BA_HW3

Problem 1: CitiBike anomaly detection & neighborhood usage

```
url = "https://raw.githubusercontent.com/jcbonilla/BusinessAnalytics/master/BAData/JC-20
1709-citibike-tripdata.csv"
citibike = read.csv(url)
citibike = as.data.frame(citibike)
library(R.basic)
library(lubridate)
citibike$tripdur_z = zscore(citibike$tripduration, na.rm = TRUE)
age <- function(dob, age.day = today(), units = "years", floor = TRUE) {
  calc.age = interval(dob, age.day) / duration(num = 1, units = units)
  if (floor) return(as.integer(floor(calc.age)))
  return(calc.age)
}
citibike$birth.year = as.Date(citibike$birth.year, "%Y")
citibike$age = age(citibike$birth.year)
citibike$age_z = zscore(citibike$age, na.rm = TRUE)
```

```
ab_trip = subset(citibike, tripdur_z > 3, select = c(1, 2, 3, 4, 5, 8, 9, 12, 13, 15, 16
, 17))
head(ab_trip)
```

```
##      tripduration           starttime            stoptime start.station.id
## 1026        70054 2017-09-01 20:25:16 2017-09-02 15:52:50             3186
## 1188      2181628 2017-09-02 08:44:17 2017-09-27 14:44:46             3202
## 2528        84632 2017-09-04 11:37:34 2017-09-05 11:08:07             3196
## 2838        55482 2017-09-04 16:17:30 2017-09-05 07:42:12             3281
## 2863        50578 2017-09-04 16:35:17 2017-09-05 06:38:16             3186
## 2873        58210 2017-09-04 16:42:37 2017-09-05 08:52:48             3272
##       start.station.name end.station.id  end.station.name bikeid
## 1026       Grove St PATH           3203     Hamilton Park  26301
## 1188        Newport PATH           3217      Bayside Park  29542
## 2528      Riverview Park           3196    Riverview Park  29293
## 2838 Leonard Gordon Park           3215     JSQ Don't Use  29597
## 2863       Grove St PATH           3205 JC Medical Center  26315
## 2873        Jersey & 3rd           3186     Grove St PATH  26281
##        usertype gender  tripdur_z age
## 1026 Subscriber      2   5.487326  41
## 1188   Customer      0 172.693381  NA
## 2528   Customer      0   6.641692  NA
## 2838   Customer      2   4.333435  33
## 2863   Customer      0   3.945109  NA
## 2873   Customer      0   4.549453  NA
```

```
nrow(ab_trip)
```

```
## [1] 32
```

```
ab_age = subset(citibike, age_z > 3, select = c(1, 2, 3, 4, 5, 8, 9, 12, 13, 15, 17, 18
))
head(ab_age)
```

```
##     tripduration           starttime            stoptime start.station.id
## 338          157 2017-09-01 09:42:49 2017-09-01 09:45:27             3276
## 433          813 2017-09-01 11:35:48 2017-09-01 11:49:21             3267
## 482          261 2017-09-01 12:25:52 2017-09-01 12:30:14             3183
## 660           96 2017-09-01 15:33:05 2017-09-01 15:34:41             3214
## 695          240 2017-09-01 15:26:15 2017-09-01 15:30:15             3183
## 709          362 2017-09-01 16:36:48 2017-09-01 16:42:51             3183
##     start.station.name end.station.id end.station.name bikeid   usertype
## 338   Marin Light Rail           3186    Grove St PATH  26279 Subscriber
## 433       Morris Canal           3199     Newport Pkwy  29232 Subscriber
## 482     Exchange Place           3267     Morris Canal  29445 Subscriber
## 660   Essex Light Rail           3267     Morris Canal  26287 Subscriber
## 695     Exchange Place           3214 Essex Light Rail  26168 Subscriber
## 709     Exchange Place           3199     Newport Pkwy  29228 Subscriber
##     gender age    age_z
## 338      2  89 4.888346
## 433      1  76 3.594738
## 482      1  76 3.594738
## 660      1  76 3.594738
## 695      1  76 3.594738
## 709      2  79 3.893263
```

```
nrow(ab_age)
```

```
## [1] 119
```

```
age_80 = subset(citibike, age > 80)
nrow(age_80)
```

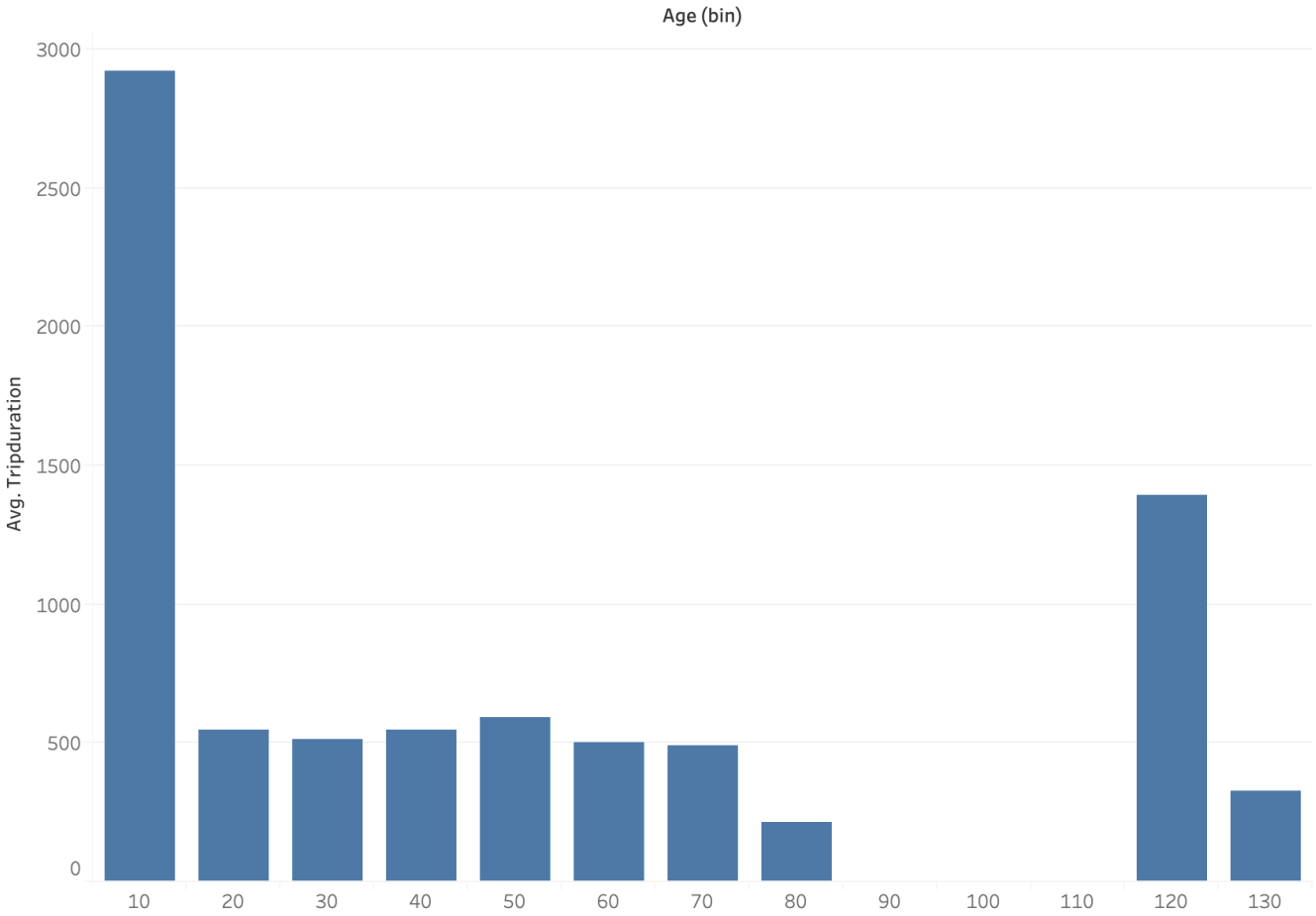
```
## [1] 37
```

There are 119 clients with age greater than 70, which contribute about 3.6% of total clients. And there are 37 clients with age greater than 80, which contribut about 1.1% of total clients. Alought senior clients does not contribute large part of clients, citibike should not ignore the corresponding policy regarding these age groups. Citibike should give safety instruction to senior clients before their rides such as wearing halmet and following speed limitation. Also, providing recommendate corresponding injuring insurance to senior clients can minimize the cost in case of injuring during rides.

```
nrow(ab_trip)
```

```
## [1] 32
```

Citibike should pay more attentaion on trips with anomalies. Specifically, there are 32 trips with during more than 12 hours. Although this is not a large amout, citibike should keep monitoring such long trips and consider about charge additional penalty fee to these trips. Having too many such long trips is harmful for the inventory.



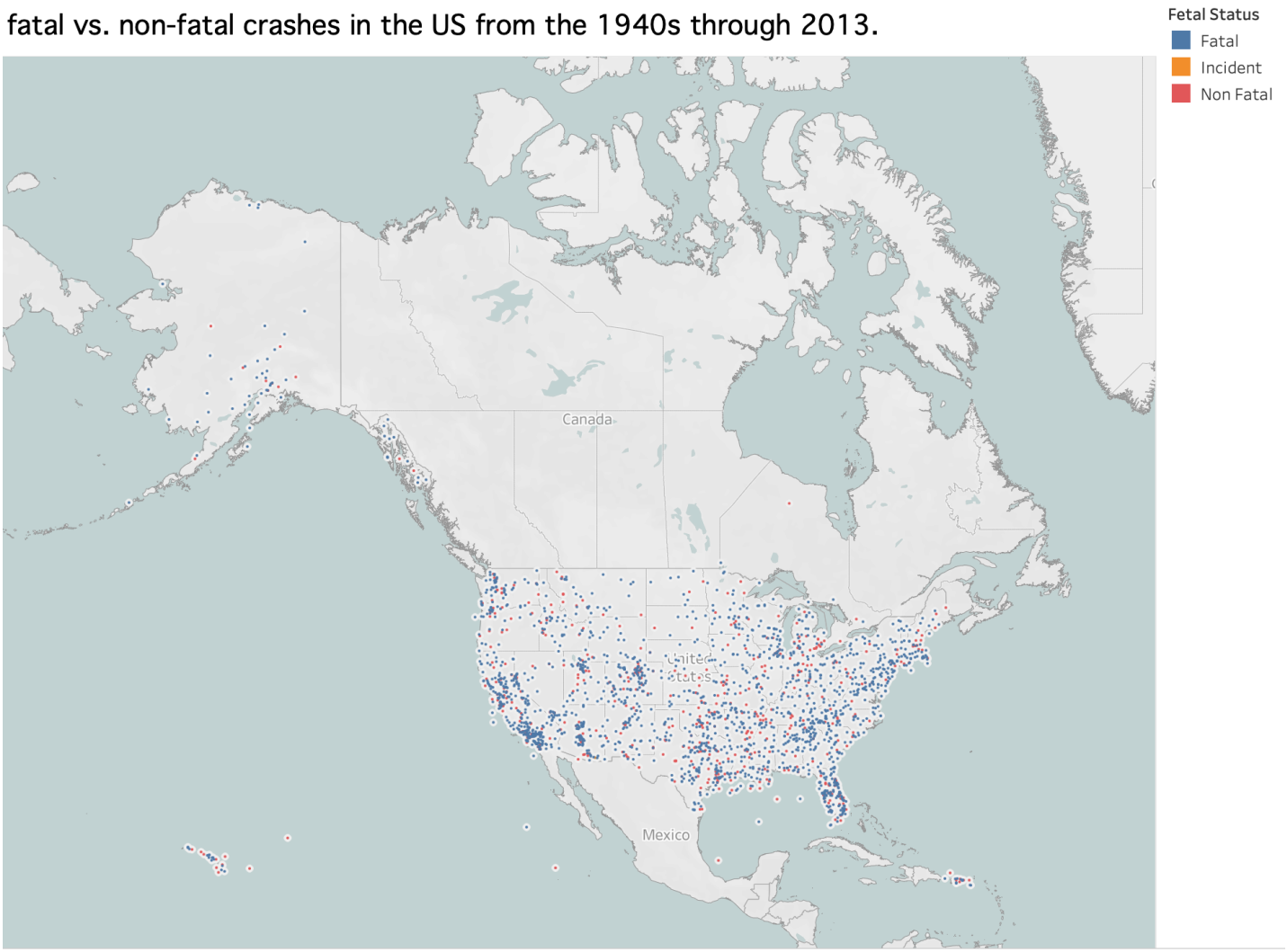
Also, the average trip duration for clients age greater or equals to 120 and age lower than 20 are extremely high, so we should consider about ignore these data because of the number of clients from these age groups only contribute a little percentage of total clients.

Problem 2: Aviation Accidents

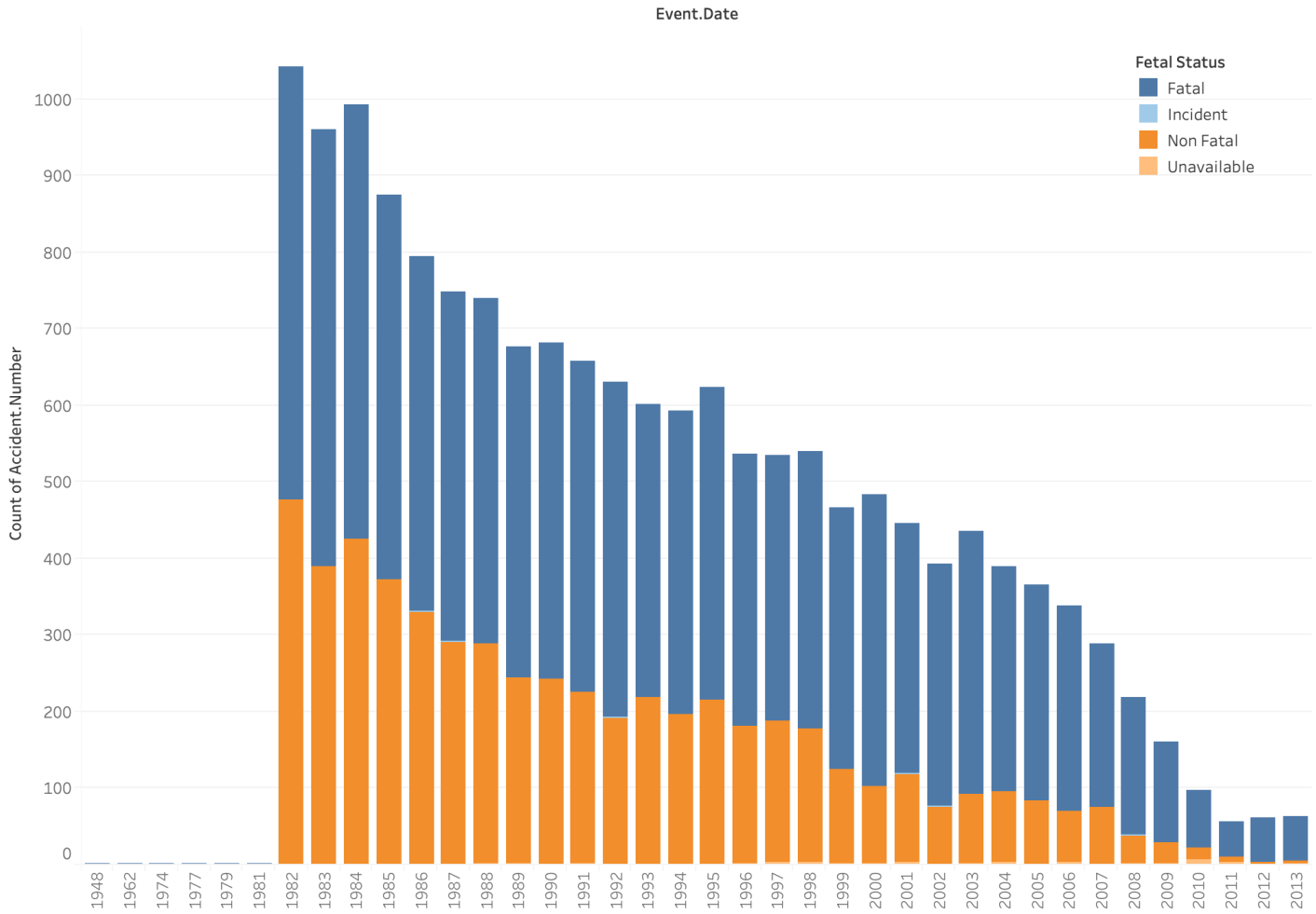
```
url2 = "https://raw.githubusercontent.com/jcbonilla/BusinessAnalytics/master/BAData/aviation.csv"
aviation = read.csv(url2)
aviation = as.data.frame(aviation)
aviation$fetal_status = gsub("[0-9]*", "", aviation$Injury.Severity)
aviation$fetal_status = gsub("[[:punct:]]", " ", aviation$fetal_status)

aviation$Date_Year = substring(aviation$Event.Date, 8, 11)
aviation$Date_Year = as.Date(aviation$Date_Year, "%Y")
write.csv(aviation, "/Users/zianzhang/Desktop/nyu/business_analytics/HW3/aviation_rev.csv")
```

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



fatal vs. non-fatal crashes in the US



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

Problem 3: Retail Targets