**Problem 1**

Dimension: 866,117 observations in 15 variables

> as.data.frame(table(data$start.station.id))

Var1 Freq

…

147 362 2972

…

> as.data.frame(table(data$start.station.id[data$end.station.id==152]))

Var1 Freq

147 438 3

**Problem 2**

> data2 <- subset(data, data$birth.year !="" & data$gender !=0)

758,458 rows left

> as.data.frame(table(bike$gender))

Var1 Freq

1 1 583648

2 2 174810

**Problem 3**

> Mode <- function(x) {

+ ux <- unique(x)

+ ux[which.max(tabulate(match(x, ux)))]

+ }

> Mode(bike$bikeid)

[1] 20758

> as.data.frame(table(bike$bikeid[bike$bikeid == 20758]))

Var1 Freq

1 20758 276

**Problem 4**

> DurationCalculator <- function(dataset,sex){

+ median(dataset$tripduration[dataset$gender == sex])

+ }

> DurationCalculator(dataset=bike, sex=2)

[1] 694

**Problem 5**

(oldies = M1, boomers = M2, millenials = M3)

> DurationCalculator(dataset=oldies, sex=1)

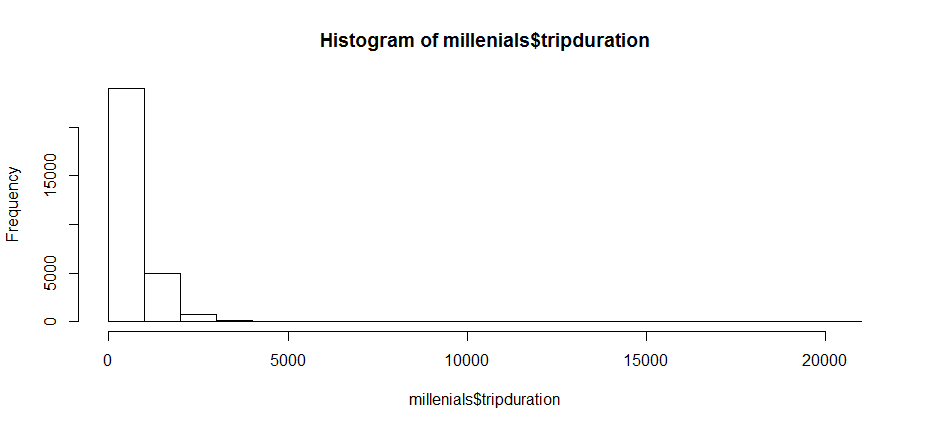
[1] 621

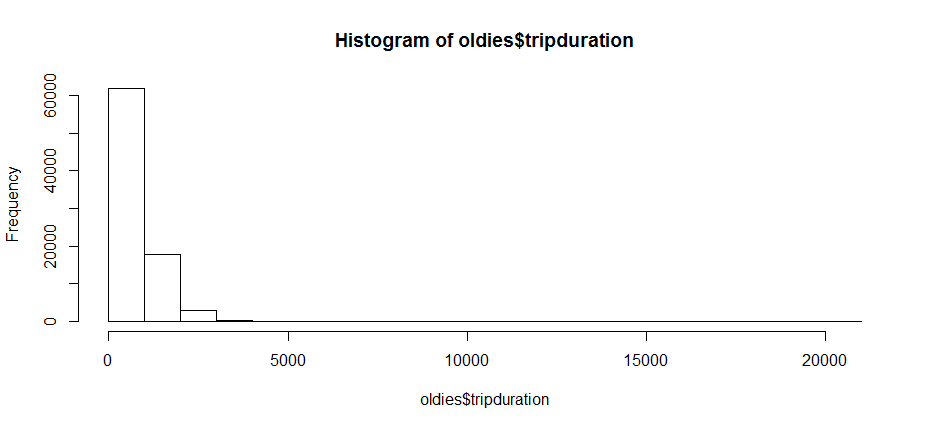
> DurationCalculator(dataset=boomers, sex=1)

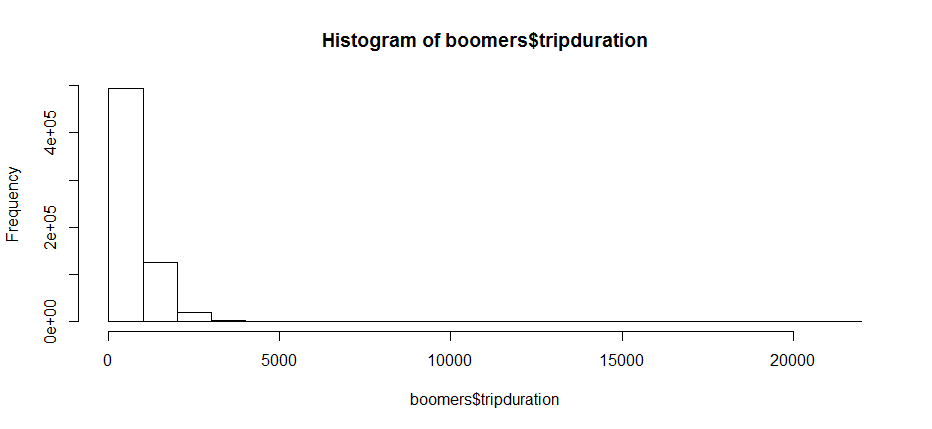
[1] 577

> DurationCalculator(dataset=millenials, sex=1)

[1] 520

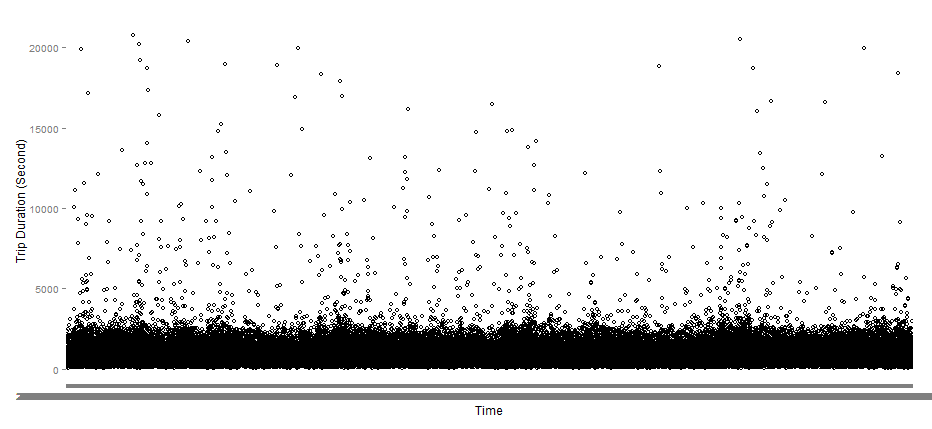
****

****

****

**Problem 6**

> ggplot(oldies, aes(x=oldies$starttime, y=oldies$tripduration)) + geom\_point(shape=1) + xlab('Time') + ylab('Trip Duration (Second)')



> ggplot(boomers, aes(x=boomers$starttime, y=boomers$tripduration)) + geom\_point(shape=1) + xlab('Time') + ylab('Trip Duration (Second)')

> ggplot(millenals, aes(x=millenials$starttime, y=millenials$tripduration)) + geom\_point(shape=1) + xlab('Time') + ylab('Trip Duration (Second)')

Getting memory allocation issues

Error: cannot allocate vector of size 2.0 Mb

In addition: Warning messages:

1: In validGrob.grob(g) :

Reached total allocation of 4010Mb: see help(memory.size)

2: In validGrob.grob(g) :

Reached total allocation of 4010Mb: see help(memory.size)

3: In validGrob.grob(g) :

Reached total allocation of 4010Mb: see help(memory.size)

4: In validGrob.grob(g) :

Reached total allocation of 4010Mb: see help(memory.size)

**Problem 7**

> bike$group <- "M1"

> bike$group[as.numeric(as.character(bike$birth.year)) > 1960] <- "M2"

> bike$group[as.numeric(as.character(bike$birth.year)) > 1990] <- "M3"

> View(bike)

> bike$tripdurationmean[bike$group == "M1"] <- mean(oldies$tripduration)

> View(bike)

> bike$tripdurationmean[bike$group == "M2"] <- mean(boomers$tripduration)

> bike$tripdurationmean[bike$group == "M3"] <- mean(millenials$tripduration)

First checking homoskedacity:

> bartlett.test(bike$tripduration,bike$group)

Bartlett test of homogeneity of variances

data: bike$tripduration and bike$group

Bartlett's K-squared = 72.2664, df = 2, p-value < 2.2e-16

> qchisq(0.950, 3)

[1] 7.814728

Chi-squared < Barlett K-squared: reject the null hypothesis of homogeneity. This isn’t good but I’m not 100% on the chi-squared threshold. I believe qchisq(.95,3) returns the test statistic for a 95% confidence test of three variables. If that interpretation is correct, our assumption for anova is violated. At any rate, I conducted the mechanical analysis as follows:

> fit = lm(formula = bike$tripduration ~ bike$group)

> anova(fit)

Analysis of Variance Table

Response: bike$tripduration

Df Sum Sq Mean Sq F value Pr(>F)

bike$group 2 2.5391e+08 126957426 221.74 < 2.2e-16 \*\*\*

Residuals 758455 4.3426e+11 572560

---

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

Since P <<<<<< .05 we can reject the hypothesis that the means are equal (assuming our test assumptions weren’t violated without heteroskedacity).