**ACCIDENT DATA ANALYSIS:**

# Introduction

The Fatal Accidents 2007 dataset consists all fatal accidents on public roads reported to the national highway transportation safety administration. I am using R for analyzing dataset by using different graphical interpretations and prediction model to find the solutions for the following research questions.

1.Fatalities by month, day of week, hour, state

2.Crash counts by Roadway function class, Route, Relation to road, Speed limit, light conditions,

3.Pedestrians involved in accident, Number of hit and run cases in accidents

4.Which type of accidents are more frequent in different road types

5.Accidents by alignment and number of lanes, Surrounding conditions and traffic controls functioning,

weather conditions and roadway traffic flow.

6.Predict fatalities by different characteristic of accident data.

# Dataset

The Fatality Analysis Reporting System (FARS) contains data on all vehicle crashes in the United States that occur on a public roadway and involve a fatality. The Fatal accident dataset downloaded from https://wiki.csc.calpoly.edu/datasets/wiki/HighwayAccidents. It has 32248 instances and 55 attributes. I used 25 variables.



1.STATE: State in U.S

2.MONTH: Month of the year

3. HOUR: Hour of the day

4.VE\_TOTAL: Vehicles involved in accident

5. PERSONS: Persons involved in accident

6. PEDS: Persons which are not occupants of motor vehicle involved in accident

7. ROAD\_FNC: Function class of road

8. ROUTE: Route Type

9.MAN\_COLL: Manner of collision

10.REL\_ROAD: Relation to road

11.TRAF\_FLO: Traffic way flow

12. NO\_LANES: Number of lanes

13.SP\_LIMIT: Speed limit

14.ALIGNMENT: Road way alignment

15.PAVE\_TYP: Road way surface type

16:SUR\_COND:Road way surrounding conditions

17:T\_CONT\_F: Traffic controls functioning

18.HIT\_RUN: Hit and run

19:LGT\_COND: Light condition

20:WEATHER1: Weather condition.

21:C\_M\_ZONE: Construction and maintenance zone

22:SCH\_BUS: School bus related vehicle

23:FATALS: Number of fatalities

24:DAY\_WEEK: Day of week

25:DRUNK\_DR: drunk and driver

# Approach

## Step 1: Data cleaning:

Created a subset with variables using for this project. Creating factors of variable and removing the unknown and null values.

## Step 2: Creating plots:

Generating plots by using ggplot2 library.

## Step 3: Perform Regression analysis:

Selecting best subset of variables to perform regression analysis by using regularized linear regression method. Split data into two parts train and test. Creating multivariate regression model by using train data set and test this model with test dataset.

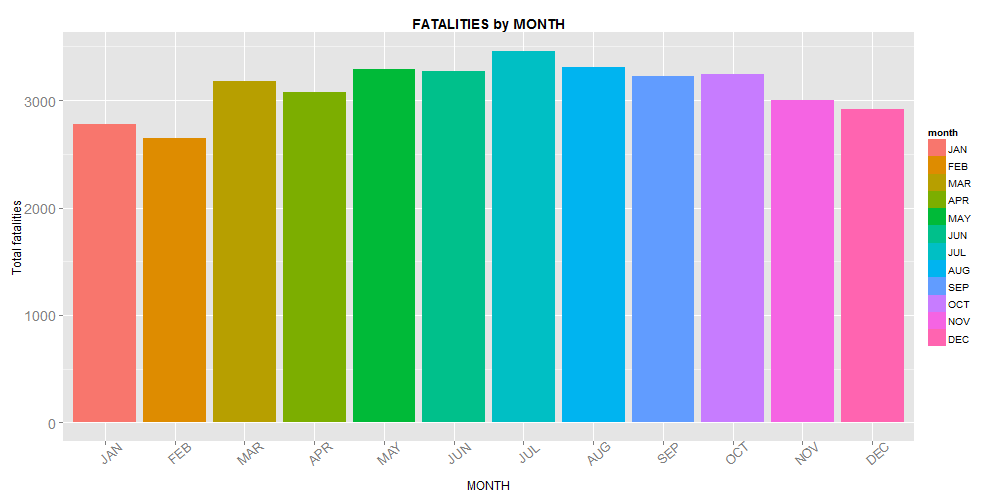
**RESULTS:**

**1.Which month of year have highest fatalities?**

Summary of Month Variable:

JAN FEB MAR APR MAY JUN JUL AUG SEP OCT NOV DEC

2493 2384 2855 2780 2992 2960 3142 2996 2910 2985 2732 2623

****

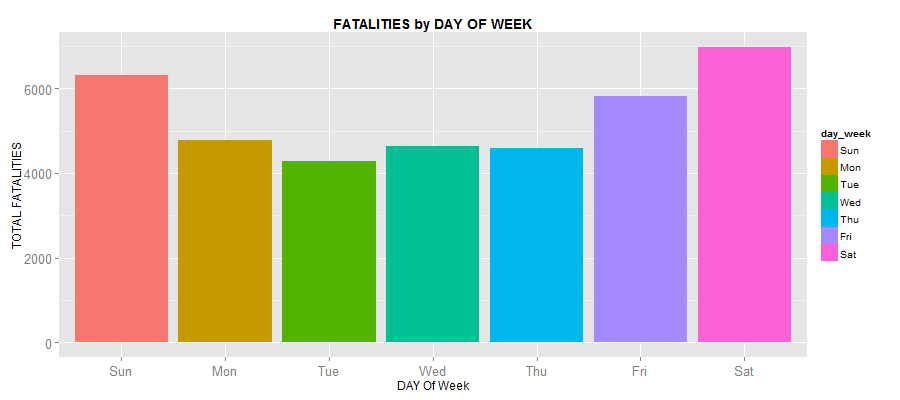
Fatalities are more in July. BY seeing this we can say accidents are more in summer than winter.

**2.Which Day of week have highest fatalities**

Summary of DAYOFWEEK Variable:

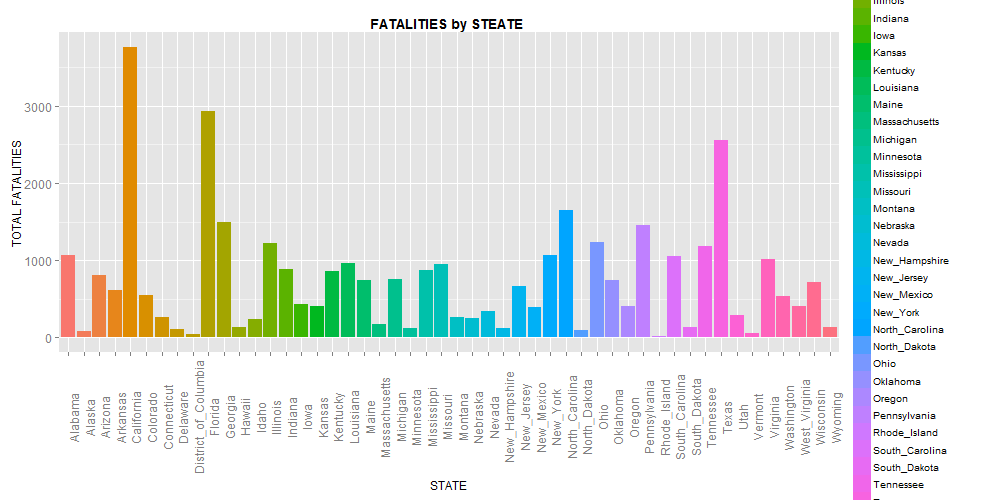
Sun Mon Tue Wed Thu Fri Sat

5617 4355 3953 4225 4206 5243 6253



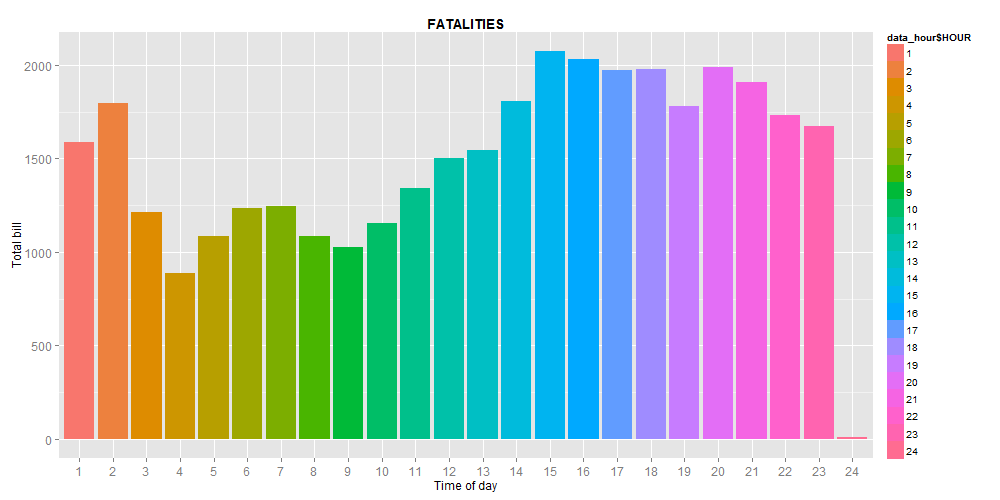
Fatalities are more in Saturday, Sunday, Friday. Accidents are more in Weekends.

**3.Which state of U.S have highest fatalities?**



From this we can see California, Florida and Texas has highest fatalities.

**4.Which Hour of day has highest fatalities?**



From this we can say accidents are more in evenings and less in morning.

**5.Number of accidents by road function class**

1.Rural Principal Arterial-Interstate

2.Rural Principal Arterial-Other

3.Rural Minor Arterial

4.Rural Major Collector

5.Rural Minor Collector

6.Rural Local Road or Street

9.Rural Unknown

11.Urban Principal Arterial - Interstate

12.Urban Principal Arterial -Other Freeways or Expressways

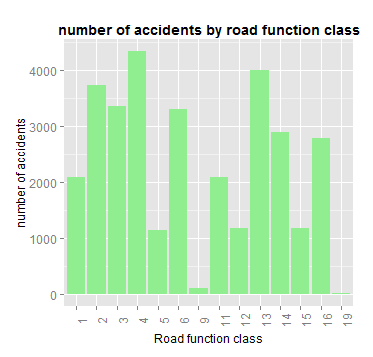
13.Urban Other Principal Arterial

14.Urban Minor Arterial

15.Urban Collector

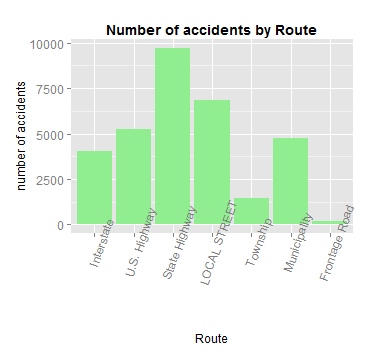
16.Urban Local Road or Street

19-Urban Unknown



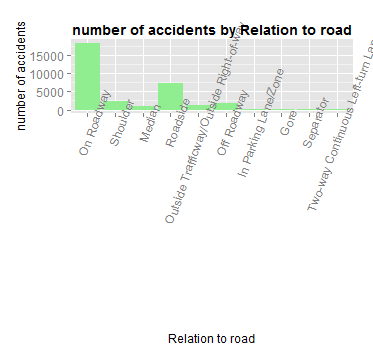
Accidents are more in Rural Major collector, Urban Other Principal Arterial, Rural Principal Arterial-Other, Rural Minor Arterial, Rural Local Road or Street, Urban Other Principal Arterial, Urban Minor Arterial, Urban Local Road or Street.

**6.Number of accidents by Route:**

****

Accidents are more in State Highways

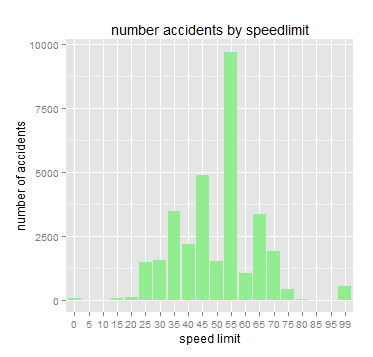
**7.Number of accidents by relation to road:**

****

Accidents are more in State Highways

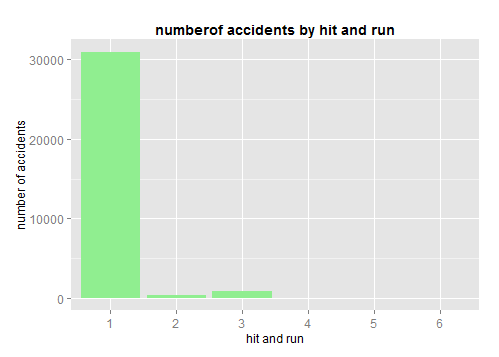
Accidents are more in On Roadway

**8.Number of accidents by speed limit:**

****

Accidents are more at 55 speed limit

**9.Number of hit and run in accidents:**

****

1.No Hit-and-Run

2.Hit Motor Vehicle

3.Hit Pedestrian

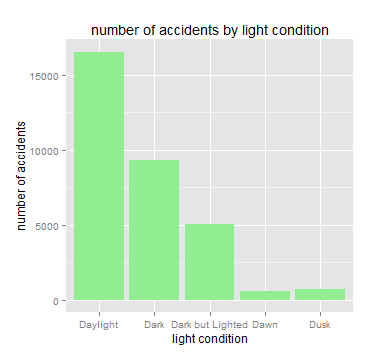
4.Hit Parked Vehicle

5.Driver Leaves Scene after Non-Collision Event

6.Hit-and-Run, Other Involved Person Left Scene

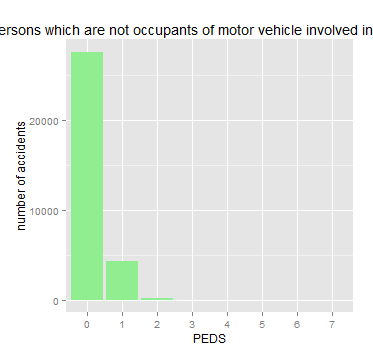
almost all accidents are not Hit and run cases .

**10.Number of accidents by light condition:**

****

Accidents are more in Daylight

**11.** **persons which are not occupants of motor vehicle involved in accident**

****

Very less pedestrians involved in accidents.

**12.** **which type of accidents are more frequent in different road types**

PAVE\_TYP

MAN\_COLL Concrete Blacktop Brick Slag Dirt

Not Collision with Motor Vehicle 1543 17172 6 330 176

Front-to-Rear 349 1704 0 1 1

Front-to-Front 213 3224 0 15 15

Front-to-Side, Same Direction 51 378 0 1 0

Front-to-Side, Opposite Direction 128 1510 0 3 3

Front-to-Side, Right Angle 379 3923 0 36 6

Front-to-Side/Angle-Direction Not Specified 2 123 0 0 1

Sideswipe - Same Direction 64 365 0 0 0

Sideswipe - Opposite Direction 20 375 0 0 0

Rear-to-Side 11 54 0 0 0

Rear-to-Rear 1 1 0 0 0

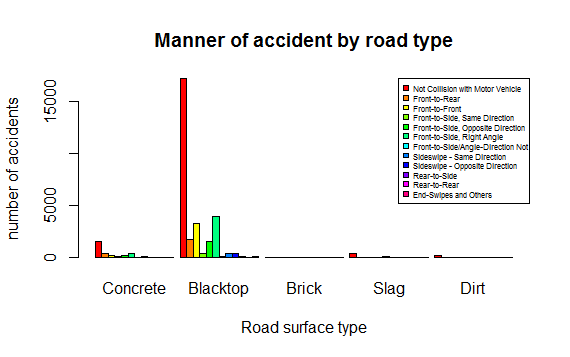
End-Swipes and Others 6 53 0 1 2

This is significant because p value is less than 0.05

Pearson's Chi-squared test

data: tbl

X-squared = 465.01, df = 44, p-value < 2.2e-16



Accidents are more in Blacktop road surface. More accidents are not collision with motor vehicle and front to side(right angle),Front to Front collisions are more.

**13.Accidents by alignment and number of lanes:**

no. of lanes

alignment Straight Curved

1 174 283

2 16673 7609

3 1941 379

4 3435 611

5 451 64

6 445 28

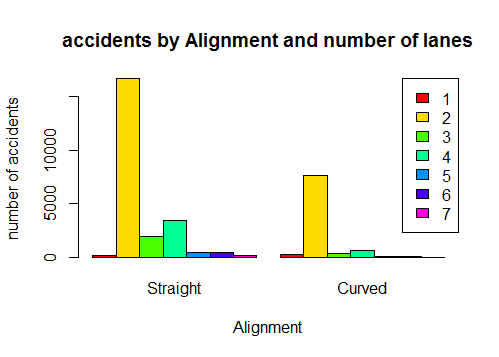
7 136 17

Chi-square test:

Pearson's Chi-squared test

data: tbl

X-squared = 1086, df = 6, p-value < 2.2e-16



Accidents are more in straight single lane and curved double lanes roads.

**14.** **Accidents by surrounding conditions and traffic controls functioning.**

T\_CONT\_F

SUR\_COND Dry Wet snow ice gravel

No Controls 21249 2881 502 443 52

Device Not Functioning 20 4 1 0 0

Functioning Improperly 21 2 0 0 0

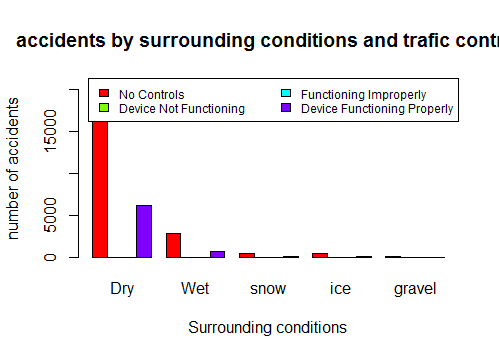
Device Functioning Properly 6209 746 65 42 9

chi-square test:

Pearson's Chi-squared test

data: tbl

X-squared = 102.09, df = 12, p-value < 2.2e-16



Accidents are more in Dry with no traffic signals.

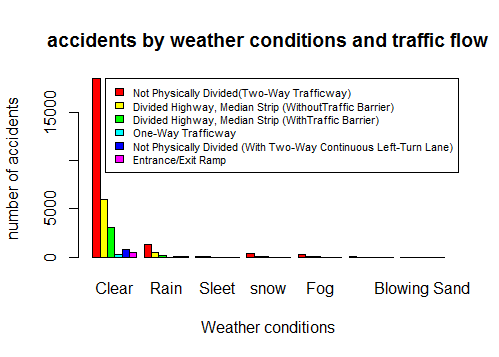
**15.** **accidents by weather conditions and traffic flow**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **TRAFFIC FLOW** | | | | | | |
| **WEATHER1** | **Clear** | **Rain** | **Sleet** | **snow** | **Fog** | **Severe Crosswinds** | **Blowing Sand** |
| Not Physically Divided(Two-Way Trafficway) | 18454 | 1316 | 67 | 416 | 247 | 46 | 10 |
| Divided Highway, Median Strip (WithoutTraffic Barrier) | 6011 | 467 | 38 | 118 | 56 | 18 | 3 |
| Divided Highway, Median Strip (WithTraffic Barrier) | 3062 | 193 | 15 | 41 | 35 | 4 | 0 |
| One-Way Trafficway | 244 | 15 | 0 | 0 | 1 | 0 | 1 |
| Not Physically Divided (With Two-Way Continuous Left-Turn Lane) | 802 | 54 | 2 | 4 | 11 | 1 | 0 |
| Entrance/Exit Ramp | 454 | 31 | 0 | 1 | 7 | 1 | 0 |

Pearson's Chi-squared test

data: tbl

X-squared = 71.948, df = 30, p-value = 2.65e-05



Accidents are more in clear weather and not physically divided Two way traffic ways.

**16.Predict fatalities:**

Summary of prediction meodel:

Call:

lm(formula = FATALS ~ VE\_TOTAL + PERSONS + PEDS + MAN\_COLL +

SP\_LIMIT + DRUNK\_DR, data = train)

Residuals:

Min 1Q Median 3Q Max

-3.7043 -0.1303 -0.0558 -0.0085 6.3945

Coefficients: Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.898720 0.010282 87.411 < 2e-16 \*\*\*

VE\_TOTAL -0.042630 0.004519 -9.434 < 2e-16 \*\*\*

PERSONS 0.072018 0.001504 47.899 < 2e-16 \*\*\*

PEDS -0.071628 0.006161 -11.627 < 2e-16 \*\*\*

MAN\_COLLFront-to-Rear -0.026040 0.011301 -2.304 0.02122 \*

MAN\_COLLFront-to-Front 0.098751 0.008949 11.035 < 2e-16 \*\*\*

MAN\_COLLFront-to-Side,Same Direction-0.018145 0.020733 -0.875 0.38150

MAN\_COLLFront-to-Side,Opposite Direction 0.034942 0.011652 2.999 0.00271 \*\*

MAN\_COLLFront-to-Side, Right Angle -0.004556 0.008386 -0.543 0.58688

MAN\_COLLFront-to-Side/Angle-Direction Not Specified 0.048630 0.036771 1.322 0.18601

MAN\_COLLSideswipe - Same Direction -0.059248 0.021057 -2.814 0.00490 \*\*

MAN\_COLLSideswipe-Opposite Direction-0.003122 0.021407 -0.146 0.88404

MAN\_COLLRear-to-Side 0.087553 0.054307 1.612 0.10693

MAN\_COLLRear-to-Rear -0.181951 0.370102 -0.492 0.62299

MAN\_COLLEnd-Swipes and Others -0.033808 0.053262 -0.635 0.52560

SP\_LIMIT 0.001462 0.000166 8.811 < 2e-16 \*\*\*

DRUNK\_DR 0.047311 0.003968 11.923 < 2e-16 \*\*\*

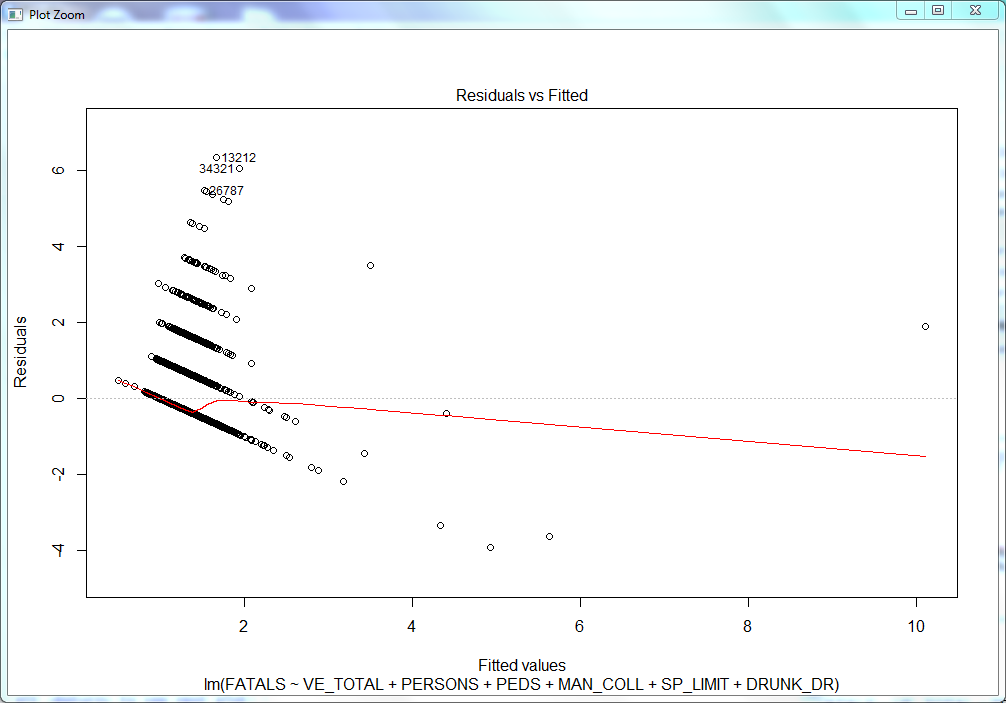
---

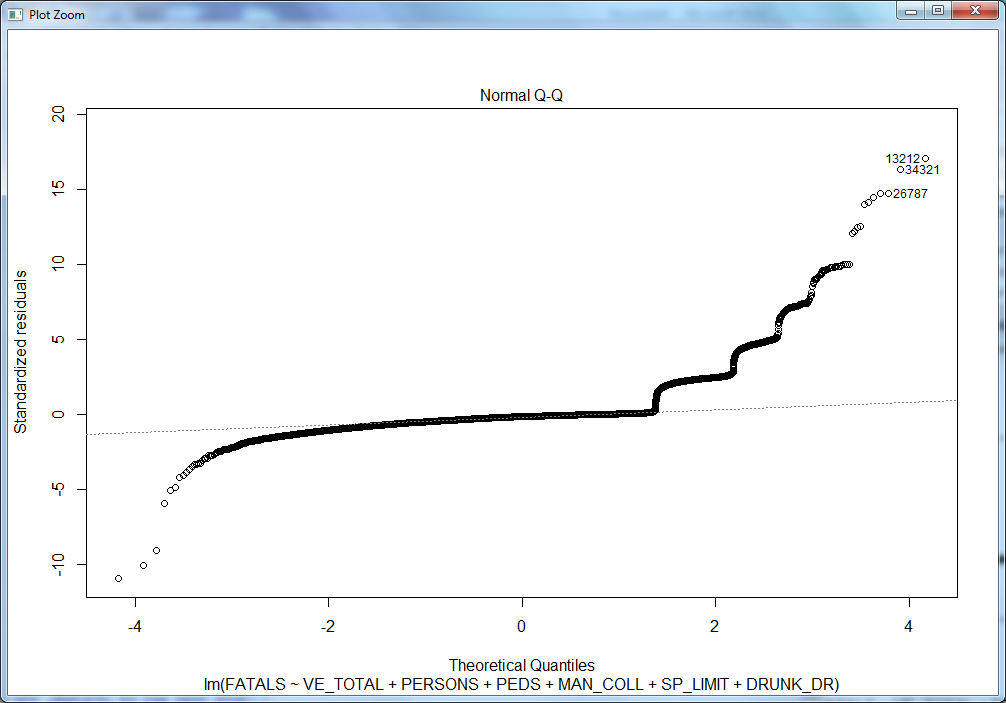
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

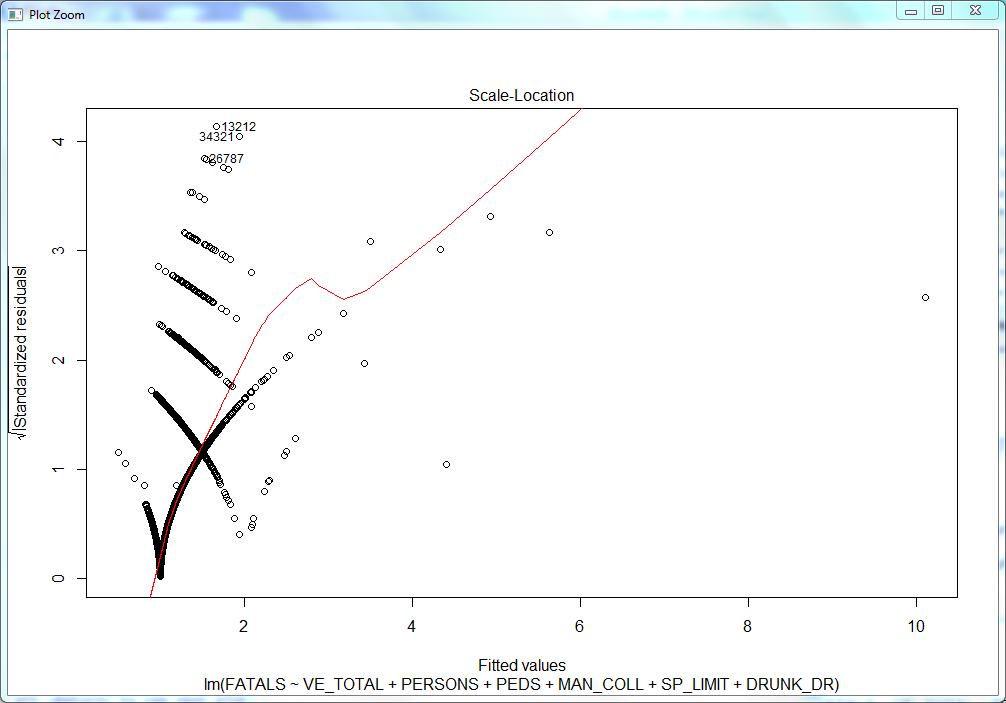
Residual standard error: 0.3701 on 25779 degrees of freedom

Multiple R-squared: 0.115, Adjusted R-squared: 0.1144

F-statistic: 209.3 on 16 and 25779 DF, p-value: < 2.2e-16

****

****

****

****

R Square of 13% percent suggests that Number of fatalities shows 13% of Variance explained by the Linear Model.

Intercept suggests if there are no accidents, fatality is 0.877, hypothetically wrong. But for any accident within specified conditions, fatality sums up by 0.877.

From the Coefficients, it depicts that -ve coefficients indicate that fatalities will be less when coefficients are -ve fatalities increases with higher +ve coefficients.

For eg: Front to rear collision decreases the changes of fatalities by 0.034.

Front to rear, Front to Side Same Direction, From to Side Right Angle, Sideswipe Opposite Direction has little effect on the number of fatalities and it decreases the effect of number of fatalities.

Rear to Rear accident has very less number of fatalities as explained by the coefficient at -0.38. As per regression model it has very less number of fatalities.

Every increase in pedestrians, it decreases the number of fatalities by -0.06. Each vehicle involvement decreases the number of fatalities by 0.028. But these conditions as derived by linear regression model is not agreed to confirm that fatalities decreases by increase in pedestrians and number of vehicles. These situations should sum up with different conditions.

Number of persons involved in an accident increases the fatality rate by 0.071.

Front to Front, Front to side, rear to side increases the fatality rate.

Speed Limit and Drunken Drive will also increase the fatality rate.

**RMSE:**

**sqrt(sum((prediction[,"fit"] - test$FATALS)^2)/nrow(test))**

[1] 0.3742685

**errors:**

errors <- prediction[,"fit"] - test$FATALS

hist(errors)

Histogram of errors:

****

**relative change:**

rel\_change <- 1 - ((test$FATALS - abs(errors)) / test$FATALS)

table(rel\_change<0.10)["TRUE"] / nrow(test)

TRUE

0.5871318

**Confusion matrix:**

Reference

Prediction 1 2 3 4 5 7 12

1 5909 437 75 21 6 1 1

2 0 0 0 0 0 0 0

3 0 0 0 0 0 0 0

4 0 0 0 0 0 0 0

5 0 0 0 0 0 0 0

7 0 0 0 0 0 0 0

12 0 0 0 0 0 0 0

**Github link : https://github.com/himabindu-kunani/Accident-Data-Analysis/tree/master**