DATA ANALYTICS – 4027

LAB-6

Name: Hari Krishna P

Reg No: 19BCE7675

DATE: 02/11/2021

Contents:

- **Datasets**

Submitted to:

Prof . Hari Seetha

Ex 6

Questions:

1. Use t student dataset
    a. write a function for calculating the mean of the scores.

```
> lapply(StudentsPerformance[, 6:8], mean)
$`math score`
[1] 66.089

$`reading score`
[1] 69.169

$`writing score`
[1] 68.054
```

    b. Write a function to compute std.deviation of the scores.

```
> lapply(StudentsPerformance[, 6:8], sd)
$`math score`
[1] 15.16308

$`reading score`
[1] 14.60019

$`writing score`
[1] 15.19566

>
```

2. Find (min, max, mean, variance, SD, range of numeric columns.

Min:

```
> lapply(StudentsPerformance[, 6:8], min)
$`math score`
[1] 0

$`reading score`
[1] 17

$`writing score`
[1] 10
```

Max

```
> lapply(StudentsPerformance[, 6:8], max)
$`math score`
[1] 100

$`reading score`
[1] 100

$`writing score`
[1] 100
```

Mean:

```
> lapply(StudentsPerformance[, 6:8], mean)
$`math score`
[1] 66.089

$`reading score`
[1] 69.169

$`writing score`
[1] 68.054
```

Variance:

```
> lapply(StudentsPerformance[, 6:8], var)
$`math score`
[1] 229.919

$`reading score`
[1] 213.1656

$`writing score`
[1] 230.908
```

Range:

```
> lapply(StudentsPerformance[, 6:8], range)
$`math score`
[1]   0 100

$`reading score`
[1]  17 100

$`writing score`
[1]  10 100
```

Sd:

```
> lapply(StudentsPerformance[, 6:8], sd)
$`math score`
[1] 15.16308

$`reading score`
[1] 14.60019

$`writing score`
[1] 15.19566

>
```

Another Method To Find Is:

Max:

```
> max(StudentsPerformance$`math score`)
[1] 100
> max(StudentsPerformance$`reading score`)
[1] 100
> max(StudentsPerformance$`writing score`)
[1] 100
>
```

Min:

```
> min(StudentsPerformance$`math score`)
[1] 0
> min(StudentsPerformance$`writing score`)
[1] 10
> min(StudentsPerformance$`reading score`)
[1] 17
>
```

Variance:

```
> var(StudentsPerformance$`math score`)
[1] 229.919
> var(StudentsPerformance$`reading score`)
[1] 213.1656
> var(StudentsPerformance$`writing score`)
[1] 230.908
>
```

SD:

```
> sd(StudentsPerformance$`math score`, na.rm=TRUE)
[1] 15.16308
> sd(StudentsPerformance$`writing score`, na.rm=TRUE)
[1] 15.19566
> sd(StudentsPerformance$`reading score`, na.rm=TRUE)
[1] 14.60019
>
```

Range:

```
> range(StudentsPerformance$`math score`)
[1]   0 100
> range(StudentsPerformance$`reading score`)
[1]  17 100
> range(StudentsPerformance$`writing score`)
[1]  10 100
>
```

3. The summary values should be in a single data frame with the following columns: variable name, mean, sd, minimum, and maximum.

```
> summary(StudentsPerformance)
   gender          race/ethnicity     parental level of education    lunch          test preparation course   math score     reading score  
 Length:1000        Length:1000        Length:1000                 Length:1000        Length:1000             Min.   :  0.00   Min.   : 17.00  
 Class :character   Class :character   Class :character            Class :character   Class :character        1st Qu.: 57.00   1st Qu.: 59.00  
 Mode  :character   Mode  :character   Mode  :character            Mode  :character   Mode  :character        Median : 66.00   Median : 70.00  
                                                                                                              Mean   : 66.09   Mean   : 69.17  
                                                                                                              3rd Qu.: 77.00   3rd Qu.: 79.00  
                                                                                                              Max.   :100.00   Max.   :100.00  

 writing score   
 Min.   : 10.00  
 1st Qu.: 57.75  
 Median : 69.00  
 Mean   : 68.05  
 3rd Qu.: 79.00  
 Max.   :100.00  
> 
```

4. Select students who are Female grade.

```
> x1 = StudentsPerformance[StudentsPerformance$gender=="female",,drop=FALSE]
> x1
# A tibble: 518 x 8
   gender `race/ethnicity` `parental level of education` lunch       `test preparation course` `math score` `reading score` `writing score`
   <chr>  <chr>            <chr>                         <chr>       <chr>                           <dbl>           <dbl>           <dbl>
 1 female group B          bachelor's degree             standard    none                               72              72              74
 2 female group C          some college                  standard    completed                          69              90              88
 3 female group B          master's degree               standard    none                               90              95              93
 4 female group B          associate's degree            standard    none                               71              83              78
 5 female group B          some college                  standard    completed                          88              95              92
 6 female group B          high school                   free/reduced none                              38              60              50
 7 female group B          high school                   standard    none                               65              81              73
 8 female group A          master's degree               standard    none                               50              53              58
 9 female group C          some high school              standard    none                               69              75              78
10 female group B          some high school              free/reduced none                              18              32              28
# ... with 508 more rows
> 
```

5. Find the duplicate records and count them.

```
> StudentsPerformance[duplicated(StudentsPerformance$'race/ethnicity')&duplicated(StudentsPerformance$gender)&duplicated(StudentsPerformance$'parental level o
f education'),]
# A tibble: 993 x 8
   gender `race/ethnicity` `parental level of education` lunch       `test preparation course` `math score` `reading score` `writing score`
   <chr>  <chr>            <chr>                         <chr>       <chr>                           <dbl>           <dbl>           <dbl>
 1 male   group C          some college                  standard    none                               76              78              75
 2 female group B          associate's degree            standard    none                               71              83              78
 3 female group B          some college                  standard    completed                          88              95              92
 4 male   group B          some college                  free/reduced none                              40              43              39
 5 female group B          high school                   free/reduced none                              38              60              50
 6 male   group C          associate's degree            standard    none                               58              54              52
 7 male   group D          associate's degree            standard    none                               40              52              43
 8 female group B          high school                   standard    none                               65              81              73
 9 male   group A          some college                  standard    completed                          78              72              70
10 female group A          master's degree               standard    none                               50              53              58
# ... with 983 more rows
> 
> sum(duplicated(StudentsPerformance$gender)&duplicated(StudentsPerformance$`race/ethnicity`)&duplicated(StudentsPerformance$`parental level of education`))
[1] 993
> 
```

6. Remove Duplicate Rows based on gender

```
> StudentsPerformance[!duplicated(StudentsPerformance$gender),]
# A tibble: 2 x 8
  gender `race/ethnicity` `parental level of education` lunch       `test preparation course` `math score` `reading score` `writing score`
  <chr>  <chr>            <chr>                         <chr>       <chr>                           <dbl>           <dbl>           <dbl>
1 female group B          bachelor's degree             standard    none                               72              72              74
2 male   group A          associate's degree            free/reduced none                              47              57              44
> 
```

7. Sort the data on descending order of writing_score

```
> x1 <- StudentsPerformance[order(-StudentsPerformance$`writing score`),]
> x1
# A tibble: 1,000 x 8
   gender `race/ethnicity` `parental level of education` lunch       `test preparation course` `math score` `reading score` `writing score`
   <chr>  <chr>            <chr>                         <chr>       <chr>                          <dbl>           <dbl>           <dbl>
 1 female group D           master's degree               standard    none                              87             100             100
 2 female group E           bachelor's degree             standard    completed                         99             100             100
 3 female group C           bachelor's degree             standard    completed                         96             100             100
 4 female group D           some high school              standard    completed                         97             100             100
 5 female group D           master's degree               free/reduced completed                        85              95             100
 6 female group D           high school                   standard    completed                         88              99             100
 7 female group E           bachelor's degree             standard    none                             100             100             100
 8 female group E           bachelor's degree             free/reduced completed                        92             100             100
 9 female group E           master's degree               standard    completed                         94              99             100
10 female group D           bachelor's degree             free/reduced completed                        93             100             100
# ... with 990 more rows
> |
```

8. Select variables that start with " m".

```
> Data1 <- select(StudentsPerformance,starts_with("m"))
> Data1
# A tibble: 1,000 x 1
   `math score`
          <dbl>
 1           72
 2           69
 3           90
 4           47
 5           76
 6           71
 7           88
 8           40
 9           64
10           38
# ... with 990 more rows
```

9. Find the sum of math_score group by gender

```
> StudentsPerformance %>% group_by(gender,sum(`math score`))
# A tibble: 1,000 x 9
# Groups:   gender, sum(`math score`) [2]
   gender `race/ethnicity` `parental level of education` lunch       `test preparation course` `math score` `reading score` `writing score` `sum(\`math sco~
   <chr>  <chr>            <chr>                         <chr>       <chr>                          <dbl>           <dbl>           <dbl>            <dbl>
 1 female group B           bachelor's degree             standard    none                              72              72              74            66089
 2 female group C           some college                  standard    completed                         69              90              88            66089
 3 female group B           master's degree               standard    none                              90              95              93            66089
 4 male   group A           associate's degree            free/reduced none                             47              57              44            66089
 5 male   group C           some college                  standard    none                              76              78              75            66089
 6 female group B           associate's degree            standard    none                              71              83              78            66089
 7 female group B           some college                  standard    completed                         88              95              92            66089
 8 male   group B           some college                  free/reduced none                             40              43              39            66089
 9 male   group D           high school                   free/reduced completed                        64              64              67            66089
10 female group B           high school                   free/reduced none                             38              60              50            66089
# ... with 990 more rows
```