Problem 1

```
>
df1=data.frame(Name=c('James','Paul','Richards','Marico','Samantha','Ravi','Rag
hu',
+                      '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
> install.packages("dplyr")
> 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 sort the data set's observations by the name of the state, and sums the number of sales per state.

Problem 2

a)

```
> df2 <- read.csv("WorldCupMatches.csv")
> dim(df2)
[1] 852  20
```

There are 852 rows and 20 columns in the data set.

b)

```
> summary(df2)
      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
Assistant.1
 Min.   :  2000   Min.   :0.0000       Min.   :0.0000       Length:852
Length:852
 1st Qu.: 30000   1st Qu.:0.0000       1st Qu.:0.0000       Class :character
Class :character
 Median : 41580   Median :0.0000       Median :0.0000       Mode  :character
Mode  :character
 Mean   : 45165   Mean   :0.7089       Mean   :0.4284
 3rd Qu.: 61375   3rd Qu.:1.0000       3rd Qu.:1.0000
 Max.   :173850   Max.   :6.0000       Max.   :5.0000
 NA's   :2
 Assistant.2           RoundID            MatchID          Home.Team.Initials
Away.Team.Initials
 Length:852         Min.   :     201   Min.   :       25   Length:852
Length:852
 Class :character   1st Qu.:     262   1st Qu.:     1189   Class :character
Class :character
 Mode  :character   Median :     337   Median :     2191   Mode  :character
Mode  :character
                    Mean   :10661773   Mean   : 61346868
                    3rd Qu.:  249722   3rd Qu.: 43950059
                    Max.   :97410600   Max.   :300186515
```

c)

```
> length(unique(df2$City))
[1] 151
```

Matches were held at 151 different locations

d)

```
> mean(df2$Attendance, na.rm = TRUE)
[1] 45164.8
```

The mean attendance is 45,164.8

e)

```
> df2 %>% group_by(Home.Team.Name) %>% summarise(sum_goals =
sum(Home.Team.Goals))
```

```
# A tibble: 78 × 2
   Home.Team.Name sum_goals
   <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)

```
> aggregate(df2$Attendance, by=list(df2$Year), FUN=mean, na.rm = T)
   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
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 55374.91
```

There seems to be a noticeable decline in attendance in 1934 and 1938, which is when World War II began. Attendance also seems to spike in 1990 and 1994, which is shortly after the Cold War ended.

Problem 3

a)

```
> df3 <- read.csv("metabolite.csv")
>
> df3 %>% group_by(Label) %>% summarise(count = length(Label))
```

```
# A tibble: 2 × 2
  Label     count
  <chr>     <int>
1 Alzheimer    35
2 Healthy      34
```

There are 35 Alzheimer's patients in the data set.

b)

```
> sum(is.na(df3))
[1] 432
> col_NA <- rep(NA, times = ncol(df3))
>
> for(j in 1:ncol(df3)) {
+  col_NA[j] <- sum(is.na(df3[ ,j]))
+ }
> col_NA
  [1]  0  0  0  0  0  0  0 20  1  0  0 20  0  0  1 62 69  0  0  0  0 60  0  2
0  0  0  0  0  0  1
 [32]  0  0  0  1  0  2  0  1  0  2  2  1  0  0  7  0  0  0  8  2  0  0  0  0
1  0  5  2  0  4  2
 [63]  1  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 47
0  0  0  0  0  0  0
 [94]  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 10
0  0  0  0  0  0  0
[125]  0  0  0  0  0  0  0 52 19  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0
0  0  0  0  0  0  0
[156]  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  1  1  1  1  1  1  1  1  1
1  1  1  1  2  2  1
[187]  1  1  1  1  1  1
> sum(col_NA)
[1] 432
```

c)

```
> df3.1 <- df3[is.na(df3['Dopamine'])==F, ]
> head(df3.1$Dopamine, 10)
 [1] 0.233 0.234 0.231 0.244 0.233 0.225 0.240 0.239 0.231 0.236
```

d)

```
> df3.1[['c4.OH.Pro']][is.na(df3.1[['c4.OH.Pro']])] <- median(df3.1$c4.OH.Pro,
na.rm = T)
> head(df3.1$c4.OH.Pro, 10)
 [1] 0.236 0.199 0.199 0.215 0.186 0.185 0.215 0.237 0.215 0.192
```