Data_607_Project_2_Sangeetha_Sasikumar_Untidy_Dataset_Grades

Enid Roman

2022-10-08

ABOUT THE DATASET:

This dataset was created by Sangeetha Sasikumar.

Not sure where Sangeetha got the information from but it looks like grades based on gender and age.

```
# Upload the libraries needed.

library(tidyr)
library(tidyverse)
```

What age and gender has the highest average of them all?

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v dplyr   1.0.9
## v tibble  3.1.8     v stringr 1.4.1
## v readr   2.1.2     v forcats 0.5.2
## v purrr   0.3.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
library(ggplot2)
```

```
# Import the data from github.
# Link is provided to the csv file below:

#https://github.com/enidroman/data_607_data_aquisition_and_management_project/blob/main/Sangeetha%20Sas

urlfile <-'https://raw.githubusercontent.com/enidroman/data_607_data_aquisition_and_management_project/

grades <- read.csv(urlfile)
grades
```

```
##       Names   Age   Grades Gender
## 1     Sally     6 85.00000      1
## 2   Michael   "7" 80.66667      0
## 3 Elizabeth   "6" 92.00000      1
## 4   Anthony     7 93.50000      0
## 5     Mary  6.111 90.00010      1
## 6    Steven     7 91.00000      0
```

DATA CLEANING AND TRANSFORMATION

In observing the dataset I see that:

1. **Age column needs to be converted from character to integer.**

2. **Grades need to be converted from double to integer.**

3. **Gender should be converted from integer to character.**

4. **The special character from Age Column needs to be removed.**

5. **The decimals in Grades Columns needs to removed.**

6. **The 1 for Female and 0 for Male needs to be renamed to Female and Male.**

```
# Mutate function to remove the quotes from the integers in columns Age.

quotes <- grades %>%
  mutate(across(
    everything(),
    ~ map_chr(.x, ~ gsub("\"", "", .x))
  ))
quotes
```

7. **I don't think the dataset needs to be converted to pivot wide or pivot long.**

```
##       Names   Age   Grades Gender
## 1     Sally     6       85      1
## 2   Michael     7 80.66667      0
## 3 Elizabeth     6       92      1
## 4   Anthony     7     93.5      0
## 5     Mary  6.111  90.0001      1
## 6    Steven     7       91      0
```

```r
# Round the decimal to 0 for Age and Grades columns to remove the decimals.

decimal <- quotes
decimal$Age<-round(as.numeric(decimal$Age), 0)
decimal$Grades<-round(as.numeric(decimal$Grades), 0)
decimal
```

```
##       Names Age Grades Gender
## 1     Sally   6     85      1
## 2   Michael   7     81      0
## 3 Elizabeth   6     92      1
## 4   Anthony   7     94      0
## 5     Mary    6     90      1
## 6    Steven   7     91      0
```

```r
# Checked to make sure that all columns is in the proper Class.

numbers <- decimal
numbers$Age <- as.integer(numbers$Age)                # First column is a double.
numbers$Grades <- as.integer(numbers$Grades)          # Second column is a double.
numbers$Gender <- as.character(numbers$Gender)        # Third column is an integer.
sapply(numbers, class)
```

```
##       Names         Age      Grades      Gender
## "character"   "integer"   "integer" "character"
```

```r
# Transform 1 to Female and 0 to Male.

gender <- numbers
gender$Gender[gender$Gender == 1] <- "female"
gender$Gender[gender$Gender == 0] <- "male"
gender
```

```
##       Names Age Grades Gender
## 1     Sally   6     85 female
## 2   Michael   7     81   male
## 3 Elizabeth   6     92 female
## 4   Anthony   7     94   male
## 5     Mary    6     90 female
## 6    Steven   7     91   male
```

```r
# Rearranged the columns.

gender <- gender[, c("Names", "Gender", "Age", "Grades")]
gender
```

```
##       Names Gender Age Grades
## 1     Sally female   6     85
## 2   Michael   male   7     81
## 3 Elizabeth female   6     92
## 4   Anthony   male   7     94
## 5     Mary  female   6     90
## 6    Steven   male   7     91
```

Analysis

No analysis was requested on the discussion but I created my own analysis.

```
# Summary of each column.

summary(gender)
```

I see the dataframe is 6 rows in length. Names and Gender columns is class as characters. The Age column Min is 6.0, 1st Quarter is 6.0., Median 6.5, Mean is 6.5, 3rd Quarter is 7.0, the Max is 7.0. The Grades column Min is 81.0, 1st Quarter is 86.25, Median 90.50, Mean is 88.83, 3rd Quarter is 91.75, the Max is 94.00.

```
##     Names              Gender              Age          Grades
##  Length:6           Length:6           Min.   :6.0   Min.   :81.00
##  Class :character   Class :character   1st Qu.:6.0   1st Qu.:86.25
##  Mode  :character   Mode  :character   Median :6.5   Median :90.50
##                                        Mean   :6.5   Mean   :88.83
##                                        3rd Qu.:7.0   3rd Qu.:91.75
##                                        Max.   :7.0   Max.   :94.00
```

```
# Aggregate Function to compute summary statistic for subsets of the data, Grades by gender(Male and Fe

group_mean <- aggregate(x = gender$Grades,
                    by = list(gender$Gender),
                    FUN = mean)

colnames(group_mean) <- c("Gender", "Mean")

group_mean
```

Female has a higher average then the male.

```
##   Gender     Mean
## 1 female 89.00000
## 2   male 88.66667
```

```
# Aggregate Function to compute summary statistic for subsets of the data, Age by Gender(Female and Mal

group_mean <- aggregate(x = gender$Grades,
                    by = list(gender$Age),
                    FUN = mean)

colnames(group_mean) <- c("Age", "Mean")

group_mean
```

6 years old had a higher average then 7 year old.

```
##   Age     Mean
## 1   6 89.00000
## 2   7 88.66667
```

```
# Aggregate function to aggregate the sum to summarize the data frame based on the two variables, Outco

list_aggregate <- aggregate(gender$Grades, by = list(gender$Gender, gender$Age), FUN = sum)

colnames(list_aggregate) <- c("Gender", "Age", "Mean")

list_aggregate
```

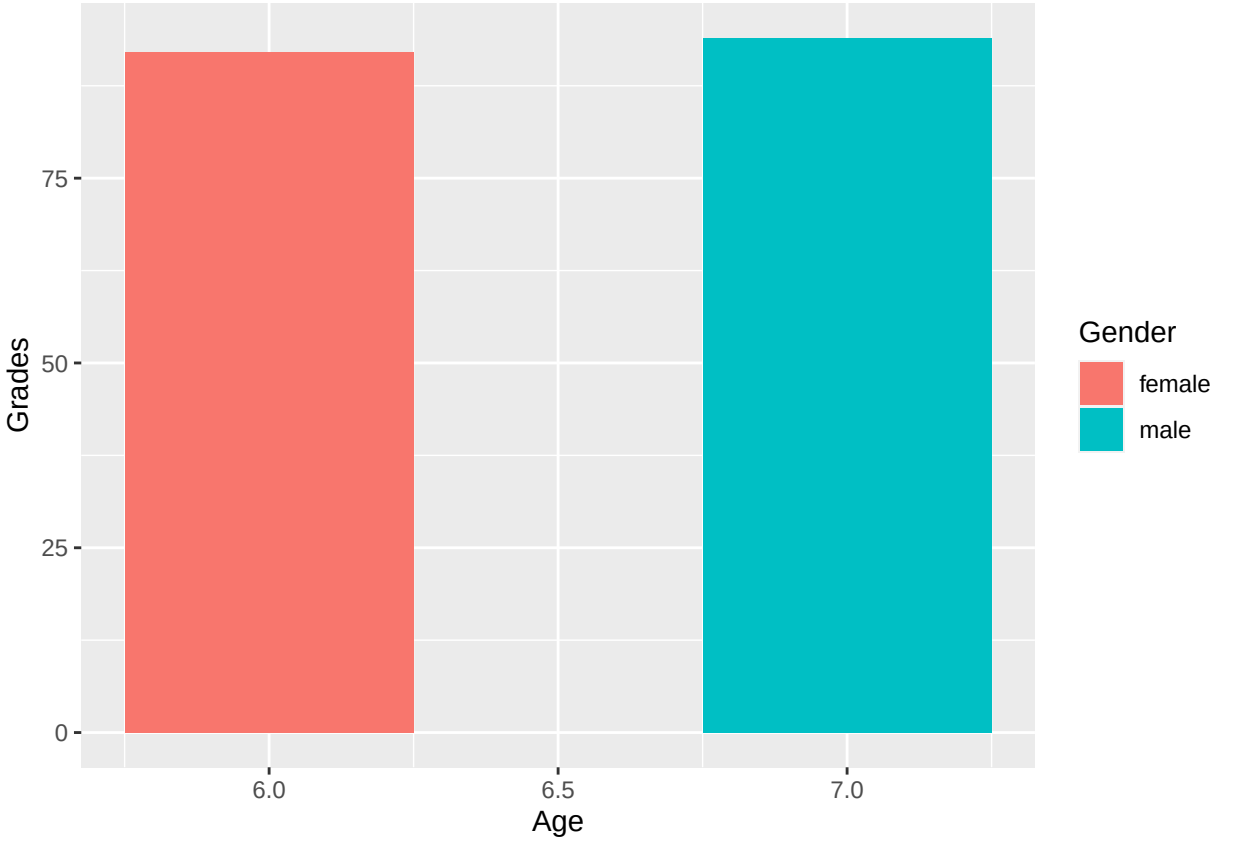
6 year old females had higher average then 7 year old male.

```
##   Gender Age Mean
## 1 female   6  267
## 2   male   7  266
```

```
# Bar graph by age, gender, and grades.

ggplot(gender, aes(x = Age, y = Grades, fill = Gender)) +
  geom_col(width = 0.5, position = "dodge")
```

Not much of a difference. Slightly a little higher in grades for the 7 year old male then female.



CONCLUSION

In my analysis I observed that in this dataset 6 year old females had higher average then 7 year old male. I would of prefer to know more information regarding the dataset and for the dataset to have more observations. It would of been nice to have to take it apart or convert it from long to wide or vice versa. Did alot of data cleaning.