Traffic Violation Data Analysis

Here is the cleanup code for this data set. We produced a new csv file by removing all rows with NA's. This may cause bias, but since we are analyzing speeding and location of the incident, observations missing these data are not useful.

```
citationDf = read.csv("~/OpenDataDay/ItsATrap/data/highway_patrol_data.csv")

#Convert factors to numeric for speed quantities (mph)
citationDf$Alleged_Speed = as.numeric(levels(citationDf$Alleged_Speed)[citationDf$Alleged_Speed])
```

```
## Warning: NAs introduced by coercion
```

```
citationDf$Posted_Speed = as.numeric(levels(citationDf$Posted_Speed)[citationDf$Posted_Speed])
```

```
## Warning: NAs introduced by coercion
```

```
#Cut out non-speeding related violations
speedingDf = citationDf[which(citationDf$Violation_Category == "SPEED - UNSAFE" |
                                citationDf$Violation_Category == "SPEEDING - OVER LIMITS"),]

#Cut out missing speeds or below-the-limit speeds
speedingDfNoNA = speedingDf[which(!(is.na(speedingDf$Alleged_Speed) | is.na(speedingDf$Posted_Speed))),]
speedingDfNoNA = speedingDfNoNA[which(speedingDfNoNA$Alleged_Speed > speedingDfNoNA$Posted_Speed),]

#Convert lat/lng into numerics and remove missing locations
speedingDfNoNA$Latitude = as.numeric(levels(speedingDfNoNA$Latitude)[speedingDfNoNA$Latitude])
```

```
## Warning: NAs introduced by coercion
```

```
speedingDfNoNA$Longitude = as.numeric(levels(speedingDfNoNA$Longitude)[speedingDfNoNA$Longitude])
```

```
## Warning: NAs introduced by coercion
```

```
speedingDfNoNA = speedingDfNoNA[which(!(is.na(speedingDfNoNA$Longitude) | is.na(speedingDfNoNA$Latitude

#Remove 0 mph posted speeds
speedingDfNoNA = speedingDfNoNA[which(speedingDfNoNA$Posted_Speed > 0),]

#Create location/riskFactor csv
mapsDf = speedingDfNoNA[,c("Latitude", "Longitude")]
meanSpeeding = mean(speedingDfNoNA$Alleged_Speed - speedingDfNoNA$Posted_Speed)
logistic = function(x) 1/(1+exp((x-meanSpeeding)/9))
riskFactors = logistic(speedingDfNoNA$Alleged_Speed - speedingDfNoNA$Posted_Speed)
```

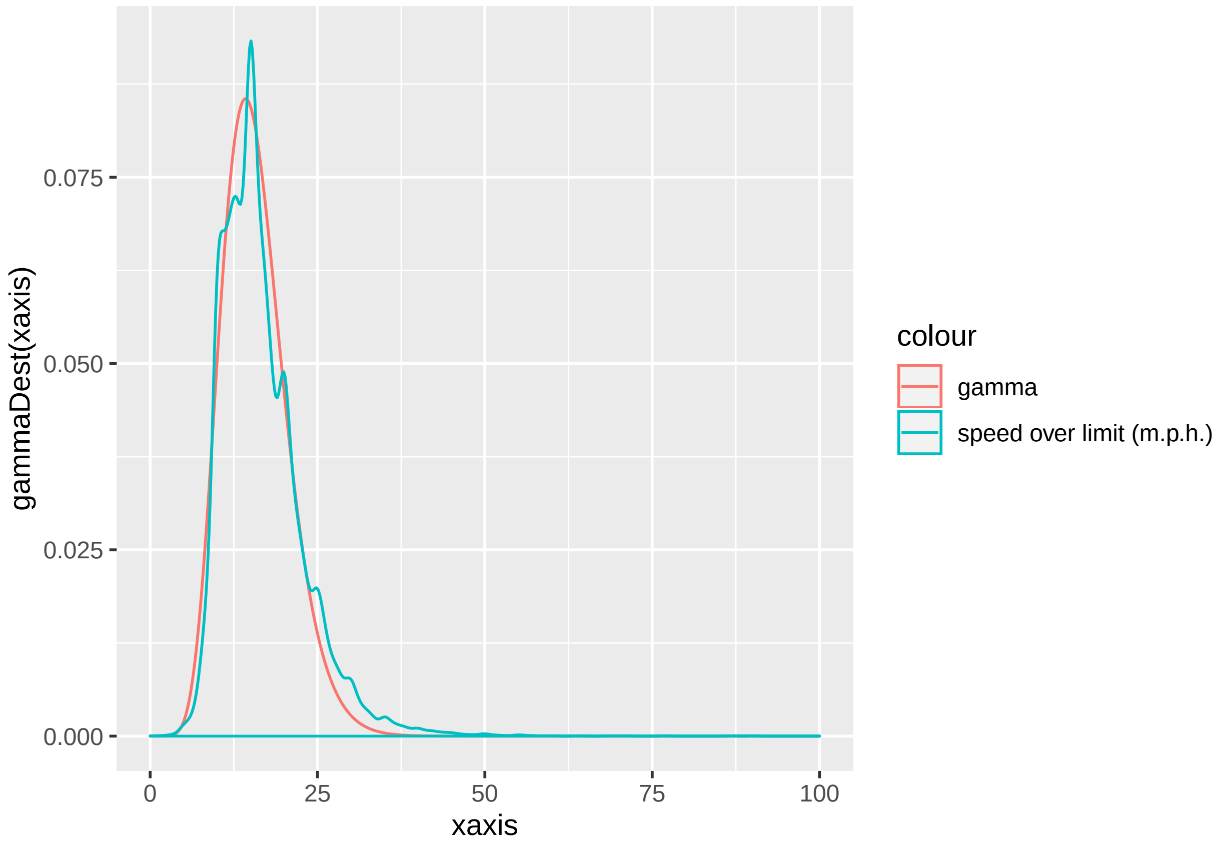
```
mapsDf$riskFactor = riskFactors

# #Create final csv for google maps heatmap
# write.csv(mapsDf, "mapsDf.csv")
```

We have noticed that the distribution of speeds looks much like a gamma distribution when corrected by the speed limit. A potential explanation is that the incidence of police stops increases with polynomial growth, while the incidence of a certain speed occuring decays expoinentially.

This causes an initial growth

```
a=10.5
b=1.5
gammaDest = function(x) 1/(gamma(a)*b^a)*x^(a-1)*exp(-x/b)
xaxis = seq(0,100,0.05)
p = ggplot() + geom_line(aes(x=xaxis, y=gammaDest(xaxis), colour = "gamma")) +
  geom_density(aes(x=speedingDfNoNA$Alleged_Speed-speedingDfNoNA$Posted_Speed, colour = "speed over lim
plot(p)
```


gammaDest(xaxis)
xaxis
colour
gamma
speed over limit (m.p.h.)