Using GLMs to identify statistically significant causes for lethality in UK trafic accidents.

```r
library(tidyverse)
```

First we load the data and clean it.

Data source: https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data

```r
# load data and turn most columns to factors
df <- read_csv("./Accidents_categorical.csv") %>%
  mutate_if(is.character, as.factor) %>%
  mutate_if(is.numeric, as.factor)

# columns that we will drop for performance / relevance reasons.
drop <- c("Accident_Index", "Latitude", "Longitude", "Hour_of_Day", "Datetime", "Driver_IMD_Decile", "R
          "Engine_CC")
df <- df[, !(colnames(df) %in% drop)]

# transform some sparse numerical columns to factors with a few levels.
df$Age_of_Driver <- as.factor(df$Age_of_Driver)
df$Age_of_Vehicle <- as.factor(df$Age_of_Vehicle)
df$Day_of_Month <- as.factor(df$Day_of_Month)
```

Now we will fit a binomial GLM and get the fitted coefficient estimates to analyse them data with python.

```r
# fit glm
model <- glm(Accident_Severity ~ ., data=df, family=binomial())
sum.mary <- summary(model)
sum.mary$coefficients
write.csv(sum.mary$coefficients, "coeff.csv")
```

Now in python we process the data to get the top 16 largest yet significant coefficients (p-value < 0.05).

```python
# load data
df = pd.read_csv("./coeff.csv")

# rename some columns
df = df.rename(columns={"Unnamed: 0": "coef", "Pr(>|z|)":"p"})

# reindex by p-value
df.index = df.p

# drop unnecessary columns
df = df.drop(["Std. Error", "z value"], axis=1)

# drop all coefficients that aren't significant.
# change sign since our labels for lethality are inverted
df = -df[df.index < 0.05]

# get top 16
```

```
df_pos = df[df.Estimate > 0].nlargest(16, "Estimate")

# drop the last unnecessary columns and trivially reindex
df_pos = df_pos.drop(["p"], axis=1)
df_pos.index = range(len(df_pos))

# now we can save this as simple table
# and visuallise with any tool we wish.
df_pos.to_csv("./accidents.csv")
```