BA_HW4_EDA

Yiming Ge

September 30, 2020

###PROBLEM 1: CitiBike anomaly detection & neighborhood usage get data

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(stringr)
url1<-('https://raw.githubusercontent.com/jcbonilla/BusinessAnalytics/master/BAData/JC-201709-citibike-
Bikedata<-read.csv(url1,header = TRUE,stringsAsFactors = FALSE)
head(Bikedata)
```

```
##   tripduration           starttime            stoptime start.station.id
## 1          364 2017-09-01 00:02:01 2017-09-01 00:08:05             3183
## 2          357 2017-09-01 00:08:12 2017-09-01 00:14:09             3187
## 3          432 2017-09-01 00:10:12 2017-09-01 00:17:24             3195
## 4          934 2017-09-01 00:10:11 2017-09-01 00:25:46             3272
## 5          932 2017-09-01 00:10:16 2017-09-01 00:25:48             3272
## 6          414 2017-09-01 00:15:32 2017-09-01 00:22:26             3186
##   start.station.name start.station.latitude start.station.longitude
## 1     Exchange Place                40.71625               -74.03346
## 2          Warren St                40.72112               -74.03805
## 3            Sip Ave                40.73074               -74.06378
## 4       Jersey & 3rd                40.72333               -74.04595
## 5       Jersey & 3rd                40.72333               -74.04595
## 6      Grove St PATH                40.71959               -74.04312
##   end.station.id end.station.name end.station.latitude end.station.longitude
## 1           3276 Marin Light Rail             40.71458              -74.04282
## 2           3199     Newport Pkwy             40.72874              -74.03211
## 3           3280      Astor Place             40.71928              -74.07126
## 4           3207      Oakland Ave             40.73760              -74.05248
```

```
## 5           3207      Oakland Ave             40.73760             -74.05248
## 6           3480     WS Don't Use              0.00000               0.00000
##   bikeid   usertype birth.year gender
## 1  29670 Subscriber       1989      1
## 2  26163 Subscriber       1980      1
## 3  26273 Subscriber       1988      1
## 4  26297 Subscriber       1991      1
## 5  29247 Subscriber       1993      2
## 6  29589   Customer       NULL      0
```

What anomalies are detectable with tripduration and the age of the user?

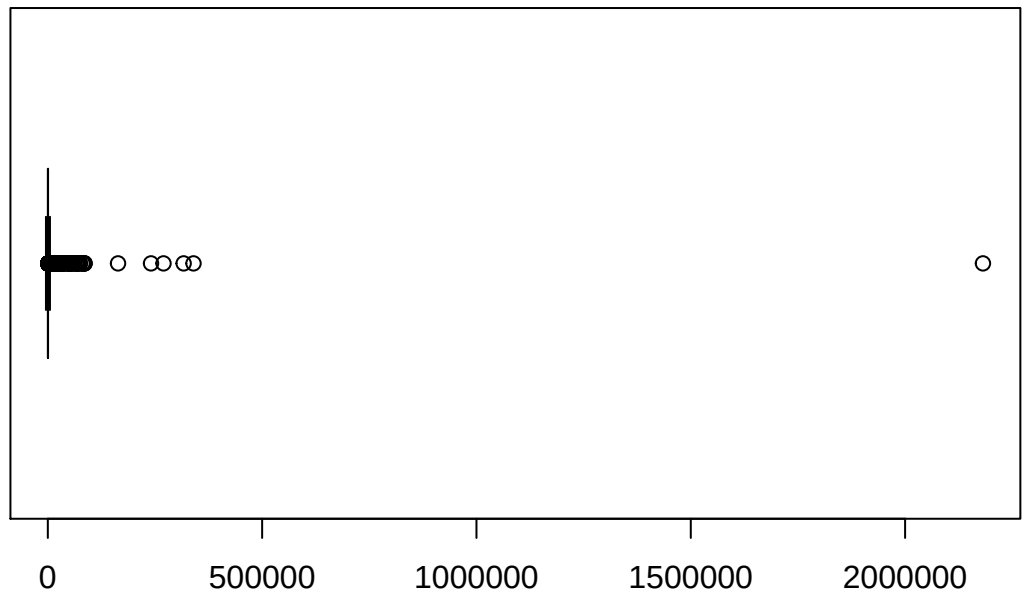
```
summary(Bikedata$tripduration)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##      61.0     238.0     355.0     756.9     610.0 2181628.0
```

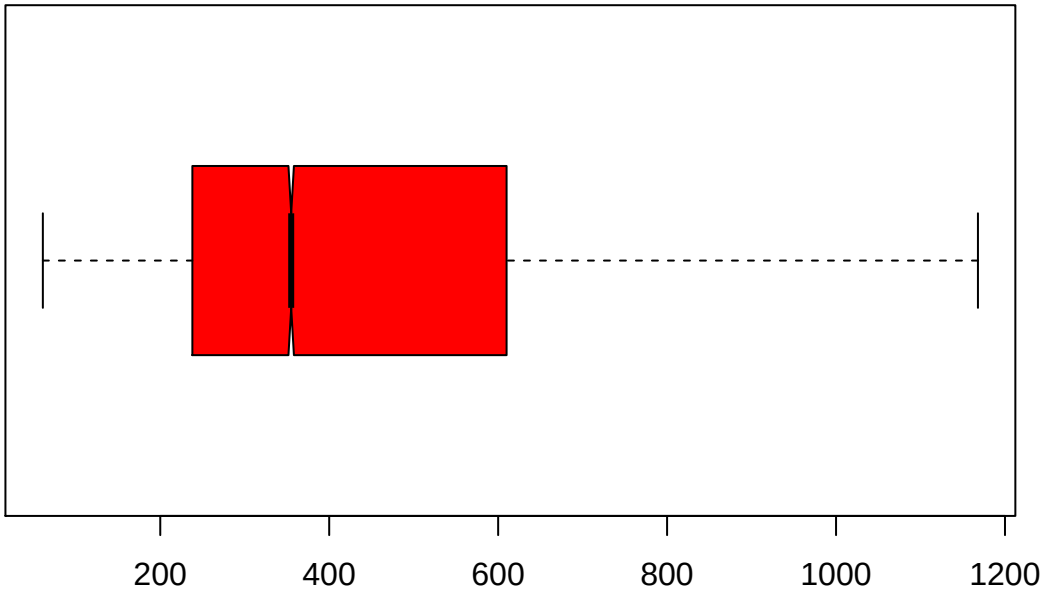
```
sd(Bikedata$tripduration)
```

```
## [1] 12628.57
```

```
boxplot(Bikedata$tripduration,col = 'red',horizontal = TRUE,notch = TRUE,outline = TRUE)
```



```
boxplot(Bikedata$tripduration,col = 'red',horizontal = TRUE,notch = TRUE,outline = FALSE)
```



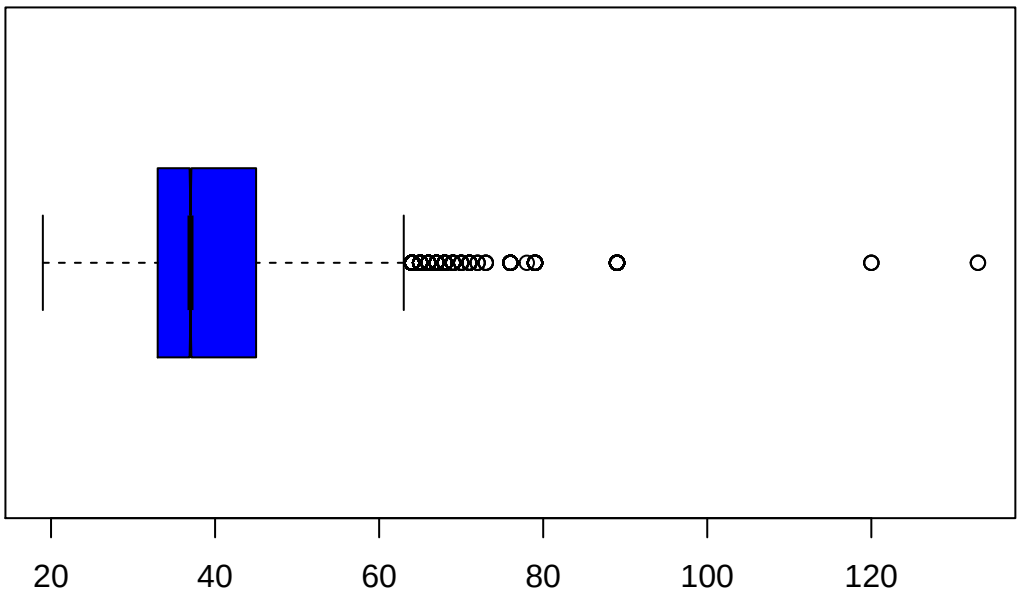

```
Bikedata<-Bikedata[-which(Bikedata$birth.year =='NULL'),]#drop null birth year data
Bikedata$age<-2020-suppressWarnings(as.numeric(Bikedata$birth.year))#calculate the age
summary(Bikedata$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   19.00   33.00   37.00   39.87   45.00  133.00
```

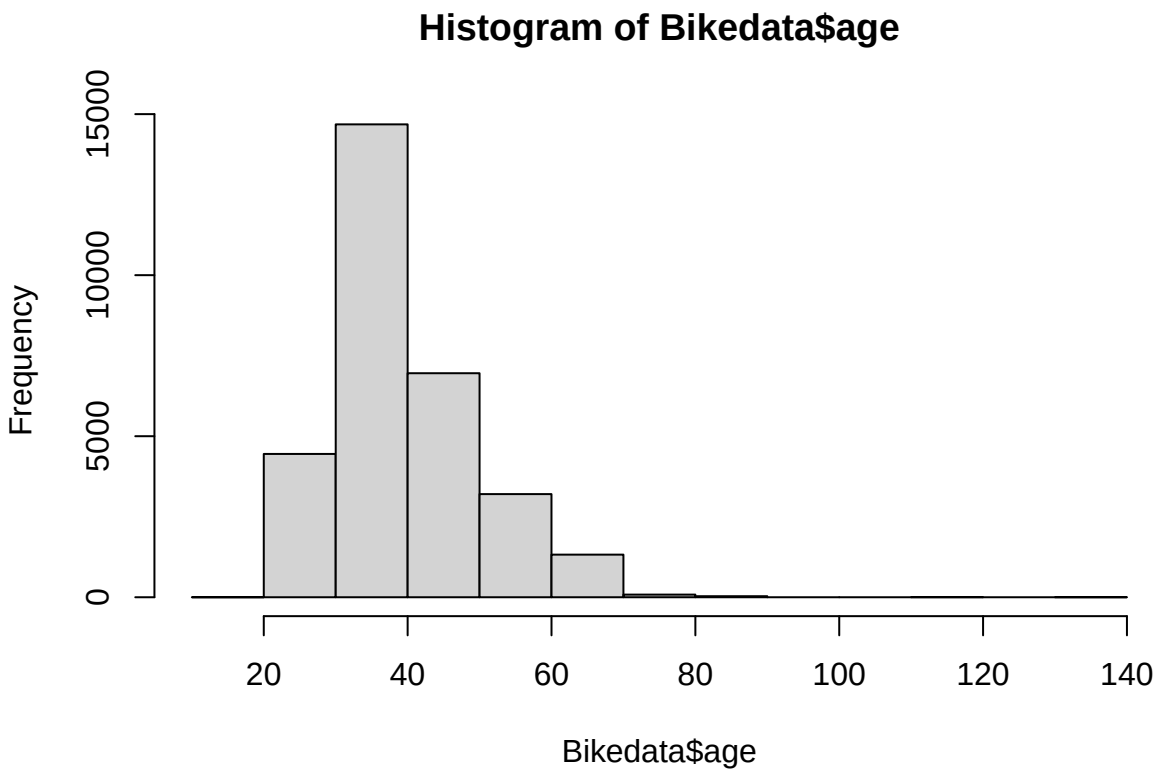
```
sd(Bikedata$age)
```

```
## [1] 10.04942
```

```
boxplot(Bikedata$age,col = 'blue',horizontal = TRUE,notch = TRUE,outline = TRUE,na.rm=TRUE)
```




```
hist(Bikedata$age)
```



Which neighborhoods have the highest demand in traffic usage? Assume from 6am to 9am morning Assume from 5pm to 8pm evening Assume from 11am to 2pm alternative time

```
Bikedata$StartHourTime<-as.numeric(substring(Bikedata$starttime,12,13))#get 'hour' time
morning<-subset(Bikedata,StartHourTime>5&StartHourTime<10)
evening<-subset(Bikedata,StartHourTime>16&StartHourTime<21)
Atime<-subset(Bikedata,StartHourTime>10&StartHourTime<15)
```

morning start location frequency

```
mNeighbor<-table(morning$start.station.name) %>% as.data.frame() %>% arrange(desc(Freq))
head(mNeighbor,3)#top3 neighborhood name
```

```
##            Var1 Freq
## 1 Hamilton Park 1208
## 2  Morris Canal  566
## 3  Brunswick St  368
```

evening start location frequency

```
eNeighbor<-table(evening$start.station.name) %>% as.data.frame() %>% arrange(desc(Freq))
head(eNeighbor,3)#top3 neighborhood name
```

```
##             Var1 Freq
## 1  Grove St PATH 2183
## 2 Exchange Place 1075
## 3        Sip Ave  763
```

Alternative time start location frequency

```
aNeighbor<-table(Atime$start.station.name) %>% as.data.frame() %>% arrange(desc(Freq))
head(aNeighbor,3)#top3 neighborhood name
```

```
##             Var1 Freq
## 1  Grove St PATH  454
## 2  Hamilton Park  353
## 3 Exchange Place  300
```

recommendations: 1.From anomalies analysis in (1), we found the variance of trip duration is very high and most trip duration is between 200 to 600.The outlier is very huge which is 2181628.Some operation mistakes may cause this error.In addition,we also found some outliers in age which are lager than 100 years old. The wrong customer information may cause this error. I suggest the company to check and repair its data collecting system. 2.From geographic usage in (2), I give the top 3 high frequency neighborhood in different time. In the morning, I recommend to put more bikes in Hamilton Park, Morris Canal,Brunswick St. In the evening, I recommend to put more bikes in Grove St PATH,Exchange Place,Sip Ave. In the Alternative time, I recommend to put more bikes in Grove St PATH,Hamilton Park,Exchange Place.

###PROBLEM 2: Aviation Accidents get data

```
url2 = 'https://raw.githubusercontent.com/jcbonilla/BusinessAnalytics/master/BAData/aviation.csv'
Avidata<-read.csv(url2, header=TRUE,stringsAsFactors=FALSE)
head(Avidata)
```

```
##          Event.Id Investigation.Type Accident.Number   Event.Date
## 1 20130607X70213           Accident      CEN13FA326   06/07/2013
## 2 20130607X04715           Accident      ERA13FA273   06/06/2013
## 3 20130531X43432           Accident     WPR13FA254B   05/31/2013
## 4 20130531X43432           Accident     WPR13FA254A   05/31/2013
## 5 20130530X14133           Accident      CEN13WA307   05/29/2013
## 6 20130528X64403           Accident      WPR13FA244   05/28/2013
##                  Location         Country Latitude  Longitude Airport.Code
## 1              Baker, LA   United States  30.57639  -91.13694
## 2         Manchester, KY   United States  37.13278  -83.75639
## 3             Anthem, AZ   United States  33.86472 -112.20139        KDVT
## 4             Anthem, AZ   United States  33.86472 -112.20139        KDVT
## 5  Gap Aerodrome, France          France        NA         NA
## 6       Mountainaire, AZ   United States  35.08278 -111.66778         FLG
##                     Airport.Name Injury.Severity Aircraft.Damage
## 1                                      Fatal(1)       Destroyed
## 2                                      Fatal(3)       Destroyed
## 3  Phoenix Deer Valley Airport        Fatal(4)       Destroyed
## 4  Phoenix Deer Valley Airport        Fatal(4)       Destroyed
## 5                                      Fatal(1)       Destroyed
## 6    Flagstaff Pulliam Airport        Fatal(2)       Destroyed
##   Aircraft.Category Registration.Number                      Make
## 1         Airplane              N510LD           HAWKER BEECHCRAFT
## 2       Helicopter              N114AE     BELL HELICOPTER TEXTRON
## 3         Airplane              N2459K                      CESSNA
## 4         Airplane              N327PA                       PIPER
## 5         Airplane               N68XM                       PIPER
## 6         Airplane              N999PK   RAYTHEON AIRCRAFT COMPANY
```

```
##            Model Amateur.Built Number.of.Engines     Engine.Type
## 1        B200GT            No                  2     Turbo Prop
## 2        206L-1            No                  1    Turbo Shaft
## 3          172S            No                  1  Reciprocating
## 4     PA-28-181            No                  1  Reciprocating
## 5  PA-28RT-201T            No                  1  Reciprocating
## 6           A36            No                  1  Reciprocating
##               FAR.Description Schedule Purpose.of.Flight Air.Carrier
## 1  Part 91: General Aviation                   Personal
## 2  Part 91: General Aviation                Positioning
## 3  Part 91: General Aviation              Instructional
## 4  Part 91: General Aviation              Instructional
## 5                    Unknown
## 6  Part 91: General Aviation                   Personal
##   Total.Fatal.Injuries Total.Serious.Injuries Total.Minor.Injuries
## 1                    1                     NA                   NA
## 2                    3                     NA                   NA
## 3                    4                     NA                   NA
## 4                    4                     NA                   NA
## 5                    1                     NA                   NA
## 6                    2                     NA                   NA
##   Total.Uninjured Weather.Condition Broad.Phase.of.Flight Report.Status
## 1              NA              VMC                          Preliminary
## 2              NA              VMC                          Preliminary
## 3              NA              VMC                          Preliminary
## 4              NA              VMC                          Preliminary
## 5              NA                                               Foreign
## 6              NA              VMC                          Preliminary
##   Publication.Date  X
## 1      06/14/2013  NA
## 2      06/12/2013  NA
## 3      06/10/2013  NA
## 4      06/10/2013  NA
## 5      06/04/2013  NA
## 6      06/05/2013  NA
```

Analysis of fatal vs. non-fatal crashes in the US from the 1940s through 2013.

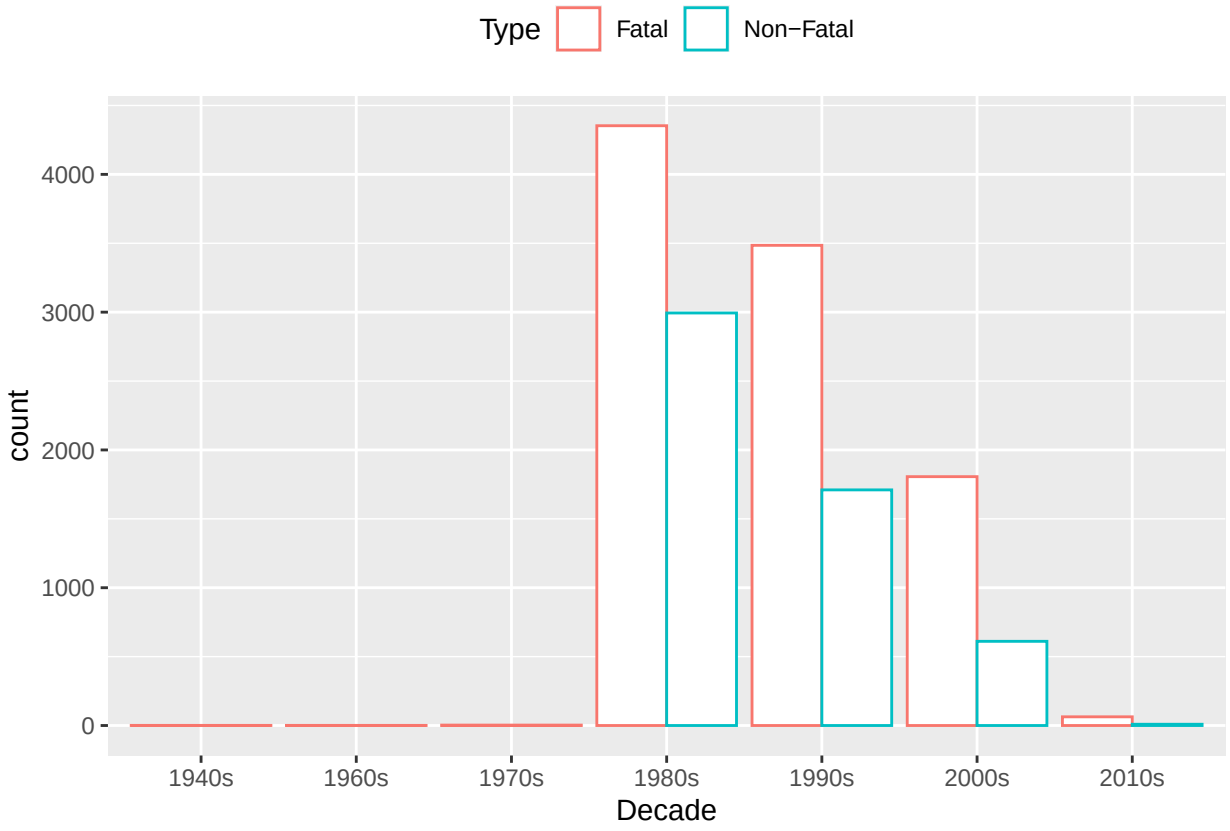
```
SelectAvi<-select(Avidata,Event.Date,Country,Injury.Severity)
SelectAvi<-SelectAvi[-which(SelectAvi$Injury.Severity == ' Unavailable '),]#drop unavailable
SelectAvi$Type<-ifelse(SelectAvi$Injury.Severity %in% ' Non-Fatal ','Non-Fatal','Fatal')
head(SelectAvi)
```

```
##     Event.Date         Country Injury.Severity  Type
## 1  06/07/2013   United States        Fatal(1)  Fatal
## 2  06/06/2013   United States        Fatal(3)  Fatal
## 3  05/31/2013   United States        Fatal(4)  Fatal
## 4  05/31/2013   United States        Fatal(4)  Fatal
## 5  05/29/2013          France        Fatal(1)  Fatal
## 6  05/28/2013   United States        Fatal(2)  Fatal
```

```
USA <- SelectAvi[which(SelectAvi$Country == ' United States ' ),]
USA$Year<-as.numeric(substring(USA$Event.Date,8,11))#This takes string and only keeps the characters be
USA$Decade<-cut(USA$Year,seq(1940,2020,10),labels = c('1940s','1950s','1960s','1970s','1980s','1990s','
```

Draw the plot

```
ggplot(USA, aes(x=Decade, color=Type)) + stat_count(fill="white", position="dodge")+theme(legend.positi
```



Countries with most incidents

```
SelectAviFatal<-subset(SelectAvi,Type == 'Fatal')
numextract <- function(string){
  str_extract(string, "\\-*\\d+\\.*\\d*")
}
SelectAviFatal$Number<-as.numeric(numextract(SelectAviFatal$Injury.Severity))
Death<-aggregate(Number~Country,data=SelectAviFatal,sum)
sort_Death<-Death[order(-Death$Number),]
head(sort_Death,1)#show rank1 death number country
```

```
##             Country Number
## 121  United States   21898
```

```
incidents<-table(SelectAviFatal$Country) %>% as.data.frame() %>% arrange(desc(Freq))
head(incidents,1)#show rank1 incidents country
```

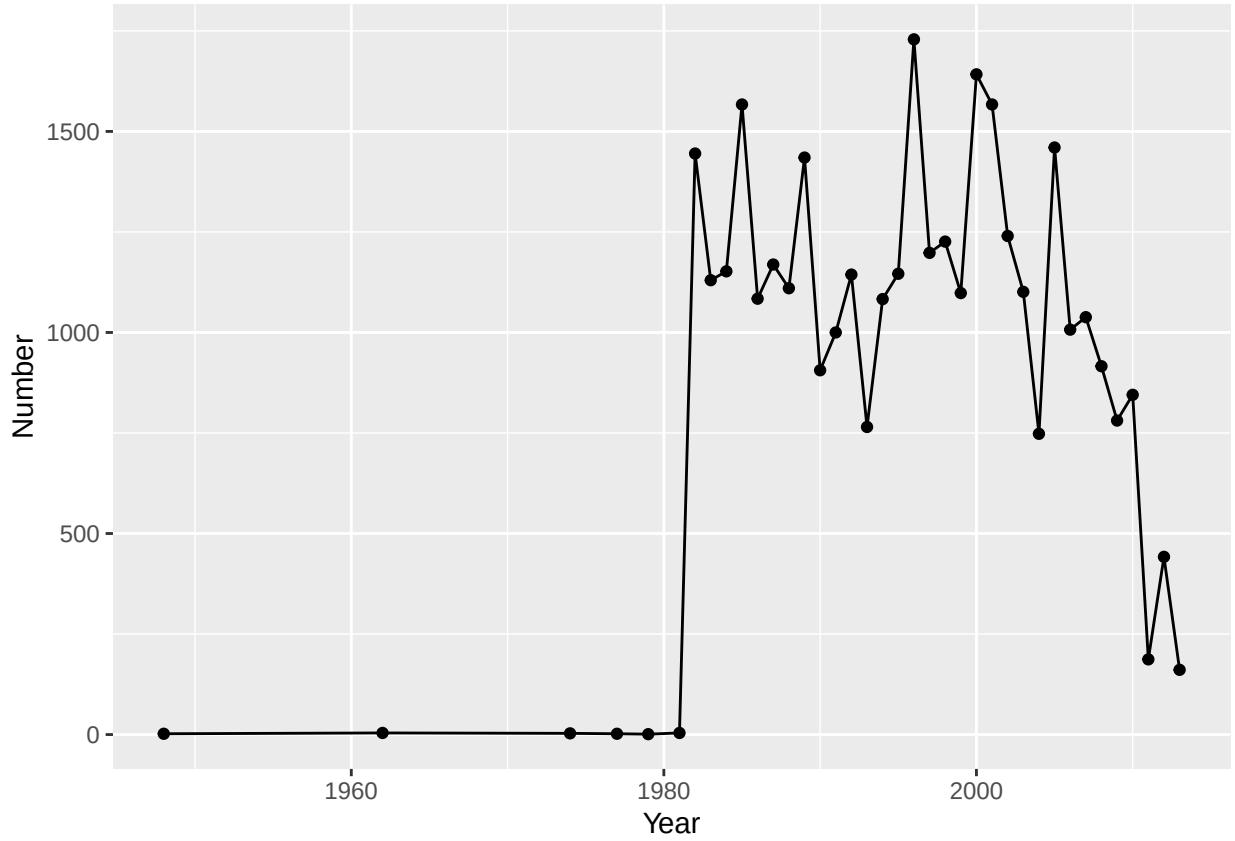
```
##              Var1 Freq
## 1  United States  9712
```

Historical deaths by year

```r
SelectAviFatal$Year<-as.numeric(substring(SelectAviFatal$Event.Date,8,11))
Deaths_by_year<-aggregate(Number~Year,data = SelectAviFatal,sum)
Deaths_by_year
```

```
##    Year Number
## 1  1948      2
## 2  1962      4
## 3  1974      3
## 4  1977      2
## 5  1979      1
## 6  1981      4
## 7  1982   1445
## 8  1983   1130
## 9  1984   1152
## 10 1985   1567
## 11 1986   1084
## 12 1987   1169
## 13 1988   1110
## 14 1989   1435
## 15 1990    906
## 16 1991   1000
## 17 1992   1144
## 18 1993    765
## 19 1994   1083
## 20 1995   1146
## 21 1996   1729
## 22 1997   1198
## 23 1998   1226
## 24 1999   1098
## 25 2000   1642
## 26 2001   1567
## 27 2002   1240
## 28 2003   1101
## 29 2004    748
## 30 2005   1460
## 31 2006   1007
## 32 2007   1038
## 33 2008    916
## 34 2009    781
## 35 2010    845
## 36 2011    187
## 37 2012    442
## 38 2013    161
```

```r
ggplot(data=Deaths_by_year, aes(x=Year, y=Number, group=1)) +
  geom_line()+
  geom_point()
```



###PROBLEM 3: Retail Targets get data

```
url3 = 'https://raw.githubusercontent.com/jcbonilla/BusinessAnalytics/master/BAData/HDLData.csv'
HDLdata<-read.csv(url3, header=TRUE,stringsAsFactors=FALSE)
head(HDLdata)
```

```
##   areaname county state r1 r2 Lcount HDcount pop_2000 pop_2010 income_2000
## 1  Autauga   1001    AL  6  3      1       1    43671    54571       48458
## 2  Baldwin   1003    AL  6  3      2       2   140415   182265       47028
## 3  Barbour   1005    AL  6  3      0       0    29038    27457       31877
## 4     Bibb   1007    AL  6  3      0       0    20826    22915       37230
## 5   Blount   1009    AL  6  3      0       0    51024    57322       41573
## 6  Bullock   1011    AL  6  3      0       0    11714    10914       23990
##   income_2010 pct_U18_2000 pct_U18_2010 pctcollege_2000 pctcollege_2010
## 1       63458         28.6         26.8            18.0            21.8
## 2       57447         24.4         23.0            23.1            26.6
## 3       40109         25.4         21.9            10.9            12.4
## 4       51951         25.4         22.7             7.1            11.1
## 5       53807         25.4         24.6             9.6            12.3
## 6       33763         26.1         22.3             7.7             9.0
##   ownhome_2000 ownhome_2010 density_2000 density_2010 pctwhite_2000
## 1         80.8         75.4         73.3         91.8          80.7
## 2         79.5         72.5         88.0        114.6          87.1
## 3         73.1         66.8         32.8         31.0          51.3
## 4         80.2         75.6         33.4         36.8          76.7
## 5         83.4         80.6         79.0         88.9          95.1
```

```
## 6         74.5         68.8         18.7         17.5          25.3
##   pctwhite_2010 pctblack_2000 pctblack_2010
## 1          78.5          22.1          17.7
## 2          85.7          12.2           9.4
## 3          48.0          44.3          46.9
## 4          75.8          24.2          22.0
## 5          92.6           1.5           1.3
## 6          23.0          65.4          70.2
```

```
summary(HDLdata)
```

```
##    areaname             county         state                r1
##  Length:3146        Min.   : 1001   Length:3146        Min.   :1.000
##  Class :character   1st Qu.:18180   Class :character   1st Qu.:4.000
##  Mode  :character   Median :29180   Mode  :character   Median :5.000
##                     Mean   :30404                      Mean   :5.193
##                     3rd Qu.:45084                      3rd Qu.:7.000
##                     Max.   :56045                      Max.   :9.000
##
##        r2            Lcount           HDcount           pop_2000
##  Min.   :1.000   Min.   : 0.0000   Min.   : 0.0000   Min.   :      0
##  1st Qu.:2.000   1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.:  11081
##  Median :3.000   Median : 0.0000   Median : 0.0000   Median :  24544
##  Mean   :2.669   Mean   : 0.5423   Mean   : 0.6205   Mean   :  89369
##  3rd Qu.:3.000   3rd Qu.: 1.0000   3rd Qu.: 0.0000   3rd Qu.:  61728
##  Max.   :4.000   Max.   :20.0000   Max.   :48.0000   Max.   :9500000
##
##     pop_2010        income_2000     income_2010      pct_U18_2000
##  Min.   :      0   Min.   :    0   Min.   :     0   Min.   : 2.00
##  1st Qu.:  11066   1st Qu.:35752   1st Qu.: 44914   1st Qu.:23.70
##  Median :  25837   Median :40674   Median : 51348   Median :25.30
##  Mean   :  98052   Mean   :42073   Mean   : 53234   Mean   :25.53
##  3rd Qu.:  66528   3rd Qu.:46548   3rd Qu.: 59254   3rd Qu.:27.20
##  Max.   :9800000   Max.   :97225   Max.   :140286   Max.   :46.60
##                                                     NA's   :9
##   pct_U18_2010   pctcollege_2000 pctcollege_2010  ownhome_2000
##  Min.   : 0.00   Min.   : 0.0    Min.   : 0.00   Min.   : 0.00
##  1st Qu.:21.40   1st Qu.:11.2    1st Qu.:12.90   1st Qu.:70.50
##  Median :23.30   Median :14.5    Median :16.60   Median :75.30
##  Mean   :23.42   Mean   :16.5    Mean   :18.69   Mean   :73.82
##  3rd Qu.:25.10   3rd Qu.:19.3    3rd Qu.:22.00   3rd Qu.:79.10
##  Max.   :41.60   Max.   :63.7    Max.   :69.50   Max.   :89.90
##  NA's   :3
##   ownhome_2010    density_2000      density_2010      pctwhite_2000
##  Min.   : 0.00   Min.   :    0.0   Min.   :    0.00   Min.   : 4.50
##  1st Qu.:68.70   1st Qu.:   16.9   1st Qu.:   16.82   1st Qu.:76.80
##  Median :73.60   Median :   42.4   Median :   45.10   Median :91.10
##  Mean   :72.17   Mean   :  243.1   Mean   :  259.08   Mean   :84.42
##  3rd Qu.:77.30   3rd Qu.:  104.3   3rd Qu.:  113.55   3rd Qu.:96.70
##  Max.   :89.80   Max.   :66834.6   Max.   :69467.50   Max.   :99.70
##                                                       NA's   :9
##  pctwhite_2010   pctblack_2000   pctblack_2010
##  Min.   : 2.70   Min.   : 0.00   Min.   : 0.000
##  1st Qu.:75.25   1st Qu.: 0.50   1st Qu.: 0.400
```

```
##  Median :89.10   Median : 2.10   Median : 2.000
##  Mean   :82.89   Mean   : 9.32   Mean   : 8.883
##  3rd Qu.:95.50   3rd Qu.:11.10   3rd Qu.:10.200
##  Max.   :99.20   Max.   :85.80   Max.   :85.700
##  NA's   :3       NA's   :9       NA's   :3
```

compute the NE region as follows: Maine(ME), New York(NY), New Jersey(NJ), Vermont(VT), Massachusetts(MA), Rhode Island(RI), Connecticut(CT), New Hampshire(NH), and Pennsylvania(PA)

```
NEregion <- c('ME','NY','NJ','VT','MA','RI','CT','NH','PA')
HDLdata$Region<-ifelse(HDLdata$state %in% NEregion,'NE','NON-NE')
head(HDLdata)
```

```
##   areaname county state r1 r2 Lcount HDcount pop_2000 pop_2010 income_2000
## 1  Autauga   1001    AL  6  3      1       1    43671    54571       48458
## 2  Baldwin   1003    AL  6  3      2       2   140415   182265       47028
## 3  Barbour   1005    AL  6  3      0       0    29038    27457       31877
## 4     Bibb   1007    AL  6  3      0       0    20826    22915       37230
## 5   Blount   1009    AL  6  3      0       0    51024    57322       41573
## 6  Bullock   1011    AL  6  3      0       0    11714    10914       23990
##   income_2010 pct_U18_2000 pct_U18_2010 pctcollege_2000 pctcollege_2010
## 1       63458         28.6         26.8            18.0            21.8
## 2       57447         24.4         23.0            23.1            26.6
## 3       40109         25.4         21.9            10.9            12.4
## 4       51951         25.4         22.7             7.1            11.1
## 5       53807         25.4         24.6             9.6            12.3
## 6       33763         26.1         22.3             7.7             9.0
##   ownhome_2000 ownhome_2010 density_2000 density_2010 pctwhite_2000
## 1         80.8         75.4         73.3         91.8          80.7
## 2         79.5         72.5         88.0        114.6          87.1
## 3         73.1         66.8         32.8         31.0          51.3
## 4         80.2         75.6         33.4         36.8          76.7
## 5         83.4         80.6         79.0         88.9          95.1
## 6         74.5         68.8         18.7         17.5          25.3
##   pctwhite_2010 pctblack_2000 pctblack_2010 Region
## 1          78.5          22.1          17.7 NON-NE
## 2          85.7          12.2           9.4 NON-NE
## 3          48.0          44.3          46.9 NON-NE
## 4          75.8          24.2          22.0 NON-NE
## 5          92.6           1.5           1.3 NON-NE
## 6          23.0          65.4          70.2 NON-NE
```

Use a variable for each 10-year percent change. Demographics include: Population, Income, Density, Ownhome

```
HDLdata_total<-aggregate(HDLdata[c("pop_2000", "pop_2010","income_2000","income_2010","density_2000","d
pop_change<-(HDLdata_total$pop_2010-HDLdata_total$pop_2000)/HDLdata_total$pop_2000
income_change<-(HDLdata_total$income_2010-HDLdata_total$income_2000)/HDLdata_total$income_2000
density_change<-(HDLdata_total$density_2010-HDLdata_total$density_2000)/HDLdata_total$density_2000
ownhome_change<-(HDLdata_total$ownhome_2010-HDLdata_total$ownhome_2000)/HDLdata_total$ownhome_2000
changeOne<-rbind(pop_change,income_change,density_change,ownhome_change)
colnames(changeOne) <- c("NE","NON-NE")
changeOne
```

```
##                           NE      NON-NE
## pop_change      0.032034597  0.11246624
## income_change   0.287486322  0.26328387
## density_change  0.034120198  0.08514444
## ownhome_change -0.007108451 -0.02341877
```

Use a variable for each 10-year percent change. Demographics include: Percentage of U18, college, white, black

```
HDLdata$U18Change<-HDLdata$pct_U18_2010-HDLdata$pct_U18_2000
HDLdata$collegeChange<-HDLdata$pctcollege_2010-HDLdata$pctcollege_2000
HDLdata$whiteChange<-HDLdata$pctwhite_2010-HDLdata$pctwhite_2000
HDLdata$blackChange<-HDLdata$pctblack_2010-HDLdata$pctblack_2000
HDLdata_mean<-aggregate(HDLdata[c('U18Change','collegeChange','whiteChange','blackChange')],by=list(Reg
HDLdata_mean
```

```
##   Region U18Change collegeChange whiteChange blackChange
## 1     NE -2.578341      3.244700   -2.294470  -0.2165899
## 2 NON-NE -2.073425      2.103209   -1.453767  -0.4362671
```