

# Week 2 Exercises

James Buchholz

November 6, 2022

Please complete all exercises below. You may use stringr, lubridate, or the forcats library.

Place this at the top of your script: library(stringr) library(lubridate) library(forcats)

## Exercise 1

Read the sales\_pipe.txt file into an R data frame as sales.

```
library(stringr)
library(lubridate)

## Warning: package 'lubridate' was built under R version 4.2.2
##
## Attaching package: 'lubridate'
## The following objects are masked from 'package:base':
##
##      date, intersect, setdiff, union
library(forcats)

sales_pipe <- "~/GitHub/DSE5002/Week_2/Data/sales_pipe.txt"

sales <- read.table(sales_pipe,
                    header = T,
                    fill = T,
                    sep = "|",
                    fileEncoding = 'WINDOWS-1252')
```

## Exercise 2

You can extract a vector of columns names from a data frame using the colnames() function. Notice the first column has some odd characters. Change the column name for the FIRST column in the sales date frame to Row.ID.

**Note:** You will need to assign the first element of colnames to a single character.

```
colnames(sales)[1] <- "Row.ID"
```

## Exercise 3

Convert both Ship.Date and Order.Date to date vectors within the sales data frame. What is the number of days between the most recent order and the oldest order? How many years is that? How many weeks?

**Note:** Use lubridate

```
install.packages('lubridate')

## Warning: package 'lubridate' is in use and will not be installed
library(lubridate)

sales$Ship.Date <- mdy(sales$Ship.Date)
inherits(sales$Ship.Date, c("Date"))

## [1] TRUE

sales$Order.Date <- as.Date(sales$Order.Date, format = "%m/%d/%Y")
inherits(sales$Order.Date, c("Date"))

## [1] TRUE

min_date <- min(sales$Order.Date)
max_date <- max(sales$Order.Date)

difftime(max_date, min_date, units = 'days')

## Time difference of 1455 days

year_diff <- time_length(difftime(max_date, min_date), 'years')
year_diff

## [1] 3.983573

difftime(max_date, min_date, units = 'weeks')

## Time difference of 207.8571 weeks
```

## Exercise 4

What is the average number of days it takes to ship an order?

```
list_of_days <- as.integer(format(as.Date(sales$Ship.Date), "%d"))
mean(list_of_days)

## [1] 15.57785
```

## Exercise 5

How many customers have the first name Bill? You will need to split the customer name into first and last name segments and then use a regular expression to match the first name bill. Use the length() function to determine the number of customers with the first name Bill in the sales data.

```
name_split <- stringr::str_split(string = sales$Customer.Name, pattern = " ", n = Inf, simplify = T)
bill_list <- name_split[name_split %in% "Bill"]
length(bill_list)

## [1] 30
```

## Exercise 6

How many mentions of the word 'table' are there in the Product.Name column? **Note you can do this in one line of code**

```
length(grep("table", sales$Product.Name))
```

```
## [1] 156
```

## Exercise 7

Create a table of counts for each state in the sales data. The counts table should be ordered alphabetically from A to Z.

```
state_table <- table(sales$State)
state_table
```

```
##
##      Alabama      Arizona      Arkansas      California      Colorado
##          22          83          24          790          66
## Connecticut Delaware      Florida      Georgia      Idaho
##          30          34          134          68          7
## Illinois      Indiana      Iowa      Kansas      Kentucky
##         202          42          10          7          54
## Louisiana      Maryland Massachusetts      Michigan      Minnesota
##          19          43          36          107          38
## Mississippi      Missouri      Montana      Nebraska      Nevada
##          13          16          8          18          17
## New Hampshire      New Jersey      New Mexico      New York North Carolina
##          8          41          11          392          73
## Ohio      Oklahoma      Oregon      Pennsylvania      Rhode Island
##         192          26          36          246          14
## South Carolina South Dakota      Tennessee      Texas      Utah
##          20          2          69          386          24
## Virginia      Washington West Virginia      Wisconsin      Wyoming
##          75          150          4          35          1
```

## Exercise 8

Create an alphabetically ordered barplot for each sales Category in the State of Texas.

```
category_table <- table(sales$Category, sales$State)
category_table
```

```
##
##      Alabama Arizona Arkansas California Colorado Connecticut
## Furniture          4          20          2          182          20          2
## Office Supplies    13          44          16          481          33          20
## Technology         5          19          6          127          13          8
##
##      Delaware Florida Georgia Idaho Illinois Indiana Iowa Kansas
## Furniture          5          34          10          3          51          4          1          2
## Office Supplies    20          76          46          2          120          33          7          3
## Technology         9          24          12          2          31          5          2          2
##
```

```
##      Kentucky Louisiana Maryland Massachusetts Michigan Minnesota
## Furniture      14         1      13           4      18         5
## Office Supplies 36        10      22          26      71        27
## Technology     4         8       8           6      18         6
##
##      Mississippi Missouri Montana Nebraska Nevada New Hampshire
## Furniture      4         2       1           1       3         4
## Office Supplies 6         9       5          13      11         4
## Technology     3         5       2           4       3         0
##
##      New Jersey New Mexico New York North Carolina Ohio Oklahoma
## Furniture      8         0      88          18      37         5
## Office Supplies 26        7     234          40     107        15
## Technology     7         4      70          15      48         6
##
##      Oregon Pennsylvania Rhode Island South Carolina South Dakota
## Furniture      7         47          4           2         1
## Office Supplies 23        152          8          14         1
## Technology     6         47          2           4         0
##
##      Tennessee Texas Utah Virginia Washington West Virginia
## Furniture     16      80   3       21          33         1
## Office Supplies 45     225  16      42          86         3
## Technology     8      81   5       12          31         0
##
##      Wisconsin Wyoming
## Furniture     11       1
## Office Supplies 16       0
## Technology     8       0
```

```
cat_nums <- c(80, 225, 81)
cat_name <- c("Furniture", "Office Supplies", "Technology")
cat_table_texas <- data.frame(cat_name, cat_nums)
cat_table_texas
```

```
##      cat_name cat_nums
## 1 Furniture      80
## 2 Office Supplies 225
## 3 Technology     81
```

```
barplot(cat_nums,
        main = "Each sales Category in the State of Texas",
        xlab = "Category",
        ylab = "Number of instances",
        names.arg = cat_name)
```

## Each sales Category in the State of Texas



### Exercise 9

Find the average profit by region. **Note:** You will need to use the `aggregate()` function to do this. To understand how the function works type `?aggregate` in the console.

```
avg_prof_by_reg <- aggregate(Profit ~ Region, data = sales, mean)
avg_prof_by_reg
```

```
##      Region  Profit
## 1 Central 21.57428
## 2   East 21.24482
## 3  South 31.99898
## 4   West 36.97532
```

### Exercise 10

Find the average profit by order year. **Note:** You will need to use the `aggregate()` function to do this. To understand how the function works type `?aggregate` in the console.

```
order_year <- format(as.Date(sales$Order.Date, format = "%m/%d/%Y"), "%Y")
avg_prof_by_year <- aggregate(Profit ~ order_year, data = sales, mean)
avg_prof_by_year
```

```
##   order_year  Profit
## 1      2014 19.74198
## 2      2015 14.78681
```

## 3	2016 34.13981
## 4	2017 37.27987