Exploratory Analysis on Los Angeles Road Data

USC Machine Learning Team

9/10/2017

Section 1: Synopsis

The objective of this file is to perform exploratory analysis on accelerometer and GPS data in Los Angeles. This step is crucial to our understanding to the Machine Learning Project. The ultimate goal of the Machine Learning Project is to train independent models to identify different types of road inpediments. This step, exploratory data analysis, give us a sense of how to approach the main problem and picks machine learning models.

Section 2: Data Processing

Before data cleaning, we need to load the raw dataset from **los_angeles_5.csv**. We can count the number of rows in the raw dataset, and take a look at the first 10 rows.

```
if (!exists("LA5.raw")) {
    LA5.raw <- read.csv("./los_angeles_5.csv")
}
print(nrow(LA5.raw))
```

```
## [1] 2119
```

```
head(LA5.raw, 10)
```

```
##                       Date Speed Altitude Pressure        X         Y
## 1  2017-09-10 10:44:31.036  0.94 54.79278        0 0.056671 -0.032822
## 2  2017-09-10 10:44:31.091  0.55 54.99352        0 0.056671 -0.032822
## 3  2017-09-10 10:44:31.194  0.55 54.99352        0 0.064835  0.007797
## 4  2017-09-10 10:44:31.294  0.55 54.99352        0 0.078796  0.028397
## 5  2017-09-10 10:44:31.394  0.55 54.99352        0 0.058334 -0.015610
## 6  2017-09-10 10:44:31.494  0.55 54.99352        0 0.075516  0.004745
## 7  2017-09-10 10:44:31.594  0.55 54.99352        0 0.056717  0.022415
## 8  2017-09-10 10:44:31.695  0.55 54.99352        0 0.117401  0.025574
## 9  2017-09-10 10:44:31.794  0.55 54.99352        0 0.099884  0.018570
## 10 2017-09-10 10:44:31.894  0.55 54.99352        0 0.118179  0.014740
##            Z        G Latitude Longitude  Heading Magnetic.Field
## 1  -0.990891 0.993052 34.01897 -118.2889 41.89774              0
## 2  -0.990891 0.993052 34.01896 -118.2889 41.89774              0
## 3  -1.030807 1.032874 34.01896 -118.2889 41.89774              0
## 4  -1.008896 1.012367 34.01896 -118.2889 41.89774              0
## 5  -0.990509 0.992348 34.01896 -118.2889 41.89774              0
## 6  -0.978210 0.981132 34.01896 -118.2889 41.89774              0
## 7  -1.002472 1.004325 34.01896 -118.2889 41.89774              0
## 8  -1.017487 1.024556 34.01896 -118.2889 41.89774              0
## 9  -0.995087 1.000260 34.01896 -118.2889 41.89774              0
## 10 -0.993744 1.000855 34.01896 -118.2889 41.89774              0
##    Sound.Level Luminance
## 1   -120.00000         0
```

```
## 2    -31.53950         0
## 3    -33.80221         0
## 4    -27.72114         0
## 5    -30.32953         0
## 6    -33.31675         0
## 7    -30.12825         0
## 8    -31.30646         0
## 9    -32.39857         0
## 10   -29.12782         0
```

Section 3: Data Cleaning

Let's remove any missing values in the raw dataset.

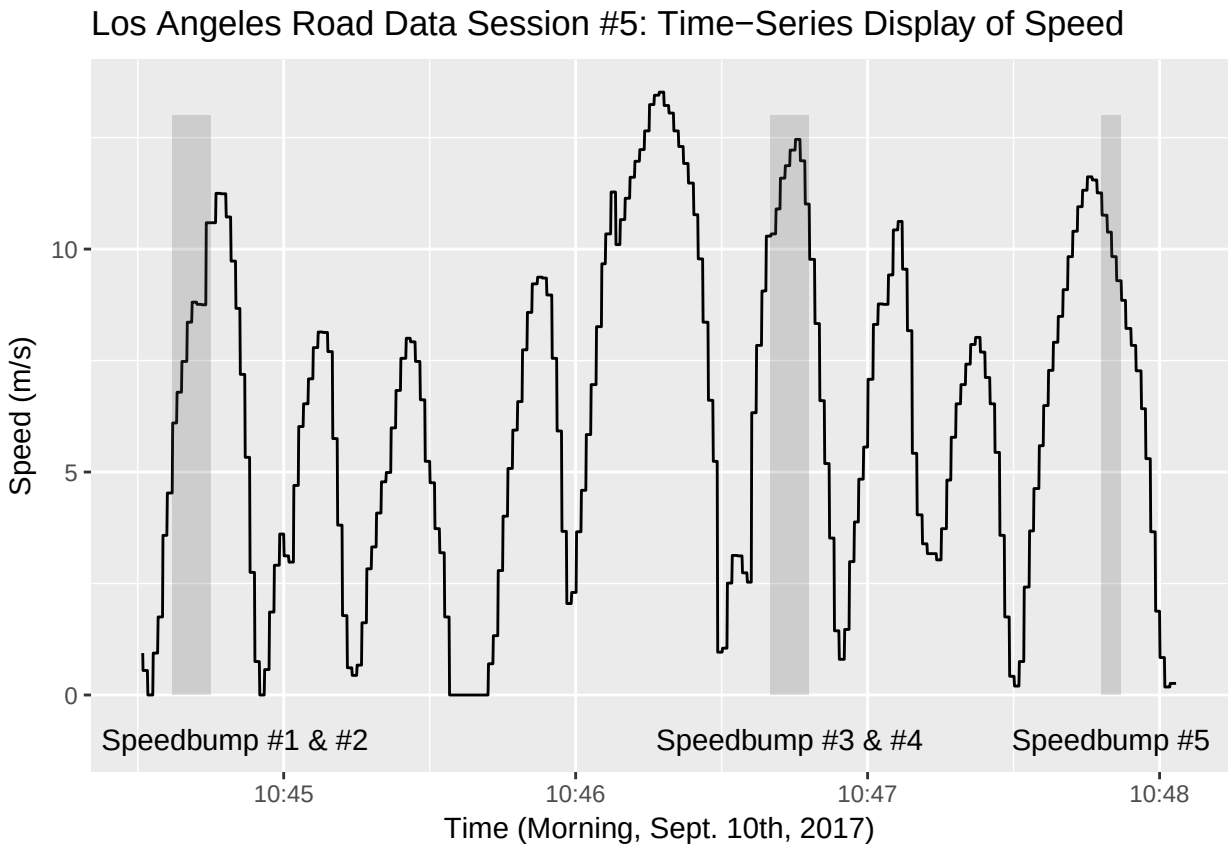
```
LA5.valid <- LA5.raw[!is.na(LA5.raw$Date) &
                       !is.na(LA5.raw$X) &
                       !is.na(LA5.raw$Y) &
                       !is.na(LA5.raw$Z), ]
LA5.valid$Date <- as.POSIXct(LA5.valid$Date, format="%Y-%m-%d %H:%M:%OS")
```

Section 4: Exploratory Data Analysis

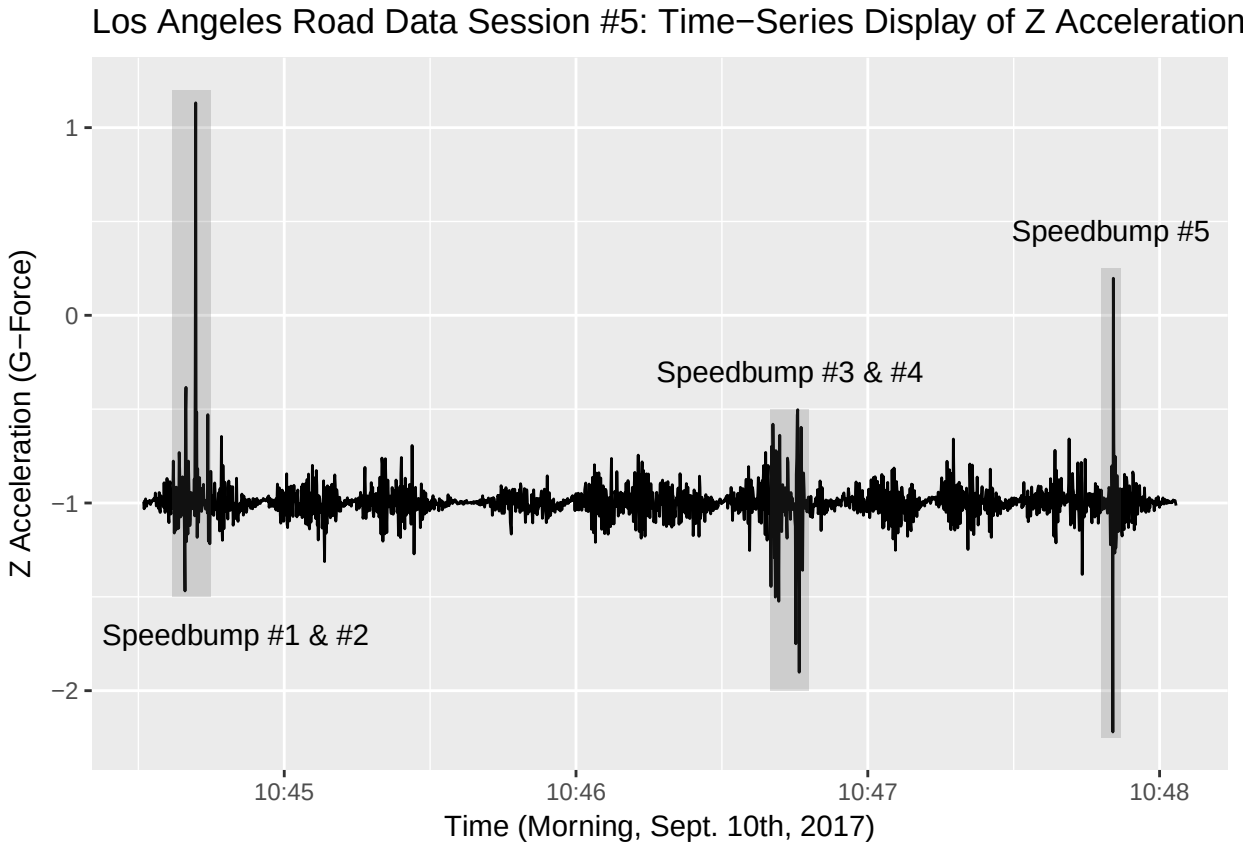
```
# visualize time-series distribution of xyz-axis accelerations
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```
plot.LA5Speed <- ggplot(LA5.valid, aes(Date, Speed)) + geom_line() + xlab("Time (Morning, Sept. 10th, 2
        ylab("Speed (m/s)") + ggtitle("Los Angeles Road Data Session #5: Time-Series Display of Speed")
plot.LA5Speed + annotate("rect", xmin = as.POSIXct("2017-09-10 10:44:37", format="%Y-%m-%d %H:%M:%OS"),
                           xmax = as.POSIXct("2017-09-10 10:44:45", format="%Y-%m-%d %H:%M:%OS"),
                           ymin = 0, ymax = 13, alpha = .2) +
        annotate("text", x = as.POSIXct("2017-09-10 10:44:50", format="%Y-%m-%d %H:%M:%OS"),
                 y = -1, label = "Speedbump #1 & #2") +
        annotate("rect", xmin = as.POSIXct("2017-09-10 10:46:40", format="%Y-%m-%d %H:%M:%OS"),
                           xmax = as.POSIXct("2017-09-10 10:46:48", format="%Y-%m-%d %H:%M:%OS"),
                           ymin = 0, ymax = 13, alpha = .2) +
        annotate("text", x = as.POSIXct("2017-09-10 10:46:44", format="%Y-%m-%d %H:%M:%OS"),
                 y = -1, label = "Speedbump #3 & #4") +
        annotate("rect", xmin = as.POSIXct("2017-09-10 10:47:48", format="%Y-%m-%d %H:%M:%OS"),
                           xmax = as.POSIXct("2017-09-10 10:47:52", format="%Y-%m-%d %H:%M:%OS"),
                           ymin = 0, ymax = 13, alpha = .2) +
        annotate("text", x = as.POSIXct("2017-09-10 10:47:50", format="%Y-%m-%d %H:%M:%OS"),
                 y = -1, label = "Speedbump #5")
```




```
plot.LA5X <- ggplot(LA5.valid, aes(Date, X)) + geom_line() + xlab("Time (Morning, Sept. 10th, 2017)") +
        ylab("X Acceleration (G-Force)") + ggtitle("Los Angeles Road Data Session #5: Time-Series Displa
plot.LA5Y <- ggplot(LA5.valid, aes(Date, Y)) + geom_line() + xlab("Time (Morning, Sept. 10th, 2017)") +
        ylab("Y Acceleration (G-Force)") + ggtitle("Los Angeles Road Data Session #5: Time-Series Displa
plot.LA5Z <- ggplot(LA5.valid, aes(Date, Z)) + geom_line() + xlab("Time (Morning, Sept. 10th, 2017)") +
        ylab("Z Acceleration (G-Force)") + ggtitle("Los Angeles Road Data Session #5: Time-Series Displa
plot.LA5Z + annotate("rect", xmin = as.POSIXct("2017-09-10 10:44:37", format="%Y-%m-%d %H:%M:%OS"),
                             xmax = as.POSIXct("2017-09-10 10:44:45", format="%Y-%m-%d %H:%M:%OS"),
                             ymin = -1.5, ymax = 1.2, alpha = .2) +
        annotate("text", x = as.POSIXct("2017-09-10 10:44:50", format="%Y-%m-%d %H:%M:%OS"),
                 y = -1.7, label = "Speedbump #1 & #2") +
        annotate("rect", xmin = as.POSIXct("2017-09-10 10:46:40", format="%Y-%m-%d %H:%M:%OS"),
                             xmax = as.POSIXct("2017-09-10 10:46:48", format="%Y-%m-%d %H:%M:%OS"),
                             ymin = -2, ymax = -0.5, alpha = .2) +
        annotate("text", x = as.POSIXct("2017-09-10 10:46:44", format="%Y-%m-%d %H:%M:%OS"),
                 y = -0.3, label = "Speedbump #3 & #4") +
        annotate("rect", xmin = as.POSIXct("2017-09-10 10:47:48", format="%Y-%m-%d %H:%M:%OS"),
                             xmax = as.POSIXct("2017-09-10 10:47:52", format="%Y-%m-%d %H:%M:%OS"),
                             ymin = -2.25, ymax = 0.25, alpha = .2) +
        annotate("text", x = as.POSIXct("2017-09-10 10:47:50", format="%Y-%m-%d %H:%M:%OS"),
                 y = 0.45, label = "Speedbump #5")
```



We can see from the above graph **Los Angeles Road Data Session #5: Time-Series Display of Z Acceleration** that there are three shaded fractions which strikes out. The overall similar pattern in these three shaded areas is the anormaly in the reading of Z acceleration. According to the test driver Ernest, in the shared areas **Speedbump #1 & #2** and **Speedbump #3 & #4**, the accelerometer read an anormaly followed by a much larger anormaly, because Ernest was stepping on the gas. Cross-referencing with the Speed graph. We find that speed might be a factor of the scale of change to Z acceleration when the vehicle hits a speedbump.

Section 5: Recommendations

Based on the findings from the Section 4, We think it will be a good start to model a logistic regression on Z acceleration to identify speedbumps. If this method shows promise, we can apply it on potholes as well.