Olen Justice

CSC 302

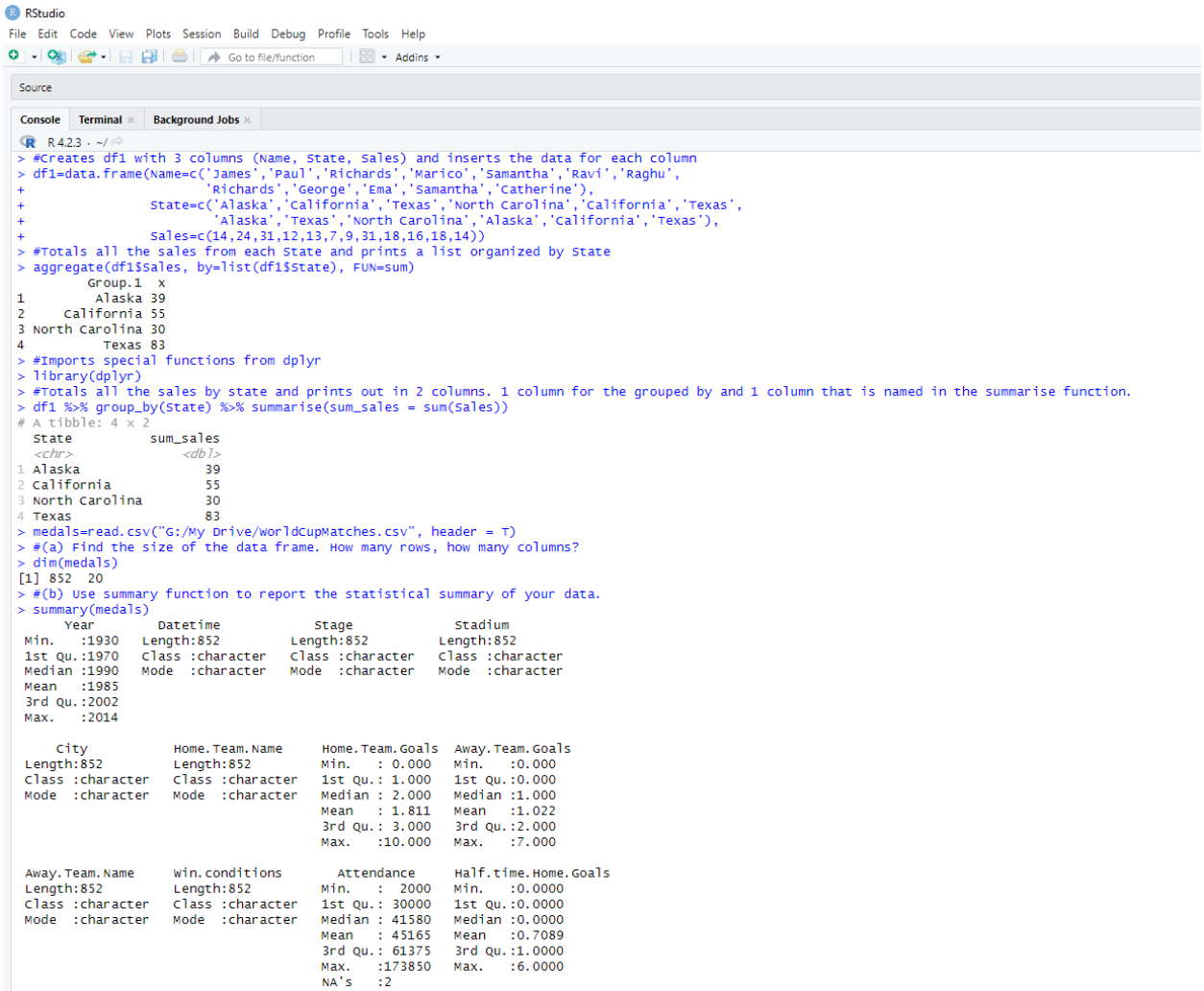
HW2

3/26/2023

Question 1). The beginning of the code creates 3 columns (Name, State, Sales) and populates the data rows for each column. The aggregate line of code sums the sales for each state and prints the list by state. The rest of the code imports dplyr special functions and then sums the sales for each state and prints a list with 2 labeled columns.

Question 2a). 852 rows and 20 columns

Question 2f). It appears the attendance trended upwards until the mid to late 60's where it has leveled off with little fluctuation since.



```
RStudio
File  Edit  Code  View  Plots  Session  Build  Debug  Profile  Tools  Help
Go to file/function   Addins
Source
Console  Terminal  Background Jobs
R 4.2.3 · ~/
> #Creates df1 with 3 columns (Name, State, Sales) and inserts the data for each column
> df1=data.frame(Name=c('James','Paul','Richards','Marico','Samantha','Ravi','Raghu',
+                       'Richards','George','Ema','Samantha','Catherine'),
+                State=c('Alaska','California','Texas','North Carolina','California','Texas',
+                        'Alaska','Texas','North Carolina','Alaska','California','Texas'),
+                Sales=c(14,24,31,12,13,7,9,31,18,16,18,14))
> #Totals all the sales from each State and prints a list organized by State
> aggregate(df1$Sales, by=list(df1$State), FUN=sum)
        Group.1  x
1        Alaska 39
2    California 55
3 North Carolina 30
4         Texas 83
> #Imports special functions from dplyr
> library(dplyr)
> #Totals all the sales by state and prints out in 2 columns. 1 column for the grouped by and 1 column that is named in the summarise function.
> df1 %>% group_by(State) %>% summarise(sum_sales = sum(Sales))
# A tibble: 4 × 2
  State          sum_sales
  <chr>              <dbl>
1 Alaska                39
2 California            55
3 North Carolina        30
4 Texas                 83
> medals=read.csv("G:/My Drive/WorldCupMatches.csv", header = T)
> #(a) Find the size of the data frame. How many rows, how many columns?
> dim(medals)
[1] 852  20
> #(b) Use summary function to report the statistical summary of your data.
> summary(medals)
      Year        Datetime            Stage             Stadium
 Min.   :1930   Length:852         Length:852         Length:852
 1st Qu.:1970   Class :character   Class :character   Class :character
 Median :1990   Mode  :character   Mode  :character   Mode  :character
 Mean   :1985
 3rd Qu.:2002
 Max.   :2014

     City           Home.Team.Name     Home.Team.Goals  Away.Team.Goals
 Length:852         Length:852         Min.   : 0.000   Min.   :0.000
 Class :character   Class :character   1st Qu.: 1.000   1st Qu.:0.000
 Mode  :character   Mode  :character   Median : 2.000   Median :1.000
                                       Mean   : 1.811   Mean   :1.022
                                       3rd Qu.: 3.000   3rd Qu.:2.000
                                       Max.   :10.000   Max.   :7.000

 Away.Team.Name     Win.conditions       Attendance     Half.time.Home.Goals
 Length:852         Length:852         Min.   :  2000   Min.   :0.0000
 Class :character   Class :character   1st Qu.: 30000   1st Qu.:0.0000
 Mode  :character   Mode  :character   Median : 41580   Median :0.0000
                                       Mean   : 45165   Mean   :0.7089
                                       3rd Qu.: 61375   3rd Qu.:1.0000
                                       Max.   :173850   Max.   :6.0000
                                       NA's   :2
```

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins

Source

Console | Terminal | Background Jobs

R 4.2.3 · ~/

```
 Half.time.Away.Goals   Referee          Assistant.1        Assistant.2
 Min.   :0.0000       Length:852         Length:852         Length:852
 1st Qu.:0.0000       Class :character   Class :character   Class :character
 Median :0.0000       Mode  :character   Mode  :character   Mode  :character
 Mean   :0.4284
 3rd Qu.:1.0000
 Max.   :5.0000

    RoundID            MatchID          Home.Team.Initials Away.Team.Initials
 Min.   :     201   Min.   :       25   Length:852         Length:852
 1st Qu.:     262   1st Qu.:     1189   Class :character   Class :character
 Median :     337   Median :     2191   Mode  :character   Mode  :character
 Mean   :10661773   Mean   : 61346868
 3rd Qu.:  249722   3rd Qu.: 43950059
 Max.   :97410600   Max.   :300186515

> #(c) Find how many unique locations olympics were held at.
> length(unique(medals$City))
[1] 151
> #(d) Find the average attendance.
> mean(medals$Attendance, na.rm = T)
[1] 45164.8
> #(e) For each Home Team, what is the total number of goals scored? (Hint: Please refer to question 1)
> library(dplyr)
> medals %>%  group_by(Home.Team.Name) %>% summarise(Goal.Totals = sum(Home.Team.Goals))
# A tibble: 78 x 2
   Home.Team.Name Goal.Totals
   <chr>                <int>
 1 Algeria                  5
 2 Angola                   0
 3 Argentina              111
 4 Australia                7
 5 Austria                 31
 6 Belgium                 27
 7 Bolivia                  1
 8 Brazil                 180
 9 Bulgaria                11
10 Cameroon                11
# i 68 more rows
# i Use `print(n = ...)` to see more rows
> #(f) What is the average number of attendees for each year?
> aggregate(medals$Attendance, by=list(medals$Year), FUN=mean)
   Group.1        x
1     1930 32808.28
2     1934 21352.94
3     1938 20872.22
4     1950 47511.18
5     1954 29561.81
6     1958 23423.14
7     1962 27911.62
8     1966 48847.97
9     1970 50124.22
10    1974 49098.76
11    1978 40678.71
12    1982 40571.60
```

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

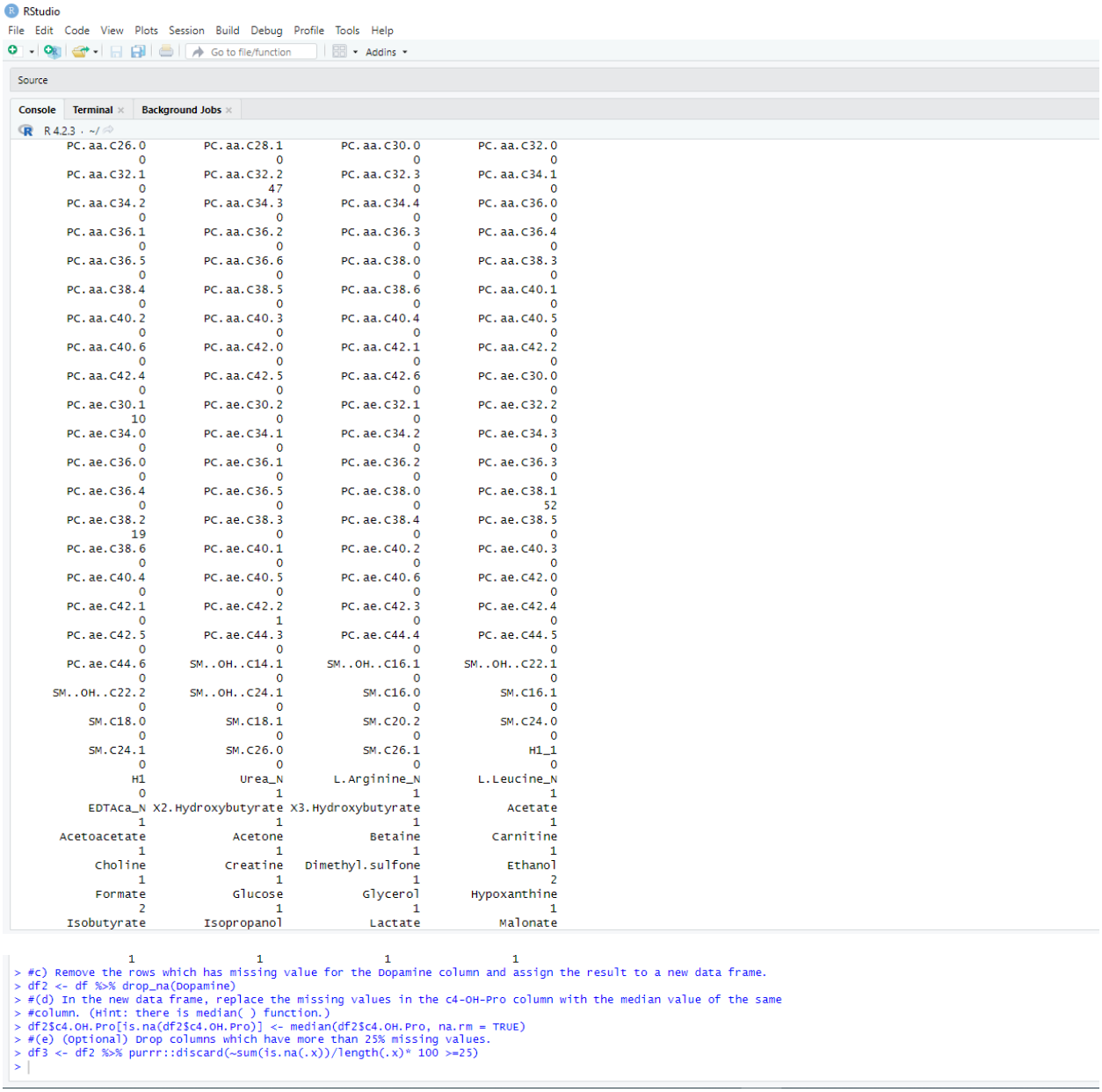
Go to file/function | Addins

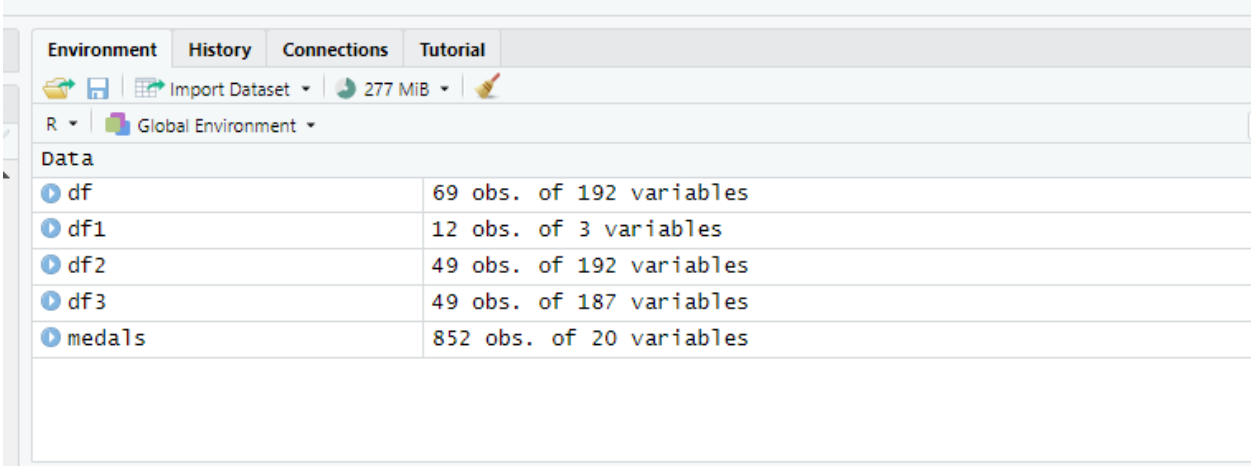
Source

Console | Terminal | Background Jobs

R 4.2.3 · ~/

```
13    1986 46039.06
14    1990 48388.75
15    1994 68991.12
16    1998 43517.19
17    2002 42268.70
18    2006 52491.23
19    2010 49669.62
20    2014       NA
> #3. Use R to read the metabolites.csv from the DATA folder on Google Drive. Then perform the followings (32 points):
> df=read.csv("G:/My Drive/metabolite.csv", header = T)
> #(a) Find how many Alzheimers patients there are in the data set. (Hint: Please refer to question 1)
> sum(df$Label == "Alzheimer")
[1] 35
> #(b) Determine the number of missing values for each column. (Hint: is.na( )
> colSums(is.na(df))
             Label                Phe                Pro                Ser
                 0                  0                  0                  0
               Thr               ADMA          alpha.AAA          c4.OH.Pro
                 0                  0                  0                 20
         Carnosine         Creatinine               DOPA           Dopamine
                 1                  0                  0                 20
         Histamine         Kynurenine            Met.SO          Nitro.Tyr
                 0                  0                  1                 62
               PEA         Putrescine          Sarcosine          Serotonin
                69                  0                  0                  0
        Spermidine           Spermine          t4.OH.Pro            Taurine
                 0                 60                  0                  2
              SDMA                 C0                C10              C10.1
                 0                  0                  0                  0
            C10.2                C12              C12.DC              C12.1
                 0                  0                  1                  0
              C14              C14.1           C14.1.OH              C14.2
                 0                  0                  1                  0
          C14.2.OH              C16.           C16.OH              C16.1
                 2                  0                  1                  0
          C16.1.OH            C16.2.           C16.2.OH                C18
                 2                  2                  1                  0
            C18.1          C18.1.OH.             C18.2                 C2
                 0                  7                  0                  0
               C3             C3.OH.              C3.1                 C4
                 0                  8                  2                  0
    C3.DC..C4.OH.              C4.1.                C5           C5.M.DC
                 0                  0                  0                  1
  C5.OH..C3.DC.M.              C5.1.           C5.1.DC       C6..C4.1.DC.
                 0                  5                  2                  0
    C5.DC..C6.OH.              C6.1.             C7.DC                 C8
                 4                  2                  1                  0
               C9     lysoPC.a.C14.0     lysoPC.a.C16.0     lysoPC.a.C16.1
                 1                  0                  0                  0
   lysoPC.a.C17.0     lysoPC.a.C18.0     lysoPC.a.C18.1     lysoPC.a.C18.2
                 0                  0                  0                  0
   lysoPC.a.C20.3     lysoPC.a.C20.4     lysoPC.a.C24.0     lysoPC.a.C26.0
                 0                  0                  0                  0
   lysoPC.a.C26.1     lysoPC.a.C28.0     lysoPC.a.C28.1        PC.aa.C24.0
                 0                  0                  0                  0
```


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Source
Console
Terminal
Background Jobs
R 4.2.3 · ~/
PC.aa.C26.0 PC.aa.C28.1 PC.aa.C30.0 PC.aa.C32.0
0 0 0 0
PC.aa.C32.1 PC.aa.C32.2 PC.aa.C32.3 PC.aa.C34.1
0 47 0 0
PC.aa.C34.2 PC.aa.C34.3 PC.aa.C34.4 PC.aa.C36.0
0 0 0 0
PC.aa.C36.1 PC.aa.C36.2 PC.aa.C36.3 PC.aa.C36.4
0 0 0 0
PC.aa.C36.5 PC.aa.C36.6 PC.aa.C38.0 PC.aa.C38.3
0 0 0 0
PC.aa.C38.4 PC.aa.C38.5 PC.aa.C38.6 PC.aa.C40.1
0 0 0 0
PC.aa.C40.2 PC.aa.C40.3 PC.aa.C40.4 PC.aa.C40.5
0 0 0 0
PC.aa.C40.6 PC.aa.C42.0 PC.aa.C42.1 PC.aa.C42.2
0 0 0 0
PC.aa.C42.4 PC.aa.C42.5 PC.aa.C42.6 PC.ae.C30.0
0 0 0 0
PC.ae.C30.1 PC.ae.C30.2 PC.ae.C32.1 PC.ae.C32.2
10 0 0 0
PC.ae.C34.0 PC.ae.C34.1 PC.ae.C34.2 PC.ae.C34.3
0 0 0 0
PC.ae.C36.0 PC.ae.C36.1 PC.ae.C36.2 PC.ae.C36.3
0 0 0 0
PC.ae.C36.4 PC.ae.C36.5 PC.ae.C38.0 PC.ae.C38.1
0 0 0 52
PC.ae.C38.2 PC.ae.C38.3 PC.ae.C38.4 PC.ae.C38.5
19 0 0 0
PC.ae.C38.6 PC.ae.C40.1 PC.ae.C40.2 PC.ae.C40.3
0 0 0 0
PC.ae.C40.4 PC.ae.C40.5 PC.ae.C40.6 PC.ae.C42.0
0 0 0 0
PC.ae.C42.1 PC.ae.C42.2 PC.ae.C42.3 PC.ae.C42.4
0 1 0 0
PC.ae.C42.5 PC.ae.C44.3 PC.ae.C44.4 PC.ae.C44.5
0 0 0 0
PC.ae.C44.6 SM..OH..C14.1 SM..OH..C16.1 SM..OH..C22.1
0 0 0 0
SM..OH..C22.2 SM..OH..C24.1 SM.C16.0 SM.C16.1
0 0 0 0
SM.C18.0 SM.C18.1 SM.C20.2 SM.C24.0
0 0 0 0
SM.C24.1 SM.C26.0 SM.C26.1 H1_1
0 0 0 0
H1 Urea_N L.Arginine_N L.Leucine_N
0 1 1 1
EDTAca_N X2.Hydroxybutyrate X3.Hydroxybutyrate Acetate
1 1 1 1
Acetoacetate Acetone Betaine Carnitine
1 1 1 1
Choline Creatine Dimethyl.sulfone Ethanol
1 1 1 2
Formate Glucose Glycerol Hypoxanthine
2 1 1 1
Isobutyrate Isopropanol Lactate Malonate
1 1 1 1
> #c) Remove the rows which has missing value for the Dopamine column and assign the result to a new data frame.
> df2 <- df %>% drop_na(Dopamine)
> #(d) In the new data frame, replace the missing values in the c4-OH-Pro column with the median value of the same
> #column. (Hint: there is median( ) function.)
> df2$c4.OH.Pro[is.na(df2$c4.OH.Pro)] <- median(df2$c4.OH.Pro, na.rm = TRUE)
> #(e) (Optional) Drop columns which have more than 25% missing values.
> df3 <- df2 %>% purrr::discard(~sum(is.na(.x))/length(.x)* 100 >=25)
>


Environment
History
Connections
Tutorial
Import Dataset
277 MiB
R
Global Environment
Data
df
69 obs. of 192 variables
df1
12 obs. of 3 variables
df2
49 obs. of 192 variables
df3
49 obs. of 187 variables
medals
852 obs. of 20 variables