

**BC2406-Analytics 1**

**Lab Quiz 2nd Submission**

**Khoo Ju Seng Bryant U1420978E**

1. **Question 1**

**Data Exploration**

The aim of doing Data Exploration is to understand the given dataset, and compare the data to the business problem. I will be using the following questions as the guidelines for Data exploration

1. Is the data sufficient?

* Firstly, run the summary function to find out if there are many NA values for dvc and ib. There are NA values in dvc and ib as shown below:

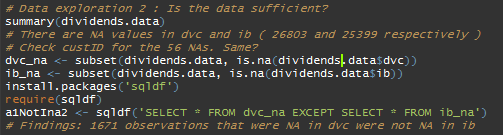


Fig. 1 Checking NA values for dvc and ib

* Using a package called ‘sqldf’, I was able to run a command that resembles sql’s select function to find out how many rows were NA in dvc but not in NA. The findings can be seen in Figure 1.

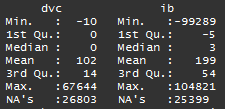


Fig 2. NA values for dvc and ib

1. Does the data provide predictive value?

* I used ggplot 2 to plot a scatter plot with a smooth line, to see the general datapoints and relationship between the two columns.

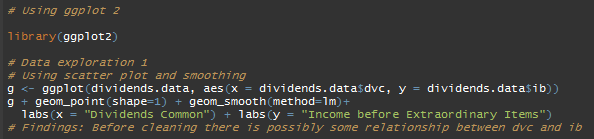


Fig. 3 R-code Using gg plot

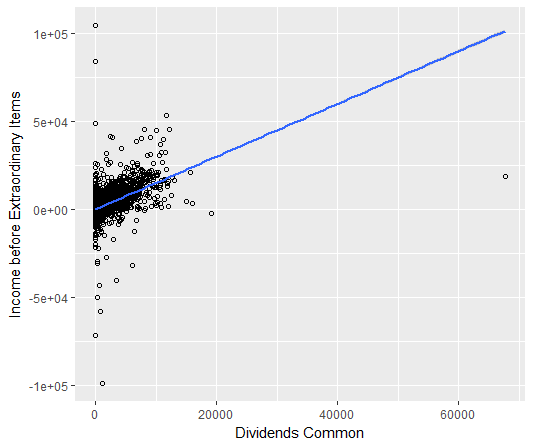


Fig 4. Scatter plot of ib against dvc

* Findings: before data cleaning, we can already see that there **is possibly** a relationship between dvc and ib. **However, there are still outliers.**

1. How is the data quality and are there any anomalies?

* After doing summary(), I noticed that there were negative dvc values.
* I ran the following code to find out how many rows were negative, and which rows are they:

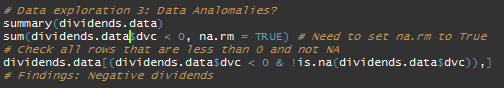


Fig 5. R-code to find explore rows

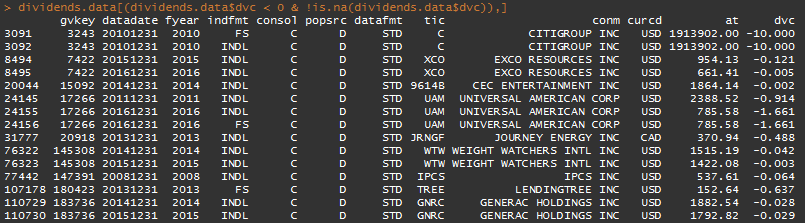


Fig 6. Results for code ran in Fig 5.

1. **Question 2**

**Data Cleaning**

1. Keep records with Stock Exchange Code = 11, 12, or 14.
   * I ran the following R-script, using the subset function on the original data *diviends.data* with the above conditions for the Stock Exchange Code (exchg) to obtain cleaned data *dividends.data.2a* for question 2a.
   * The number of records after execution is 70,161 rows.

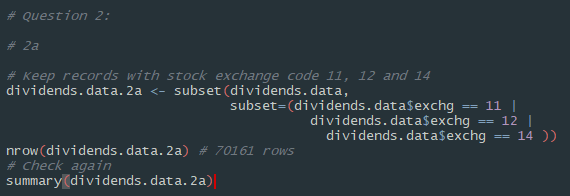


Fig 7. R-code for data cleaning in 2a

1. Keep records with currency = USD.
   * I ran the following R-script, using the subset function, based on the condition that currency is USD on the cleaned data *diviends.data.2a* to obtain cleaned data *dividends.data.2c* for question 2c.
   * The number of records after execution of 2a and 2b is 68,876 rows.

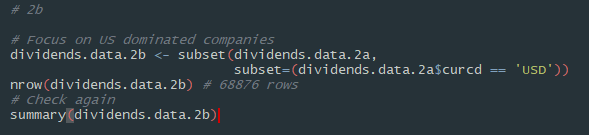


Fig 8. R-code for data cleaning in 2b

1. Drop records for SIC sector code from 6000 to 6999 and 4900 to 4999
   * I ran the following R-script, using the subset function, based on the above condition on the cleaned data *diviends.data.2b* to obtain cleaned data *dividends.data.2c* for question 2c.
   * The number of records after execution of 2a, 2b and 2c is 35,791 rows.

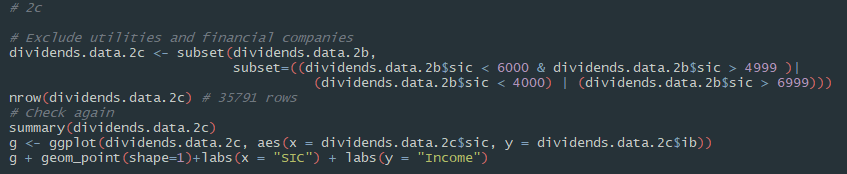


Fig 9. R-code for data cleaning in 2c

* To double check that I cleaned the correct range for SIC, in case of any mistakes in my conditions I used gg plot to plot a scatterplot. From the scatter plot, it seems that the subset function removed the right region of SIC values (highlighted in red below).

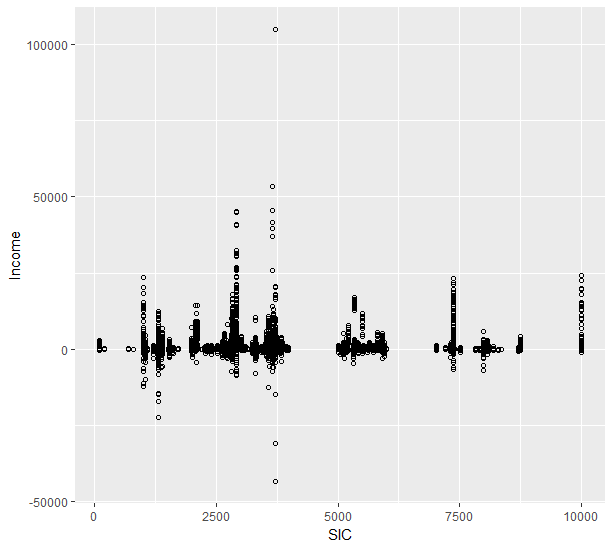


Fig 10. Scatterplot obtained from ggplot

1. Delete observations that do not report income (ib=NA)
   * I ran the following R-script, using the subset function, based on the above condition on the cleaned data *diviends.data.2c* to obtain cleaned data *dividends.data.2d* for question 2d.
   * The number of records after execution of 2a, 2b, 2c and 2d is 34,195 rows.

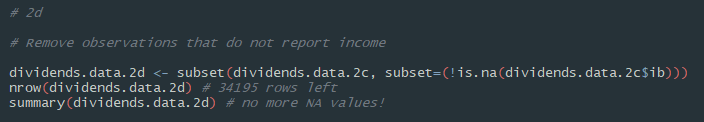


Fig 11. R-code to remove ib’s NA rows in data

1. What other data cleaning steps should be executed that I consider necessary and obvious?
   * I firstly removed NA values for dvc and negative values for dvc.
   * Also, there were no more NA values for ib.
   * I also removed NA values for a few more independent variables X (spi, idit and costat) that I felt would be used later on.

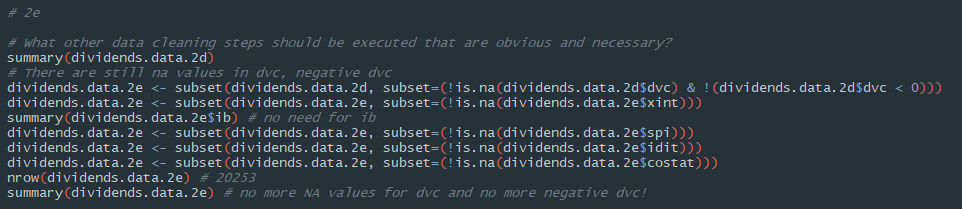


Fig 12. Remove NA and neg. dvc and NA for X variables

* Next, I realised that there are still 1355 rows is NA values in the independent variable prstkc. I ran the following code:

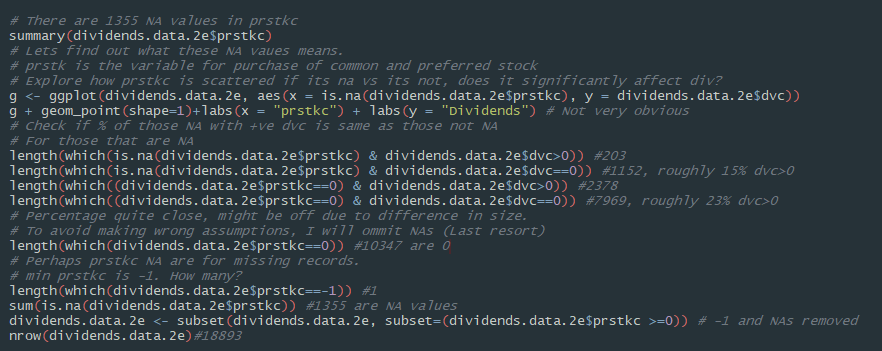


Fig 13. Removing prstkc NA and -1 values

* I firstly plotted a scatterplot to look at how the rows with NA prstkc values were spread out in terms of dividends (dvc), but the results were not very obvious

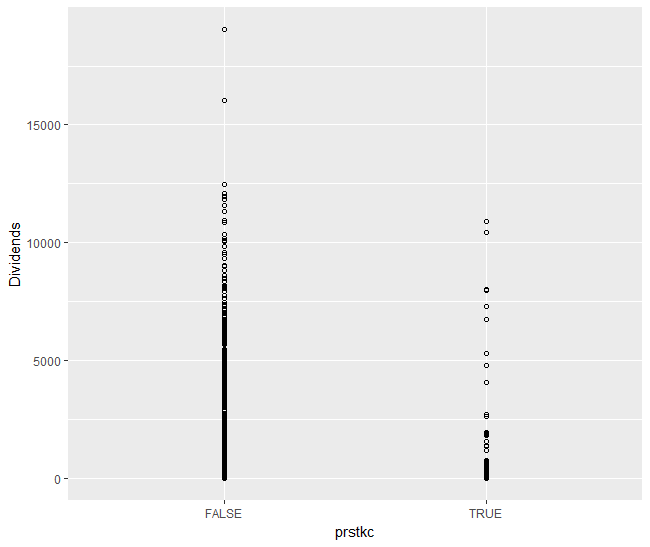


Fig 14. Dvc against is.na(prstkc)

* Hence, I decided to check if the percentage of those with NA and have +ve dvc values is the same as those that are not NA. And the percentage seems to be quite close.
* To avoid making the wrong assumptions for the NA values, I decided to omit these rows as a last resort.
* After removing values with prstkc being NA and -1, I had **18,893** rows left.
* One final data cleaning step at this stage is to convert years to factors. However, one thing that I took note of is that there is only one observation with fyear == ‘2017’. This will pose as a problem later on when I do regression and CART. However, I decided to leave the row for question 3, and it will be removed in later parts when needed.



Fig 15. Convert fyear to factors

1. **Question 3**

**Generating summary table:**

* I ran the following code to generate the summary table

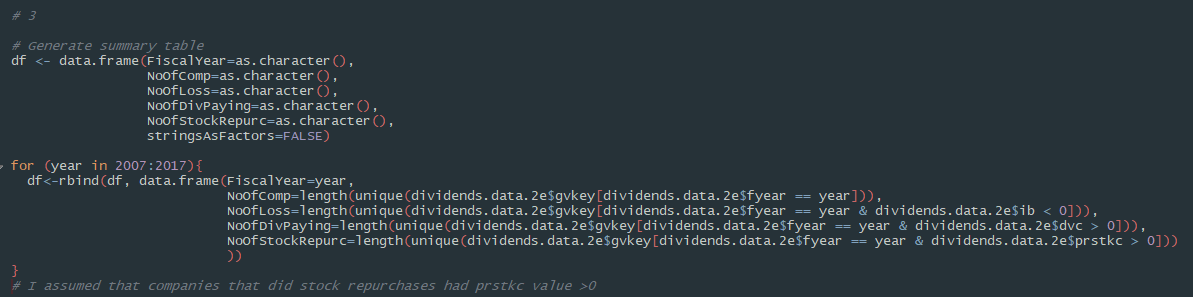


Fig 16. R-code to generate summary table

* I initialised a new dataframe and its columns.
* A for loop is used to iterate through years 2007 to 2017 and data for each year as rows.
* I assumed that stock repurchase meant positive prstkc value.
* The resulting summary table is as follows:

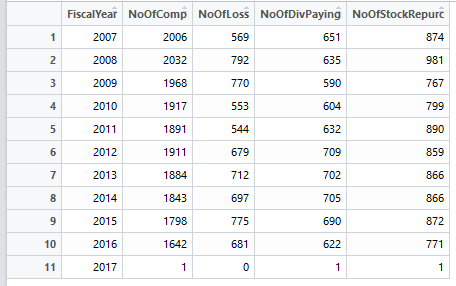


Fig 17. Summary Table (qn3)

* To observe the number of firms that are making losses, paying dividends and repurchasing stock, I found it hard to visualize from the table. **Hence instead of deciding on whether the number increased or decreased, I decided to observe the general trend over the years using ggplot.**
* Since the number of companies in each row is different, I decided to plot the following graphs.
  + Num. of Comp. that Loss / Num of Comp. against Fiscal Year
  + Num. of Div Paying Comp. / Num of Comp. against Fiscal Year
  + Num. of Comp. thatStock Repurchase / Num of Comp. against Fiscal Year
* Hence, the Y values would be a ratio ranging from 0 to 1. I will be using bar plots to represent these values.
* We also need to take note that year 2017 is only represented by 1 observation. It would be good to not take the year’s value too seriously.

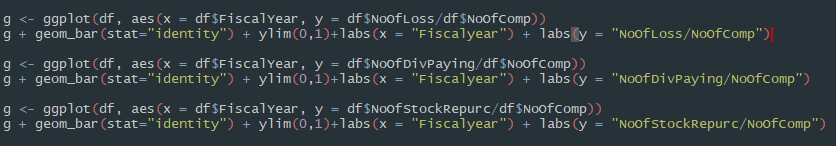


Fig 18. R-code for ggplot

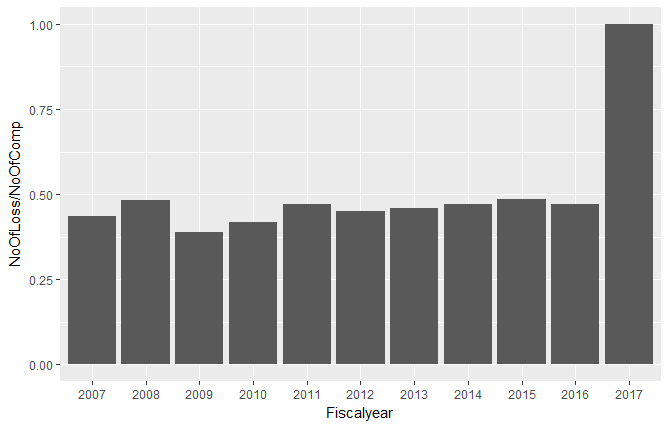


Fig 19. Ratio of companies suffering from loss

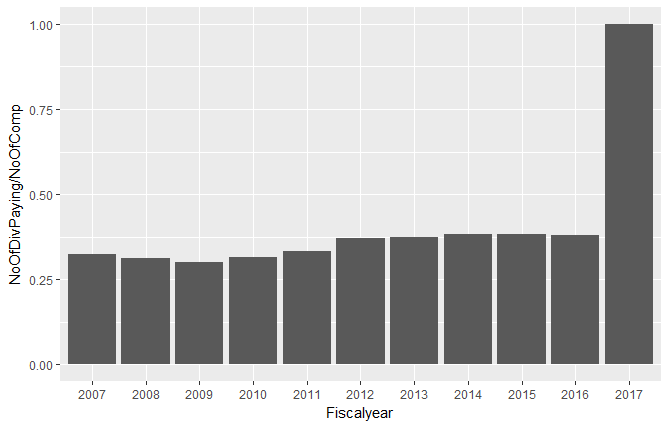


Fig 20. Ratio of companies paying dividends

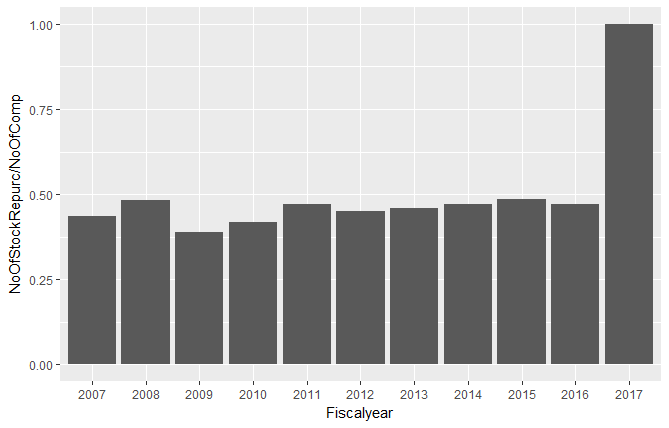


Fig 21. Ratio of companies repurchasing stock

* From figures 19 to 21, we can see that generally the ratio of companies that fall under the 3 categories remained quite stable from 2007 to 2016.
* Furthermore, I observed that there is a dip in 2009 in all of the bar plots.

1. **Question 4**