Week 2 Exercises

Amanda Frithsen

July 8, 2024

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)
```

```
##
## Attaching package: 'lubridate'

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

```
library(forcats)
```

```
sales <- read.delim("sales_pipe.txt"
                    ,stringsAsFactors=FALSE
                    ,sep = "|"
                    ,fileEncoding = "UTF-8"
                   )
```

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] "ï..Row.ID"     "Order.ID"      "Order.Date"    "Ship.Date"
##  [5] "Ship.Mode"     "Customer.ID"   "Customer.Name" "Segment"
##  [9] "Country"       "City"          "State"         "Postal.Code"
## [13] "Region"        "Product.ID"    "Category"      "Sub.Category"
## [17] "Product.Name"  "Sales"         "Quantity"      "Discount"
## [21] "Profit"
```

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

colnames(sales)
```

```
##  [1] "Row.ID"        "Order.ID"      "Order.Date"    "Ship.Date"
##  [5] "Ship.Mode"     "Customer.ID"   "Customer.Name" "Segment"
##  [9] "Country"       "City"          "State"         "Postal.Code"
## [13] "Region"        "Product.ID"    "Category"      "Sub.Category"
## [17] "Product.Name"  "Sales"         "Quantity"      "Discount"
## [21] "Profit"
```

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

```
# Convert Ship.Date to date vector
inherits(sales$Ship.Date, c("Date"))
```

```
## [1] FALSE
```

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

```
## [1] TRUE
```

```
# Convert Order.Date to date vector
inherits(sales$Order.Date, c("Date"))
```

```
## [1] FALSE
```

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

```
## [1] TRUE
```

```
# Order date of most recent order and oldest order
newest_order <- sales$Order.Date[which.max(sales$Order.Date)]
oldest_order <- sales$Order.Date[which.min(sales$Order.Date)]
```

```
# There are 1,457 days between the most recent order and the oldest order.
# There are 3 years, 11 months, 27 days between the most recent order and the oldest order.
# There are 208 weeks, 1 day between the most recent order and the oldest order.

order_span <- interval(ymd(oldest_order), ymd(newest_order))
order_span
```

```
## [1] 2014-01-03 UTC--2017-12-30 UTC
```

```
days <- as.period(order_span, unit = "days")
days
```

```
## [1] "1457d 0H 0M 0S"
```

```
years <- as.period(order_span, unit = "years")
years
```

```
## [1] "3y 11m 27d 0H 0M 0S"
```

```
weeks <- interval(ymd(oldest_order), ymd(newest_order))/weeks(1)
weeks
```

```
## [1] 208.1429
```

Exercise 4

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

```
diff_time <- interval(ymd(sales$Order.Date), ymd(sales$Ship.Date))/days(1)
average_ship_days <- sum(diff_time)/length(diff_time)
average_ship_days
```

```
## [1] 3.908482
```

```
# The average number of days it takes to ship an order is approximately 3.91 days.
```

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.

```
temp_name <- str_split_fixed(string = sales$Customer.Name,
                             pattern = " ", n = 2)
sales$Customer.First.Name <- temp_name[,1]
sales$Customer.Last.Name <- temp_name[,2]
head(sales)
```

```
##   Row.ID       Order.ID Order.Date  Ship.Date      Ship.Mode Customer.ID
## 1      1 CA-2016-152156 2016-11-08 2016-11-11   Second Class    CG-12520
## 2      2 CA-2016-152156 2016-11-08 2016-11-11   Second Class    CG-12520
## 3      3 CA-2016-138688 2016-06-12 2016-06-16   Second Class    DV-13045
## 4      4 US-2015-108966 2015-10-11 2015-10-18 Standard Class    SO-20335
## 5      5 US-2015-108966 2015-10-11 2015-10-18 Standard Class    SO-20335
## 6      6 CA-2014-115812 2014-06-09 2014-06-14 Standard Class    BH-11710
##     Customer.Name   Segment       Country            City      State
## 1     Claire Gute  Consumer United States       Henderson   Kentucky
## 2     Claire Gute  Consumer United States       Henderson   Kentucky
## 3 Darrin Van Huff Corporate United States     Los Angeles California
## 4  Sean O'Donnell  Consumer United States Fort Lauderdale    Florida
## 5  Sean O'Donnell  Consumer United States Fort Lauderdale    Florida
## 6 Brosina Hoffman  Consumer United States     Los Angeles California
##   Postal.Code Region      Product.ID        Category Sub.Category
## 1       42420  South FUR-BO-10001798       Furniture    Bookcases
## 2       42420  South FUR-CH-10000454       Furniture       Chairs
## 3       90036   West OFF-LA-10000240 Office Supplies       Labels
## 4       33311  South FUR-TA-10000577       Furniture       Tables
## 5       33311  South OFF-ST-10000760 Office Supplies      Storage
## 6       90032   West FUR-FU-10001487       Furniture  Furnishings
##                                                  Product.Name    Sales
## 1                             Bush Somerset Collection Bookcase 261.9600
## 2     Hon Deluxe Fabric Upholstered Stacking Chairs, Rounded Back 731.9400
## 3       Self-Adhesive Address Labels for Typewriters by Universal  14.6200
## 4                  Bretford CR4500 Series Slim Rectangular Table 957.5775
## 5                               Eldon Fold 'N Roll Cart System  22.3680
## 6 Eldon Expressions Wood and Plastic Desk Accessories, Cherry Wood  48.8600
##   Quantity Discount    Profit Customer.First.Name Customer.Last.Name
## 1        2     0.00   41.9136              Claire               Gute
## 2        3     0.00  219.5820              Claire               Gute
## 3        2     0.00    6.8714              Darrin           Van Huff
## 4        5     0.45 -383.0310                Sean          O'Donnell
## 5        2     0.20    2.5164                Sean          O'Donnell
## 6        7     0.00   14.1694             Brosina            Hoffman
```

```
sales <- sales[, c(1:6, 23, 22, 8:21)]
head(sales)
```

```
##   Row.ID       Order.ID Order.Date  Ship.Date      Ship.Mode Customer.ID
## 1      1 CA-2016-152156 2016-11-08 2016-11-11   Second Class    CG-12520
## 2      2 CA-2016-152156 2016-11-08 2016-11-11   Second Class    CG-12520
## 3      3 CA-2016-138688 2016-06-12 2016-06-16   Second Class    DV-13045
## 4      4 US-2015-108966 2015-10-11 2015-10-18 Standard Class    SO-20335
## 5      5 US-2015-108966 2015-10-11 2015-10-18 Standard Class    SO-20335
## 6      6 CA-2014-115812 2014-06-09 2014-06-14 Standard Class    BH-11710
##   Customer.Last.Name Customer.First.Name   Segment       Country
## 1               Gute              Claire  Consumer United States
## 2               Gute              Claire  Consumer United States
## 3          Van Huff              Darrin Corporate United States
## 4          O'Donnell                Sean  Consumer United States
## 5          O'Donnell                Sean  Consumer United States
## 6            Hoffman             Brosina  Consumer United States
##             City      State Postal.Code Region      Product.ID        Category
```

```
## 1       Henderson   Kentucky       42420  South FUR-BO-10001798       Furniture
## 2       Henderson   Kentucky       42420  South FUR-CH-10000454       Furniture
## 3     Los Angeles California       90036   West OFF-LA-10000240 Office Supplies
## 4 Fort Lauderdale    Florida       33311  South FUR-TA-10000577       Furniture
## 5 Fort Lauderdale    Florida       33311  South OFF-ST-10000760 Office Supplies
## 6     Los Angeles California       90032   West FUR-FU-10001487       Furniture
##   Sub.Category                                                  Product.Name
## 1    Bookcases                                Bush Somerset Collection Bookcase
## 2       Chairs      Hon Deluxe Fabric Upholstered Stacking Chairs, Rounded Back
## 3       Labels        Self-Adhesive Address Labels for Typewriters by Universal
## 4       Tables                    Bretford CR4500 Series Slim Rectangular Table
## 5      Storage                                   Eldon Fold 'N Roll Cart System
## 6  Furnishings Eldon Expressions Wood and Plastic Desk Accessories, Cherry Wood
##      Sales Quantity Discount    Profit
## 1 261.9600        2     0.00   41.9136
## 2 731.9400        3     0.00  219.5820
## 3  14.6200        2     0.00    6.8714
## 4 957.5775        5     0.45 -383.0310
## 5  22.3680        2     0.20    2.5164
## 6  48.8600        7     0.00   14.1694
```

```
named_bill <- sales[sales$Customer.First.Name == "Bill", ]
length(named_bill)
```

```
## [1] 22
```

```
length(unique(named_bill$Customer.Last.Name))
```

```
## [1] 6
```

```
# There are 22 customers with the first name Bill.
# There are 6 unique customers with the first name Bill.
```

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**

```
mentions_of_table <- length(str_subset(sales$Product.Name, regex(" table", ignore_case = TRUE)))
mentions_of_table
```

```
## [1] 246
```

```
# I included the method below as I was not fully trusting my first method and wanted to double.  I did

sales$Product.Name <- str_to_lower(sales$Product.Name)
mentions_of_table_confirmed <- sum(str_detect(sales$Product.Name, " table"))
mentions_of_table_confirmed
```

```
## [1] 246
```

```
# There are 246 mentions of the word 'table' in the Product.Name column.
```

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.

```
sales$State <- factor(sales$State)
is.factor(sales$State)
```

```
## [1] TRUE
```

```
levels(sales$State)
```

```
##  [1] "Alabama"              "Arizona"              "Arkansas"
##  [4] "California"           "Colorado"             "Connecticut"
##  [7] "Delaware"             "District of Columbia" "Florida"
## [10] "Georgia"              "Idaho"                "Illinois"
## [13] "Indiana"              "Iowa"                 "Kansas"
## [16] "Kentucky"             "Louisiana"            "Maine"
## [19] "Maryland"             "Massachusetts"        "Michigan"
## [22] "Minnesota"            "Mississippi"          "Missouri"
## [25] "Montana"              "Nebraska"             "Nevada"
## [28] "New Hampshire"        "New Jersey"           "New Mexico"
## [31] "New York"             "North Carolina"       "North Dakota"
## [34] "Ohio"                 "Oklahoma"             "Oregon"
## [37] "Pennsylvania"         "Rhode Island"         "South Carolina"
## [40] "South Dakota"         "Tennessee"            "Texas"
## [43] "Utah"                 "Vermont"              "Virginia"
## [46] "Washington"           "West Virginia"        "Wisconsin"
## [49] "Wyoming"
```

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

```
##
##              Alabama              Arizona             Arkansas
##                   28                  119                   22
##           California             Colorado          Connecticut
##                  993                   90                   50
##             Delaware District of Columbia              Florida
##                   47                    1                  186
##              Georgia                Idaho             Illinois
##                   79                    9                  286
##              Indiana                 Iowa               Kansas
##                   74                   11                   16
##             Kentucky            Louisiana                Maine
##                   64                   18                    4
##             Maryland        Massachusetts             Michigan
##                   63                   71                  142
```

```
##           Minnesota         Mississippi            Missouri
##                  41                  27                  37
##             Montana            Nebraska              Nevada
##                   2                  26                  24
##       New Hampshire          New Jersey          New Mexico
##                   9                  58                  11
##            New York      North Carolina        North Dakota
##                 555                 117                   7
##                Ohio            Oklahoma              Oregon
##                 211                  38                  56
##        Pennsylvania        Rhode Island      South Carolina
##                 312                  25                  28
##        South Dakota           Tennessee               Texas
##                   9                  88                 460
##                Utah             Vermont            Virginia
##                  27                  10                  80
##          Washington       West Virginia           Wisconsin
##                 254                   4                  38
##             Wyoming
##                   1
```

Exercise 8

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

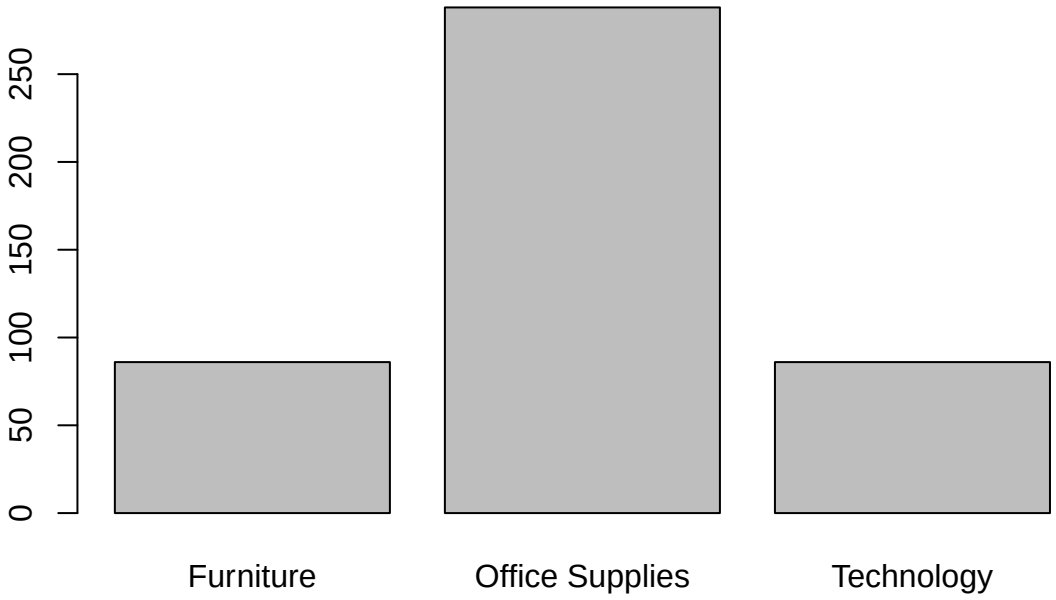
```
sales$Category <- factor(sales$Category)
is.factor(sales$Category)
```

```
## [1] TRUE
```

```
levels(sales$Category)
```

```
## [1] "Furniture"       "Office Supplies" "Technology"
```

```
Texas <- sales[sales$State == "Texas",]
barplot(table(Texas$Category))
```



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.**

```
average_profit_by_region <- aggregate(sales$Profit, list(Region = sales$Region), mean)
average_profit_by_region
```

```
##    Region        x
## 1 Central 20.46822
## 2    East 29.91937
## 3   South 11.27720
## 4    West 32.77000
```

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.**

```
sales$Order.Year <- year(sales$Order.Date)

average_profit_by_year <- aggregate(sales$Profit, list(Year = sales$Order.Year), mean)
average_profit_by_year
```

```
##   Year        x
## 1 2014 32.24582
## 2 2015 21.58676
## 3 2016 30.10960
## 4 2017 21.31825
```