Data Analysis Workflow

RZ

6/17/2021

Load require libraries

```
library(readr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Check the data

```
setwd("~/Downloads/Dataquest/R")
data <- read.csv('book_reviews.csv')
print(dim(data))
```

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

```
print(nrow(data))
```

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

```
print(ncol(data))
```

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

```
#Coloumn names
print(colnames(data))
```

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

```
# Types of each columns
col_type = c()
for (i in colnames(data)){
  col_type <- c(col_type, typeof(i))
}
print(col_type)
```

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

```
# Unique value in each columns
for (i in colnames(data)){
  print(i)
```

```
  print(unique(data[[i]]))
}
```

```
## [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
```

Data Cleaning

```
#Check the data with missing and remove missing data
data_nona <- data %>% filter(!is.na(review))

#Show the dimension of the data set
dim(data_nona)
```

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

Deal with the inconsistent data

```
data_ab <- data_nona %>% mutate(
  state = case_when(
    state =='California' ~ 'CA',
    state == 'New York' ~ 'NY',
    state == 'Texas' ~ 'TX',
    state == 'Florida' ~ 'FL',
    TRUE ~ state #ignore already abbreviation
  )
)
```

Create a new column with numerical rate

```
data_ab <- data_ab %>% 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)
)
```

Analyze the data

```
data_gp <- data_ab %>% group_by(book) %>% summarise(
  sum_1 = sum(price),
  cou_1 = n() #Number of observations in current group
)
head(data_gp)
```

```
## # A tibble: 5 x 3
##   book                                sum_1 cou_1
##   <chr>                               <dbl> <int>
## 1 Fundamentals of R For Beginners    14636.   366
## 2 R For Dummies                       5772.   361
## 3 R Made Easy                         7036.   352
## 4 Secrets Of R For Advanced Students 18000    360
## 5 Top 10 Mistakes R Beginners Make   10646.   355
```