Summary Stats on STAT 440 Kaggle Module 1

Chris Sobczak

Sat Sept 12th 2020

Contents

Dependancies

```
library(tidyverse)
library(xtable)
```

Importing Data

```
base = read.delim(file = 'baseline.txt', header = T, sep = ',')
summary(base)
```

```
##        Id            duration
##  Min.   :  1.00   Min.   :3.94
##  1st Qu.: 50.75   1st Qu.:3.94
##  Median :100.50   Median :3.94
##  Mean   :100.50   Mean   :3.94
##  3rd Qu.:150.25   3rd Qu.:3.94
##  Max.   :200.00   Max.   :3.94
```

```
str(base)
```

```
## ’data.frame’:    200 obs. of  2 variables:
##  $ Id      : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ duration: num  3.94 3.94 3.94 3.94 3.94 3.94 3.94 3.94 3.94 3.94 ...
```

Not sure what the baseline text file is for.

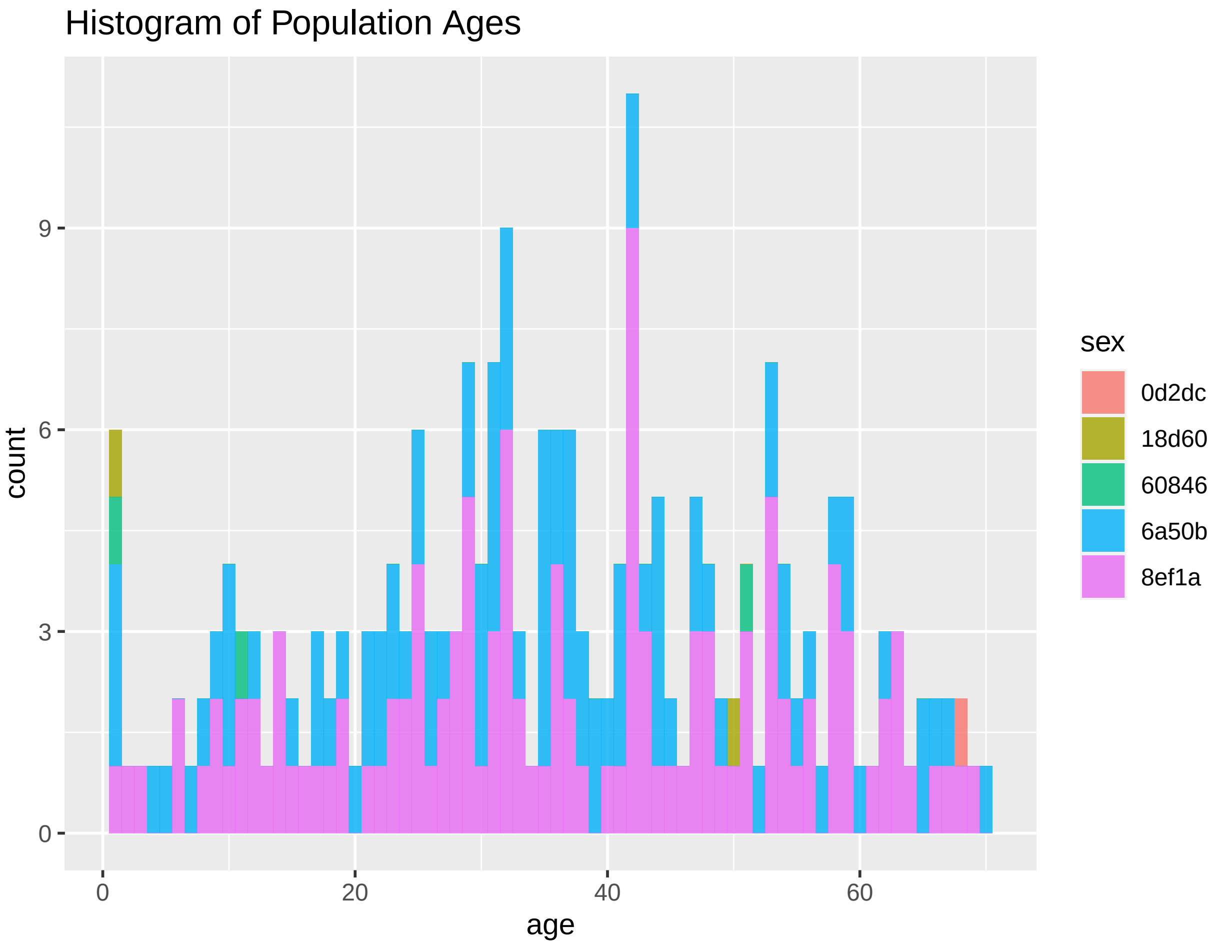
```r
train = read.delim(file = 'train.txt', header = T, sep = ',')
test = read.delim(file = 'test.txt', header = T, sep = ',')
train = mutate( .data = train,
        confirmed   =   as.Date(x = confirmed, format = "%m.%d.%Y"),
        symptoms    =   as.factor(symptoms),
        outcome     =   as.factor(outcome),
        sex         =   as.factor(sex),
        city        =   as.factor(city),
        province    =   as.factor(province),
        country     =   as.factor(country),
        age         =   as.numeric(as.factor(age))
)
test = mutate(  test,
        confirmed   =   as.Date(x = confirmed, format = "%m.%d.%Y"),
        symptoms    =   as.factor(symptoms),
        sex         =   as.factor(sex),
        city        =   as.factor(city),
        province    =   as.factor(province),
        country     =   as.factor(country),
        age         =   as.numeric(as.factor(age))
)
summary(train)
```

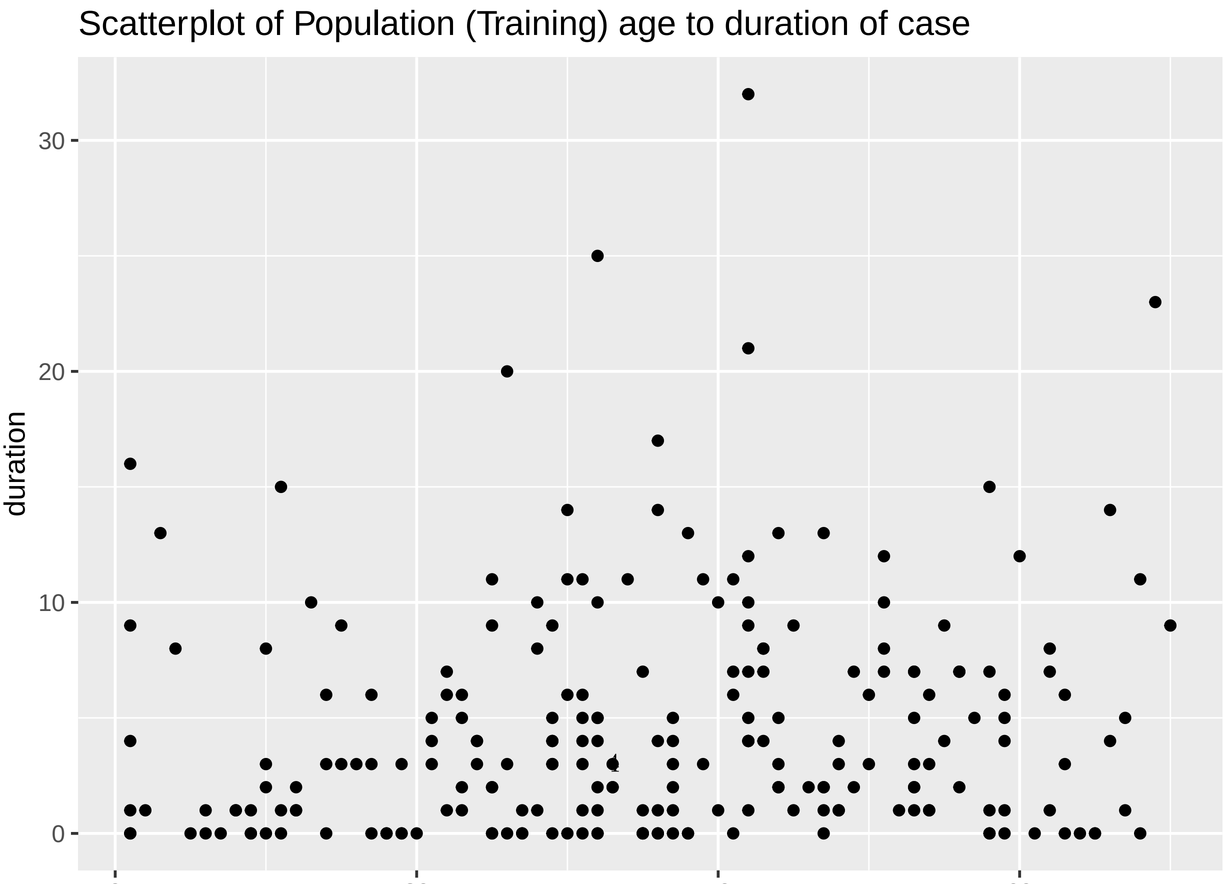
```
##       age          sex           city        province      country          V1              confirmed
##  Min.   : 1.0   0d2dc:  1   98444  : 36   56888  : 32   55b87  :130   Length:219         Min.   :2020
##  1st Qu.:24.0   18d60:  2   de0cb  : 14   0acea  : 24   2d394  : 32   Class :character   1st Qu.:2020
##  Median :36.0   60846:  3   6498e  : 12   5dfd3  : 19   867cc  : 27   Mode  :character   Median :2020
##  Mean   :35.6   6a50b: 92   bff0e  : 10   0b8ad  : 15   94e65  :  9                      Mean   :2020
##  3rd Qu.:48.5   8ef1a:121   5b9f8  :  6   e9491  : 15   8ce1f  :  7                      3rd Qu.:2020
##  Max.   :70.0               3dd07  :  5   d6364  : 13   348dd  :  3                      Max.   :2020
##                             (Other):136   (Other):101   (Other): 11                      NA’s   :146
##     duration
##  Min.   : 0.000
##  1st Qu.: 1.000
##  Median : 3.000
##  Mean   : 4.584
##  3rd Qu.: 7.000
##  Max.   :32.000
##
```

```r
summary(test)
```

```
##        Id              age            sex           city        province     country          V1
##  Min.   :  1.00   Min.   : 1.00   0d2dc:  2   98444  : 31   56888  :35   55b87  :122   Length:200
##  1st Qu.: 50.75   1st Qu.:18.00   18d60:  1   6498e  : 12   0b8ad  :22   867cc  : 32   Class :charac
##  Median :100.50   Median :31.00   60846:  1   3dd07  :  9   5dfd3  :22   2d394  : 23   Mode  :charac
##  Mean   :100.50   Mean   :31.89   6a50b: 81   de0cb  :  9   0acea  :21   94e65  :  8
##  3rd Qu.:150.25   3rd Qu.:44.00   8ef1a:115   bff0e  :  8   e9491  :20   8ce1f  :  4
##  Max.   :200.00   Max.   :65.00               533b8  :  6   d6364  :12   348dd  :  3
##                                               (Other):125   (Other):68   (Other):  8
##     duration
##  Min.   : 0.00
##  1st Qu.: 1.00
##  Median : 3.00
##  Mean   : 3.94
##  3rd Qu.: 6.00
##  Max.   :26.00
##
```

Some Plots


Histogram of Population Ages
count
9
6
3
0
0
20
40
60
age
sex
0d2dc
18d60
60846
6a50b
8ef1a


Scatterplot of Population (Training) age to duration of case
duration
30
20
10
0