Guided Project Solutions: Creating An Efficient Data Analysis Workflow

Husen Wahyu

3/7/2021

Load Library and Dataset

```
library(tidyverse)
```

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

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.2     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

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

```
book_reviews <- read.csv("book_reviews.csv")
```

Page 2 Mengenal Data

```
# How big is dataset?
dim(book_reviews)
```

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

```
sprintf("there are %s rows and %s columns", dim(book_reviews)[1], dim(book_reviews)[2])
```

```
## [1] "there are 2000 rows and 4 columns"
```

```
# What are the column names?
colnames(book_reviews)
```

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

```r
# What are the types of each of the columns?
types <- c()
for (c in colnames(book_reviews)) {
  types <- c(types, typeof(book_reviews[[c]]))
}

# What are the unique values are present in each of the columns?
for (c in colnames(book_reviews)){
  print(c)
  print(unique(book_reviews[[c]]))
  print("")
}
```

```
## [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] ""
## [1] "review"
## [1] "Excellent" "Fair"      "Poor"      "Great"     NA          "Good"
## [1] ""
## [1] "state"
## [1] "TX"         "NY"         "FL"         "Texas"      "California"
## [6] "Florida"    "CA"         "New York"
## [1] ""
## [1] "price"
## [1] 19.99 15.99 50.00 29.99 39.99
## [1] ""
```

Page 3 Dealing with Missing Data (Alternatif 1 Hapus Row)

```r
complete_book_reviews <- book_reviews %>%
  filter(!is.na(review),
         !is.na(book),
         !is.na(state),
         !is.na(price))

dim(complete_book_reviews)
```

```
## [1] 1794    4
```

Page 4 Dealing with incosistent data

```r
unique(complete_book_reviews[["state"]])
```

```
## [1] "TX"         "NY"         "FL"         "Texas"      "Florida"
## [6] "CA"         "California" "New York"
```

```
complete_book_reviews <-  complete_book_reviews %>%
  mutate(
    state_cor =
      case_when(
        state == "Texas" ~ "TX",
        state == "New York" ~ "NY",
        state == "Florida" ~ "FL",
        state == "California" ~ "CA",
        TRUE ~ state
      )

  )
```

Page 5 Mengubah Data

```
complete_book_reviews <- complete_book_reviews %>%
  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)
  )
```

Page 6 Mencari Profitable Book

we will find the most profitable book by finding the most number of money generated by the sales of the book. Which means, total purchases x price.

```
complete_book_reviews %>%
  group_by(book) %>%
  summarise(
    puchased = n(),
    mean_price = mean(price),
    sales = n() * mean(price)
  ) %>%
  arrange(-sales)
```

```
## # A tibble: 5 x 4
##   book                                puchased mean_price  sales
##   <chr>                                  <int>      <dbl>  <dbl>
## 1 Secrets Of R For Advanced Students       360       50   18000
## 2 Fundamentals of R For Beginners          366       40.0 14636.
## 3 Top 10 Mistakes R Beginners Make         355       30.0 10646.
```

```
## 4 R Made Easy                          352       20.0  7036.
## 5 R For Dummies                        361       16.0  5772.
```

The most profitable book is Secret of R for Advanced Students