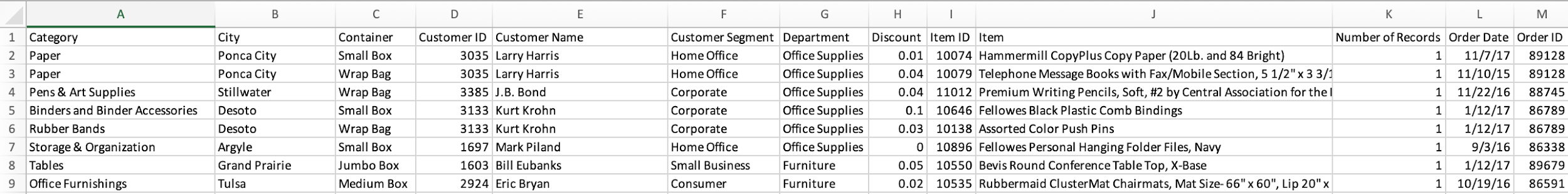
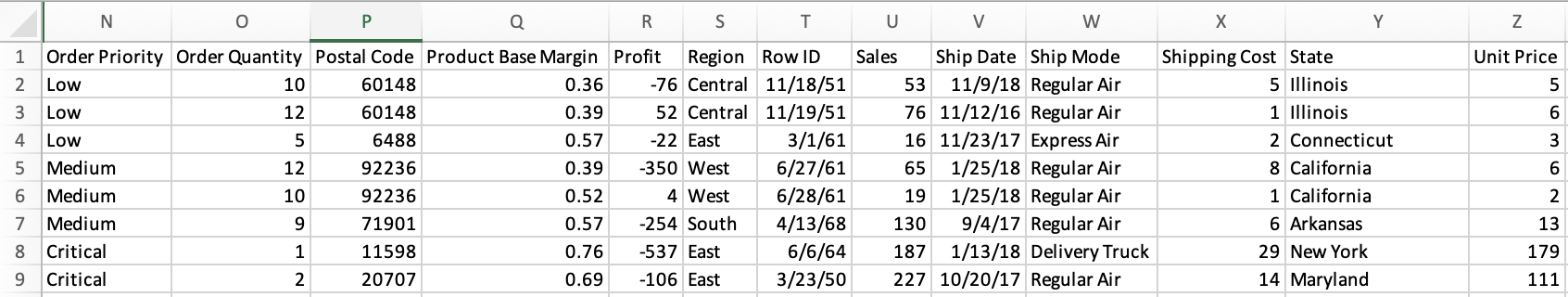
# **R for Data Science – Chapters 1 to 7 Summative Lab**

## *Analyzing Superstore Data*



For this lab, we’ll be using a fictional dataset containing information from orders at a superstore. This same dataset was used in your early days of learning Tableau last semester. The following screenshots includes a sample of the dataset, including all column names:





For each exercise below, paste the code and any outputs that it creates below each question.

1. Import the Superstore.xlsx file.

data <- read\_excel("Superstore.xlsx")

1. Create an R script for the exercises below. Call it “Superstore Lab – YourLastName.R”.

Source: Trust me

1. Is this dataset “tidy” according to the definition above figure 5.1 (the usual data science meaning)?

Yes, as it follows the prerequisites listed in Chapter 5.

1. Use the clean\_names function in the janitor package to make all column names snake case (Shipping Cost -> shipping\_cost).

data = janitor::clean\_names(data)

1. Keep only the following columns: “name”, “category”, “department”, “discount”, “order\_date”, “order\_priority”, “order\_quantity”, “profit”, “region”, “sales”, “ship\_mode”, “shipping\_cost”, “state”.

data= data[c('customer\_name', 'category', 'department', 'discount', 'order\_date', 'order\_priority', 'order\_quantity', 'profit', 'region', 'sales', 'ship\_mode', 'shipping\_cost', 'state')]

1. Create a new column called “expenses” that uses the formula “sales” – “profit”. Save this new column to the existing dataframe.

data=transform(data, expenses = sales-profit)

1. Create a new column called “large\_discount” that returns “Yes” if the discount > 0.10 and “No” otherwise.

data =mutate(data, large\_discount = ifelse((discount >0.1), "Yes", "No"))

1. You’ve been informed that a customer with a middle name is making fraudulent orders. Determine the number of customers that have used a middle name in their order. Do not count the same customer twice.

data2 <- data |>

seperate(customer\_name, " ",

into = c("1","2","3"),

remove=FALSE)

data2 |> view()

data2 <- data2 |>

drop\_na(3)

data2|> view()

919 people have a middle name.

1. Remove all customers with a middle name from the dataframe and save the change.

data <- data |>

separate(customer\_name, " ",

into = c("n1","n2","n3"),

remove=FALSE)

data <- data |>

filter(is.na(n3))

data|> view()

1. Use relocate to display the categorical columns first and save this change to the dataframe.

data= data |>

relocate('customer\_name', 'category', 'department', 'order\_date', 'region', 'ship\_mode', 'state')

1. Do each of the following and save the results to a new dataframe.
   1. For the “East” region, calculate the sum of the sales for each state.\

east\_sales <- data %>%

filter(region == "East") %>%

group\_by(state) %>%

summarise(sales = sum(sales))

* 1. For each state, calculate the average profit by department.

state\_profit <- data %>%

group\_by(state, department) %>%

summarise(AvgProfit = mean(profit))

* 1. For each month, calculate the total order quantity and sort it from highest to lowest.

data$Month <- month(data$order\_date)

data<- data |>

separate(order\_date, "-",

into = c("year","month","day"),

remove=FALSE)

data$month <- month(data$order\_date)

month\_quantity <- data %>%

group\_by(month) %>%

summarise(total\_quantity = sum(order\_quantity)) %>%

arrange(desc(total\_quantity))

* 1. For each department, calculate the average, lowest, and highest discount along with the number of data points used to calculate each.

dept\_discount <- data %>%

group\_by(department) %>%

summarise(avg\_discount = mean(discount), min\_discount = min(discount), max\_discount = max(discount), N = n())

1. Create visuals to answer the following. It may be helpful to do the calculations and first save those to a new dataframe:
   1. How does the total profit change by month and year based on department? In other words, your plot should have month, year on the x-axis, total profit on the y-axis, and each department should be a separate bar or line series.

ggplot(

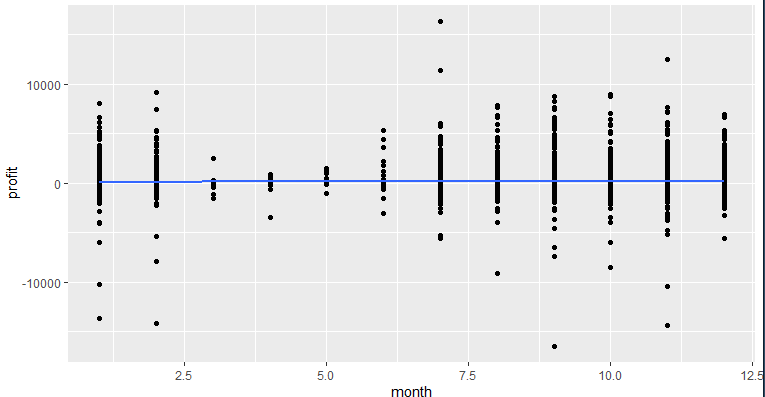
data = data,

mapping = aes(x = month,year, y = profit)

) +

geom\_point(mapping = aes()) +

geom\_smooth(method = "lm")



During the spring, profits are often lower, with profits increasing during the summer and winter months.

* 1. What proportion of orders have a given order\_priority by month? By department?

plot1 <- data %>%

group\_by(month, order\_priority) %>%

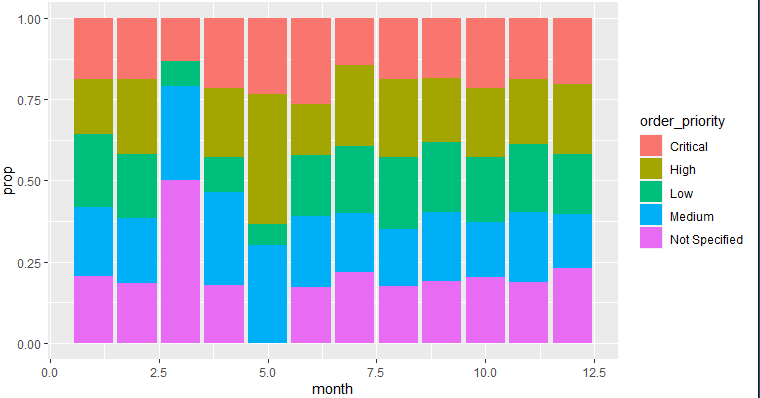
summarise(n=n()) %>%

mutate(prop=n/sum(n)) %>%

ggplot(aes(x=month,y=prop,fill=order\_priority))+

geom\_bar(stat="identity",position="fill")

plot1



plot2 <- data %>%

group\_by(department, order\_priority) %>%

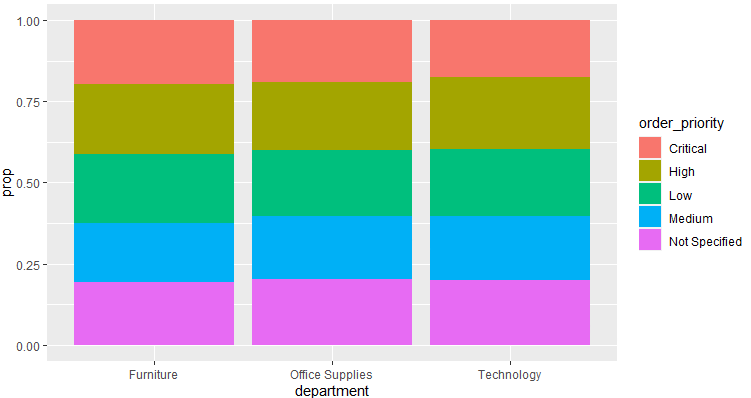
summarise(n=n()) %>%

mutate(prop=n/sum(n)) %>%

ggplot(aes(x=department,y=prop,fill=order\_priority))+

geom\_bar(stat="identity",position="fill")

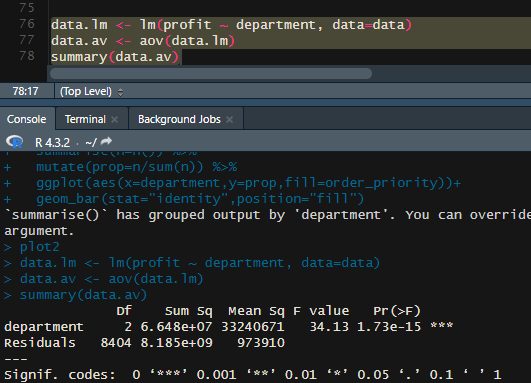
plot2



1. Create a hypothesis about the data and run an appropriate statistical test of your choosing in R to test your assertion.

Is there a difference on profit based on department?

If profit and department are compared, then there should be a difference.



because p<0.05, there is no significant difference.