sales2019

Jay Parikh

30/12/2020

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.4     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
```

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

```
library(lubridate)
```

```
##
## Attaching package: ’lubridate’
```

```
## The following objects are masked from ’package:base’:
##
##     date, intersect, setdiff, union
```

```
sales <- read_csv("sales2019.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   date = col_character(),
##   user_submitted_review = col_character(),
##   title = col_character(),
##   total_purchased = col_double(),
##   customer_type = col_character()
## )
```

```
dim(sales)
```

```
## [1] 5000    5
```

```
glimpse(sales)
```

```
## Rows: 5,000
## Columns: 5
## $ date                 <chr> "5/22/19", "11/16/19", "6/27/19", "2011-06-19...
## $ user_submitted_review <chr> "it was okay", "Awesome!", "Awesome!", "Aweso...
## $ title                <chr> "Secrets Of R For Advanced Students", "R For ...
## $ total_purchased      <dbl> 7, 3, 1, 3, NA, 1, 5, NA, 7, 1, 7, NA, 3, 2, ...
## $ customer_type        <chr> "Business", "Business", "Individual", "Indivi...
```

```
#step 1: get rid of missing data

sales <- sales %>% filter(!(is.na(user_submitted_review)))
dim(sales)
```

```
## [1] 4115    5
```

```
sales_mean <- sales %>% filter(!(is.na(total_purchased))) %>% pull(total_purchased) %>% mean
#pull function selects a column in a data frame and transforms it into a vector

sales <- sales %>% mutate(total_purchased = if_else(is.na(total_purchased), sales_mean, total_purchased)

unique(sales$user_submitted_review)
```

```
## [1] "it was okay"
## [2] "Awesome!"
## [3] "Hated it"
## [4] "Never read a better book"
## [5] "OK"
## [6] "The author's other books were better"
## [7] "A lot of material was not needed"
## [8] "Would not recommend"
## [9] "I learned a lot"
```

```
is_positive <- function(sentence) {
  case_when(
    str_detect(sentence, "okay") ~ TRUE,
    str_detect(sentence, "Awesome") ~ TRUE,
    str_detect(sentence, "Never") ~ TRUE,
    str_detect(sentence, "learned") ~ TRUE,
    str_detect(sentence, "OK") ~ TRUE,
    TRUE ~ FALSE # if none of the above
  )
}

sales <- sales %>% mutate(is_positive = unlist(map(user_submitted_review, is_positive)))

sales <- sales %>% mutate(pre_or_post = if_else(mdy(date) < ymd("2019/07/01"), "Pre", "Post"))
```

```
## Warning: Problem with ‘mutate()‘ input ‘pre_or_post‘.
## i  1604 failed to parse.
## i Input ‘pre_or_post‘ is ‘if_else(mdy(date) < ymd("2019/07/01"), "Pre", "Post")‘.
## Warning: 1604 failed to parse.
```

```
sales_title <- sales %>% group_by(pre_or_post, title) %>% summarise(books_sold = sum(total_purchased))
```

```
## ‘summarise()‘ regrouping output by ’pre_or_post’ (override with ‘.groups‘ argument)
```

```
sales_customer <- sales %>% group_by(pre_or_post, customer_type) %>% summarise(books_sold = sum(total_pu
```

```
## ‘summarise()‘ regrouping output by ’pre_or_post’ (override with ‘.groups‘ argument)
```

```
sales_is_positive <- sales %>% group_by(pre_or_post) %>% summarise(is_positive = sum(is_positive))
```

```
## ‘summarise()‘ ungrouping output (override with ‘.groups‘ argument)
```