Week 1 Lesson 2

In this lesson, we’ll begin by introducing data basics in the context of business analytics. We will work with the data sets available on the course webpage, in Modules/Week 1/ Data sets for Weeks 1-6.

### Exercise 1.

Read the description of bike sharing data set that could be found on the course webpage, in Modules/Week 1/ Data sets for Weeks 1-6. It might help to load the data in R Studio using the Quarto file “Start\_working\_with\_bike\_data.qmd” provided along with the data to start working on the data. Answer the following questions.

* 1. What is an observational unit (also called case or observation) in this data set?
  2. How many observations and how many variables does this data frame have?
  3. How were the data collected? Did one have to ask for customer consent to collect the data?
  4. What type of sampling technique does the data collection method for this data set resemble? Choose the one among simple random sampling, stratified, cluster, convenience sampling.
  5. How many variables are in the data set?
  6. Identify each variable as numerical or categorical. If numerical, state whether the variable is discrete or continuous. If categorical, state whether the variable is ordinal.   
     Note: In R, the names of data types differ from “numerical”/ “categorical”. For example, numerical data might be shown as integers (int) or doubles (dbl), while categorical data might be shown as factors (fct) or characters (chr). When you use functions like glimpse(bike), you can see these types.  
     When you import data into R, it tries to guess the data type automatically. This might not always match the actual type of the data. For instance, if you load a file where the values 1, 2, 3 represent colors like red, blue, and green, R might treat the corresponding variable as an integer by default rather than as a categorical variable, which would be the correct type.
  7. Who rented the most bikes, members or nonmembers?   
     To answer this question, think, what variables in the data set you need to answer the question. What summary of the variables you must calculate and how to calculate this summary in R.

Hint: one can use function summarise from tidyverse package to calculate summaries of the data. For example, one can run  
bike%>%

summarise(rented\_member = sum(member))

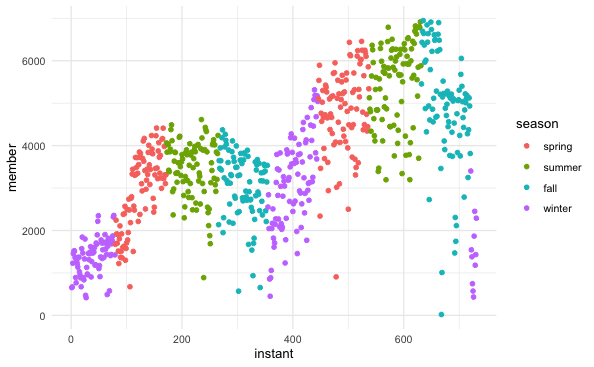
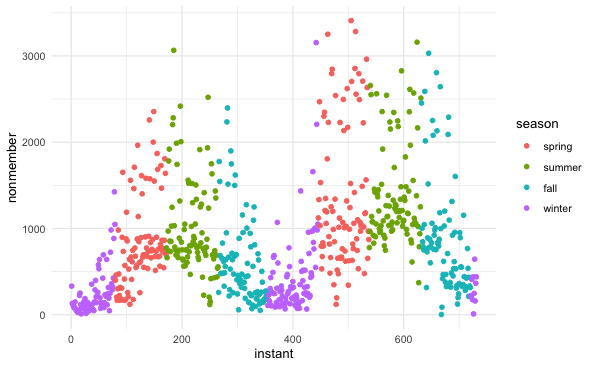
to calculate the sum of variable member.

Alternatively, one can calculate the same using

sum(bike$member).

e. What percentage of bikes were rented by member on the 1st of January 2011? To calculate the percentage, one can create a new variable in the data set using the function mutate. The variable should represent the proportion of bikes rented by members divided by the percentage of total rents.

* 1. Look at the plots below. What do they show? What variables have been used to construct these plots? What happened with the sales for members and nonmembers?



### Exercise 2.

In Exercise 2, we look at the Charles book club data set available on the course webpage, in Modules/Week 1/ Data sets for Weeks 1-6. It might help to load the data in R Studio using the Quarto file “Start\_working\_with\_book\_club\_data.qmd” provided along with the data to start working on the data. Answer the following questions.

1. What type of variable are variables Gender, M, R, F? What data types are they coded as in R data set?
2. What proportion of members have bought “The Art History of Florence”?

The following code might be useful in calculating the proportion:

book%>%

summarise(proportion\_bought = sum(Florence)/nrow(book)).

1. What percentage of women men buy the book “The Art History of Florence”? What percentage of men buy the book?

book %>%

count(Gender, Florence) %>% # Step 1: Count the combinations

group\_by(Gender) %>% # Step 2: Group by Gender

mutate(percentage\_of\_each\_gender = n / sum(n) \* 100) %>% #Step 3: Calculate percentage

ungroup()

1. Look at the bar plot below. Based on the plots, do women or men buy the book to a higher degree? Do your conclusions agree with your answer to c)?

A graph of a couple of red squares

Description automatically generated with medium confidence

1. What variables and what summaries of these were needed to construct the bar plot above?
2. Provide several suggestions of how to improve the bar plot above.