MPA 634  
Data Science for Managers  
Midterm II: Winter 2019

# I. Definitions and Concepts

1. Compare and contrast lists, tibbles, and data frames.  
     
   **list**: In R, vectors are one dimensional holders of information. Vectors have slots or positions in which information and be stored. The information can be referenced by specifying its slot number or name. Although not everyone agrees with Hadley Wickham, he states that there are two kinds of vectors:

* atomic vectors
* lists or recursive vectors

In most programming languages, lists are heterogeneous. This means that they can hold information of a variety of different classes: logical, integer, double, character, etc. The objects in a list can also have a variety of lengths. Lists can even contain other lists. Using lists within lists is a common practice.

**tibble**: A tibble is a collection of atomic vectors that are arranged in a list. Each of the atomic vectors must have the same length in a tibble. This means that tibbles are rectangular lists. This means that a tibble is a special case of a list. This is a data frame that has been “tweaked” by Hadley Wickham.  
  
**data frame**: A special type of list used to store data. Each of the atomic vectors in the list must have the same length. This makes the list rectangular. This is the traditional way of storing data in R.  
  
**Compare and contrast lists, tibbles, and data frames**: As already mentioned, data frames and tibbles are both lists. Because all lists do not contain just atomic vectors of the same length, lists in general do not satisfy the requirements for being a tibble.

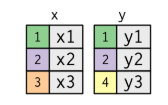
1. Define factors and then explain how they are related to order and level. Use an example to illustrate your answer.

Factor variables usually can have a fixed and known set of values. Factor variables can be represented as characters and integers. An example of a factor would be the grades that are awarded in a class.

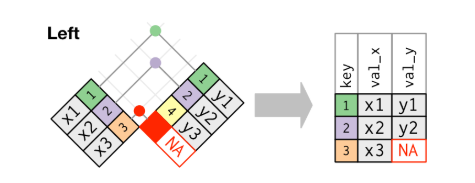
Factors are preferred when a character variable has a reasonably small number of values. Each category in a factor is called a level. Factors have the advantage that they can be ordered.

Factors allow character information to be stored very efficiently on a computer. Each level of a factor can be assigned a code. The code for each observation can then be efficiently substituted for an entire string. Factors are also helpful when we are trying to find data entry errors.

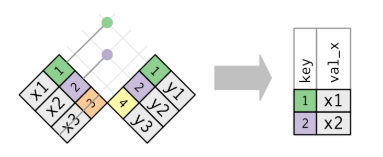
1. Compare and contrast left\_join with semi\_join. Sketch tibbles x and y to illustrate your answer.  
     
   Consider the following two tibbles:



left\_join: This is a mutating join since it adds information to tibble x from tibble y wherever the keys match. If there is not match for the primary key value in tibble x with the foreign key value in tibble y, then a missing information NA value is entered.



semi\_join: This is a filtering join so it only includes the rows in tibble x that have a match between primary key in x and the foreign key in y. In the above example, this gives the following

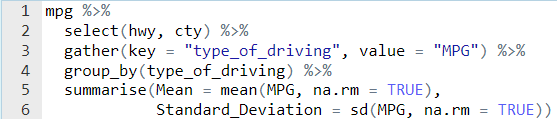


Comparison and contrast:

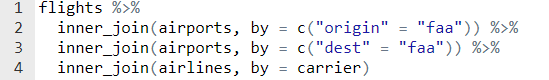
1. Two classes of time spans are durations and periods. Compare and contrast them.  
     
   Durations:  
     
     
     
     
     
     
   Periods:  
     
     
     
     
     
     
   Comparison and contrast:
2. Use an example to illustrate the difference between wide and long tibbles.

# II. Line by Line Code Interpretation

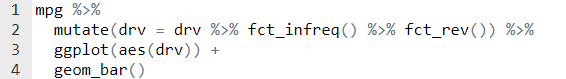
1. Recall that in the mpg tibble, there are two variables hwy and cty that report highway and city mileage respectively for each vehicle.



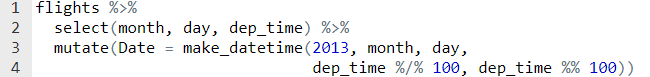
1. Recall that flights, airports, and airlines are three different related tibbles that are part of the nycflights13 library.



1. Recall that the mpg tibble contains a variable called drv which reports the type of drive train for each vehicle. Vehicles can be front-wheel drive, rear-wheel drive, and four-wheel drive.



1. The flights tibble contain the variables month, day, and departure time for each flight.



1. The mpg tibble contains the variable hwy which is miles per gallon for highway driving.

