Assignment 2

**Part 1**

**Survival Analysis with Plot ly using (Kidney Dataset)**

**URL -** [**https://plot.ly/~agarwal.ro/**](https://plot.ly/~agarwal.ro/)

**Description –**

Data on the recurrence times to infection, at the point of insertion of the catheter, for kidney patients using portable dialysis equipment. Catheters may be removed for reasons other than infection, in which case the observation is censored. Each patient has exactly 2 observations. This data has often been used to illustrate the use of random effects (frailty) in a survival model. However, one of the males (id 21) is a large outlier, with much longer survival than his peers. If this observation is removed no evidence remains for a random subject effect.

**Regression with frailty in Male**

This graph explains the frailty ratio w.r.t men in kidney dataset which is obtained by taking

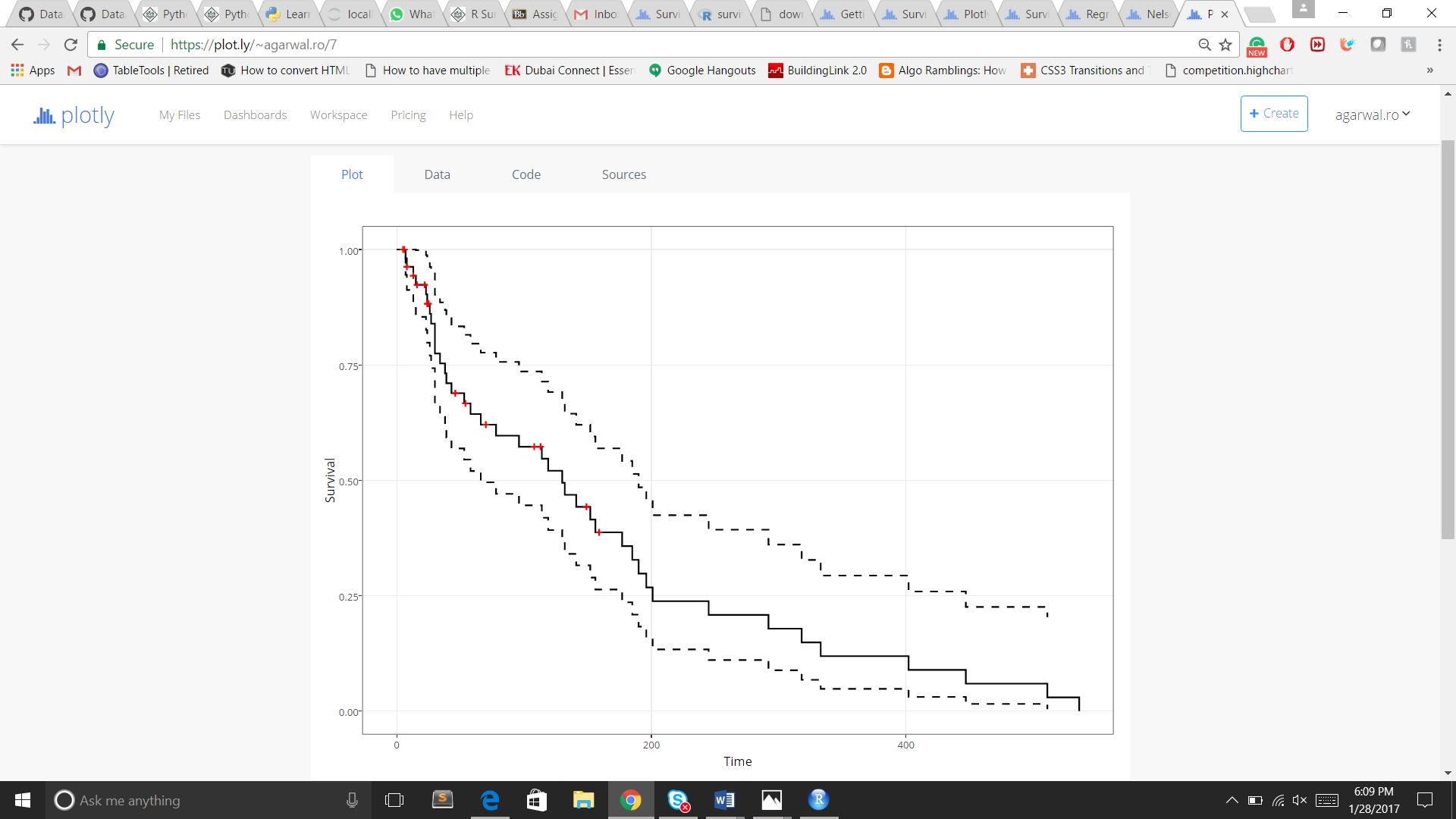
1. Time
2. Sex as 1 (male)
3. Status

Formula (kidney.surv <- Surv(time[sex==1], status[sex==1]))

Age is also taken into consider while analyzing.



**Regression with frailty in Female**



Frailty Regression Kaplan-Meier estimate for both Male and Female

This is the Regression graph which is calculated by using Kaplan-Meier formula

kidney.surv <- Surv(time[sex==2], status[sex==2])

kidney.surv

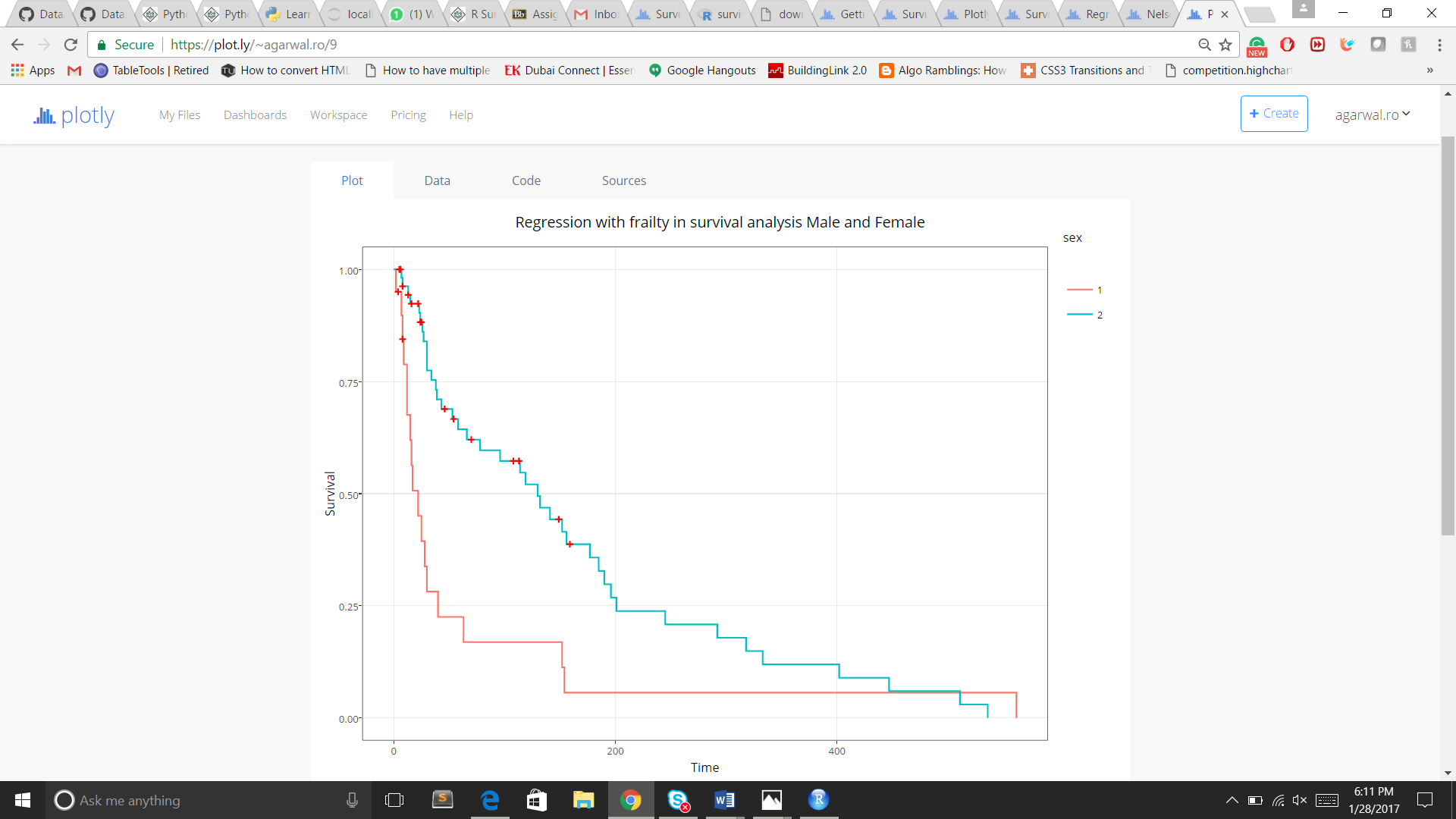
surv.fit <- survfit(kidney.surv~1)

surv.fit

summary(surv.fit)

A <-plot(surv.fit, main='Kaplan-Meier estimate with 95% confidence bounds',

xlab='time', ylab='Regression with frailty IN Female')



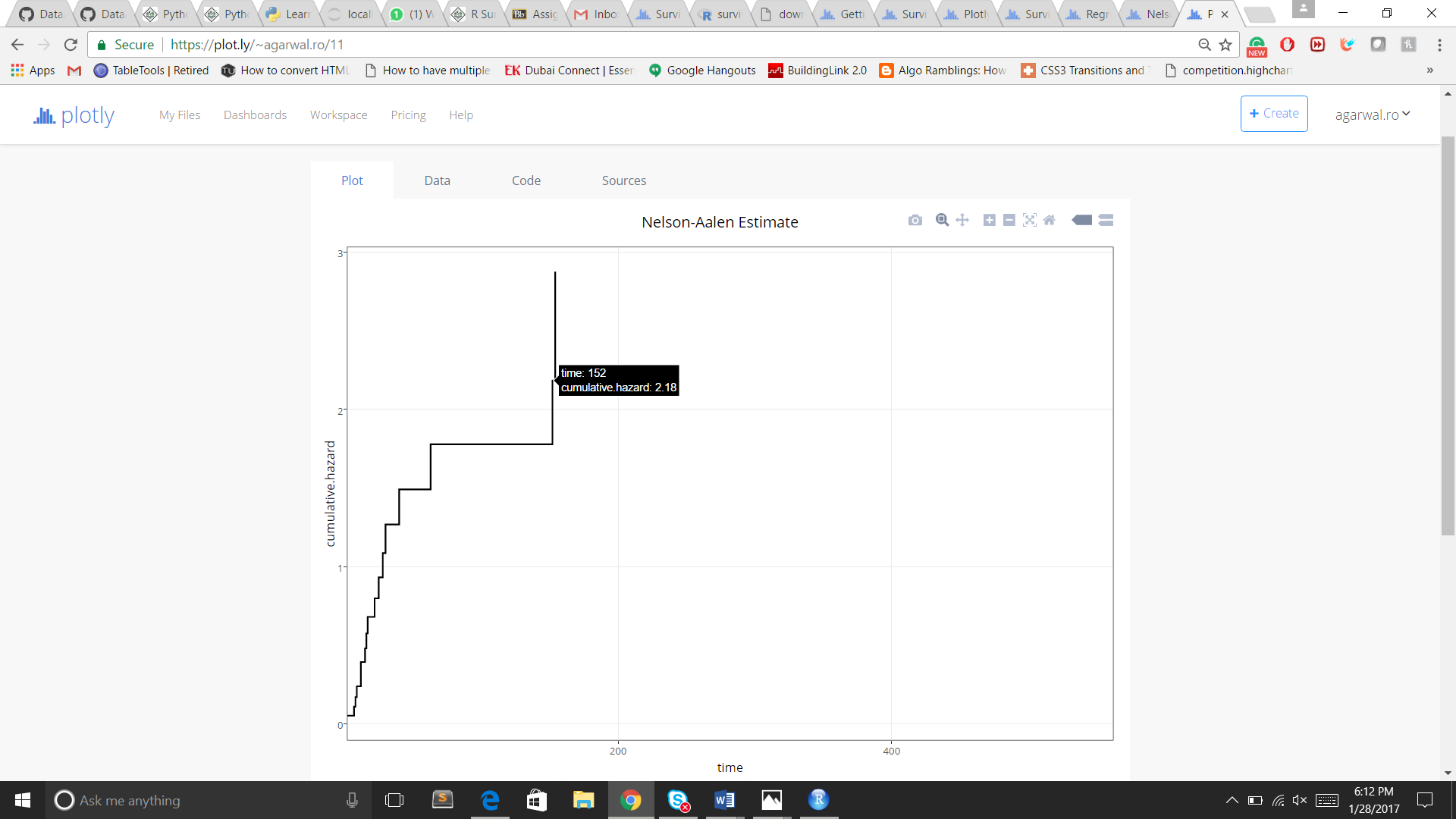
Nelson-Aalen Estimate for Male

This graph predicts the hazard occur and calculated using cumulative hazard function using the Nelson-Aalen estimator, defined as:

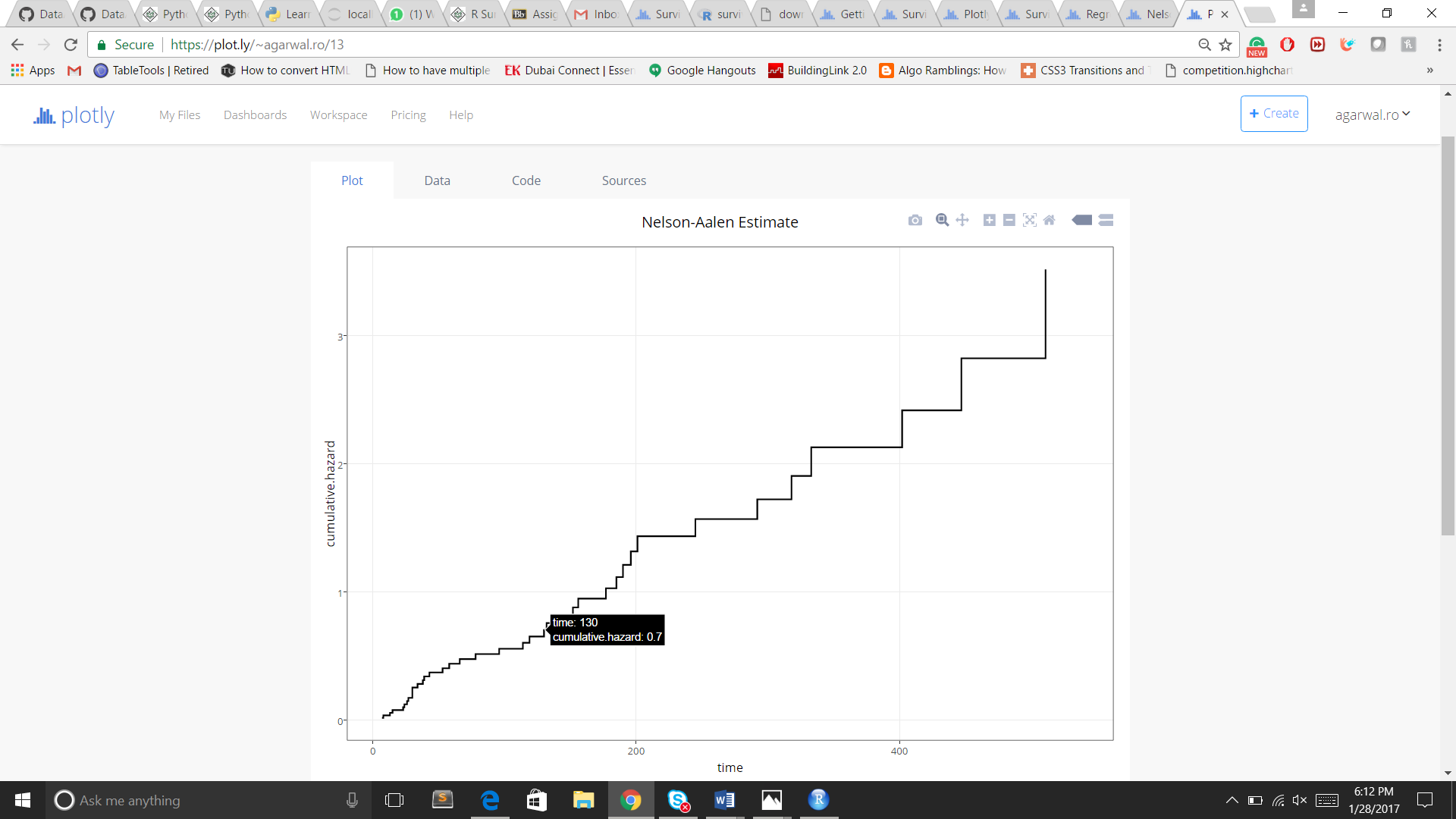
Λ^(t)=∑ti≤tdini

Λ^(t)=∑ti≤tdini

where di is the number of deaths at time ti and ni is the number of susceptible individuals. Both R and Python modules use the same estimator. However, in R we will use the -log of the Fleming and Harrington estimator, which is equivalent to the Nelson-Aalen.



Nelson-Aalen Estimate for Female



All these Analysis is done for finding cancer rate in Male and Female

Other formulas that can be used in obtaining this analysis are -:

kfit <- coxph(Surv(time, status)~ age + sex + disease + frailty(id), kidney)

kfit0 <- coxph(Surv(time, status)~ age + sex + disease, kidney)

kfitm1 <- coxph(Surv(time,status) ~ age + sex + disease +

frailty(id, dist='gauss'), kidney)

**Part 2**

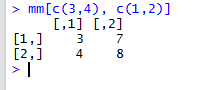
R Lab Code

a. 0-Intro.R

Exercise: select 2x2 subsection from the "bottom left" of matrix mm

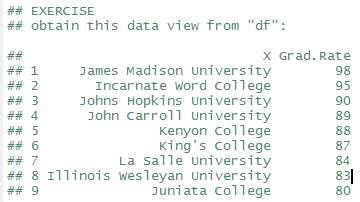
Solution: mm[c(3,4), c(1,2)]

O/P



b. Data.R

Exercise :



Solution -:

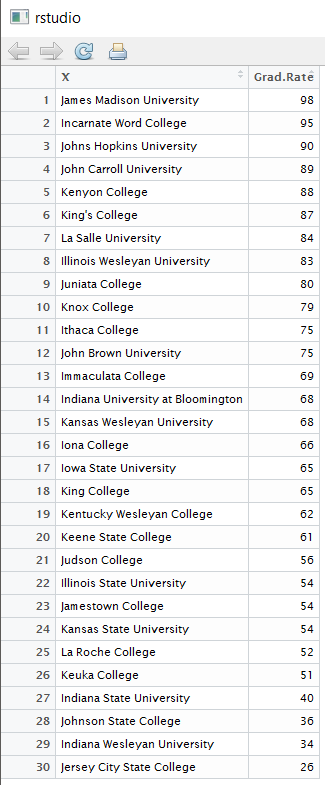
## [your code here] #this will give you all grad rate in descending order

data <- select(df, X, Grad.Rate)

result <- arrange(data, desc(Grad.Rate))

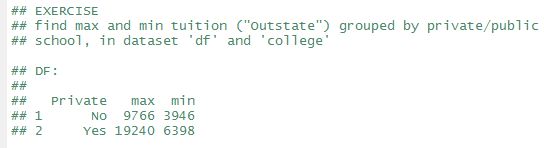
View(result)

O/P -:



Exercise :

DF



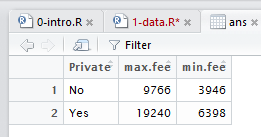
Solution -:

**ans <- group\_by(df, Private)**

**ans <- summarise(ans, max.fee= max(Outstate), min.fee = min(Outstate))**

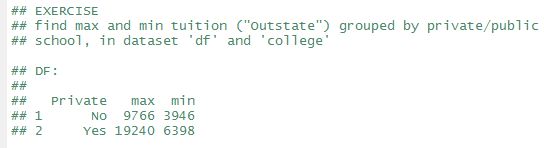
**View(ans)**

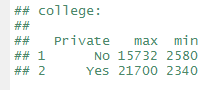
**O/P:**



Exercise :

College





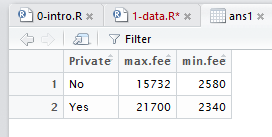
Solution -:

**ans1 <- group\_by(college, Private)**

**ans1 <- summarise(ans1, max.fee= max(Outstate), min.fee = min(Outstate))**

**View(ans1)**

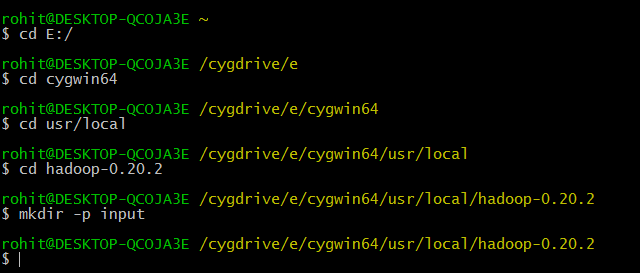
**O/P:**



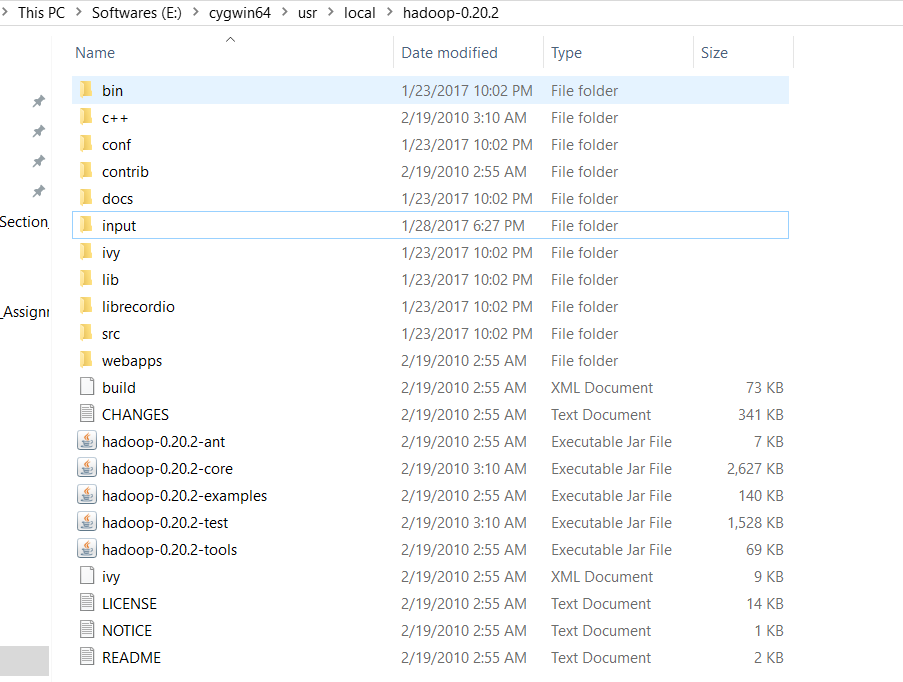
**Part 3**

**Hadoop Lab 1**

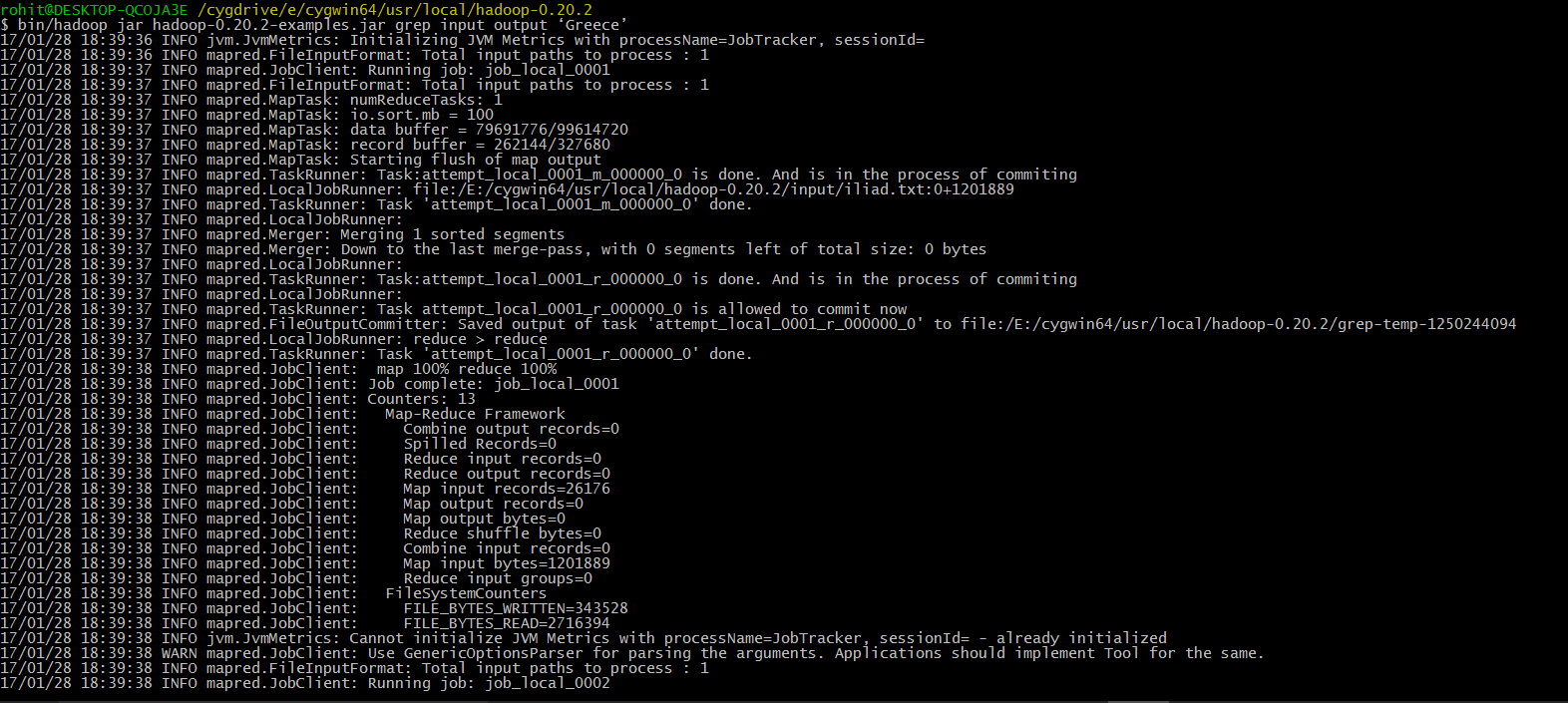
1. Command to write



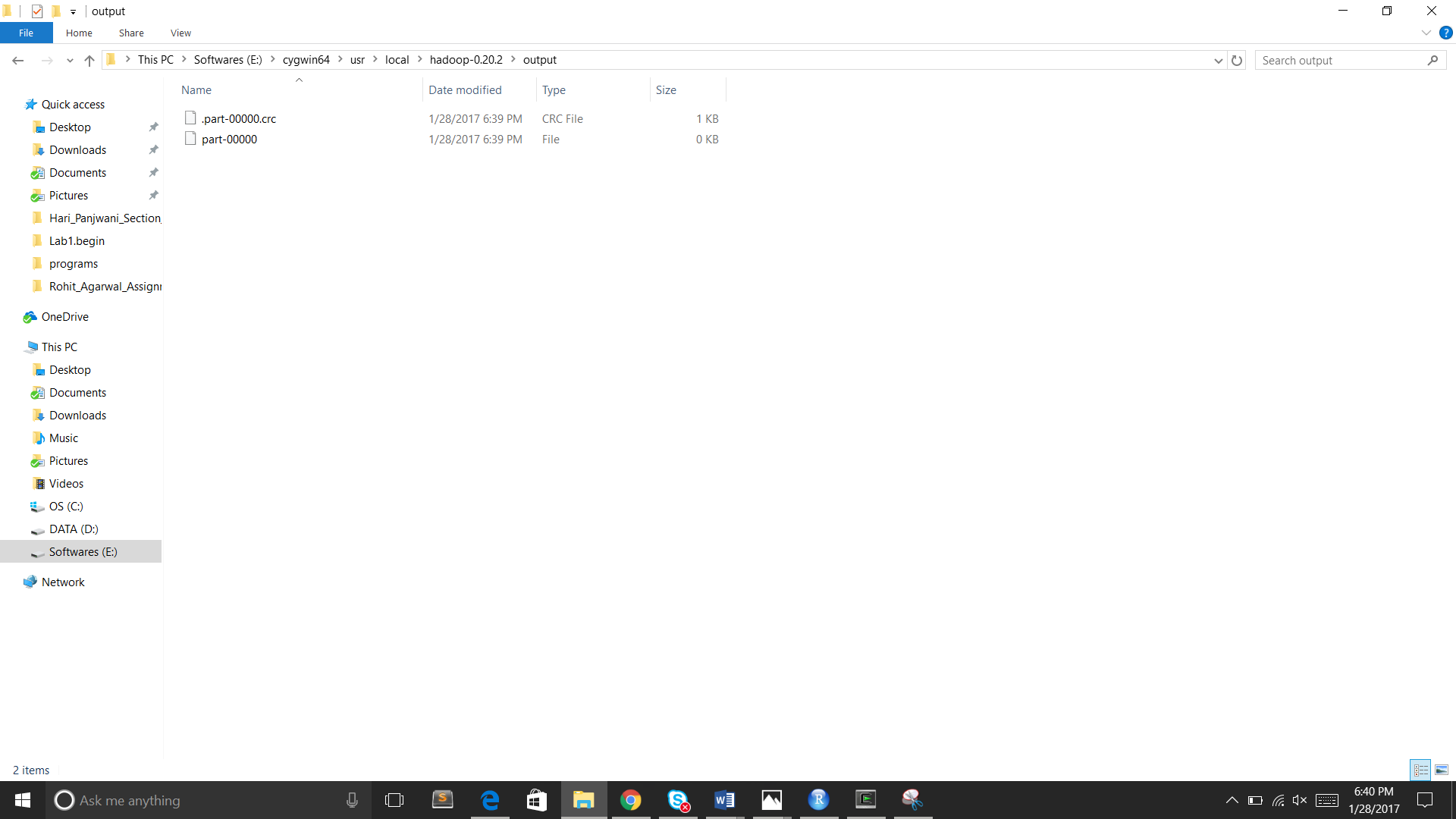
1. Output



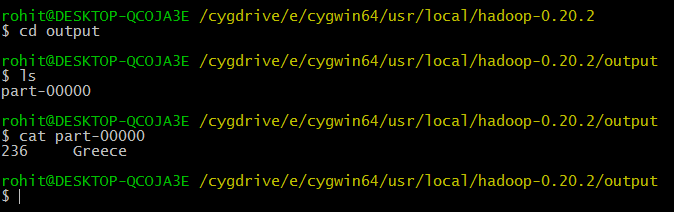
1. Now add illiad.txt file inside this folder.
2. After running pre define library prep



1. Output



1. We can see output folder with 2 files created.
2. If we open the part-00000 file will see number of occurrence of CAT word.



1. It shows Greece word is occurred 236 times in illiad.txt.