---
title: "Creating an efficient data analysis workflow"
output:
  pdf_document: default
  html_document: default

Introduction

We are presented with a dataset from a book store to determine which book from a category was most profitable from a campaign. This dataset can be obtained from
https://data.world/dataquest/book-reviews

Load tidyverse and dataset

```{r}
library(tidyverse)

bookdf <- read.csv("book_reviews.csv")
view(bookdf)
```
Reviewing the data set
```{r}
dim(bookdf)
colnames(bookdf)

typeof(bookdf[["book"]])
typeof(bookdf$review)
typeof(bookdf[["state"]])
typeof(bookdf[["price"]])

unique(bookdf[["book"]])
unique(bookdf[["review"]])
unique(bookdf[["state"]])
unique(bookdf[["price"]])
```
Following an extensive review of the dataset, the dataset has 2,000 rows and 4 columns.
The column names for the dataset are, book: this represents the name of the book; review represents the review rank the book received; state represents the location the book was bought; and price represents the cost of the book.
The columns are of character types except for the price column which is of double.

Reviewing the unique values in each column, there are 5 distinct books in the dataset: R made easy, R for dummies, Secrets of R for advanced students, top 10 mistakes R beginners make, and fundamentals of R for beginners; the reviews column have 5 entries, excellent, fair, poor, great and good, however, there are null entries in this column, so we have to plan on handling null values; the state entry has Texas, New York, Florida, California entries. Some state entries are inconsistent. We need to harmonize all state entries to 2 state codes. The prices range from 15.99 as the lowest entry to 50.00 as the hightest entry in the price column.

Handling missing values
```{r}
bookdf_complete <- na.omit(bookdf)

view(bookdf_complete)
```
Eliminating the null values eliminated about 200 entries.

Handling State naming conventions
```{r}
bookdf_complete <- bookdf_complete%>%
  mutate(
    state = case_when(
      state == "Texas" ~ "TX",
      state == "California" ~ "CA",
      state == "Florida" ~ "FL",
      state == "New York" ~ "NY",
      TRUE ~ state
    )
  )

view(bookdf_complete)
```
Handling the Review column
```{r}
bookdf_complete <- bookdf_complete %>%
  mutate(
    review_num = case_when(
      review == "Poor" ~ 1,
      review == "Fair" ~ 2,
      review == "Good" ~ 3,
      review == "Great" ~ 4,
      review == "Excellent" ~ 5
    ),
    is_high_review = if_else(review_num >=4, TRUE, FALSE)
      )
view(bookdf_complete)
```

determining the most profitable book
The most profitable book could either be the book with the highest count sold or the book with the highest price (but we do not know the margin on each book)
```{r}
bookdf_complete <- bookdf_complete %>%
  mutate(
    book_num = case_when(
      book == "R Made Easy" ~ 1,
      book == "R For Dummies" ~ 1,
      book == "Secrets Of R For Advanced Students" ~ 1,
      book == "Top 10 Mistakes R Beginners Make" ~ 1,
      book == "Fundamentals of R For Beginners" ~ 1,
    )
  )
```

```{r}
summary_book <- bookdf_complete %>%
  group_by(book) %>%
  summarize(
    sum_book = sum(book_num)

  )%>%
  arrange(-sum_book)

summary_book
```

Conclusion
The objective was to determine which book was the most profitable
book for the company. After cleaning the dataset, I determined that the metric which aligned most to the objective was to figure out which book had the most sale.
Fundamentals of R for Beginners had both a high review rating and it out-sold all other books.