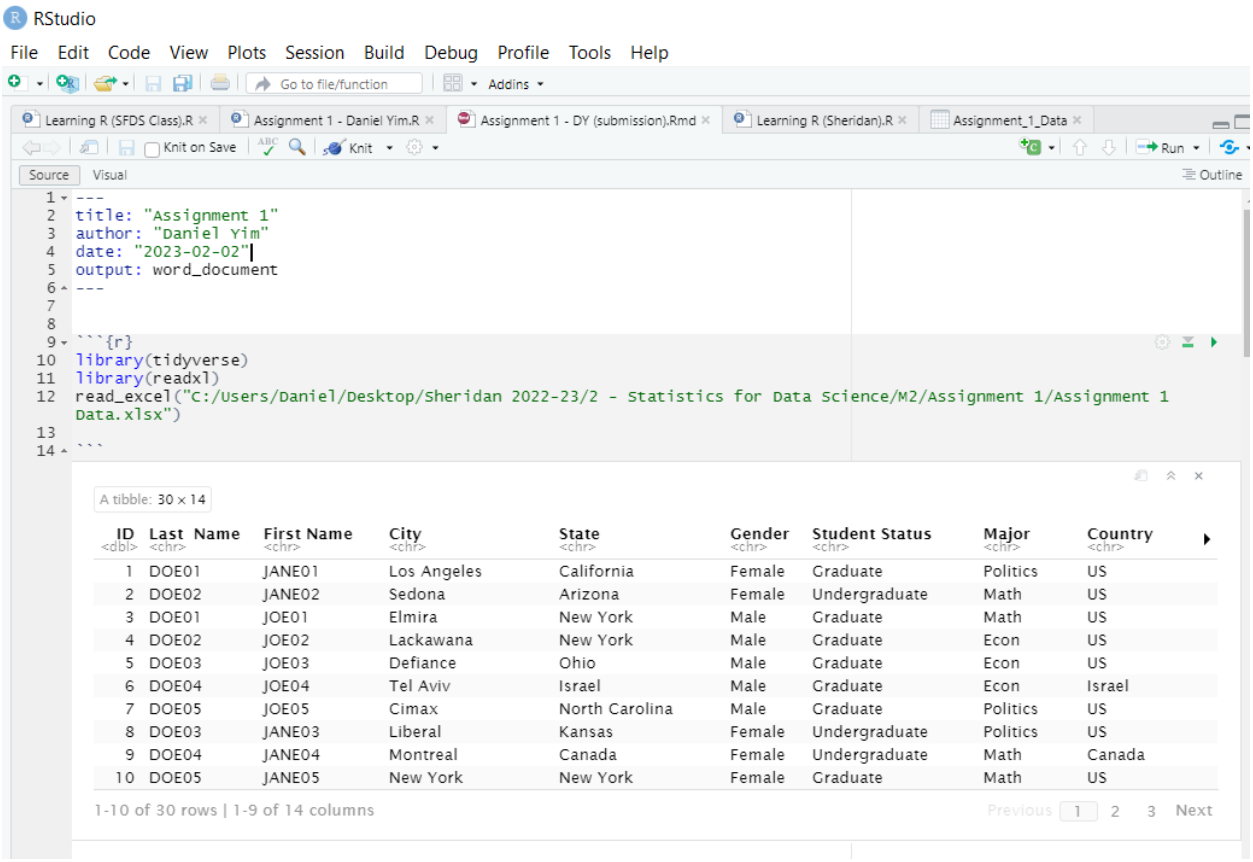
---
title: "Assignment 1"
author: "Daniel Yim"
date: "2023-02-02"
output: word_document

```{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")
```



Part 1

1. There are 14 variables

```{r}
ls(Assignment_1_Data)
```

```
 [1] "Age"                          "Average score (grade)"        "City"
 [4] "Country"                      "First Name"                   "Gender"
 [7] "Height (in)"                  "ID"                           "Last Name"
[10] "Major"                        "Newspaper readership (times/wk)" "SAT"
[13] "State"                        "Student Status"
```

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 <chr>	n <int>
Female	15
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 <chr>	n <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 <chr> | sum(`Newspaper readership (times/wk)`) <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))
```
```
```


```
       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
```

A tibble: 1 × 6

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

1 row

A tibble: 1 × 6

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

A tibble: 1 × 3

var(newspaper) <dbl>	sd(newspaper) <dbl>	Range (newspaper) <dbl>
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 × 3]

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

2 rows