Car driver safety data

Xiaoyan Tang

Table 1.

City has most passenger car

PHY\_CITY num

1 ORLANDO 591

2 LOS ANGELES 442

3 HOUSTON 366

4 MIAMI 282

5 SAN DIEGO 277

6 BROOKLYN 221

7 SAN FRANCISCO 185

8 ATLANTA 183

9 LAS VEGAS 161

10 PHOENIX 158

Table 2: Companies that has most of the car accidents

LEGAL\_NAME num

1 FIRST STUDENT INC 865

2 DURHAM SCHOOL SERVICES LP 326

3 NEW YORK CITY TRANSIT 274

4 REGIONAL TRANSPORTATION DISTRICT 172

5 FIRST TRANSIT INC 165

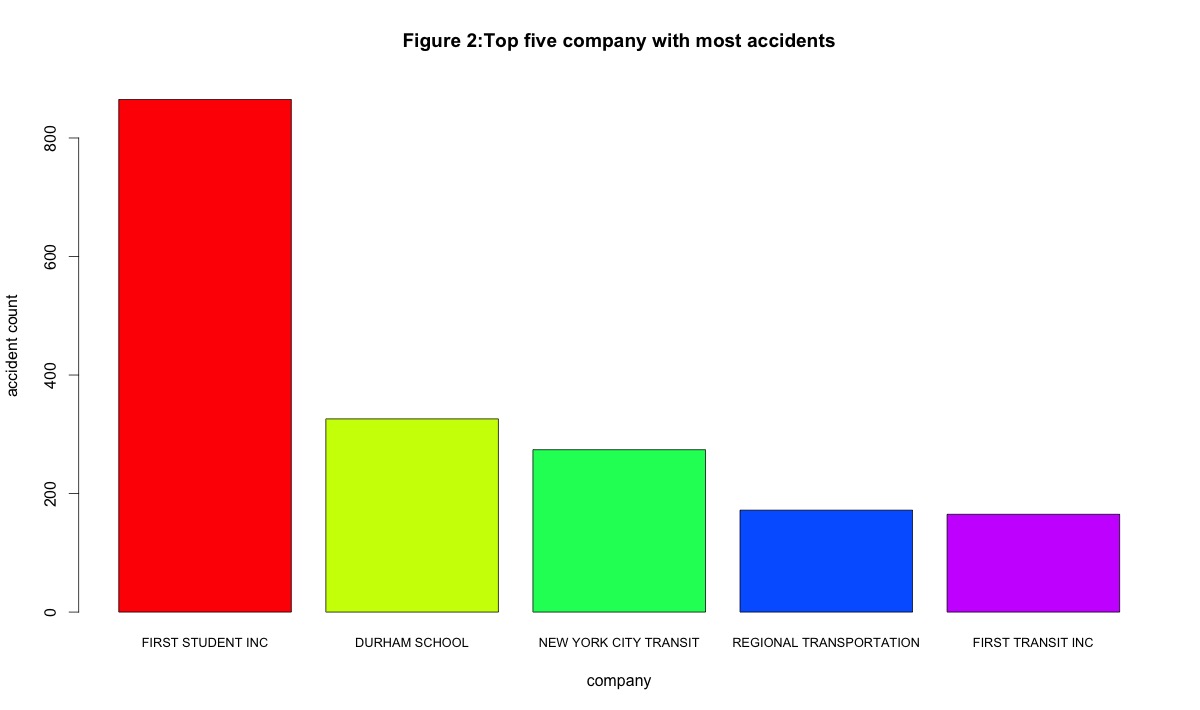
6 NEW JERSEY TRANSIT CORPORATION 143

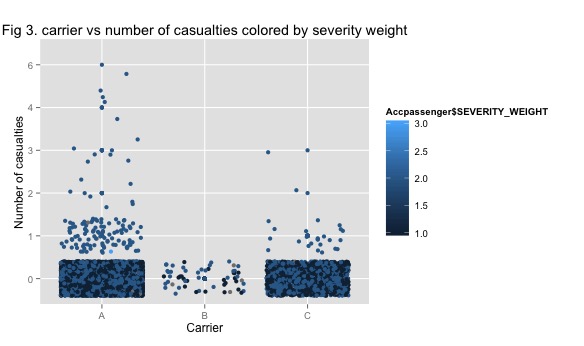
7 GREYHOUND LINES INC 126

8 SOUTHEASTERN PENNSYLVANIA TRANSPORTATION AUTHORITY SEPTA 110

9 ILLINOIS CENTRAL SCHOOL BUS LLC 57

10 APPLE BUS COMPANY 48





R> sqldf("SELECT PHY\_CITY, COUNT(Report\_number ) as num FROM Accpassenger GROUP BY PHY\_CITY ORDER BY num DESC limit 10")

PHY\_CITY num

1 CINCINNATI 1059

2 BROOKLYN 388

3 LISLE 338

4 DALLAS 203

5 DENVER 177

6 PHILADELPHIA 169

7 NEWARK 145

8 ORLANDO 75

9 HOUSTON 63

10 JOLIET 57

R> #We found that CINCINNATI has most of the accidents reported with commericial car#

R> names(Accpassenger)

[1] "DOT\_NUMBER" "LEGAL\_NAME" "DBA\_NAME" "CARRIER\_OPERATION"

[5] "HM\_FLAG" "PC\_FLAG" "PHY\_STREET" "PHY\_CITY"

[9] "PHY\_STATE" "PHY\_ZIP" "PHY\_COUNTRY" "MAILING\_STREET"

[13] "MAILING\_CITY" "MAILING\_STATE" "MAILING\_ZIP" "MAILING\_COUNTRY"

[17] "TELEPHONE" "FAX" "EMAIL\_ADDRESS" "MCS150\_DATE"

[21] "MCS150\_MILEAGE" "MCS150\_MILEAGE\_YEAR" "ADD\_DATE" "OIC\_STATE"

[25] "NBR\_POWER\_UNIT" "DRIVER\_TOTAL" "REPORT\_NUMBER" "REPORT\_SEQ\_NO"

[29] "REPORT\_DATE" "REPORT\_STATE" "FATALITIES" "INJURIES"

[33] "TOW\_AWAY" "HAZMAT\_RELEASED" "TRAFFICWAY\_DESC" "ACCESS\_CONTROL\_DESC"

[37] "ROAD\_SURFACE\_CONDITION\_DESC" "WEATHER\_CONDITION\_DESC" "LIGHT\_CONDITION\_DESC" "VEHICLE\_ID\_NUMBER"

[41] "VEHICLE\_LICENSE\_NUMBER" "VEHICLE\_LICENSE\_STATE" "SEVERITY\_WEIGHT" "TIME\_WEIGHT"

[45] "CITATION\_ISSUED\_DESC" "SEQ\_NUM" "accident\_severity"

R> #data description

R> #DOT\_NUMBER - Unique USDOT Number of the Motor Carrier

R> #LEGAL\_NAME - Legal name of a carrier

R> #DBA\_NAME - Carrier's Doing-Business-As name

R> #CARRIER\_OPERATION - Codes identifying carriers' type of Operation; A = Interstate, B = Intrastate Hazmat, C = Intrastate Non-Hazmat

R> #HM\_FLAG - Carrier is subject to placardable HM threshold ( Y = Yes, N = No)

R> #PC\_FLAG - Carrier is subject to passengercarrier Threshold (Y = Yes, N = No)

R> #MCS150\_MILEAGE - Vehicle Mileage Traveled (VMT) reported on the carrier's MCS-150 form

R> #MCS150\_MILEAGE\_YEAR - Year for which VMT was reported

R> #ADD\_DATE - Date when carrier information was added to MCMIS Database System

R> #OIC\_STATE - FMCSA State office with oversight for this carrier

R> #NBR\_POWER\_UNIT - Number of power units reported

R> #DRIVER\_TOTAL - Number of drivers reported

R> #Report\_number - Unique state report number for the incident

R> #Report\_seq\_no - Sequence number for each vehicle involved in a crash

R> #DOT\_Number - Unique number assigned to a company by the DOT

R> #Report\_Date - The date a incident occurred

R> #Report\_State - State abbreviation

R> #Fatalities - Total number of fatalities reported in the crash

R> #Injuries - Total number of injuries reported in the crash

R> #Tow\_Away - 'Y' indicates that a vehicle involved in the crash was towed from the scene.

R> #Hazmat\_released - 'Y' indicates that hazardous materials were released at the time of a crash

R> #Trafficway\_Desc - Description of the trafficway

R> #Access\_Control\_Desc - Description of the access control

R> #Road\_surface\_Condition\_Desc - Description of the road surface condition

R> #Weather\_Condition\_Desc - Description of the weather condition

R> #Light\_Condition\_Desc - Description of the light condition

R> #Vehicle\_ID\_Number - Vehicle Identification number (VIN)

R> #Vehicle\_License\_number - Vehicle license number

R> #Vehicle\_license\_state - vehicle license state

R> #Severity\_Weight - The severity weight that is assigned to the incident

R> #Time\_weight - the time weight that is assigned to the incident

R> #citation\_issue\_desc - Description of the citation issue

R> #seq\_num - Sequence number

R> # I will partition the data, first 5000 for training and the rest 2694 will be left for fitting#

R> #data tranfer, know how many years there are for the car accidents#

R> Accpassenger$MCS150\_MILEAGE\_YEAR=2017-Accpassenger$MCS150\_MILEAGE\_YEAR

R> train=Accpassenger[1:5000,]

R> test=Accpassenger[5000:7695,]

R> levels(train$WEATHER\_CONDITION\_DESC)

[1] "" "Blowing Sand Soil Dirt Or Snow" "Fog" "No Adverse Conditions"

[5] "Other" "Rain" "Severe Crosswinds" "Sleet Hail"

[9] "Snow" "Unknown"

R> levels(train$ROAD\_SURFACE\_CONDITION\_DESC)

[1] "" "Dry" "Ice" "Other" "SandMudDirtOil Or Gravel"

[6] "Slush" "Snow" "Unknown" "Water(Standing Moving)" "Wet"

R> levels(train$LIGHT\_CONDITION\_DESC)

[1] "" "Dark - Lighted" "Dark - Not Lighted" "Dark - Unknown Roadway Lighting"

[5] "Dawn" "Daylight" "Dusk" "Other"

[9] "Unknown"

R> # continous varibles are listed as follows#

R> #MCS150\_MILEAGE,DRIVER\_TOTAL,Injuries,NBR\_POWER\_UNIT,DRIVER\_TOTAL

R> #Set up dummy varibles#

R> # Set up 3 dummy variables for carrier operation

R> train$inter= ifelse((train$CARRIER\_OPERATION=='A'), 1, 0)

R> train$Intra= ifelse((train$CARRIER\_OPERATION=='B'), 1, 0)

R> train$intranon= ifelse((train$CARRIER\_OPERATION=='C'), 1, 0)

R> #accident#

R> train$FATALITIES[train$FATALITIES>0] = 1

R> class(train$FATALITIES)

[1] "numeric"

R> #HM\_FLAG#

R> train$HM\_FLAG1=ifelse((train$HM\_FLAG=='Y'),1,0)

R> train$HM\_FLAG0=ifelse(train$HM\_FLAG=='N',1,0)

R> #PC\_FLAG#

R> train$PC\_FLAG1=ifelse((train$PC\_FLAG=='Y'),1,0)

R> train$PC\_FLAG0=ifelse((train$PC\_FLAG=='N'),1,0)

R> #define factorial data#

R> train$WEATHER\_CONDITION\_DESC <- factor(train$WEATHER\_CONDITION\_DESC)

R> train$ROAD\_SURFACE\_CONDITION\_DESC<-factor(train$ROAD\_SURFACE\_CONDITION\_DESC)

R> train$LIGHT\_CONDITION\_DESC<-factor(train$LIGHT\_CONDITION\_DESC)

R> train$Time\_weight<-factor(train$TIME\_WEIGHT)

R> train$SEVERITY\_WEIGHT<-factor(train$SEVERITY\_WEIGHT)

R> # We use binary logistic regression for fitting#

R> mod0<-glm(FATALITIES~MCS150\_MILEAGE+MCS150\_MILEAGE\_YEAR+INJURIES+NBR\_POWER\_UNIT+DRIVER\_TOTAL+inter+Intra+HM\_FLAG1+train$WEATHER\_CONDITION\_DESC+ROAD\_SURFACE\_CONDITION\_DESC+LIGHT\_CONDITION\_DESC, family=binomial,data=train)

Warning message:

glm.fit: fitted probabilities numerically 0 or 1 occurred

R> summary(mod0)

Call:

glm(formula = FATALITIES ~ MCS150\_MILEAGE + MCS150\_MILEAGE\_YEAR +

INJURIES + NBR\_POWER\_UNIT + DRIVER\_TOTAL + inter + Intra +

HM\_FLAG1 + train$WEATHER\_CONDITION\_DESC + ROAD\_SURFACE\_CONDITION\_DESC +

LIGHT\_CONDITION\_DESC, family = binomial, data = train)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.007 -0.184 -0.157 -0.099 3.375

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -4.93e+01 5.27e+03 -0.01 0.993

MCS150\_MILEAGE 1.69e-09 2.53e-09 0.67 0.504

MCS150\_MILEAGE\_YEAR 2.84e-02 6.39e-02 0.44 0.657

INJURIES 1.13e-01 1.99e-02 5.70 1.2e-08

NBR\_POWER\_UNIT -7.64e-05 3.66e-05 -2.09 0.037

DRIVER\_TOTAL 1.84e-05 2.45e-05 0.75 0.454

inter 1.03e+00 5.48e-01 1.87 0.061

Intra -1.60e+01 1.69e+03 -0.01 0.992

HM\_FLAG1 9.49e-01 1.05e+00 0.90 0.366

train$WEATHER\_CONDITION\_DESCBlowing Sand Soil Dirt Or Snow 1.62e-01 3.71e+03 0.00 1.000

train$WEATHER\_CONDITION\_DESCFog 1.48e+01 2.86e+03 0.01 0.996

train$WEATHER\_CONDITION\_DESCNo Adverse Conditions 1.42e+01 2.86e+03 0.00 0.996

train$WEATHER\_CONDITION\_DESCOther 1.47e+01 2.86e+03 0.01 0.996

train$WEATHER\_CONDITION\_DESCRain 1.46e+01 2.86e+03 0.01 0.996

train$WEATHER\_CONDITION\_DESCSevere Crosswinds -1.32e+00 5.95e+03 0.00 1.000

train$WEATHER\_CONDITION\_DESCSleet Hail 1.65e+01 2.86e+03 0.01 0.995

train$WEATHER\_CONDITION\_DESCSnow 1.42e+01 2.86e+03 0.00 0.996

train$WEATHER\_CONDITION\_DESCUnknown -6.43e-01 3.14e+03 0.00 1.000

ROAD\_SURFACE\_CONDITION\_DESCDry 1.54e+01 2.14e+03 0.01 0.994

ROAD\_SURFACE\_CONDITION\_DESCIce -8.55e-01 2.32e+03 0.00 1.000

ROAD\_SURFACE\_CONDITION\_DESCOther -5.57e-01 4.09e+03 0.00 1.000

ROAD\_SURFACE\_CONDITION\_DESCSandMudDirtOil Or Gravel -1.30e+00 6.10e+03 0.00 1.000

ROAD\_SURFACE\_CONDITION\_DESCSlush -1.36e-01 2.88e+03 0.00 1.000

ROAD\_SURFACE\_CONDITION\_DESCSnow 1.45e+01 2.14e+03 0.01 0.995

ROAD\_SURFACE\_CONDITION\_DESCUnknown 5.44e-01 3.56e+03 0.00 1.000

ROAD\_SURFACE\_CONDITION\_DESCWater(Standing Moving) -1.20e+00 5.17e+03 0.00 1.000

ROAD\_SURFACE\_CONDITION\_DESCWet 1.53e+01 2.14e+03 0.01 0.994

LIGHT\_CONDITION\_DESCDark - Lighted 1.54e+01 3.87e+03 0.00 0.997

LIGHT\_CONDITION\_DESCDark - Not Lighted 1.59e+01 3.87e+03 0.00 0.997

LIGHT\_CONDITION\_DESCDark - Unknown Roadway Lighting -5.95e-01 4.64e+03 0.00 1.000

LIGHT\_CONDITION\_DESCDawn 1.52e+01 3.87e+03 0.00 0.997

LIGHT\_CONDITION\_DESCDaylight 1.43e+01 3.87e+03 0.00 0.997

LIGHT\_CONDITION\_DESCDusk 1.52e+01 3.87e+03 0.00 0.997

LIGHT\_CONDITION\_DESCOther -1.12e+00 5.83e+03 0.00 1.000

LIGHT\_CONDITION\_DESCUnknown -1.28e-01 5.74e+03 0.00 1.000

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 868.14 on 4965 degrees of freedom

Residual deviance: 772.05 on 4931 degrees of freedom

(34 observations deleted due to missingness)

AIC: 842.1

Number of Fisher Scoring iterations: 18

R> mod1<-glm(FATALITIES~MCS150\_MILEAGE+MCS150\_MILEAGE\_YEAR+INJURIES+NBR\_POWER\_UNIT+DRIVER\_TOTAL+inter+Intra+HM\_FLAG1+train$WEATHER\_CONDITION\_DESC+ROAD\_SURFACE\_CONDITION\_DESC+LIGHT\_CONDITION\_DESC, family=binomial(link="cloglog"),data=train)

Warning message:

glm.fit: fitted probabilities numerically 0 or 1 occurred

R> summary(mod1)

Call:

glm(formula = FATALITIES ~ MCS150\_MILEAGE + MCS150\_MILEAGE\_YEAR +

INJURIES + NBR\_POWER\_UNIT + DRIVER\_TOTAL + inter + Intra +

HM\_FLAG1 + train$WEATHER\_CONDITION\_DESC + ROAD\_SURFACE\_CONDITION\_DESC +

LIGHT\_CONDITION\_DESC, family = binomial(link = "cloglog"),

data = train)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.964 -0.184 -0.160 -0.100 3.354

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -4.91e+01 5.07e+03 -0.01 0.992

MCS150\_MILEAGE 1.70e-09 2.45e-09 0.70 0.487

MCS150\_MILEAGE\_YEAR 2.32e-02 6.33e-02 0.37 0.714

INJURIES 9.19e-02 1.60e-02 5.75 8.8e-09

NBR\_POWER\_UNIT -7.60e-05 3.56e-05 -2.13 0.033

DRIVER\_TOTAL 1.83e-05 2.38e-05 0.77 0.441

inter 1.02e+00 5.42e-01 1.89 0.059

Intra -1.58e+01 1.63e+03 -0.01 0.992

HM\_FLAG1 8.80e-01 1.02e+00 0.86 0.390

train$WEATHER\_CONDITION\_DESCBlowing Sand Soil Dirt Or Snow 2.11e-01 3.57e+03 0.00 1.000

train$WEATHER\_CONDITION\_DESCFog 1.49e+01 2.75e+03 0.01 0.996

train$WEATHER\_CONDITION\_DESCNo Adverse Conditions 1.42e+01 2.75e+03 0.01 0.996

train$WEATHER\_CONDITION\_DESCOther 1.47e+01 2.75e+03 0.01 0.996

train$WEATHER\_CONDITION\_DESCRain 1.47e+01 2.75e+03 0.01 0.996

train$WEATHER\_CONDITION\_DESCSevere Crosswinds -1.31e+00 5.72e+03 0.00 1.000

train$WEATHER\_CONDITION\_DESCSleet Hail 1.66e+01 2.75e+03 0.01 0.995

train$WEATHER\_CONDITION\_DESCSnow 1.44e+01 2.75e+03 0.01 0.996

train$WEATHER\_CONDITION\_DESCUnknown -5.61e-01 3.02e+03 0.00 1.000

ROAD\_SURFACE\_CONDITION\_DESCDry 1.53e+01 2.04e+03 0.01 0.994

ROAD\_SURFACE\_CONDITION\_DESCIce -8.85e-01 2.23e+03 0.00 1.000

ROAD\_SURFACE\_CONDITION\_DESCOther -5.71e-01 3.93e+03 0.00 1.000

ROAD\_SURFACE\_CONDITION\_DESCSandMudDirtOil Or Gravel -1.18e+00 5.93e+03 0.00 1.000

ROAD\_SURFACE\_CONDITION\_DESCSlush -2.72e-01 2.76e+03 0.00 1.000

ROAD\_SURFACE\_CONDITION\_DESCSnow 1.45e+01 2.04e+03 0.01 0.994

ROAD\_SURFACE\_CONDITION\_DESCUnknown 5.82e-01 3.45e+03 0.00 1.000

ROAD\_SURFACE\_CONDITION\_DESCWater(Standing Moving) -1.32e+00 4.96e+03 0.00 1.000

ROAD\_SURFACE\_CONDITION\_DESCWet 1.50e+01 2.04e+03 0.01 0.994

LIGHT\_CONDITION\_DESCDark - Lighted 1.53e+01 3.74e+03 0.00 0.997

LIGHT\_CONDITION\_DESCDark - Not Lighted 1.57e+01 3.74e+03 0.00 0.997

LIGHT\_CONDITION\_DESCDark - Unknown Roadway Lighting -6.13e-01 4.47e+03 0.00 1.000

LIGHT\_CONDITION\_DESCDawn 1.51e+01 3.74e+03 0.00 0.997

LIGHT\_CONDITION\_DESCDaylight 1.42e+01 3.74e+03 0.00 0.997

LIGHT\_CONDITION\_DESCDusk 1.51e+01 3.74e+03 0.00 0.997

LIGHT\_CONDITION\_DESCOther -1.09e+00 5.60e+03 0.00 1.000

LIGHT\_CONDITION\_DESCUnknown -1.63e-02 5.54e+03 0.00 1.000

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 868.14 on 4965 degrees of freedom

Residual deviance: 774.70 on 4931 degrees of freedom

(34 observations deleted due to missingness)

AIC: 844.7

Number of Fisher Scoring iterations: 18

R code:

library(ggplot2)

library(clue)

library(knitr)

library(sqldf)

options(prompt="R> ", digits=4, show.signif.stars=FALSE)

options("repos"="http://cran.us.r-project.org")

install.packages("ResourceSelection",repos="http://cran.us.r-project.org") #Hosmer-Lemeshow GOF statistic

install.packages("lmtest", repos="http://cran.us.r-project.org")

install.packages("ROCR",repos="http://cran.us.r-project.org") #to construct ROC curve

library("ResourceSelection")

library(astsa)

library(nlme)

library(lmtest)

library(car)

library(MASS)

library(aod)

library(Rcpp)

# working with a large dataset#

setwd("/Users/xiaoyantang/Downloads/Inspection\_2017Sep")

ffdf1 <- read.table(file = "2017Sep\_Inspection.txt", header = TRUE, sep = ",")

head(ffdf1,row=5)

#crash report#

ffdf2 <-read.table(file = "2017Sep\_crash.txt", header = TRUE, sep = ",")

head(ffdf2,row=5)

#censos data#

ffdfcen<-read.table(file = "FMCSA\_CENSUS1\_2017Sep.txt", header = TRUE, sep = ",")

#passenger data#

passenger<-read.table(file = "SMS\_AB\_Pass\_2017Sep.txt", header = TRUE, sep = ",")

#Subset passenger use data in the sensor dataset#

cenpassenger=subset(ffdfcen,ffdfcen$PC\_FLAG=="Y")

#There are 33416 passenger in total#

# I want to know which City has the most passenger car, just top 10#

sqldf("SELECT PHY\_CITY, COUNT(DOT\_NUMBER) as num FROM cenpassenger GROUP BY PHY\_CITY ORDER BY num DESC limit 10")

#accident#

ffdf2$accident\_severity[ffdf2$FATALITIE>0] = "Fatal Accident"

ffdf2$accident\_severity[ffdf2$FATALITIE==0] = "Slight Accident"

ggplot(ffdf2,aes(y=factor(ffdf2$FATALITIE),x=accident\_severity,color=ffdf2$ROAD\_SURFACE\_CONDITION\_DESC))+geom\_point()+geom\_jitter()+xlab("severity")+ylab("Number of casualties")+ggtitle("Fig 2. severity vs number of casualties colored by road surface")

# It seems that most of the accients are happening under dry ice condtion

# I also merged two table using inner join so that conses passnger with car accidents will show#

Accpassenger<-merge(cenpassenger, ffdf2, by = "DOT\_NUMBER")

#Then what I find is that some viechle are always in accidents, I should count how many cars are like that#

sqldf("SELECT LEGAL\_NAME, COUNT(REPORT\_NUMBER) as num FROM Accpassenger GROUP BY DOT\_NUMBER ORDER BY num DESC limit 10")

#These compnaies had a lot of accidents,should be banned#

company = data.frame(name= c("FIRST STUDENT INC","DURHAM SCHOOL","NEW YORK CITY TRANSIT","REGIONAL TRANSPORTATION","FIRST TRANSIT INC"),Count=c(865,326,274,172,165))

barplot( height= company$Count, names.arg = company$name,main="Figure 2:Top five company with most accidents",col=rainbow(5),ylim=c(0,900),xlab="company", ylab="accident count",cex.names=0.8)

ggplot(Accpassenger,aes(y=factor(Accpassenger$FATALITIE),x=Accpassenger$CARRIER\_OPERATION,color=Accpassenger$SEVERITY\_WEIGHT))+geom\_point()+geom\_jitter()+xlab("Carrier")+ylab("Number of casualties")+ggtitle("Fig 3. carrier vs number of casualties colored by severity weight")

# We can find that among A = Interstate, B = Intrastate Hazmat, C = Intrastate Non-Hazmat, B has the least casualties

#I also want to know which city has most of the car accident#

sqldf("SELECT PHY\_CITY, COUNT(Report\_number ) as num FROM Accpassenger GROUP BY PHY\_CITY ORDER BY num DESC limit 10")

#We found that CINCINNATI has most of the accidents reported with commericial car#

names(Accpassenger)

#data description

#DOT\_NUMBER - Unique USDOT Number of the Motor Carrier

#LEGAL\_NAME - Legal name of a carrier

#DBA\_NAME - Carrier's Doing-Business-As name

#CARRIER\_OPERATION - Codes identifying carriers' type of Operation; A = Interstate, B = Intrastate Hazmat, C = Intrastate Non-Hazmat

#HM\_FLAG - Carrier is subject to placardable HM threshold ( Y = Yes, N = No)

#PC\_FLAG - Carrier is subject to passengercarrier Threshold (Y = Yes, N = No)

#MCS150\_MILEAGE - Vehicle Mileage Traveled (VMT) reported on the carrier's MCS-150 form

#MCS150\_MILEAGE\_YEAR - Year for which VMT was reported

#ADD\_DATE - Date when carrier information was added to MCMIS Database System

#OIC\_STATE - FMCSA State office with oversight for this carrier

#NBR\_POWER\_UNIT - Number of power units reported

#DRIVER\_TOTAL - Number of drivers reported

#Report\_number - Unique state report number for the incident

#Report\_seq\_no - Sequence number for each vehicle involved in a crash

#DOT\_Number - Unique number assigned to a company by the DOT

#Report\_Date - The date a incident occurred

#Report\_State - State abbreviation

#Fatalities - Total number of fatalities reported in the crash

#Injuries - Total number of injuries reported in the crash

#Tow\_Away - 'Y' indicates that a vehicle involved in the crash was towed from the scene.

#Hazmat\_released - 'Y' indicates that hazardous materials were released at the time of a crash

#Trafficway\_Desc - Description of the trafficway

#Access\_Control\_Desc - Description of the access control

#Road\_surface\_Condition\_Desc - Description of the road surface condition

#Weather\_Condition\_Desc - Description of the weather condition

#Light\_Condition\_Desc - Description of the light condition

#Vehicle\_ID\_Number - Vehicle Identification number (VIN)

#Vehicle\_License\_number - Vehicle license number

#Vehicle\_license\_state - vehicle license state

#Severity\_Weight - The severity weight that is assigned to the incident

#Time\_weight - the time weight that is assigned to the incident

#citation\_issue\_desc - Description of the citation issue

#seq\_num - Sequence number

# I will partition the data, first 5000 for training and the rest 2694 will be left for fitting#

#data tranfer, know how many years there are for the car accidents#

Accpassenger$MCS150\_MILEAGE\_YEAR=2017-Accpassenger$MCS150\_MILEAGE\_YEAR

train=Accpassenger[1:5000,]

test=Accpassenger[5000:7695,]

levels(train$WEATHER\_CONDITION\_DESC)

levels(train$ROAD\_SURFACE\_CONDITION\_DESC)

levels(train$LIGHT\_CONDITION\_DESC)

# continous varibles are listed as follows#

#MCS150\_MILEAGE,DRIVER\_TOTAL,Injuries,NBR\_POWER\_UNIT,DRIVER\_TOTAL

#Set up dummy varibles#

# Set up 3 dummy variables for carrier operation

train$inter= ifelse((train$CARRIER\_OPERATION=='A'), 1, 0)

train$Intra= ifelse((train$CARRIER\_OPERATION=='B'), 1, 0)

train$intranon= ifelse((train$CARRIER\_OPERATION=='C'), 1, 0)

#accident#

train$FATALITIES[train$FATALITIES>0] = 1

class(train$FATALITIES)

#HM\_FLAG#

train$HM\_FLAG1=ifelse((train$HM\_FLAG=='Y'),1,0)

train$HM\_FLAG0=ifelse(train$HM\_FLAG=='N',1,0)

#PC\_FLAG#

train$PC\_FLAG1=ifelse((train$PC\_FLAG=='Y'),1,0)

train$PC\_FLAG0=ifelse((train$PC\_FLAG=='N'),1,0)

#define factorial data#

train$WEATHER\_CONDITION\_DESC <- factor(train$WEATHER\_CONDITION\_DESC)

train$ROAD\_SURFACE\_CONDITION\_DESC<-factor(train$ROAD\_SURFACE\_CONDITION\_DESC)

train$LIGHT\_CONDITION\_DESC<-factor(train$LIGHT\_CONDITION\_DESC)

train$Time\_weight<-factor(train$TIME\_WEIGHT)

train$SEVERITY\_WEIGHT<-factor(train$SEVERITY\_WEIGHT)

# We use binary logistic regression for fitting#

mod0<-glm(FATALITIES~MCS150\_MILEAGE+MCS150\_MILEAGE\_YEAR+INJURIES+NBR\_POWER\_UNIT+DRIVER\_TOTAL+inter+Intra+HM\_FLAG1+train$WEATHER\_CONDITION\_DESC+ROAD\_SURFACE\_CONDITION\_DESC+LIGHT\_CONDITION\_DESC, family=binomial,data=train)

summary(mod0)

mod1<-glm(FATALITIES~MCS150\_MILEAGE+MCS150\_MILEAGE\_YEAR+INJURIES+NBR\_POWER\_UNIT+DRIVER\_TOTAL+inter+Intra+HM\_FLAG1+train$WEATHER\_CONDITION\_DESC+ROAD\_SURFACE\_CONDITION\_DESC+LIGHT\_CONDITION\_DESC, family=binomial(link="cloglog"),data=train)

summary(mod1)