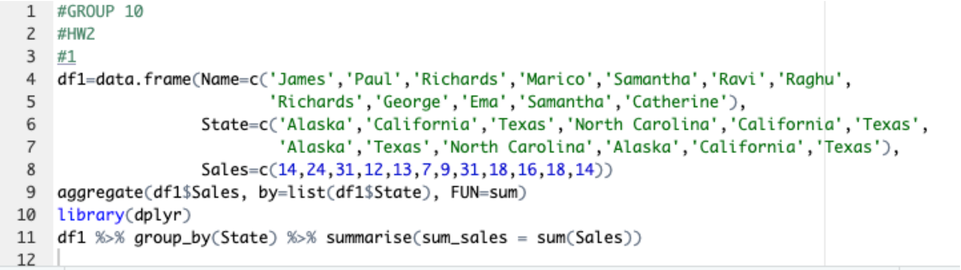
Group 10

HW2

3/11/25

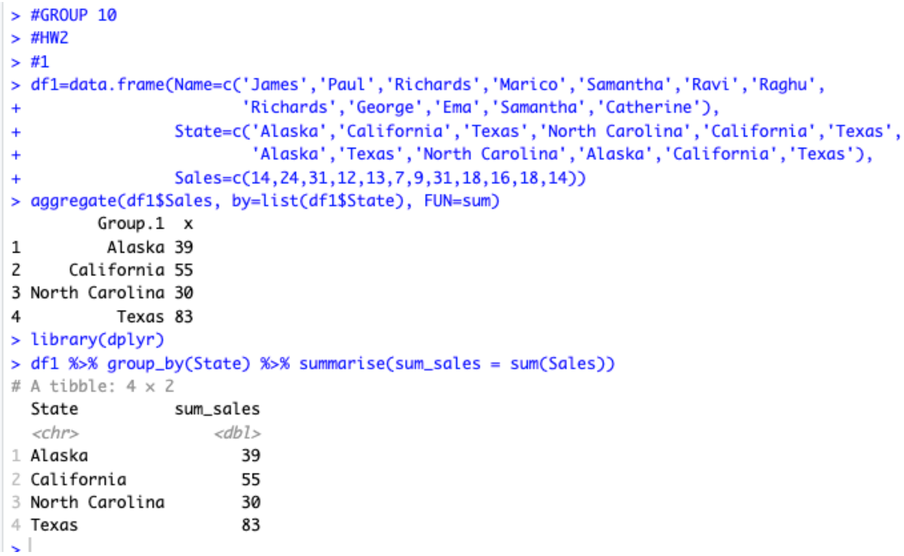
Github repository: https://github.com/bsurgalski/surgalski.git

1. Copy & pasted code from Canvas:



```
#GROUP 10
#HW2
#1
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))
aggregate(df1$Sales, by=list(df1$State), FUN=sum)
library(dplyr)
df1 %>% group_by(State) %>% summarise(sum_sales = sum(Sales))

```

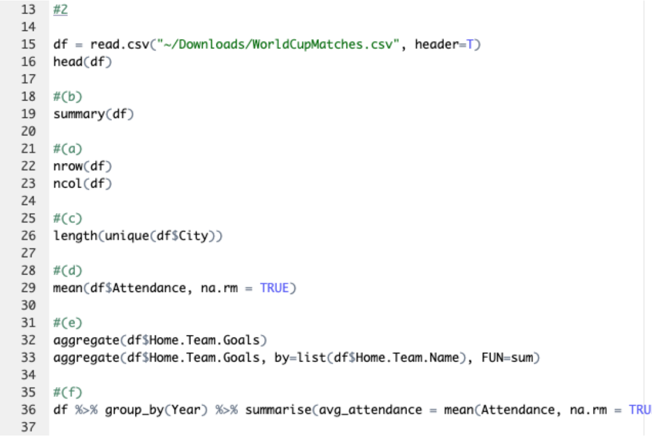
Output:



```
> #GROUP 10
> #HW2
> #1
> 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))
> 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
> library(dplyr)
> 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
> |
```

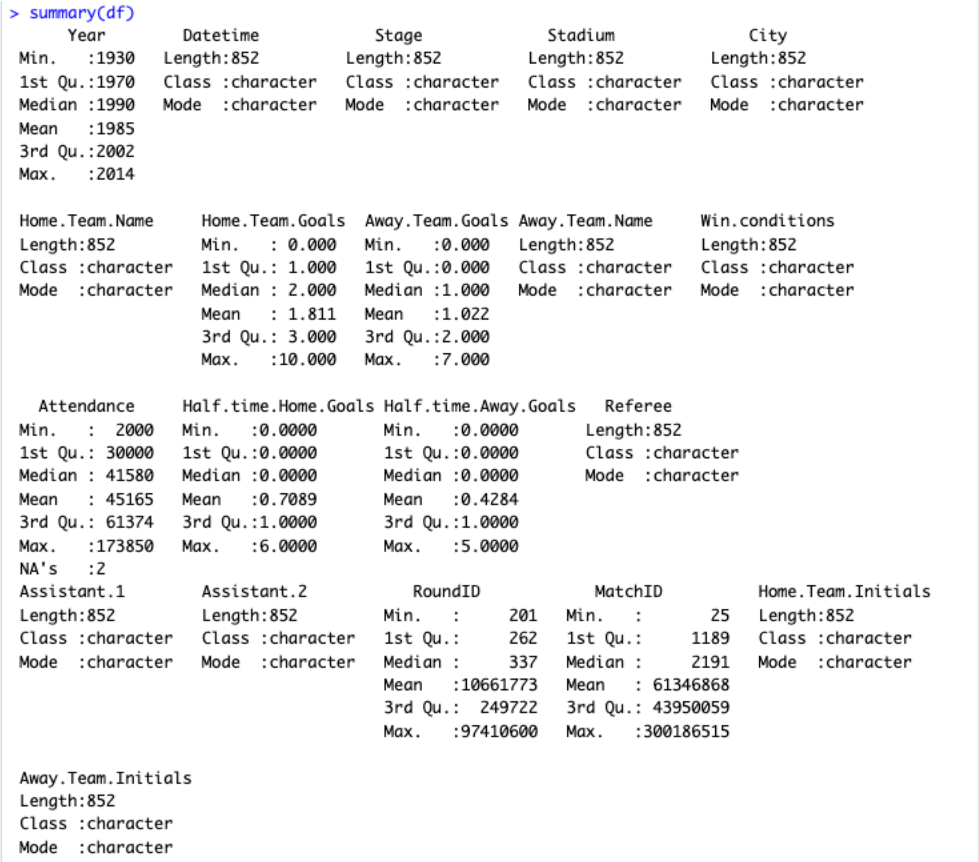
These lines of code first give you a data frame with states and sales in each state with columns labeled “Group.1” and “x”. Then the code gives you the states and sales in each state with columns labeled “State” and “sum_sales”.

2. Code input for (a) – (f)

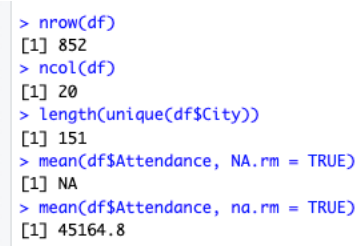


```r
#2

df = read.csv("~/Downloads/WorldCupMatches.csv", header=T)
head(df)

#(b)
summary(df)

#(a)
nrow(df)
ncol(df)

#(c)
length(unique(df$City))

#(d)
mean(df$Attendance, na.rm = TRUE)

#(e)
aggregate(df$Home.Team.Goals)
aggregate(df$Home.Team.Goals, by=list(df$Home.Team.Name), FUN=sum)

#(f)
df %>% group_by(Year) %>% summarise(avg_attendance = mean(Attendance, na.rm = TRUE))

```

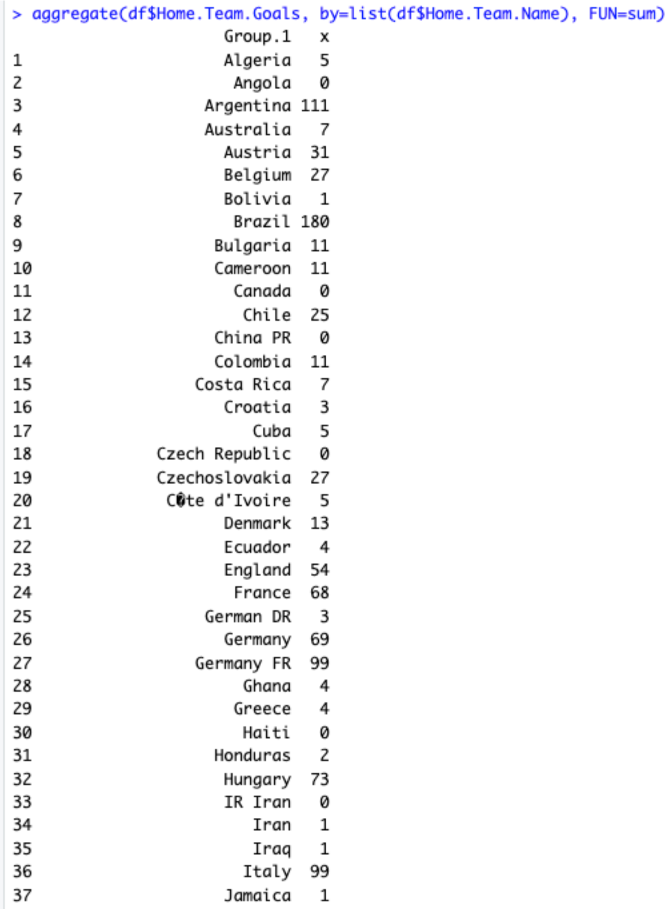
Output (a) – (f)



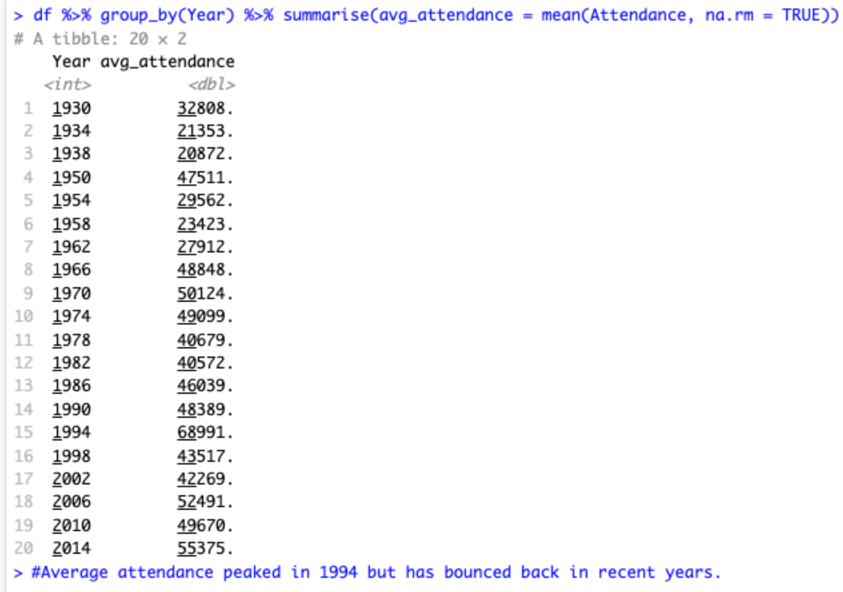
```
> summary(df)
     Year        Datetime            Stage             Stadium              City          
 Min.   :1930   Length:852         Length:852         Length:852         Length:852        
 1st Qu.:1970   Class :character   Class :character   Class :character   Class :character  
 Median :1990   Mode  :character   Mode  :character   Mode  :character   Mode  :character  
 Mean   :1985                                                                              
 3rd Qu.:2002                                                                              
 Max.   :2014                                                                              

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

   Attendance     Half.time.Home.Goals Half.time.Away.Goals   Referee         
 Min.   :  2000   Min.   :0.0000       Min.   :0.0000       Length:852        
 1st Qu.: 30000   1st Qu.:0.0000       1st Qu.:0.0000       Class :character  
 Median : 41580   Median :0.0000       Median :0.0000       Mode  :character  
 Mean   : 45165   Mean   :0.7089       Mean   :0.4284                         
 3rd Qu.: 61374   3rd Qu.:1.0000       3rd Qu.:1.0000                         
 Max.   :173850   Max.   :6.0000       Max.   :5.0000                         
 NA's   :2                                                                    
 Assistant.1        Assistant.2           RoundID            MatchID          Home.Team.Initials
 Length:852         Length:852         Min.   :     201   Min.   :       25   Length:852        
 Class :character   Class :character   1st Qu.:     262   1st Qu.:     1189   Class :character  
 Mode  :character   Mode  :character   Median :     337   Median :     2191   Mode  :character  
                                       Mean   :10661773   Mean   : 61346868                     
                                       3rd Qu.:  249722   3rd Qu.: 43950059                     
                                       Max.   :97410600   Max.   :300186515                     

 Away.Team.Initials
 Length:852        
 Class :character  
 Mode  :character  
```



```
> nrow(df)
[1] 852
> ncol(df)
[1] 20
> length(unique(df$City))
[1] 151
> mean(df$Attendance, NA.rm = TRUE)
[1] NA
> mean(df$Attendance, na.rm = TRUE)
[1] 45164.8
```

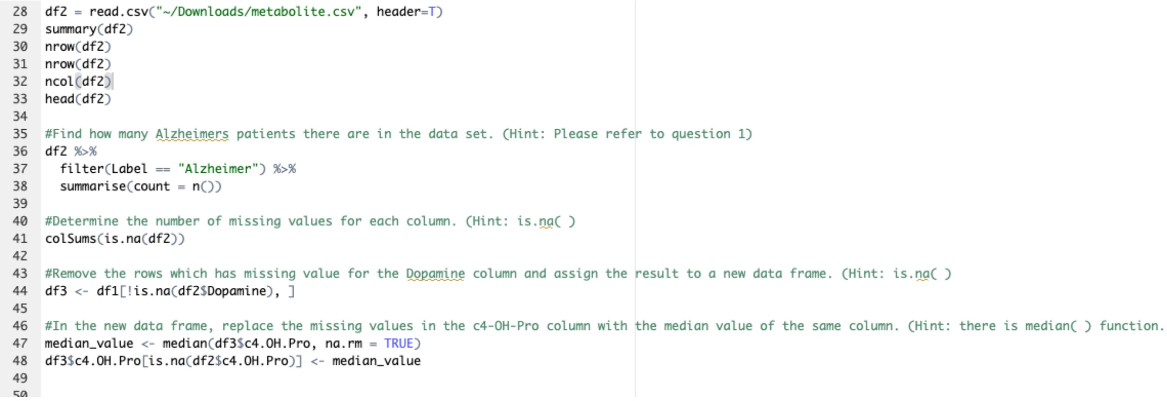


```
> aggregate(df$Home.Team.Goals, by=list(df$Home.Team.Name), FUN=sum)
                      Group.1   x
1                     Algeria   5
2                      Angola   0
3                   Argentina 111
4                   Australia   7
5                     Austria  31
6                     Belgium  27
7                     Bolivia   1
8                      Brazil 180
9                    Bulgaria  11
10                   Cameroon  11
11                     Canada   0
12                      Chile  25
13                   China PR   0
14                   Colombia  11
15                 Costa Rica   7
16                    Croatia   3
17                       Cuba   5
18             Czech Republic   0
19             Czechoslovakia  27
20              C�te d'Ivoire   5
21                    Denmark  13
22                    Ecuador   4
23                    England  54
24                     France  68
25                  German DR   3
26                    Germany  69
27                 Germany FR  99
28                      Ghana   4
29                     Greece   4
30                      Haiti   0
31                   Honduras   2
32                    Hungary  73
33                    IR Iran   0
34                       Iran   1
35                       Iraq   1
36                      Italy  99
37                    Jamaica   1
```



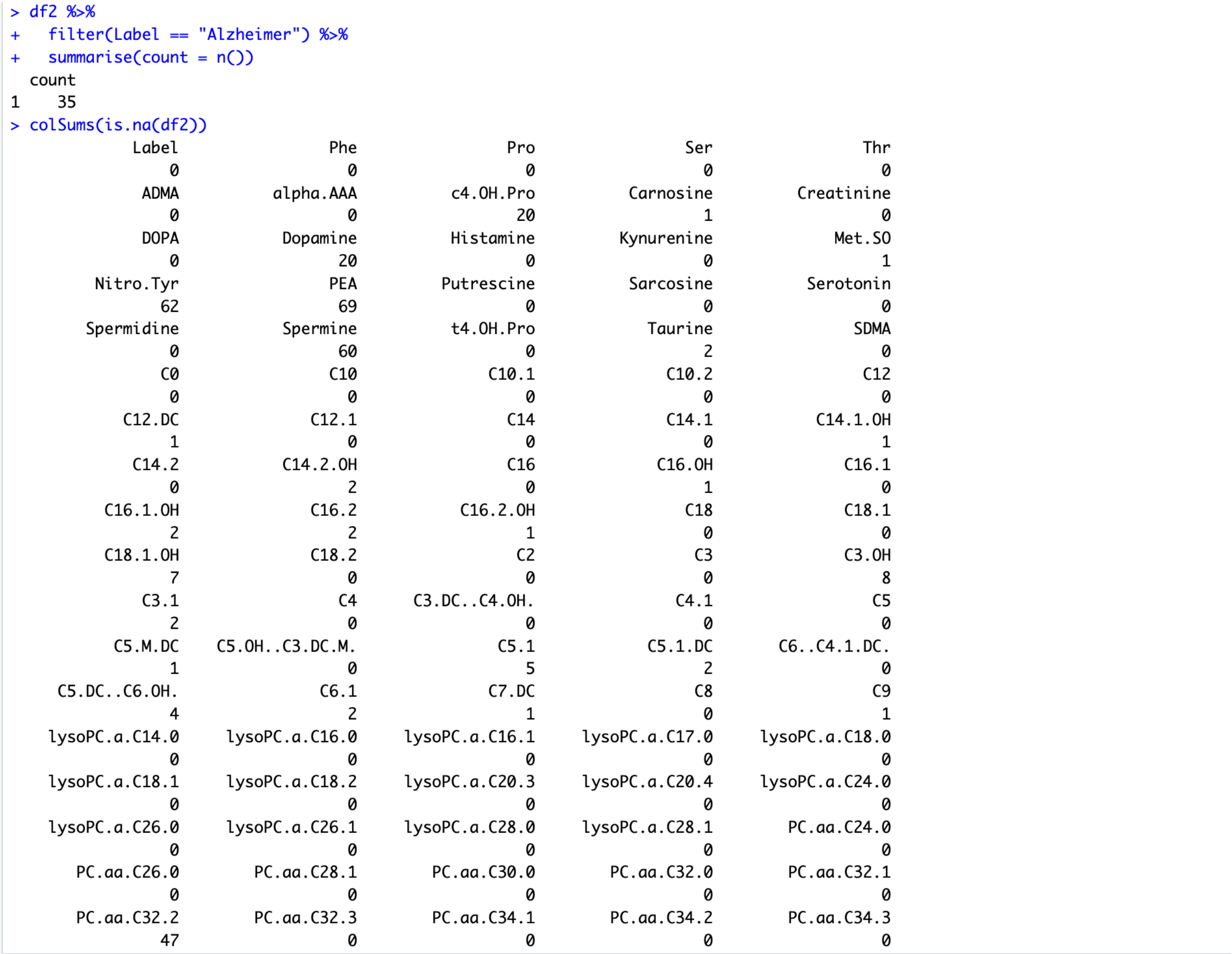
```
> df %>% group_by(Year) %>% summarise(avg_attendance = mean(Attendance, na.rm = TRUE))
# A tibble: 20 × 2
    Year avg_attendance
   <int>          <dbl>
 1  1930         32808.
 2  1934         21353.
 3  1938         20872.
 4  1950         47511.
 5  1954         29562.
 6  1958         23423.
 7  1962         27912.
 8  1966         48848.
 9  1970         50124.
10  1974         49099.
11  1978         40679.
12  1982         40572.
13  1986         46039.
14  1990         48389.
15  1994         68991.
16  1998         43517.
17  2002         42269.
18  2006         52491.
19  2010         49670.
20  2014         55375.
> #Average attendance peaked in 1994 but has bounced back in recent years.
```

3. Code input (a) – (e)

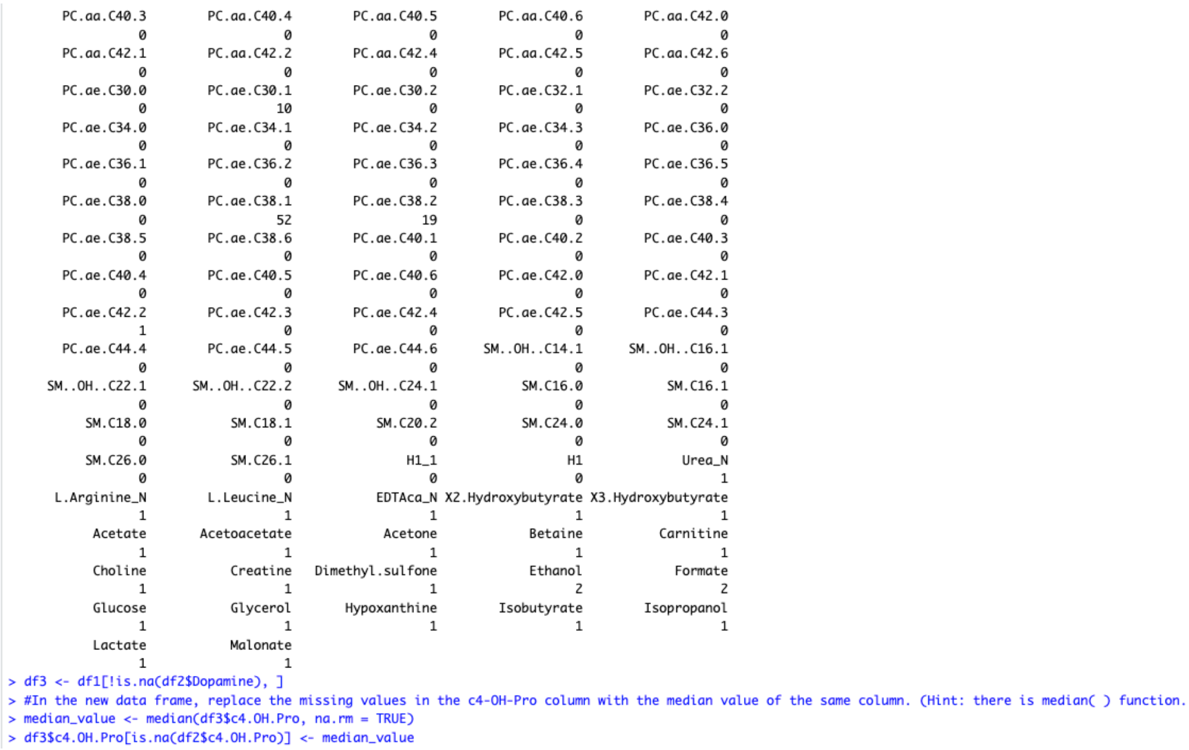


```
df2 = read.csv("~/Downloads/metabolite.csv", header=T)
summary(df2)
nrow(df2)
nrow(df2)
ncol(df2)
head(df2)

#Find how many Alzheimers patients there are in the data set. (Hint: Please refer to question 1)
df2 %>%
  filter(Label == "Alzheimer") %>%
  summarise(count = n())

#Determine the number of missing values for each column. (Hint: is.na( )
colSums(is.na(df2))

#Remove the rows which has missing value for the Dopamine column and assign the result to a new data frame. (Hint: is.na( )
df3 <- df1[!is.na(df2$Dopamine), ]

#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.
median_value <- median(df3$c4.OH.Pro, na.rm = TRUE)
df3$c4.OH.Pro[is.na(df2$c4.OH.Pro)] <- median_value
```

Output (a) – (e)



```
> df2 %>%
+   filter(Label == "Alzheimer") %>%
+   summarise(count = n())
  count
1    35
> colSums(is.na(df2))
             Label                Phe                Pro                Ser                Thr
                 0                  0                  0                  0                  0
              ADMA          alpha.AAA          c4.OH.Pro          Carnosine         Creatinine
                 0                  0                 20                  1                  0
              DOPA           Dopamine          Histamine         Kynurenine             Met.SO
                 0                 20                  0                  0                  1
         Nitro.Tyr                PEA         Putrescine          Sarcosine          Serotonin
                62                 69                  0                  0                  0
        Spermidine           Spermine          t4.OH.Pro            Taurine               SDMA
                 0                 60                  0                  2                  0
                C0                C10              C10.1              C10.2                C12
                 0                  0                  0                  0                  0
            C12.DC              C12.1                C14              C14.1           C14.1.OH
                 1                  0                  0                  0                  1
             C14.2           C14.2.OH                C16             C16.OH              C16.1
                 0                  2                  0                  1                  0
          C16.1.OH              C16.2           C16.2.OH                C18              C18.1
                 2                  2                  1                  0                  0
          C18.1.OH              C18.2                 C2                 C3              C3.OH
                 7                  0                  0                  0                  8
              C3.1                 C4      C3.DC..C4.OH.               C4.1                 C5
                 2                  0                  0                  0                  0
           C5.M.DC    C5.OH..C3.DC.M.               C5.1            C5.1.DC       C6..C4.1.DC.
                 1                  0                  5                  2                  0
     C5.DC..C6.OH.               C6.1              C7.DC                 C8                 C9
                 4                  2                  1                  0                  1
    lysoPC.a.C14.0     lysoPC.a.C16.0     lysoPC.a.C16.1     lysoPC.a.C17.0     lysoPC.a.C18.0
                 0                  0                  0                  0                  0
    lysoPC.a.C18.1     lysoPC.a.C18.2     lysoPC.a.C20.3     lysoPC.a.C20.4     lysoPC.a.C24.0
                 0                  0                  0                  0                  0
    lysoPC.a.C26.0     lysoPC.a.C26.1     lysoPC.a.C28.0     lysoPC.a.C28.1        PC.aa.C24.0
                 0                  0                  0                  0                  0
       PC.aa.C26.0        PC.aa.C28.1        PC.aa.C30.0        PC.aa.C32.0        PC.aa.C32.1
                 0                  0                  0                  0                  0
       PC.aa.C32.2        PC.aa.C32.3        PC.aa.C34.1        PC.aa.C34.2        PC.aa.C34.3
                47                  0                  0                  0                  0
```



```
       PC.aa.C40.3        PC.aa.C40.4        PC.aa.C40.5        PC.aa.C40.6        PC.aa.C42.0
                 0                  0                  0                  0                  0
       PC.aa.C42.1        PC.aa.C42.2        PC.aa.C42.4        PC.aa.C42.5        PC.aa.C42.6
                 0                  0                  0                  0                  0
       PC.ae.C30.0        PC.ae.C30.1        PC.ae.C30.2        PC.ae.C32.1        PC.ae.C32.2
                 0                 10                  0                  0                  0
       PC.ae.C34.0        PC.ae.C34.1        PC.ae.C34.2        PC.ae.C34.3        PC.ae.C36.0
                 0                  0                  0                  0                  0
       PC.ae.C36.1        PC.ae.C36.2        PC.ae.C36.3        PC.ae.C36.4        PC.ae.C36.5
                 0                  0                  0                  0                  0
       PC.ae.C38.0        PC.ae.C38.1        PC.ae.C38.2        PC.ae.C38.3        PC.ae.C38.4
                 0                 52                 19                  0                  0
       PC.ae.C38.5        PC.ae.C38.6        PC.ae.C40.1        PC.ae.C40.2        PC.ae.C40.3
                 0                  0                  0                  0                  0
       PC.ae.C40.4        PC.ae.C40.5        PC.ae.C40.6        PC.ae.C42.0        PC.ae.C42.1
                 0                  0                  0                  0                  0
       PC.ae.C42.2        PC.ae.C42.3        PC.ae.C42.4        PC.ae.C42.5        PC.ae.C44.3
                 1                  0                  0                  0                  0
       PC.ae.C44.4        PC.ae.C44.5        PC.ae.C44.6      SM..OH..C14.1      SM..OH..C16.1
                 0                  0                  0                  0                  0
     SM..OH..C22.1      SM..OH..C22.2      SM..OH..C24.1           SM.C16.0           SM.C16.1
                 0                  0                  0                  0                  0
          SM.C18.0           SM.C18.1           SM.C20.2           SM.C24.0           SM.C24.1
                 0                  0                  0                  0                  0
          SM.C26.0           SM.C26.1               H1_1                 H1             Urea_N
                 0                  0                  0                  0                  1
      L.Arginine_N        L.Leucine_N           EDTAca_N X2.Hydroxybutyrate X3.Hydroxybutyrate
                 1                  1                  1                  1                  1
           Acetate       Acetoacetate            Acetone            Betaine          Carnitine
                 1                  1                  1                  1                  1
           Choline           Creatine   Dimethyl.sulfone            Ethanol            Formate
                 1                  1                  1                  2                  2
           Glucose           Glycerol       Hypoxanthine        Isobutyrate        Isopropanol
                 1                  1                  1                  1                  1
           Lactate           Malonate
                 1                  1
> df3 <- df1[!is.na(df2$Dopamine), ]
> #In the new data frame, replace the missing values in the c4-OH-Pro column with the median value of the same column. (Hint: there is median( ) function.
> median_value <- median(df3$c4.OH.Pro, na.rm = TRUE)
> df3$c4.OH.Pro[is.na(df2$c4.OH.Pro)] <- median_value
```