# Project-1

Ayo

2022-07-29

## Importing the packages

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.1.3
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.7     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

```
## Warning: package 'ggplot2' was built under R version 4.1.3
```

```
## Warning: package 'tibble' was built under R version 4.1.3
```

```
## Warning: package 'tidyr' was built under R version 4.1.3
```

```
## Warning: package 'readr' was built under R version 4.1.3
```

```
## Warning: package 'purrr' was built under R version 4.1.3
```

```
## Warning: package 'dplyr' was built under R version 4.1.3
```

```
## Warning: package 'stringr' was built under R version 4.1.3
```

```
## Warning: package 'forcats' was built under R version 4.1.3
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(skimr)
```

```
## Warning: package 'skimr' was built under R version 4.1.3
```

## Reading the dataset

```
df <- read.csv('book_reviews.csv')
```

### Assessing the dataframe

```
dim(df)
```

```
## [1] 2000    4
```

```
colnames(df)
```

```
## [1] "book"   "review" "state"  "price"
```

### check the datatypes

```
for (col in colnames(df)) {
  print(class(df[[col]]))
  }
```

```
## [1] "character"
## [1] "character"
## [1] "character"
## [1] "numeric"
```

### Checking the distinct values

```
for (col in colnames(df)) {
  print(col)
  print(unique(df[[col]]))
}
```

```
## [1] "book"
## [1] "R Made Easy"                          "R For Dummies"
## [3] "Secrets Of R For Advanced Students" "Top 10 Mistakes R Beginners Make"
## [5] "Fundamentals of R For Beginners"
## [1] "review"
## [1] "Excellent" "Fair"      "Poor"      "Great"     NA          "Good"
## [1] "state"
## [1] "TX"         "NY"         "FL"         "Texas"      "California"
## [6] "Florida"    "CA"         "New York"
## [1] "price"
## [1] 19.99 15.99 50.00 29.99 39.99
```

```
table(df$review)
```

```
##
## Excellent      Fair      Good     Great      Poor
##       345       369       363       349       368
```

```
new_df <- df %>%
  filter(!is.na(review))
```

### Cleaning the dataset

```
clean_df <- new_df %>%
  mutate(
    state = case_when(
      state == 'California' ~ 'CA',
      state == 'New York' ~ 'NY',
      state == 'Texas' ~ 'TX',
      state == 'Florida' ~ 'FL',
      TRUE ~ state
    )
  )
```

```
clean_df <- clean_df %>%
  mutate(
    review_num = case_when(
      review == "Poor" ~ 1,
      review == "Fair" ~ 2,
      review == "Good" ~ 3,
      review == "Great" ~ 4,
      review == "Excellent" ~ 5
      ),
    high_review = if_else(review_num >= 4, TRUE, FALSE)

)
```

## Objective of analysis

### What book is the most profitable book

### The most purchased book

```
clean_df %>%
  group_by(book) %>%
  summarise(revenue = sum(price)) %>%
  arrange(-revenue)
```

```
## # A tibble: 5 x 2
##   book                                revenue
##   <chr>                                 <dbl>
## 1 Secrets Of R For Advanced Students  18000
```

```
## 2 Fundamentals of R For Beginners     14636.
## 3 Top 10 Mistakes R Beginners Make    10646.
## 4 R Made Easy                          7036.
## 5 R For Dummies                        5772.
```

```
clean_df %>%
  group_by(book) %>%
  summarise(sold = n()) %>%
  arrange(-sold)
```

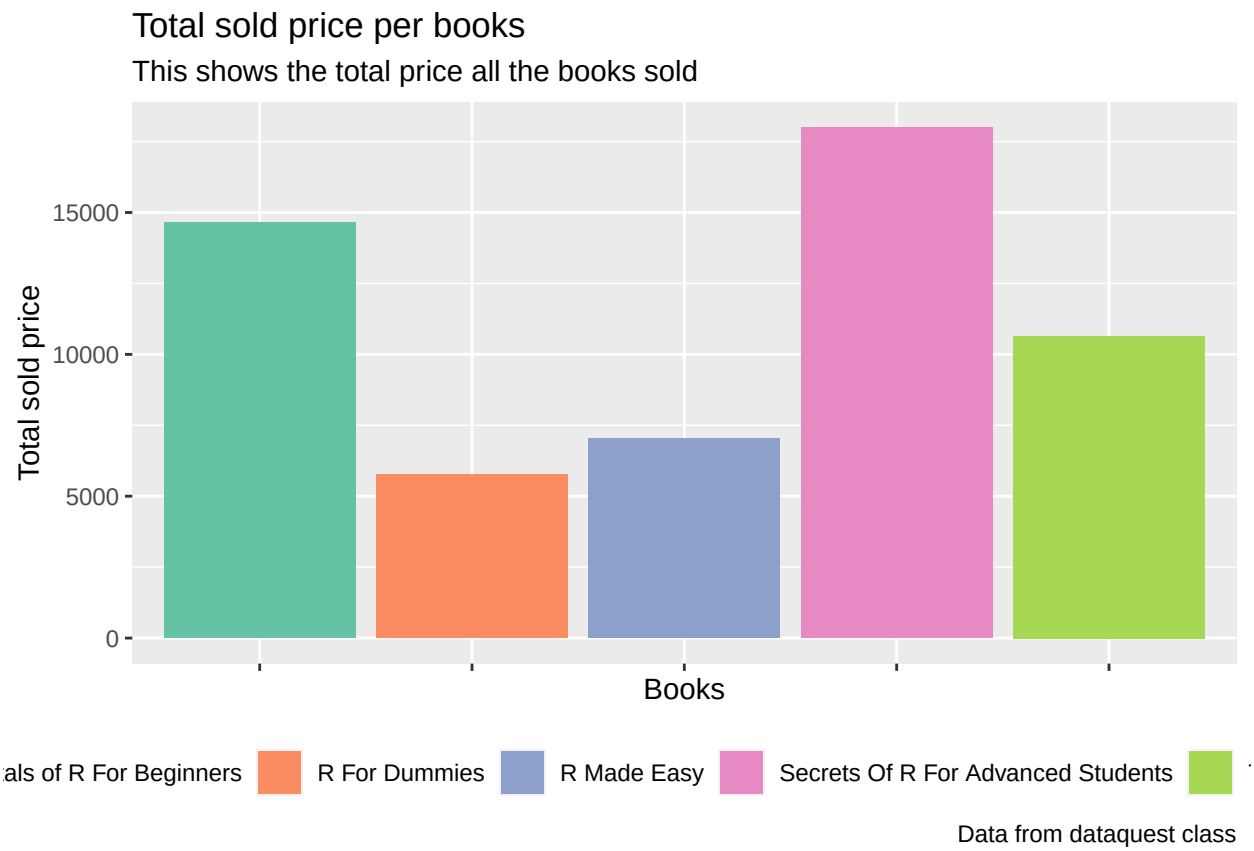
```
## # A tibble: 5 x 2
##   book                                sold
##   <chr>                              <int>
## 1 Fundamentals of R For Beginners      366
## 2 R For Dummies                        361
## 3 Secrets Of R For Advanced Students   360
## 4 Top 10 Mistakes R Beginners Make     355
## 5 R Made Easy                          352
```

Here we see the highest sold books is Fundamental of R For Beginners

## Data visualization

```
library(ggplot2)
analysis <- clean_df %>%
  group_by(book) %>%
  summarise(revenue = sum(price)) %>%
  arrange(-revenue) %>%
  ggplot(aes(x=book, y=revenue, fill=book)) + geom_col(position = 'dodge') +
  labs(title = "Total sold price per books",
       subtitle = 'This shows the total price all the books sold',
       x= "Books",
       y= 'Total sold price',
       caption= "Data from dataquest class") + theme(axis.text.x = element_blank())
```

```
analysis <- analysis + scale_fill_brewer(palette="Set2") + theme(legend.position = 'bottom')
analysis
```




From the result above we can see that the most profitable book is secrets of R for Advanced students