FIN 510 Big Data Analytics in Finance

Lab 3: Data Wrangling

Due on 09/04/2021

Transforming Data on Firm Fundamentals

The file firm.csv contains firm fundamental variables, including total assets, net income, and cash dividends, regarding 8 firms during fiscal years of 2015 to 2018.

In this lab assignment, we will manage data using the four dplyr functions: filter, arrange, select, and mutate.

DESCRIPTION OF VARIABLES FOR FIRM FUNDAMENTALS EXAMPLE	
DATADATE	01/01/2014-12/31/2019
TIC	AAPL MSFT GOOGL INTC IBM FB ORCL FJTSY
FYEAR	Fiscal year
GVKEY	Global company key
SIC	Standard industrial classification code
CONM	Company Name
AT	Total Assets
NI	Net income (loss)
DV	Cash dividends

0) Load the packages

Use library() to load tidyverse.

1) create a data frame

Load the data using read_csv() and save the result in a tibble named df. Return the first six rows, the number of rows, the number of columns, and column names using head(), nrow(), ncol(), and names(), respectively.

2) subset data

2.1) one column

Select the net income (NI) column from df using two ways: df$ and select().

2.2) three columns

Select fiscal year (FYEAR), ticker (TIC), and net income (NI) columns from df using two ways: df[,] and select().

2.3) four rows and three columns

Select the first four rows and fiscal year (FYEAR), ticker (TIC), and net income (NI) columns from df using df[,].

3) filter rows

3.1) observations that belong to AAPL in 2015

Pick rows that TIC is equal to AAPL and FYEAR is equal to 2015 using df[,] and filter().

3.2) observations that have the minimum NI value

Pick rows that NI is equal to the minimum NI value using df[,] and filter().

3.3) FYEAR, TIC, and NI columns of observations that belong to AAPL in 2015

Select FYEAR, TIC, NI columns and pick rows that TIC is equal to AAPL and FYEAR is equal to 2015 using df[,].

4) arrange rows

Do not modify df in place. Only display ordered results.

4.1) ascending order of NI

Reorder df in an ascending order of NI using arrange()

4.2) descending order of NI

Reorder df in a descending order of NI using arrange() with desc().

4.3) ascending order of TIC and FYEAR, and descending order of NI

Reorder df in an ascending order of TIC and FYEAR, and descending order of NI using arrange().

5) create new variables

5.1) return on assets

Create a new column named ROA in df which divides net income (NI) by total assets (AT).

5.2) net income bin numbers

In lecture 3, we created equal width bins and these bins have different number of observations.

In the question, we want to create bins that have an equal number of observations in each bin.

Step 1) identify thresholds of four bins that have an equal number of observations

To create four bins that have an equal number of observations in each bin, we need to use function quantile() to identify five thresholds: the minimum value, first, second, and third quantiles, and the maximum value.

The first parameter in function quantile() is the variable's name, df$NI. The second parameter is a vector of probabilities. For example, use 0.5 to represent the second quantile, because 50% of the data lies below this value.

Hint: use probs = c(0,0.25,0.5,0.75,1) to return the minimum value, first, second, and third quantiles, and the maximum value.

Save the result as bins.

Step 2) remove the names of a named vector

The name of the first element in bins is 0% and the value of the first element is 771.775. Use unname(bins) to remove the names of a named vector. For example, bins <- unname(bins) returns the values only.

Step 3) bin net income

To transform a continuous variable into a categorical variable, use function .bincode() to bin net income (NI) according to the thresholds found in step 2.

The first parameter in function .bincode() is the variable's name, df$NI. The second parameter specifies cut points, which should be bins without names. Use include.lowest = TRUE as the third parameter to include the lowest value, 771.775, in the first bin.

Save the binned net income as NI_bin. It takes a value of 1 if NI is in [771.775, 8857.75], a value of 2 if NI is in (8857.75, 12032.5], a value of 3 if NI is in (12032.5, 19871.75], and a value of 4 if NI is in (19871.75, 59531].

Use table(df$NI_bin) to return the frequency count of unique values in NI_bin. Do you have 8 observations in each bin of NI?

5.3) log of assets

Create a new column named AT_LOG in df which computes the log value of total assets (AT) using mutate(). Use head() to print the first six rows of df.

5.4) delete a variable

Remove return on assets (ROA) from df by assigning NULL to the column. Use head() to print the first six rows of df.