Introduction

We will be acting as a data analyst for a company that sells books for learning programming. This company has produced multiple books, and each has received many reviews. This company wants us to check out the sales data and see if we can extract any useful information from it.

Getting Familiar With The Data

```
# importing the csv file
library(readr)
review_df <- read_csv("book_reviews.csv")
```

```
# finding dimension of data frame
dimension_ofdataset <- dim(review_df)
print(dimension_ofdataset)
```

```
# finding the number of columns in this data frame
ncol <- ncol(review_df)
print(ncol)
```

```
# finding the number of columns in this data frame
nrow <- nrow(review_df)
print(nrow)
```

```
# finding the name of columns in this data frame
vector_cols <- colnames(review_df)
print(vector_cols)
```

```
# finding the data type of each column by using typeof() function

book_types <- c()

for (title in colnames(review_df)) {
  type <- typeof(review_df[[title]])
  book_types <- c(book_types, type)
}
print(book_types)
```

```
# learning about the a tibble columns, types and dimensions
library(tibble)
glimpse(review_df)
```

```
# finding the unique value in each of columns
for (col in colnames(review_df)) {
  print("Unique values in the column:")
  print(col)
  print(unique(review_df[[col]]))
  print("")
}
```

```
unique(review_df[['book']])
```

Handling Missing Data

Examine the data and get and understanding of which columns have data missing.

```
# getting the number of missing data
complete_books <- review_df %>%
  filter(!(is.na(review)))

print(nrow(complete_books))
```

Dealing With Inconsistent Labels

```
library(tidyverse)
complete_books <- complete_books %>%
  mutate(
    state = case_when(
      state == "New York" ~ "NY",
      state == "Florida" ~ "FL",
      state == "Texas" ~ "TX",
      state == "California" ~ "CA",
      TRUE ~ state  # retain remaining entries that are already abbreviated
      )
  )

print(unique(complete_books$state))
```

Transforming The Review Data

```
complete_books <- complete_books %>%
  mutate(
    score = case_when(
      review == "Poor" ~ 1,
      review == "Fair" ~ 2,
      review == "Good" ~ 3,
      review == "Great" ~ 4,
      review == "Excellent" ~ 5
    ),
    is_high_review = if_else(score >= 4,TRUE, False)
  )

glimpse(complete_books)
```

Analyzing Data

```
# number of books sales
book_sales <- complete_books %>%
  group_by(book) %>%
```

```
  summarise(
    sales = n(),
    revenue = sum(price),
    avg_sales = mean(revenue/sales)
  ) %>%
  arrange(-avg_sales)
```

```
head(book_sales)
```

Reporting The Results

We can see that the number of sales is similar for all books, with the minimum being 352 sales for "R Made Easy", and the maximum being 366 for "Fundamentals of R For Beginners".

The revenue from each book is much more disparate, with "Secrets Of R For Advanced Students" bringing in the most revenue of $18,000.00 and also having the highest average sales price of $50.00.

If to work book titles from lowest revenue to highest, lower revenue books will look like entry-level book to R, with increased revenue with the level of difficulty the book contains. This is an observation and requires further data for analysis.