Case Study: Airbag and Car Accidents

Alireza Mostafizi

May 15, 2018

This documnet answers to the questions regarding the *Airbag and Car Accidents* case study. I start with installign *DAAG* packages and loading the data.

```
#install.packages("DAAG")
library(DAAG)
```

```
## Loading required package: lattice
```

```
data('nassCDS')
#?nassCDS
```

This dataset includes the information regarding 26217 police-reported accidents from 1997 to 2002 where there was either property or occupant (only front seat) injury and at least one car was towed. The dataset has 15 variables including, but not limited to, estimated impact speed and the type of accident, characteristics of the vehicle such as safety and year of manufacturing, drivers characteristics such as sex and age, injury level, and if the airbags were deployed or not.

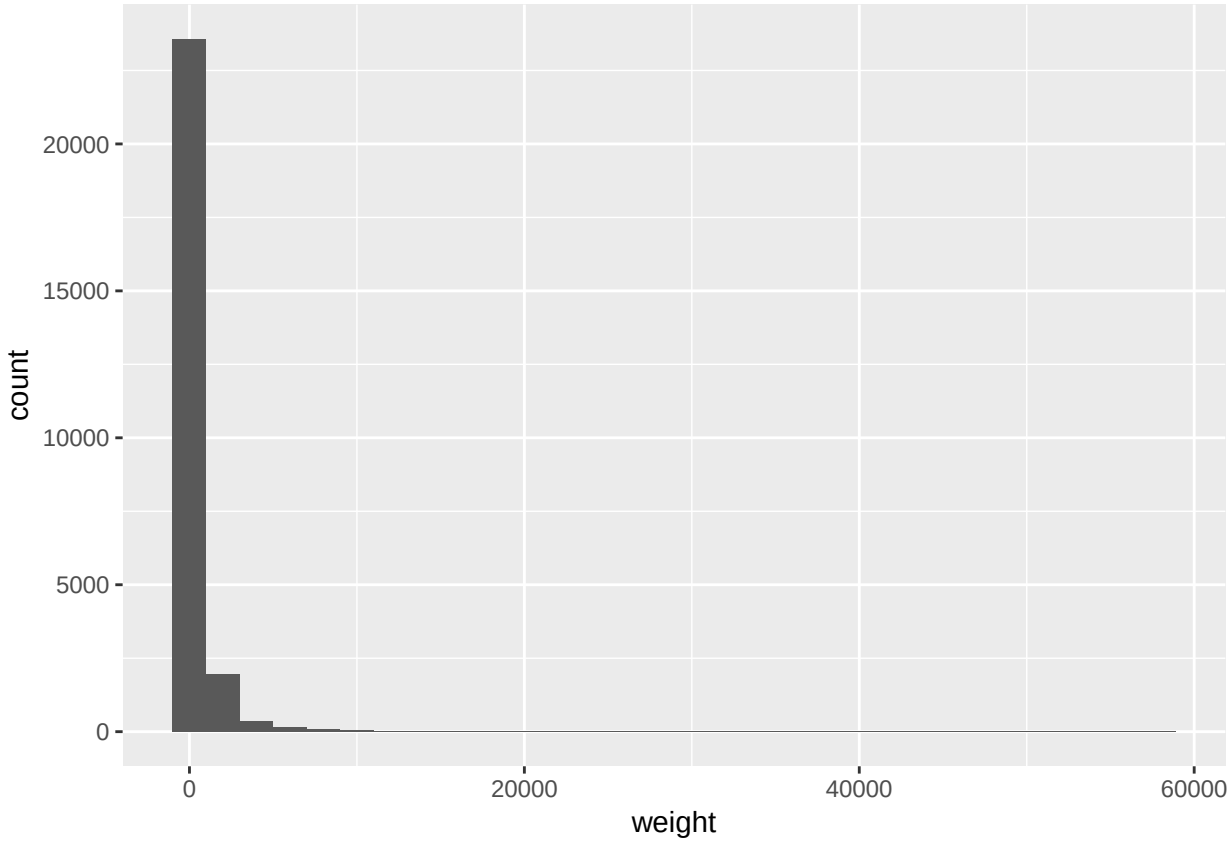
With this in mind, let's answer the questions in hand,

1. Histogram of the variable *weight*.

```
library(ggplot2)

ggplot(data = nassCDS) +
  geom_histogram(aes(weight))
```

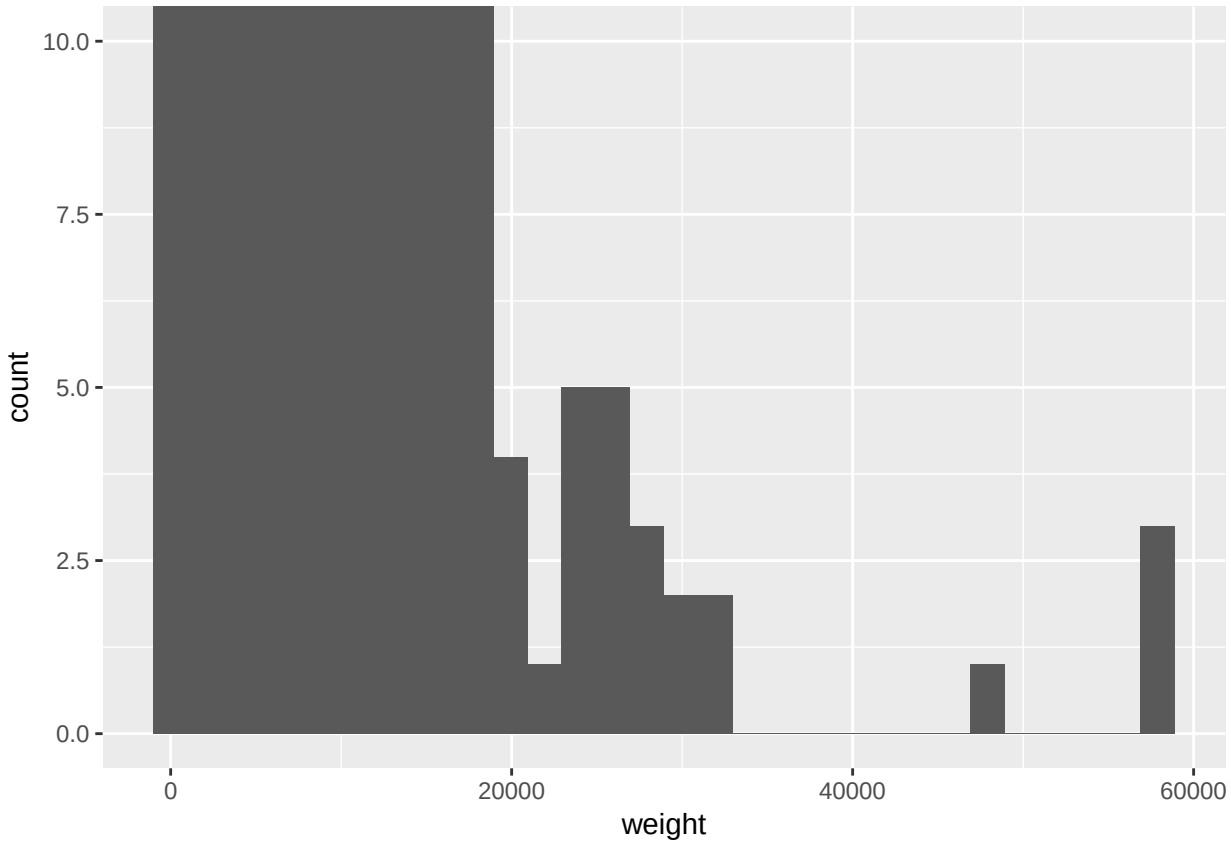
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



As the dataset help file mentions, the weughts have uncertain accuracy. Unexpectedly large x range for the historgam above shows that there are extremely large values for weight. If we zoom into the y axis and limit it from 0 to 10,

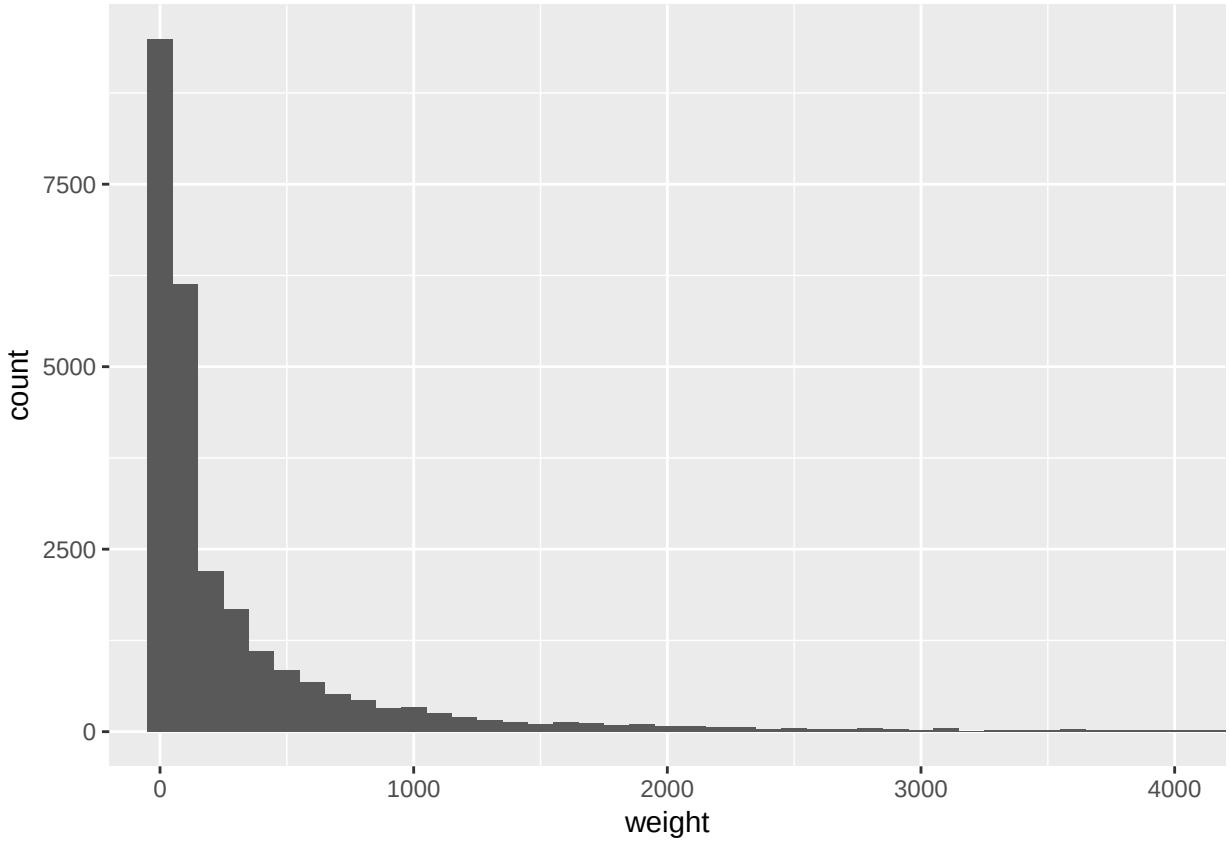
```
ggplot(data = nassCDS) +
  geom_histogram(aes(weight)) +
  coord_cartesian(ylim = c(0,10))
```

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



There are observations, although a few, with weights of roughly 5000 and 6000. And this shows that the data has uncertain accuracy. We can see the distributiuon of the weights with the following plot in a better way,

```
ggplot(data = nassCDS) +
  geom_histogram(aes(weight), binwidth = 100) +
  coord_cartesian(xlim = c(0,4000))
```



It can be seen that most of the observations have weights of lower than 500. We can plot Parallel Coordinate plot for the observations with high and low weights to investiage what differenciates these observatiosn from each other.

To do this, we have to filter the data, and add a categorical variable that classifies low and high weights. For the purpose of this assignment, I assue any weight above 4000 is high and any weight less than 100 is low. Thus, we have,

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
extreme_weight_nassCDS <- nassCDS %>%
                          filter(weight > 4000 | weight < 1) %>%
                          mutate(high = ifelse(weight > 4000, TRUE, FALSE))

#install.packages("GGally")
library(GGally)
```

```
##
```

```
## Attaching package: 'GGally'

## The following object is masked from 'package:dplyr':
##
##     nasa
```

```
ggparcoord(extreme_weight_nassCDS[order(extreme_weight_nassCDS$high),], columns=1:14,
           groupColumn = "high", scale="uniminmax", alphaLines = 0.2) +
           xlab("") + ylab("") +
           scale_colour_manual(values = c("blue","red")) +
           theme(axis.ticks.y = element_blank(),
           axis.text.y = element_blank())
```