ICA 5 Module 5 dplyr_library

word_document: default pdf_document: default —

Author: Andres Felipe Alba Hernández
Department: Electrical Engineering
Date: September 30, 2018
Course: ISYE670 Data Science for Engineers
Professor: Dr. Christine Nguyen
Northern Illinois University

Questions 1,2 and 3 can be compacted in the commands line below.

```
#Initializing
rm(list=ls()) #deleting enviroment variables
#Installing the library (only once)
#install.packages("tidyverse", lib="/home/leasanspy/DataScience_NIU/Rpackages")
#install.packages("dplyr", lib="/home/leasanspy/DataScience_NIU/Rpackages")
library(dplyr,lib.loc="/home/leasanspy/DataScience_NIU/Rpackages") #Loading the library
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

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

4) As may be observed from the commands below the income dataset describes Units sales, Revenue and Expenditure by Region, Department, and Year.

```
income <-read.csv("data_module_5.csv",header = TRUE)
summary(income)
```

```
##  Region   Department       Year        Unit_Sales        Revenue
##  A:12   Min.   :1.0   Min.   :2000   Min.   : 447.0   Min.   :3770
##  B:12   1st Qu.:1.0   1st Qu.:2001   1st Qu.: 657.2   1st Qu.:4523
##  C:12   Median :1.5   Median :2002   Median : 832.5   Median :5396
##         Mean   :1.5   Mean   :2002   Mean   : 864.6   Mean   :5401
##         3rd Qu.:2.0   3rd Qu.:2004   3rd Qu.:1084.5   3rd Qu.:6190
##         Max.   :2.0   Max.   :2005   Max.   :1339.0   Max.   :7416
##   Expenditure
##  Min.   :2050
##  1st Qu.:2836
##  Median :3655
##  Mean   :3678
##  3rd Qu.:4648
##  Max.   :5497
```

5) The only factor variable in the dataset is the Region and it has three levels {A,B,C}. I consider that Departments could be also a factor variable with levels 1 and 2.

```
str(income)
```

```
## 'data.frame':    36 obs. of  6 variables:
```

```
##  $ Region     : Factor w/ 3 levels "A","B","C": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Department : int  1 1 1 1 1 1 2 2 2 2 ...
##  $ Year       : int  2000 2001 2002 2003 2004 2005 2000 2001 2002 2003 ...
##  $ Unit_Sales : int  508 688 837 966 1143 1319 600 785 935 1089 ...
##  $ Revenue    : int  3770 4205 4878 5515 5935 6686 4002 4424 4869 5507 ...
##  $ Expenditure: int  2050 2847 3404 3986 4743 5442 2100 2802 3258 3950 ...
```

```
#income
```

6) Sample command by number and by percentage:

```
S1 <- sample_n(income,5) #five random samples
S2 <- sample_n(income,5) #other five random samples
S3 <- sample_frac(income, 0.02)
S4 <- sample_frac(income, 0.02)
```

```
print(S1)
```

```
##    Region Department Year Unit_Sales Revenue Expenditure
## 12      A          2 2005       1339    6580        5335
## 3       A          1 2002        837    4878        3404
## 27      C          1 2002        720    4926        3649
## 30      C          1 2005       1132    6414        5497
## 25      C          1 2000        447    4034        2301
```

```
print(S2)
```

```
##    Region Department Year Unit_Sales Revenue Expenditure
## 31      C          2 2000        511    4048        2201
## 18      B          1 2005       1236    7083        5323
## 30      C          1 2005       1132    6414        5497
## 28      C          1 2003        828    5345        4261
## 12      A          2 2005       1339    6580        5335
```

```
print(S3)
```

```
##    Region Department Year Unit_Sales Revenue Expenditure
## 10      A          2 2003       1089    5507        3950
```

```
print(S4)
```

```
##   Region Department Year Unit_Sales Revenue Expenditure
## 9      A          2 2002        935    4869        3258
```

7) The select function is used to select desired variables

a) Retrieve only the Department and Year columns

b) Retrieve all columns except the Expenditure column

```
select(income,Department,Year) #answer for a
```

```
##    Department Year
## 1           1 2000
## 2           1 2001
## 3           1 2002
## 4           1 2003
## 5           1 2004
## 6           1 2005
## 7           2 2000
## 8           2 2001
## 9           2 2002
## 10          2 2003
## 11          2 2004
## 12          2 2005
## 13          1 2000
## 14          1 2001
## 15          1 2002
## 16          1 2003
## 17          1 2004
## 18          1 2005
## 19          2 2000
## 20          2 2001
## 21          2 2002
## 22          2 2003
## 23          2 2004
## 24          2 2005
## 25          1 2000
## 26          1 2001
## 27          1 2002
## 28          1 2003
## 29          1 2004
## 30          1 2005
## 31          2 2000
## 32          2 2001
## 33          2 2002
## 34          2 2003
## 35          2 2004
## 36          2 2005
```

```r
select(income,-Expenditure) #answer for b
```

```
##    Region Department Year Unit_Sales Revenue
## 1       A          1 2000        508    3770
## 2       A          1 2001        688    4205
## 3       A          1 2002        837    4878
## 4       A          1 2003        966    5515
## 5       A          1 2004       1143    5935
## 6       A          1 2005       1319    6686
## 7       A          2 2000        600    4002
## 8       A          2 2001        785    4424
## 9       A          2 2002        935    4869
## 10      A          2 2003       1089    5507
## 11      A          2 2004       1214    6173
## 12      A          2 2005       1339    6580
## 13      B          1 2000        524    4017
## 14      B          1 2001        625    4665
## 15      B          1 2002        764    5272
## 16      B          1 2003        902    5975
## 17      B          1 2004       1083    6463
## 18      B          1 2005       1236    7083
## 19      B          2 2000        449    4078
## 20      B          2 2001        596    4539
## 21      B          2 2002        701    5190
## 22      B          2 2003       802    5659
## 23      B          2 2004       984    6242
## 24      B          2 2005      1159    6850
## 25      C          1 2000       447    4034
## 26      C          1 2001       550    4476
## 27      C          1 2002       720    4926
## 28      C          1 2003       828    5345
## 29      C          1 2004      1015    5928
## 30      C          1 2005      1132    6414
## 31      C          2 2000       511    4048
## 32      C          2 2001       668    4794
## 33      C          2 2002       775    5446
## 34      C          2 2003       939    6159
## 35      C          2 2004      1075    6877
## 36      C          2 2005      1216    7416
```

8) Using the help function, answer the following questions in your own words.

a) What is the ends_with() function?
b) What is the contains() function?
c) What is the matches() function?

All this functions are apply to the columns headers, basically it will select the column name that accomplish certain search using teh helpers.

For example:

a) ends_with() will pick the headers that ends with certain string that you give.
b) contains() will pick the headers that its name contain the given string, you can choose if you want to ignore or use the case differences.

c) matches() this helper help you to pick the right header by using regular expressions.
There are other helpers that can be obsered using the help for "select." In order to be more clear, I create one example for each select helper below:

```
#as_tibble(income)
summary(income)
```

```
##  Region   Department       Year        Unit_Sales       Revenue
##  A:12   Min.   :1.0   Min.   :2000   Min.   : 447.0   Min.   :3770
##  B:12   1st Qu.:1.0   1st Qu.:2001   1st Qu.: 657.2   1st Qu.:4523
##  C:12   Median :1.5   Median :2002   Median : 832.5   Median :5396
##         Mean   :1.5   Mean   :2002   Mean   : 864.6   Mean   :5401
##         3rd Qu.:2.0   3rd Qu.:2004   3rd Qu.:1084.5   3rd Qu.:6190
##         Max.   :2.0   Max.   :2005   Max.   :1339.0   Max.   :7416
##   Expenditure
##  Min.   :2050
##  1st Qu.:2836
##  Median :3655
##  Mean   :3678
##  3rd Qu.:4648
##  Max.   :5497
```

```
summary(select(income, ends_with(match = "ue",ignore.case = TRUE))) #I extract revenue
```

```
##     Revenue
##  Min.   :3770
##  1st Qu.:4523
##  Median :5396
##  Mean   :5401
##  3rd Qu.:6190
##  Max.   :7416
```

```
summary(select(income, contains("expen",ignore.case = TRUE))) #This should extract Expenditure
```

```
##   Expenditure
##  Min.   :2050
##  1st Qu.:2836
##  Median :3655
##  Mean   :3678
##  3rd Qu.:4648
##  Max.   :5497
```

```
summary(select(income, matches("Ye*"))) #This should extract year
```

```
##       Year
##  Min.   :2000
##  1st Qu.:2001
##  Median :2002
##  Mean   :2002
##  3rd Qu.:2004
##  Max.   :2005
```

9) The filter function is used to filter rows based on the criteria provided by the user. What command will filter the rows/observations where the year is 2000 or 2002?

```
filter(income, Year == "2000" | Year == "2002")
```

```
##    Region Department Year Unit_Sales Revenue Expenditure
## 1       A          1 2000        508    3770        2050
## 2       A          1 2002        837    4878        3404
## 3       A          2 2000        600    4002        2100
## 4       A          2 2002        935    4869        3258
## 5       B          1 2000        524    4017        2187
## 6       B          1 2002        764    5272        3661
## 7       B          2 2000        449    4078        2270
## 8       B          2 2002        701    5190        3237
## 9       C          1 2000        447    4034        2301
## 10      C          1 2002        720    4926        3649
## 11      C          2 2000        511    4048        2201
## 12      C          2 2002        775    5446        3072
```

10) Suppose you want to filter to keep rows where the Year is 2002, and then select the Department and Unit_Sales columns, and save it all in a new variable.

The code below execute both approaches, I use summary only to save space at printing

```
sd <- select(filter(income, Year == "2002"), Department, Unit_Sales)
summary(sd)
```

```
##    Department    Unit_Sales
##  Min.   :1.0   Min.   :701.0
##  1st Qu.:1.0   1st Qu.:731.0
##  Median :1.5   Median :769.5
##  Mean   :1.5   Mean   :788.7
##  3rd Qu.:2.0   3rd Qu.:821.5
##  Max.   :2.0   Max.   :935.0
```

```
sd_pipe <- income %>% filter(Year=="2002")%>% select(Department, Unit_Sales)
summary(sd_pipe)
```

```
##    Department    Unit_Sales
##  Min.   :1.0   Min.   :701.0
##  1st Qu.:1.0   1st Qu.:731.0
##  Median :1.5   Median :769.5
##  Mean   :1.5   Mean   :788.7
##  3rd Qu.:2.0   3rd Qu.:821.5
##  Max.   :2.0   Max.   :935.0
```

11) The arrange function is used to arrange or re-order rows by a particular column Let’s reorder using the Department values. The values should be in ascending order.

```
arrange(income, Department)
```

```
##    Region Department Year Unit_Sales Revenue Expenditure
## 1       A          1 2000        508    3770        2050
## 2       A          1 2001        688    4205        2847
## 3       A          1 2002        837    4878        3404
## 4       A          1 2003        966    5515        3986
## 5       A          1 2004       1143    5935        4743
## 6       A          1 2005       1319    6686        5442
## 7       B          1 2000        524    4017        2187
## 8       B          1 2001        625    4665        2955
## 9       B          1 2002        764    5272        3661
## 10      B          1 2003        902    5975        4308
## 11      B          1 2004       1083    6463        4740
## 12      B          1 2005       1236    7083        5323
## 13      C          1 2000        447    4034        2301
## 14      C          1 2001        550    4476        2884
## 15      C          1 2002        720    4926        3649
## 16      C          1 2003        828    5345        4261
## 17      C          1 2004       1015    5928        4924
## 18      C          1 2005       1132    6414        5497
## 19      A          2 2000        600    4002        2100
## 20      A          2 2001        785    4424        2802
## 21      A          2 2002        935    4869        3258
## 22      A          2 2003       1089    5507        3950
## 23      A          2 2004       1214    6173        4750
## 24      A          2 2005       1339    6580        5335
## 25      B          2 2000        449    4078        2270
## 26      B          2 2001        596    4539        2681
## 27      B          2 2002        701    5190        3237
## 28      B          2 2003        802    5659        3682
## 29      B          2 2004        984    6242        4113
## 30      B          2 2005       1159    6850        4632
## 31      C          2 2000        511    4048        2201
## 32      C          2 2001        668    4794        2619
## 33      C          2 2002        775    5446        3072
## 34      C          2 2003        939    6159        3572
## 35      C          2 2004       1075    6877        4289
## 36      C          2 2005       1216    7416        4694
```

```
arrange( income, desc(Department))
```

```
##    Region Department Year Unit_Sales Revenue Expenditure
## 1       A          2 2000        600    4002        2100
## 2       A          2 2001        785    4424        2802
## 3       A          2 2002        935    4869        3258
## 4       A          2 2003       1089    5507        3950
## 5       A          2 2004       1214    6173        4750
## 6       A          2 2005       1339    6580        5335
## 7       B          2 2000        449    4078        2270
## 8       B          2 2001        596    4539        2681
## 9       B          2 2002        701    5190        3237
## 10      B          2 2003        802    5659        3682
## 11      B          2 2004        984    6242        4113
## 12      B          2 2005       1159    6850        4632
## 13      C          2 2000        511    4048        2201
## 14      C          2 2001        668    4794        2619
## 15      C          2 2002        775    5446        3072
## 16      C          2 2003        939    6159        3572
## 17      C          2 2004       1075    6877        4289
## 18      C          2 2005       1216    7416        4694
## 19      A          1 2000        508    3770        2050
## 20      A          1 2001        688    4205        2847
## 21      A          1 2002        837    4878        3404
## 22      A          1 2003        966    5515        3986
## 23      A          1 2004       1143    5935        4743
## 24      A          1 2005       1319    6686        5442
## 25      B          1 2000        524    4017        2187
## 26      B          1 2001        625    4665        2955
## 27      B          1 2002        764    5272        3661
## 28      B          1 2003        902    5975        4308
## 29      B          1 2004       1083    6463        4740
## 30      B          1 2005       1236    7083        5323
## 31      C          1 2000        447    4034        2301
## 32      C          1 2001        550    4476        2884
## 33      C          1 2002        720    4926        3649
## 34      C          1 2003        828    5345        4261
## 35      C          1 2004       1015    5928        4924
## 36      C          1 2005       1132    6414        5497
```

12)The mutate function is used to create new variables that are functions of existing variables. Create a new data frame that has the Expenditure and Revenue columns, and also create a new column “profit”, which is Revenue minus Expenditure.

```
m_df <- mutate(select(income, Expenditure, Revenue), profit=Revenue-Expenditure)
m_df_pipe <- income %>% select(Expenditure,Revenue) %>% mutate(profit=Revenue-Expenditure)
m_df
```

```
##    Expenditure Revenue profit
## 1         2050    3770   1720
## 2         2847    4205   1358
## 3         3404    4878   1474
## 4         3986    5515   1529
## 5         4743    5935   1192
## 6         5442    6686   1244
## 7          2100    4002   1902
## 8          2802    4424   1622
## 9          3258    4869   1611
## 10         3950    5507   1557
## 11         4750    6173   1423
## 12         5335    6580   1245
## 13         2187    4017   1830
## 14         2955    4665   1710
## 15         3661    5272   1611
## 16         4308    5975   1667
## 17         4740    6463   1723
## 18         5323    7083   1760
## 19         2270    4078   1808
## 20         2681    4539   1858
## 21         3237    5190   1953
## 22         3682    5659   1977
## 23         4113    6242   2129
## 24         4632    6850   2218
## 25         2301    4034   1733
## 26         2884    4476   1592
## 27         3649    4926   1277
## 28         4261    5345   1084
## 29         4924    5928   1004
## 30         5497    6414    917
## 31         2201    4048   1847
## 32         2619    4794   2175
## 33         3072    5446   2374
## 34         3572    6159   2587
## 35         4289    6877   2588
## 36         4694    7416   2722
```

```
m_df_pipe
```

```
##    Expenditure Revenue profit
## 1          2050    3770   1720
## 2          2847    4205   1358
## 3          3404    4878   1474
## 4          3986    5515   1529
## 5          4743    5935   1192
## 6          5442    6686   1244
## 7          2100    4002   1902
## 8          2802    4424   1622
## 9          3258    4869   1611
## 10         3950    5507   1557
## 11         4750    6173   1423
## 12         5335    6580   1245
## 13         2187    4017   1830
## 14         2955    4665   1710
## 15         3661    5272   1611
## 16         4308    5975   1667
## 17         4740    6463   1723
## 18         5323    7083   1760
## 19         2270    4078   1808
## 20         2681    4539   1858
## 21         3237    5190   1953
## 22        3682    5659   1977
## 23        4113    6242   2129
## 24        4632    6850   2218
## 25        2301    4034   1733
## 26        2884    4476   1592
## 27        3649    4926   1277
## 28        4261    5345   1084
## 29        4924    5928   1004
## 30        5497    6414    917
## 31        2201    4048   1847
## 32        2619    4794   2175
## 33        3072    5446   2374
## 34        3572    6159   2587
## 35        4289    6877   2588
## 36        4694    7416   2722
```

Both output have similar results.

13) Let's use the group_by and summarise functions to calculate the average expenditure by Region

```
income %>% group_by(Region) %>% summarise(avg_expenditure=mean(Expenditure))
```

```
## # A tibble: 3 x 2
##   Region avg_expenditure
##   <fct>            <dbl>
## 1 A                3722.
## 2 B                3649.
## 3 C                3664.
```

```
income %>% group_by(Region) %>% summarise(standarDesviation_expenditure=sd(Expenditure))
```

```
## # A tibble: 3 x 2
##   Region standarDesviation_expenditure
##   <fct>                          <dbl>
## 1 A                              1173.
## 2 B                              1011.
## 3 C                              1079.
```

```
income %>% group_by(Region) %>% summarise(minimum_expenditure=min(Expenditure))
```

```
## # A tibble: 3 x 2
##   Region minimum_expenditure
##   <fct>                <dbl>
## 1 A                     2050
## 2 B                     2187
## 3 C                     2201
```

```
income %>% group_by(Region) %>% summarise(maximum_expenditure=max(Expenditure))
```

```
## # A tibble: 3 x 2
##   Region maximum_expenditure
##   <fct>                <dbl>
## 1 A                     5442
## 2 B                     5323
## 3 C                     5497
```

```
income %>% group_by(Region) %>% summarise(Median_of_expenditure=median(Expenditure))
```

```
## # A tibble: 3 x 2
##   Region Median_of_expenditure
##   <fct>                  <dbl>
## 1 A                      3677
## 2 B                      3672.
## 3 C                      3610.
```