FIN 510 Big Data Analytics in Finance

Lab 4: Data Exploration

Due on 09/11/2021

Summarizing 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 explore data using the five dplyr functions: filter, arrange, select, mutate, and summarize.

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 (2015, 2016, 2017, 2018)
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 and column names using head() and names(), respectively.

2) calculate summary statistics

2.1) average net income

Compute the average net income (NI) using mean().

2.2) first decile of net income

Find a value of net income such that 10% net income values are less than it and 90% are greater than it using quantile().

2.3) summary of net income

Calculate the minimum, the first quartile, the median, the mean, the third quartile, and the maximum value of net income using summary().

3) compute counts and proportions of logical values

3.1) number of rows

Count the number of observations that have dividends (DV) greater than 0 and net income (NI) greater than 10000 using sum().

3.1) proportion of rows

Find the proportion of observations that have dividends (DV) greater than 0 and net income (NI) greater than 10000 using mean().

4) summarize data with dpylr

4.1) average and maximum net income

Calculate the average value and the maximum value of net income, the number of observations, and the number of unique TIC values using summarize(). Specify the names of the summary statistics as NI_mean, NI_max, n_rows, and n_firms, respectively.

4.2) average and minimum net income by firm

Group df by TIC and save the grouped data as df_by_TIC using group_by().

Calculate the average value and the minimum value of net income by TIC, and the number of observations in each group using summarize(). Specify the names of the summary statistics as NI_mean, NI_min, and n_rows, respectively.

5) combine multiple operations with the pipe

5.1) average net income by firm

Group df by TIC, calculate average net income and save it as NI_mean, and arrange the result in a descending order of NI_mean.

5.2) most recent return on assets by firm

Create a new variable named ROA, group df by TIC, order the result in a descending order of FYEAR, keep the first observation in each group, and name it as recent_ROA.

Hint: return on assets (ROA) is defined as the ratio of net income (NI) by total assets (AT).

6) combine firm fundamentals with executives' compensation

6.1) loan executive compensation data

The file manager.csv contains executives' compensation information regarding 7 firms during fiscal years of 2015 to 2018. We used this data set in lecture 3.

DESCRIPTION OF VARIABLES FOR EXECUTIVE COMPEASTION EXAMPLE	
YEAR	Fiscal year (2015, 2016, 2017, 2018)
TICKER	AAPL MSFT GOOGL INTC IBM FB ORCL
GVKEY	Global company key
SIC	Standard Industrial Classification code
CO_PER_ROL	ID for each executive and company combination
EXECID	Executive's ID number
EXEC_FULLNAME	Executive's full name
SALARY	Salary (thousand dollars)
BONUS	Bonus (thousand dollars)
TDC1	Total compensation (salary, bonus, other annual, restricted stock grants, stock option grants, long-term incentive payouts, and all other in thousand dollars)

Load the data using read_csv() and save the result in a tibble named manager. Return the first six rows and column names using head() and names(), respectively.

6.2) frequency counts of firm

Return frequency counts of unique TIC values in df using table().

Return frequency counts of unique TICKER values in manager using table().

Do these data sets cover same firms?

6.3) frequency counts of year

Return frequency counts of unique FYEAR values in df using table().

Return frequency counts of unique YEAR values in manager using table().

Do these data sets cover same fiscal years?

6.4) inner join

In this question, we will match the two data sets based on firm and year. In the merged data set, each row includes both executives' compensation information (e.g., salary, bonus, total compensation), as well as firm fundamentals (e.g., net income, total assets, dividends).

Use inner_join() to return only matching rows from manager and df based on firm and year.

Inside inner_join(), specify manager, df, and the join criteria. Notice that firm and year have different names in manager and df. Save the resulting tibble as merged.

Hint: by = c("TICKER"="TIC","YEAR"="FYEAR") matches TICKER from manager with TIC from df, and YEAR from manager with FYEAR from df.

Return the first six rows and column names of merged using head() and names(), respectively. Notice that the TICKER column stores firm tickers, the YEAR column stores fiscal years, and TIC and FYEAR columns are no longer listed.

6.5) average salary and return on assets during 2017 and 2018

Combine the following operations with the pipe:

- Select rows with YEAR=2017 or 2018 from merged using filter().

- Create a new variable named ROA in merged, which divides net income (NI) by total assets (AT).

- Group the data by TICKER using group_by().

- Using summarize(), calculate average SALARY and average ROA, and count the number of rows in each group. Specify the names of the summary statistics as SALARY_mean, ROA_mean, and n_rows, respectively.

- Order the result in a descending order of ROA_mean using arrange().