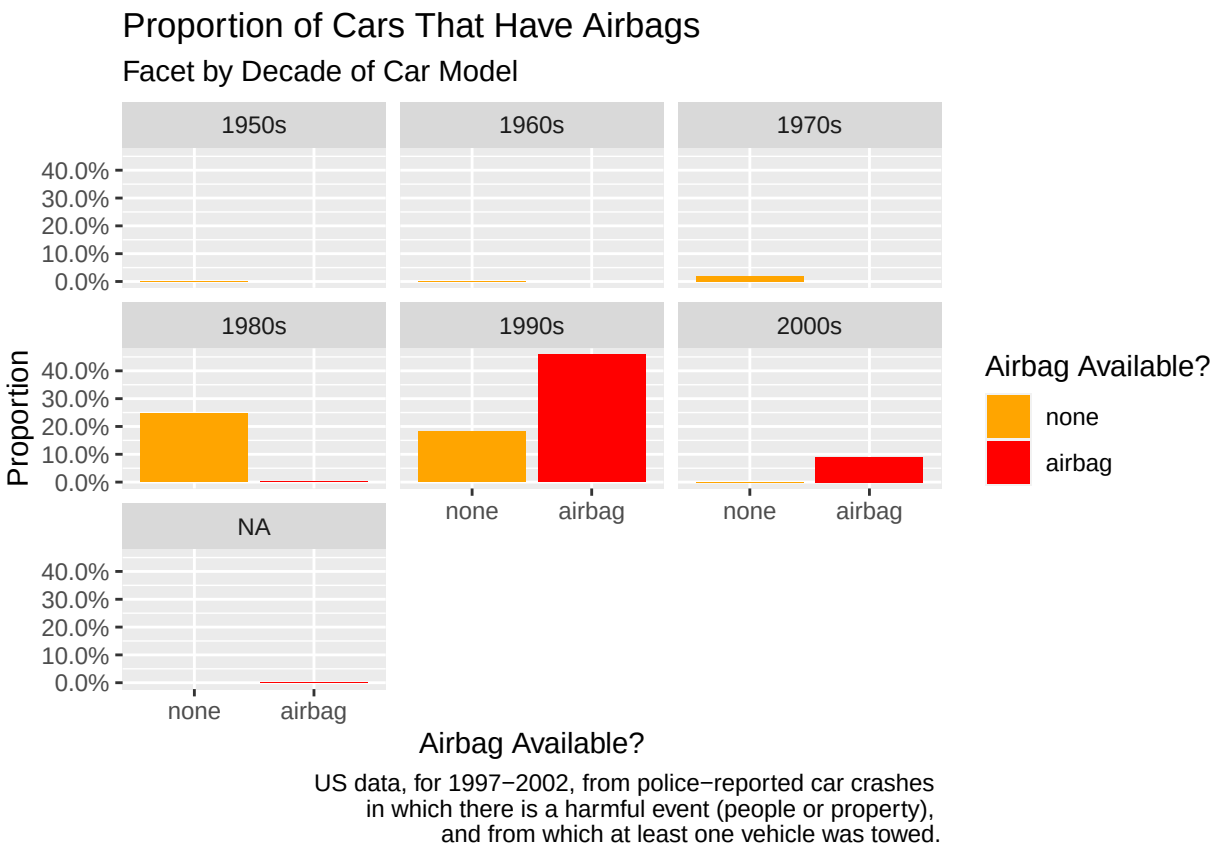



I would agree with the statement that the observation weights are of uncertain accuracy. I'm not sure what the weighting is being used for, especially when you have a range of weight values from 0 and 57871.59. You can tell from the histogram (extreme outliers beyond weight values of 4000 were excluded) that the distribution is very skewed towards the right (median = 86.99, mean = 462.78).

To investigate which cases have high weights and low weights, I facet based on a categorical factor to help alleviate some of the high density of points. Then I color coded the points based on where the weight value fell in the quantile. From the plot, you can visually see how that there is a high density of high weight values which should be considered as unusual values.

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

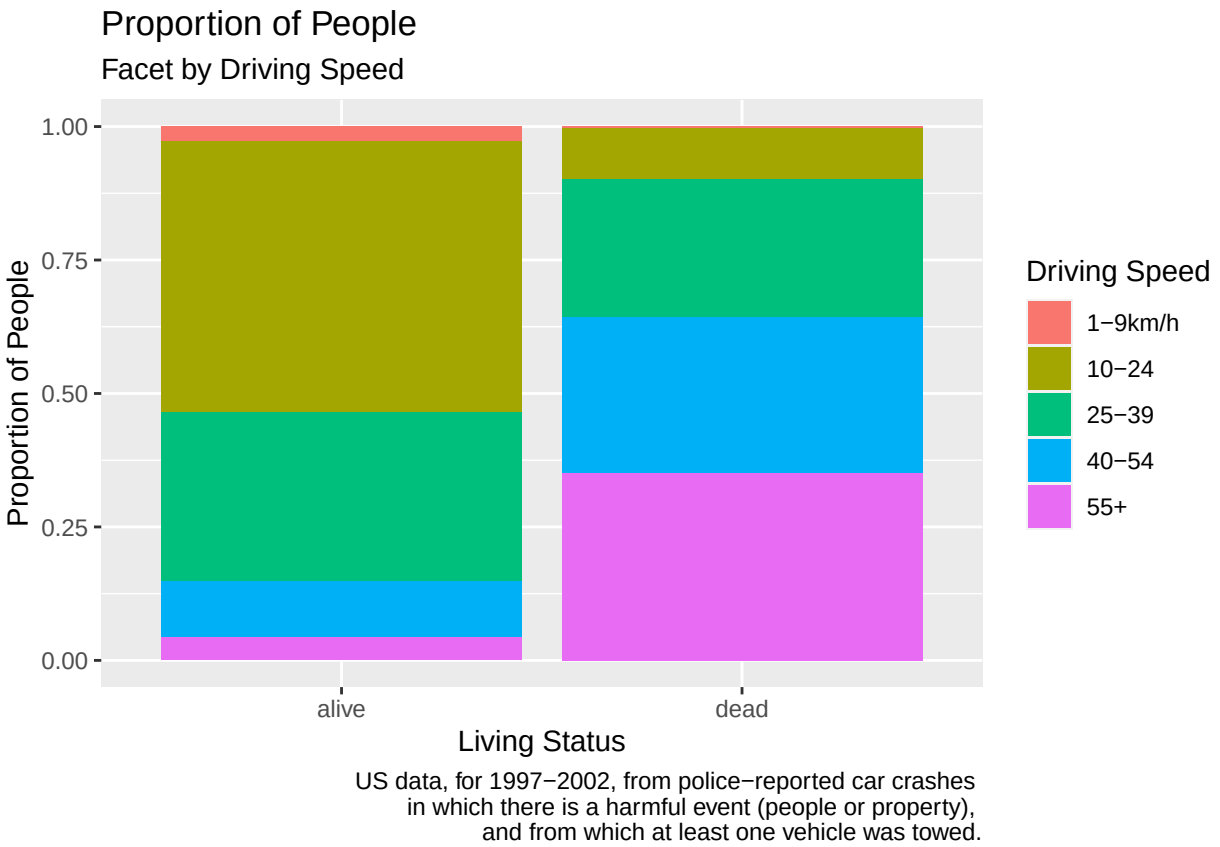
```
legend_title <- "Airbag Available?"

ggplot(accidents, aes(x=airbag)) +
  geom_bar(aes(fill=airbag)) +
facet_wrap(~decade_veh)+
   aes(y=stat(count)/sum(stat(count))) +
  scale_y_continuous(labels = scales::percent)+
    labs(title = "Proportion of Cars That Have Airbags",
         subtitle = "Facet by Decade of Car Model",
         x = "Airbag Available?",
       y = "Proportion",
       caption = "US data, for 1997-2002, from police-reported car crashes \n in which there is a harmfu
  scale_fill_manual(legend_title,values=c("orange","red"))
```




Newer vehicle models are more likely to have airbags. Note that there were some incomplete records.

3. How does death rate depend on vehicle speed?

```
legend_title2 <- "Driving Speed"
ggplot(accidents, aes(x=dead)) +
  geom_bar(position="fill", aes(fill=dvcat)) +
  labs(title = "Proportion of People",
         subtitle = "Facet by Driving Speed",
         x = "Living Status",
       y = "Proportion of People",
       caption = "US data, for 1997-2002, from police-reported car crashes \n in which there is a harmfu
    scale_fill_discrete(legend_title2)
```




The higher the driving speed, the higher the proportion of people who die.

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?

I grouped proportions of people who survived and died based on the following pairs:

- Airbags and Deployed: If a car had airbags, then I hypothesize that there would be a relationship between survival is airbags deployed properly
- Impact and Seatbelt: If occupants were belted in, then there may be less chance of injury depending on the type of impact.
- Impact and Airbag: Similarly, if there were airbags available, then there may be less chance of injury depending on the type of impact.

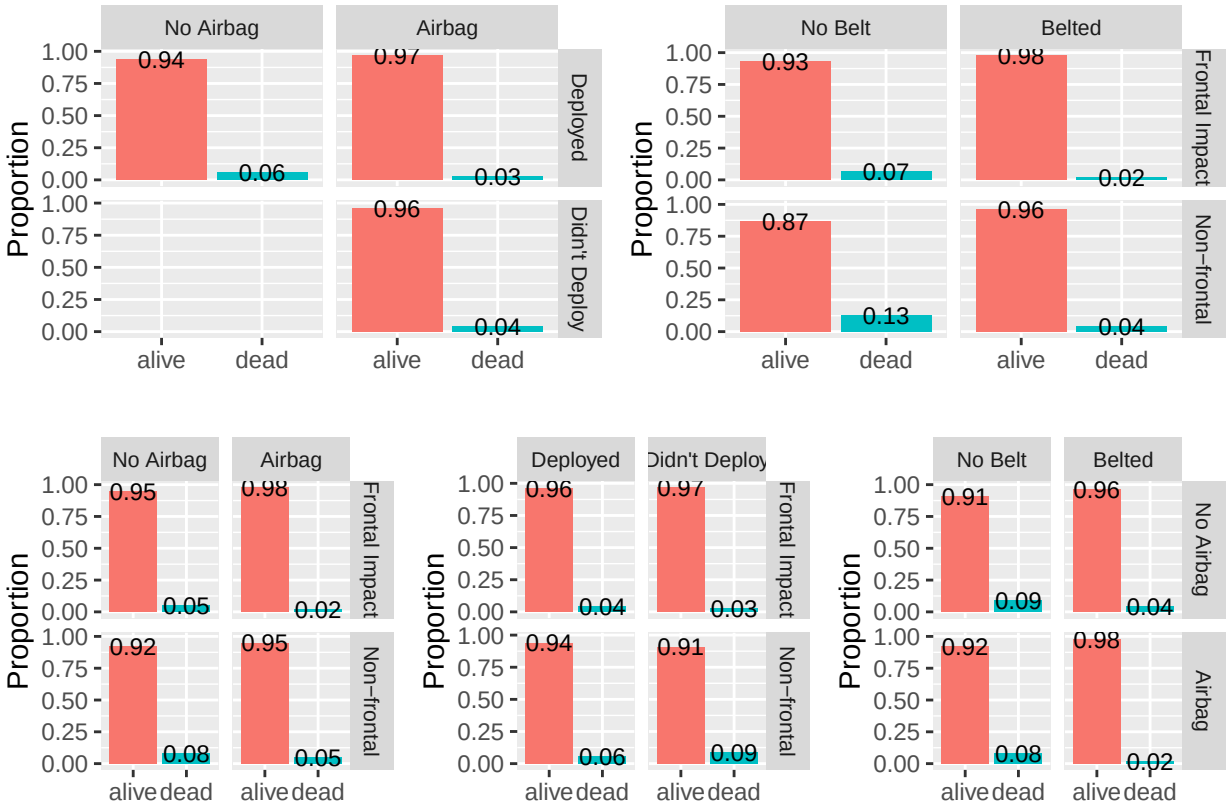
*Impact and Deploy: Perhaps the type of impact may cause an issue in if airbags were deployed. If so, then in cases where airbags didn't deploy, then there may be an increase of death.

- Airbags and Seatbelt: Does having airbags and seatbelt increase the chances of survival?

Alternative

```r
p1 <- accidents %>%
  group_by(Airbag, Deploy) %>%
  count(dead) %>%
  mutate(freq = round(n/sum(n), digits=2)) %>%
  ggplot(aes(x = dead, y = freq, fill = dead)) +
  geom_bar(stat="identity", position = 'dodge') +
  geom_text(aes(label = freq), size = 3)+
  facet_grid(vars(Deploy), vars(Airbag)) +
   theme(legend.position="none")+
  labs(y = " Proportion",
       x = "")+
  theme(strip.text.x = element_text(size = 7.5),
        strip.text.y = element_text(size = 7.5))

p2 <- accidents %>%
  group_by(Impact, Seatbelt) %>%
  count(dead) %>%
  mutate(freq = round(n/sum(n), digits=2)) %>%
  ggplot(aes(x = dead, y = freq, fill = dead)) +
  geom_bar(stat="identity", position = 'dodge') +
  geom_text(aes(label = freq), size = 3)+
  facet_grid(vars(Impact), vars(Seatbelt)) +
   theme(legend.position="none")+
  labs(y = " Proportion",
       x = "")+
  theme(strip.text.x = element_text(size = 7.5),
        strip.text.y = element_text(size = 7.5))

p3 <- accidents %>%
  group_by(Impact, Airbag) %>%
  count(dead) %>%
  mutate(freq = round(n/sum(n), digits=2)) %>%
  ggplot(aes(x = dead, y = freq, fill = dead)) +
  geom_bar(stat="identity", position = 'dodge') +
  geom_text(aes(label = freq), size = 3)+
  facet_grid(vars(Impact), vars(Airbag)) +
   theme(legend.position="none")+
  labs(y = " Proportion",
       x = "")+
  theme(strip.text.x = element_text(size = 7.5),
        strip.text.y = element_text(size = 7.5))

p4 <- accidents %>%
  group_by(Impact, Deploy) %>%
  count(dead) %>%
  mutate(freq = round(n/sum(n), digits=2)) %>%
  ggplot(aes(x = dead, y = freq, fill = dead)) +
  geom_bar(stat="identity", position = 'dodge') +
  geom_text(aes(label = freq), size = 3)+
  facet_grid(vars(Impact), vars(Deploy)) +
   theme(legend.position="none")+
  labs(y = " Proportion",
       x = "")+
```

```
  theme(strip.text.x = element_text(size = 7.5),
        strip.text.y = element_text(size = 7.5))

p5 <- accidents %>%
  group_by(Airbag, Seatbelt) %>%
  count(dead) %>%
  mutate(freq = round(n/sum(n), digits=2)) %>%
  ggplot(aes(x = dead, y = freq, fill = dead)) +
  geom_bar(stat="identity", position = 'dodge') +
  geom_text(aes(label = freq), size = 3)+
  facet_grid(vars(Airbag), vars(Seatbelt)) +
   theme(legend.position="none")+
  labs(y = " Proportion",
       x = "")+
  theme(strip.text.x = element_text(size = 7.5),
        strip.text.y = element_text(size = 7.5))

grid.arrange(arrangeGrob(p1, p2, ncol=2),
arrangeGrob(p3, p4, p5, ncol=3),
nrow=2, heights=c(1.5,1.5))
```







According to the plots above, it seems like that having your belt and airbag (assuming that it is properly deployed) results in the least proportion of deaths. It is not surprising to see the increase in proportions of death if occupants are not belted and their car has no airbag.

Another interesting finding is that the type of impact matters a lot in the effectiveness of safety measures (airbags and seatbelts). If the impacts are non-frontal, there is an even greater increase in proportion of

deaths when occupants do not have seatbelts or if their airbags didn't deploy.

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

```
accidents %>%
  filter(!is.na(decade_veh) & yearVeh >=1970) %>%
ggplot(aes(x=decade_veh, fill = sex)) +
  geom_bar(position = "dodge") +
  labs(title = "Number of Females and Males Involved In Accident",
         x = "Decade of Vehicle Model",
       y = "Count of People")+
  facet_wrap(~age) +
  theme(axis.text.x = element_text(angle = 75, hjust=1),
        legend.position = c(.85,.05))
```

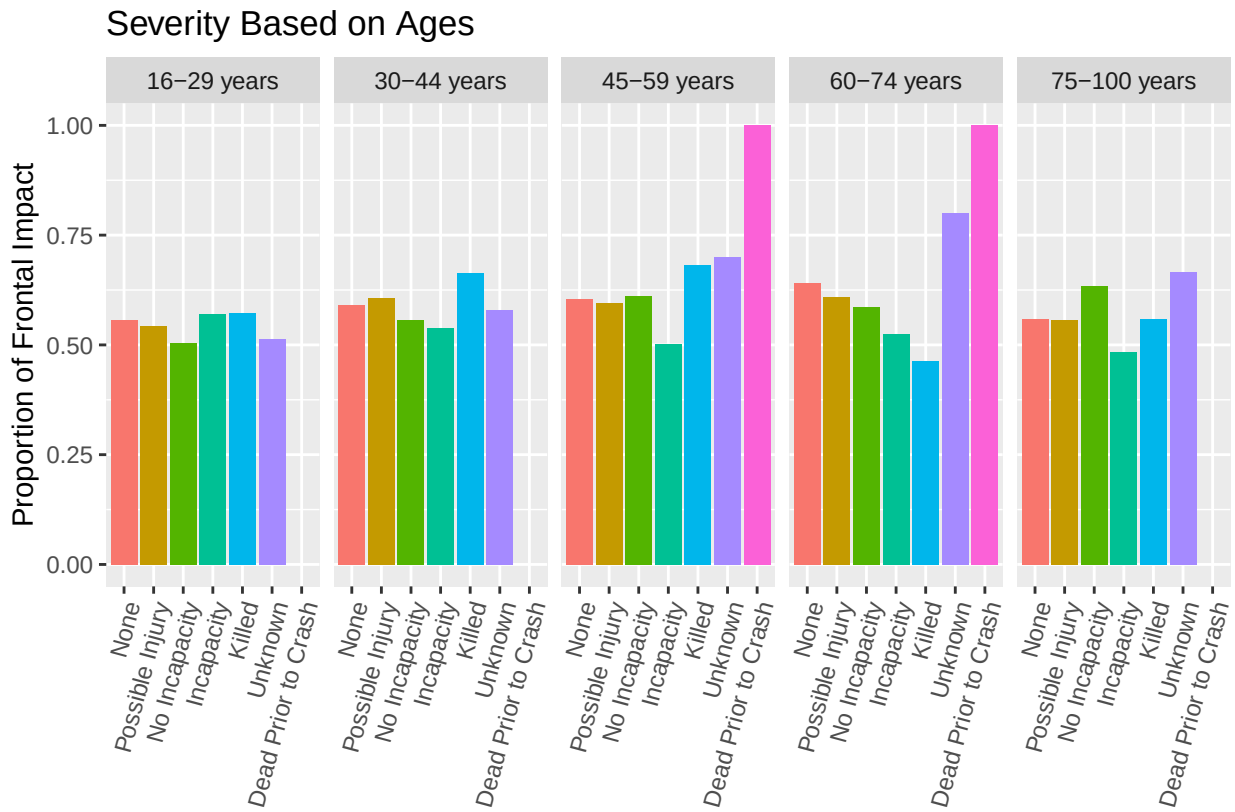

Overall, there was an increase in the number of accidents in the 1990s. The majority of people involved in accidents are those who are aged between 16-29 years of age, with more male occupants being involved.

```
accidents %>%
  filter(!is.na(Severity)) %>%
  group_by(frontal, Severity, age, dead, airbag) %>%
  summarise(N=sum(frontal)) %>%
  group_by(frontal, Severity, age) %>%
  mutate(Prop=N/sum(N)) %>%
ggplot(aes(x = Severity))+
  geom_bar(stat='identity', position = "dodge", aes(y=Prop, fill = Severity)) +
```

```
  labs(title = "Severity Based on Ages",
         x = "",
       y = "Proportion of Frontal Impact")+
  facet_grid(~age)+
   theme(axis.text.x = element_text(angle = 75, hjust=1),
         legend.position = "none")
```

```
## `summarise()` has grouped output by 'frontal', 'Severity', 'age', 'dead'. You
## can override using the `.groups` argument.
```

```
## Warning: Removed 79 rows containing missing values (geom_bar).
```



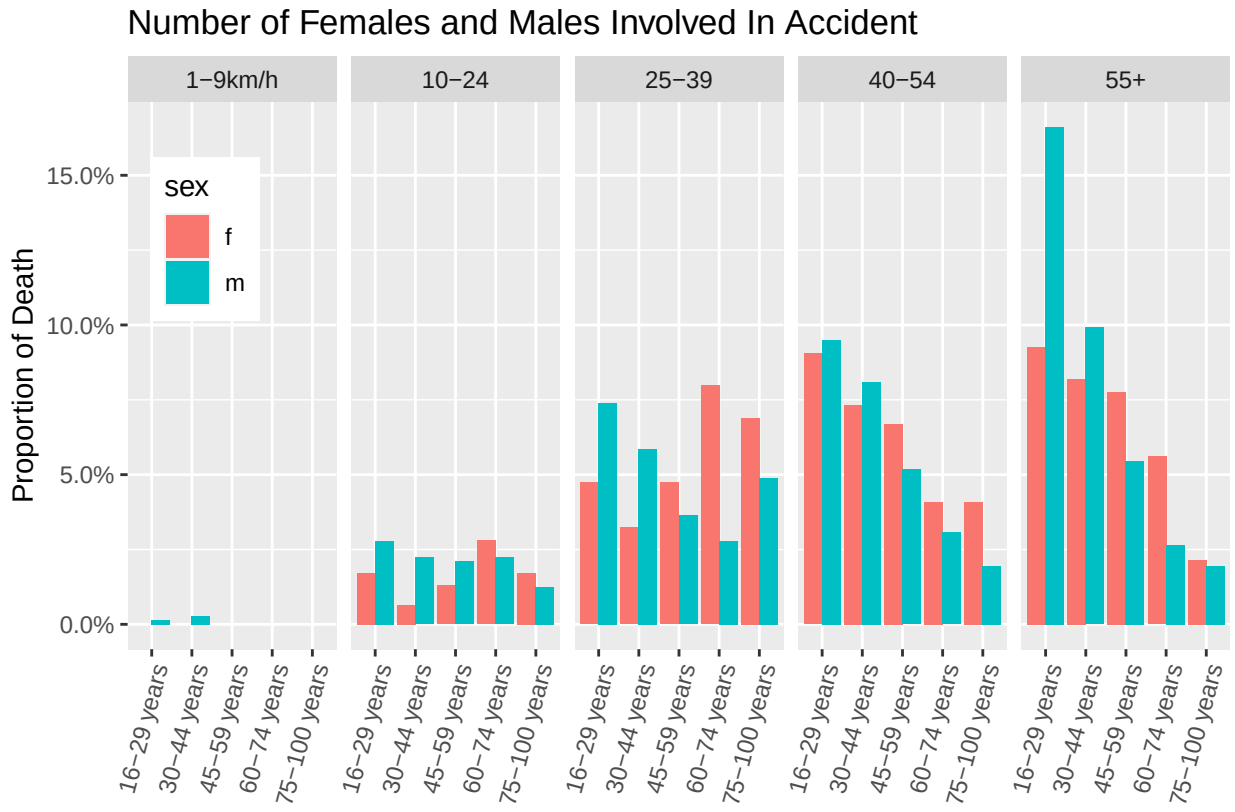

Those who are older (between 45-74 years of old) tend to suffer from other factors that lead to their death, which subsequently led them to crash their vehicles.

```
accidents %>%
  group_by(dead, sex, dvcat, age) %>%
  summarise(N=sum(Death)) %>%
  group_by(sex, dead,.drop=T) %>%
  mutate(Prop=N/sum(N)) %>%
  ggplot(aes(x=age))+
  geom_bar(stat='identity', position = "dodge", aes(y=Prop, fill = sex))+
  scale_y_continuous(labels = scales::percent)+
   labs(title = "Number of Females and Males Involved In Accident",
         x = "",
```

```
       y = "Proportion of Death")+
  facet_grid(~dvcat)+
   theme(axis.text.x = element_text(angle = 75, hjust=1),
         legend.position = c(0.07,0.75))
```

```
## `summarise()` has grouped output by 'dead', 'sex', 'dvcat'. You can override
## using the `.groups` argument.
```

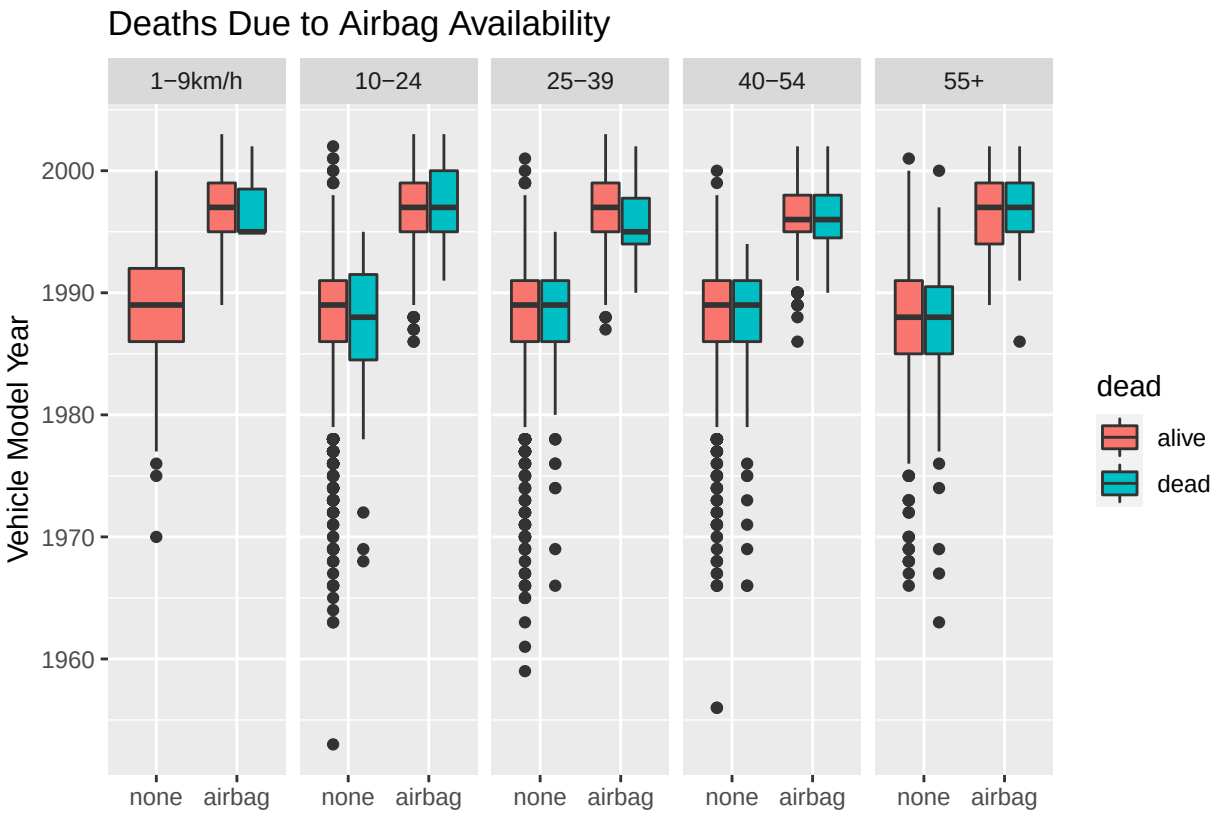
```
## Warning: Removed 50 rows containing missing values (geom_bar).
```




In general, a higher proportion of males deaths may have been caused by high speeding, particularly those between 16-44 years of age.

```
ggplot(accidents, aes(x = airbag, y=yearVeh))+
  geom_boxplot(aes(fill = dead))+
  labs(title = "Deaths Due to Airbag Availability",
         x = "",
       y = "Vehicle Model Year")+
  facet_grid(~dvcat)
```

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




Airbags only really became available between 1995 - 2000s. If airbags were available, then the number of deaths caused by high driving speed tend to be equivalent or less.