

```
---
title: "Assignment 1"
author: "Daniel Yim"
date: "2023-02-02"
output: word_document
---
```

```
```{r}
library(tidyverse)
library(readxl)
read_excel("C:/Users/Daniel/Desktop/Sheridan 2022-23/2 - Statistics for Data Science/M2/Assignment
1/Assignment 1 Data.xlsx")
```

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

Learning R (SFDS Class).R Assignment 1 - Daniel Yim.R Assignment 1 - DY (submission).Rmd Learning R (Sheridan).R Assignment\_1\_Data

Knit on Save Knit Run

Source Visual Outline

```
1 ---
2 title: "Assignment 1"
3 author: "Daniel Yim"
4 date: "2023-02-02"
5 output: word_document
6 ---
7
8
9 ```{r}
10 library(tidyverse)
11 library(readxl)
12 read_excel("C:/Users/Daniel/Desktop/Sheridan 2022-23/2 - Statistics for Data Science/M2/Assignment 1/Assignment 1
13 Data.xlsx")
14 ```
```

A tibble: 30 x 14

ID	Last Name	First Name	City	State	Gender	Student Status	Major	Country
1	DOE01	JANE01	Los Angeles	California	Female	Graduate	Politics	US
2	DOE02	JANE02	Sedona	Arizona	Female	Undergraduate	Math	US
3	DOE01	JOE01	Elmira	New York	Male	Graduate	Math	US
4	DOE02	JOE02	Lackawana	New York	Male	Graduate	Econ	US
5	DOE03	JOE03	Defiance	Ohio	Male	Graduate	Econ	US
6	DOE04	JOE04	Tel Aviv	Israel	Male	Graduate	Econ	Israel
7	DOE05	JOE05	Cimax	North Carolina	Male	Graduate	Politics	US
8	DOE03	JANE03	Liberal	Kansas	Female	Undergraduate	Politics	US
9	DOE04	JANE04	Montreal	Canada	Female	Undergraduate	Math	Canada
10	DOE05	JANE05	New York	New York	Female	Graduate	Math	US

1-10 of 30 rows | 1-9 of 14 columns

Previous 1 2 3 Next

```

Part 1

1. There are 14 variables
```{r}
ls(Assignment_1_Data)
```

[1] "Age"
[4] "Country"
[7] "Height (in)"
[10] "Major"
[13] "State"

"Average score (grade)"
"First Name"
"ID"
"Newspaper readership (times/wk)"
"Student Status"

"City"
"Gender"
"Last Name"
"SAT"

2. Format of variables are as follows, numeric and character
```{r}
str(Assignment_1_Data)
```

tibble [30 × 14] (s3: tbl_df/tbl/data.frame)
 $ ID : num [1:30] 1 2 3 4 5 6 7 8 9 10 ...
 $ Last Name : chr [1:30] "DOE01" "DOE02" "DOE01" "DOE02" ...
 $ First Name : chr [1:30] "JANE01" "JANE02" "JOE01" "JOE02" ...
 $ City : chr [1:30] "Los Angeles" "Sedona" "Elmira" "Lackawana" ...
 $ State : chr [1:30] "California" "Arizona" "New York" "New York" ...
 $ Gender : chr [1:30] "Female" "Female" "Male" "Male" ...
 $ Student Status : chr [1:30] "Graduate" "Undergraduate" "Graduate" "Graduate" ...
 $ Major : chr [1:30] "Politics" "Math" "Math" "Econ" ...
 $ Country : chr [1:30] "US" "US" "US" "US" ...
 $ Age : num [1:30] 30 19 26 33 37 25 39 21 18 33 ...
 $ SAT : num [1:30] 2263 2006 2221 1716 1701 ...
 $ Average score (grade) : num [1:30] 67 63 78 78 65 69 96 87 91 71 ...
 $ Height (in) : num [1:30] 61 64 73 68 71 67 70 62 62 66 ...
 $ Newspaper readership (times/wk): num [1:30] 5 7 6 3 6 5 5 5 6 5 ...

3. In the above results, all the variables that are in character format are categorical. ID is discrete and the rest of the numeric variables are continuous."

4. This is sample data because there are only 30 observations of a student university population which is typically in the thousands."

```

```

5. There are 15 males and 15 females
```{r}
Assignment_1_Data %>% count(Gender)
```

A tibble: 2 × 2

 Gender n
 <chr> <int>
1 Female 15
2 Male 15

2 rows

6. The average age is 25.2
```{r}
mean(Assignment_1_Data$Age)
```

[1] 25.2

```

```
7. There are 15 Graduate and 15 Undergraduate students
```

```
##{r}
Assignment_1_Data %>% count('Student Status')
```

A tibble: 2 × 2

| Student Status<br><chr> | n<br><int> |
|-------------------------|------------|
| Graduate                | 15         |
| Undergraduate           | 15         |

2 rows

```
8. The average SAT score is shown below:
```

```
##{r}
mean(Assignment_1_Data$SAT)
```

[1] 1848.9

```
9.
```

```
##{r}
library(tidyverse)
library(dplyr)
```

```
Assignment_1_Data %>%
 group_by(Gender) %>%
 summarize(sum('Newspaper readership (times/wk)'))
```

A tibble: 2 × 2

| Gender<br><chr> | sum('Newspaper readership (times/wk)')<br><dbl> |
|-----------------|-------------------------------------------------|
| Female          | 78                                              |
| Male            | 68                                              |

2 rows

```
#10. I had to split it into three pipes since there appears to be a limit
```

```
##{r}
data <- Assignment_1_Data
summary(data)
avg_score <- data$`Average score (grade)`
height <- data$`Height (in)`
newspaper <- data$`Newspaper readership (times/wk)`

data %>%
 summarize(var(Age),sd(Age),"Range (Age)" = max(Age)-min(Age),var(SAT),sd(SAT),"Range (SAT)" = max(SAT)-min(SAT))

data %>%
 summarize(var(avg_score),sd(avg_score),"Range (avg_score)" = max(height)-min(height),var(height),sd(height),"Range (height)" = max(height)-min(height))

data %>%
 summarize(var(newspaper),sd(newspaper),"Range (newspaper)" = max(newspaper)-min(newspaper))
```

```
R Console
```

```
tbl_df
1 x 6
```

```
tbl_df
1 x 6
```

```
tbl_df
1 x 3
```

| ID                    | Last Name                       | First Name       | City             | State        |
|-----------------------|---------------------------------|------------------|------------------|--------------|
| Gender                |                                 |                  |                  |              |
| Min. : 1.00           | Length:30                       | Length:30        | Length:30        | Length:30    |
| Length:30             |                                 |                  |                  |              |
| 1st Qu.: 8.25         | Class :character                | Class :character | Class :character | Class        |
| :character            | Class :character                |                  |                  |              |
| Median :15.50         | Mode :character                 | Mode :character  | Mode :character  | Mode         |
| :character            | Mode :character                 |                  |                  |              |
| Mean :15.50           |                                 |                  |                  |              |
| 3rd Qu.:22.75         |                                 |                  |                  |              |
| Max. :30.00           |                                 |                  |                  |              |
| Student Status        | Major                           | Country          | Age              | SAT          |
| Average score (grade) |                                 |                  |                  |              |
| Length:30             | Length:30                       | Length:30        | Min. :18.0       | Min. :1338   |
| Min. :63.00           |                                 |                  |                  |              |
| Class :character      | Class :character                | Class :character | 1st Qu.:19.0     | 1st Qu.:1658 |
| 1st Qu.:72.00         |                                 |                  |                  |              |
| Mode :character       | Mode :character                 | Mode :character  | Median :23.0     | Median :1817 |
| Median :79.50         |                                 |                  |                  |              |
|                       |                                 |                  | Mean :25.2       | Mean :1849   |
| Mean :80.37           |                                 |                  |                  |              |
|                       |                                 |                  | 3rd Qu.:30.0     | 3rd Qu.:2032 |
| 3rd Qu.:88.00         |                                 |                  |                  |              |
|                       |                                 |                  | Max. :39.0       | Max. :2309   |
| Max. :96.00           |                                 |                  |                  |              |
| Height (in)           | Newspaper readership (times/wk) |                  |                  |              |
| Min. :59.00           | Min. :3.000                     |                  |                  |              |
| 1st Qu.:63.00         | 1st Qu.:4.000                   |                  |                  |              |
|                       |                                 |                  |                  |              |
| Median :66.50         | Median :5.000                   |                  |                  |              |
| Mean :66.43           | Mean :4.867                     |                  |                  |              |
| 3rd Qu.:70.75         | 3rd Qu.:6.000                   |                  |                  |              |
| Max. :75.00           | Max. :7.000                     |                  |                  |              |

```
R Console
```

```
tbl_df
1 x 6
```

```
tbl_df
1 x 6
```

```
tbl_df
1 x 3
```

A tibble: 1 x 6

| var(Age) | sd(Age)  | Range (Age) | var(SAT) | sd(SAT)  | Range (SAT) |
|----------|----------|-------------|----------|----------|-------------|
| <dbl>    | <dbl>    | <dbl>       | <dbl>    | <dbl>    | <dbl>       |
| 47.2     | 6.870226 | 21          | 75686.71 | 275.1122 | 971         |

1 row

```
R Console
```

```
tbl_df
1 x 6
```

```
tbl_df
1 x 6
```

```
tbl_df
1 x 3
```

A tibble: 1 x 6

| var(avg_score) | sd(avg_score) | Range (avg_score) | var(height) | sd(height) | Range (height) |
|----------------|---------------|-------------------|-------------|------------|----------------|
| <dbl>          | <dbl>         | <dbl>             | <dbl>       | <dbl>      | <dbl>          |
| 102.2402       | 10.11139      | 16                | 21.7023     | 4.658573   | 16             |

R Console

tbl\_df  
1 x 6

tbl\_df  
1 x 6

tbl\_df  
1 x 3

A tibble: 1 x 3

|   | var(newspaper)<br><dbl> | sd(newspaper)<br><dbl> | Range (newspaper)<br><dbl> |
|---|-------------------------|------------------------|----------------------------|
| 1 | 1.636782                | 1.279368               | 4                          |

1 row

#Part 2:  
# If:  
# Age <- c(22, 25, 18, 20)  
#Name <- c("James", "Mathew", "olivia", "Stella")  
#Gender <- c("M", "M", "F", "F")  
#Then: what is the R-code for getting the following output:  
## Age Name Gender  
## 1 22 James M  
## 2 25 Mathew M

```

{r}
Age <- c(22, 25, 18, 20)
Name <- c("James", "Mathew", "olivia", "Stella")
Gender <- c("M", "M", "F", "F")

data2 <- data.frame(Age,Name,Gender)
data2[c(1,2),]

```

Description: df [2 x 3]

|   | Age<br><dbl> | Name<br><chr> | Gender<br><chr> |
|---|--------------|---------------|-----------------|
| 1 | 22           | James         | M               |
| 2 | 25           | Mathew        | M               |

2 rows