GRAPHICS PORTFOLIO ASSIGNMENT

Is 2nd August the deadliest car crash day?

Aaradhana Bharill, CM Box 3202

Introduction

This report is trying to study the validity of the Twitter story claim that 2nd August is the deadliest day for car crashes. This is a link to the Twitter story for reference: https://twitter.com/i/moments/1025000711539572737?cn=ZmxleGlibGVfcmVjc18y&refsrc=email. The source of the data used for this report is: Fatality Analysis Reporting System (FARS) from National Highway Traffic Safety Administration's (NHTSA).

Methods

First, a side-by-side boxplot of `Day` vs `Total number of deaths` was plotted.

```
ggplot(data=courses.df,mapping=aes(x=factor(day), y=TotalDeaths)) +
  geom_boxplot() +
  geom_boxplot(data=courses.df[courses.df$day=="2",],
               aes(x=factor(day), y=TotalDeaths),fill="dark blue")+
  labs(title="Car crashes",x="Day", y="Total number of deaths") +
  theme_bw()
```

Then, a side-by-side boxplot of Month vs. Total number of deaths was plotted.

```
ggplot(data=courses.df,mapping=aes(x=factor(month), y=TotalDeaths)) +
  geom_boxplot() +
  geom_boxplot(data=courses.df[courses.df$month=="August",],
                       aes(x=factor(month), y=TotalDeaths),fill="dark blue")
+
  labs(title="Car crashes", x="Month", y="Total number of deaths") +
  theme_bw()+
  theme(axis.text.x = element_text(angle=15))
```

However it was not sufficient to look at the impact of the day and month alone. The combined effect of both day and month was important since they are not independent in how they affect the total number of deaths. So, then, 12 side-by-side boxplots for each month, were plotted for Day vs. Total number of deaths.

```
ggplot(data=courses.df,mapping=aes(x=factor(day), y=TotalDeaths)) +
  geom_boxplot() +
  labs(title="Car crashes", x="Day", y="Total number of deaths") +
  facet_wrap( ~month) +
  theme_bw()+
  scale_x_discrete(breaks=c(0,5,10,15,20,25,30))
```

Results

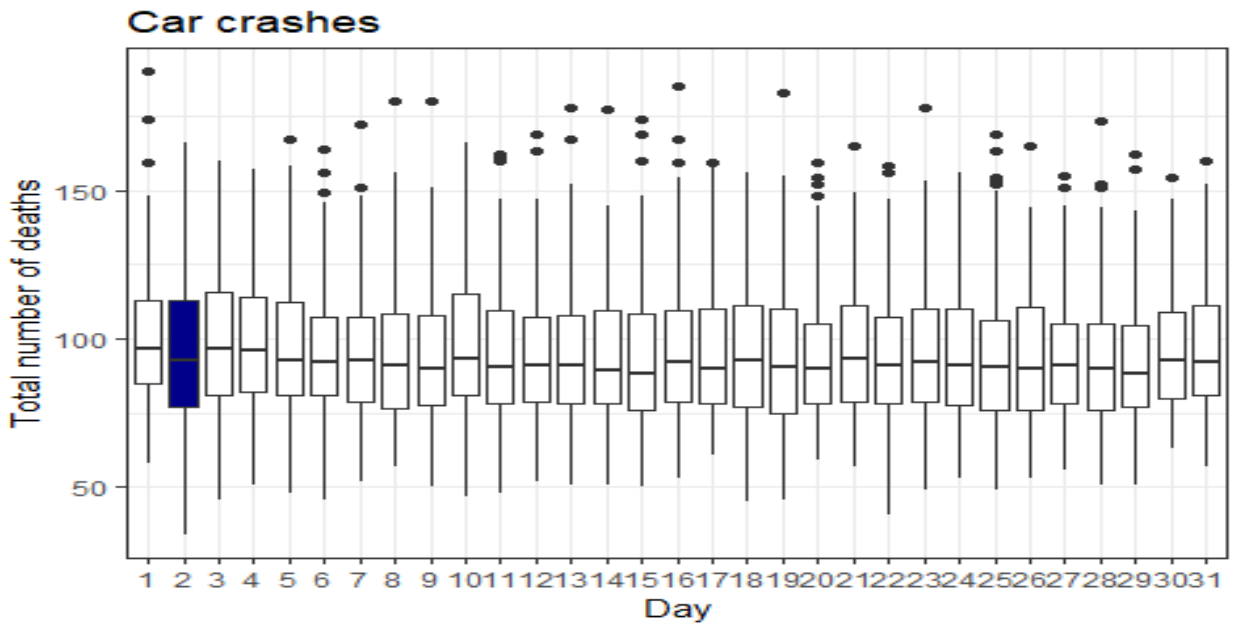

Figure 1: Car crashes: day vs. total number of deaths.

In figure 1, while the mean and quartiles of the second day of every month, doesn't seem to be higher than most other days, the upper whisker is the highest or seemingly at least the second highest after the day 10.

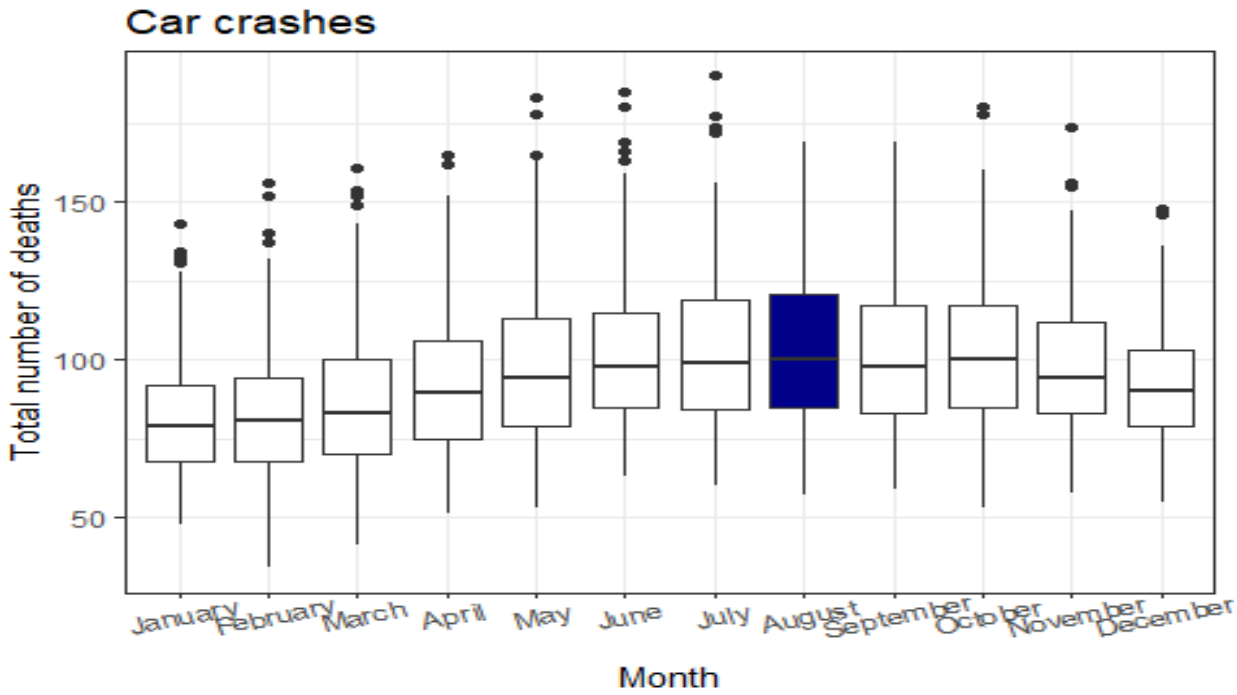

Figure 2: Car crashes: month vs. total number of deaths.

In figure 2, August has the highest mean and higher quartiles, and the highest upper whisker or seemingly second highest after September.

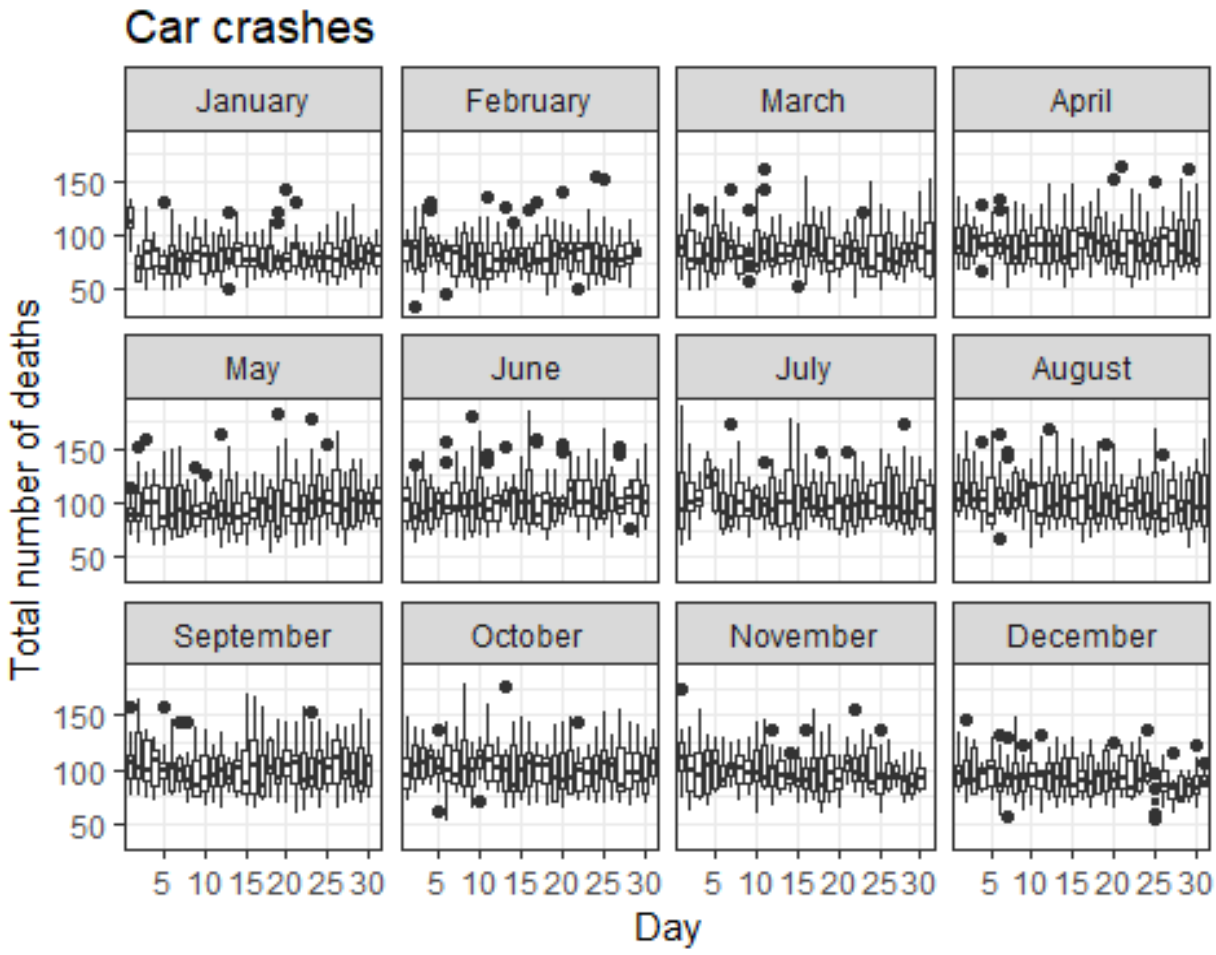

Figure 3: Car crashes: day vs. total number of deaths for each month.

In figure 3, it is apparent that 2nd August doesn't have the highest mean or quartiles or whiskers or even outliers and it really doesn't stand out in any way at all in terms of total number of deaths compared to the other days.

It seems like whoever wrote the twitter story, made their conclusion from the fact that day 2 has the highest boxplot whisker and August has the highest boxplot whisker too. However, it is clear from figure 3 that when the month and day are put together, 2nd August doesn't stand out in any way compared to the 365 other days of the year, hence it is not the deadliest car crash day.

Appendix

```
#Load required packages
library(plyr)
library(ggplot2)

#Read into data frame
courses.dff <- read.csv(file="FARS.csv", header=TRUE, stringsAsFactors=FALSE)

#remove rows with missing data (weird first row)
courses.df<- courses.dff[complete.cases(courses.dff), ]

#change month factor levels to be better understandable
courses.df <- mutate(courses.df, month = factor(month,
                    levels= c(1,2,3,4,5,6,7,8,9,10,11,12),
                    labels=c("January","February","March","April","May","June"
,
                             "July","August","September","October","November",
"December")))
#Boxplot day vs. total number of deaths
ggplot(data=courses.df,mapping=aes(x=factor(day), y=TotalDeaths)) +
  geom_boxplot() +
  geom_boxplot(data=courses.df[courses.df$day=="2",],
               aes(x=factor(day), y=TotalDeaths),fill="dark blue")+
  labs(title="Car crashes",x="Day", y="Total number of deaths") +
  theme_bw()

#Boxplot month vs. total number of deaths
ggplot(data=courses.df,mapping=aes(x=factor(month), y=TotalDeaths)) +
  geom_boxplot() +
  geom_boxplot(data=courses.df[courses.df$month=="August",],
                       aes(x=factor(month), y=TotalDeaths),fill="dark blue")
+
  labs(title="Car crashes", x="Month", y="Total number of deaths") +
  theme_bw()+
  theme(axis.text.x = element_text(angle=15))

#Boxplot day vs. total number of deaths for each month
ggplot(data=courses.df,mapping=aes(x=factor(day), y=TotalDeaths)) +
  geom_boxplot() +
  labs(title="Car crashes", x="Day", y="Total number of deaths") +
  facet_wrap( ~month) +
  theme_bw()+
  scale_x_discrete(breaks=c(0,5,10,15,20,25,30))
```