NEW YORK INSTITUTE OF TECHNOLOGY

College of Engineering & Computing Sciences

Business Data Analytics Project

(Employee Attrition Data & Finance Data)

Prepared for:

Porfessor Amba Sekhar

Prepared by:

Shukun Liu (1300107)

Buyun Li (1298535)

Xiaoliang Zhang (1300019)

Data Analytics Practicum (BUSIE 702)

July 11st, 2022

Table of Contents

CASE 1 - HR Employee Attrition

CASE 2 - Finance Practicum Data

CASE 1 - HR Employee Attrition

1.0 Background

The high employee turnover rate has emerged as one of the major issues plaguing management hierarchys as the competition for human resources has been more intensed and employee are considering more factors about their career and life. An organisation loses a lot of money due to employee turnout. According to an American Management Association (AMA) research on replacing an employee, on the one hand, when an individual quits their position, the resources they produced and carried with them are gone. The cost is at least equal to 30% of their yearly income, and for jobs where there is a skills shortage, the cost can be up to 1.5 times as much as their annual compensation. Since skilled workers make up the majority of departing employees, rivals may cut costs by hiring them back, which helps them greatly but puts pressure on the market. Additionally, the resigning personnel have suffered losses due to their own actions.

Factors that affect employee turnover can be classified into four categories: personal factors, organizational factors, individual-rrganization fit factors, external environmental factors. The examples of factors of these four categories are shown in table below.

Categories	Examples
Personal Factor	Personal characteristics, work motivation and sense of achievement.
Organizational Factor	Salary and benefits, promotion, training, company benefits, working conditions.
Individual-rrganization Fit Factor	Corporate culture, interpersonal relationships.
External Environmental Factor	Labour market conditions, job opportunities outside the organisation, employment situation.

Since employee resignation is not good for individuals and companies, why should employees resign? What then are the motivating elements for employee turnover? How can these variables be measured? Managing the behaviour among the organization that is the key to reducing employee turnover is the answer to these questions.

1.1 Project Aims

In case 1, we tried to predict employees left or not under different conditions. In our data preprocessing, we used one-hot-encode for all the categorical variables and tried some special data engineering to expand our training data. Moreover, we fitted our data with different models, and selected the best three models to investigate further.

1.2 Introduction of Dataset

This dataset was taken from professor Amba Sekhar which is used in BUSIE 702 course at NYIT. This dataset includes 14,999 observations on 11 variables with total 164,989 value. Table 1-1 shown below was the sample data we used for this case. The Response Variable is 'left'. The Explanatory variables were 'satisfaction_level', 'last_evaluation', 'number_project', 'average_montly_hours', 'time_spend_company', 'work_accident', 'promotion_last_5years', 'is_smoker' and 'salary'. We had 4 character variables ('left', 'is_smoker', 'department' and 'salary'), the remaining 7 variables are numerical variables.

satisfaction_ level	last_ evaluation	number_ project	average_ montly_hours	time_spend_ company	work_ accident	left	promotion_ last_5years	is_ smoker	department	salary
0.38	0.53	2	157	3	0	yes	0	*NA*	sales	low
0.8	0.86	5	262	6	0	yes	0	yes	sales	medium
0.11	0.88	7	272	4	0	yes	0	*NA*	sales	medium
0.72	0.87	5	223	5	0	yes	0	*NA*	sales	low
0.37	0.52	2	*NA*	*NA*	0	yes	0	no	sales	low

Table 1-1. Preview of the HR Employee Attrition dataset

2.0 Data Cleaning

First, the missing data (*NA* value)of the whole dataset needs to be identified. According to table 1-2 below, there are 3 columns has *NA* values, 'average_montly_hours', 'time_spend_company' and 'is_smoker'. Since each variable has only 14,999 data, but 'Is_smoker' is missing 14,764 data, there are too many missing data, so this variable will be deleted directly. Under this dataset is large enough, and 'average_montly_hours', 'time_spend_company' are also numerical variable, those *NA* value filled in with their means.

satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company
0	0	0	368	151
work_accident	left	promotion_last_5years	is_smoker	department
0	0	0	14764	0
salary				
0				

Table 1-2. Number of missing data in HR Employee Attrition dataset.

Then, we converted the character column 'left', 'department' and 'salary' to factor, also converted the numeric column 'work_accident' and 'promotion_last_5years' to factor. Also, we noticed that there were some factor levels in the column. We have 'department' and 'salary' with 10 levels and 3 levels, 'left' have two levels with 'yes' and 'no'. 'work_accident' and 'promotion_last_5years' has two levels 1 and 0, then we exchange and rename 1 and 0 as 1 = 'YES' and 0 = 'NO'. This pre-processing prepares the model for subsequent training and allows for more accurate prediction and analysis.

3.0 Exploratory Data Analysis (EDA)

Firstly, exploring the data in this dataset is to find the relationships between variables. Since in this report we are mainly focused whether employees leave their jobs or not, the main focus on the variable 'left'. After converting the factor variable to numerical variable, we can use correlation function to check each variables correlation. The table 1-3 is the correlation value about the 'left'. Larger numbers indicate stronger relationships, with positive indicating positive correlation and negative numbers indicating negative correlation.

(Variable Name)	left
satisfaction_level	-0.388374983
last_evaluation	0.006567120
number_project	0.023787185
average_montly_hours	0.070678890
time_spend_company	0.143317003
work_accident	-0.154621634
left	1.000000000
promotion_last_5years	-0.061788107
department	0.009935740
salary	-0.001293717

Table 1-3. Correlation value of left.

The figure 1-1 is the Correlation Heatmap of HR dataset. It is same with table 1-3. The darker the color indicates a stronger relationship, with red indicating a positive correlation and blue a negative correlation. From table 1-3 and figure 1-1, It can be seen, the strongest relationship with 'left' is 'satisfaction_level', followed by 'work_accident', then 'time_spend_company', and then 'average_montly_hours' and so on.

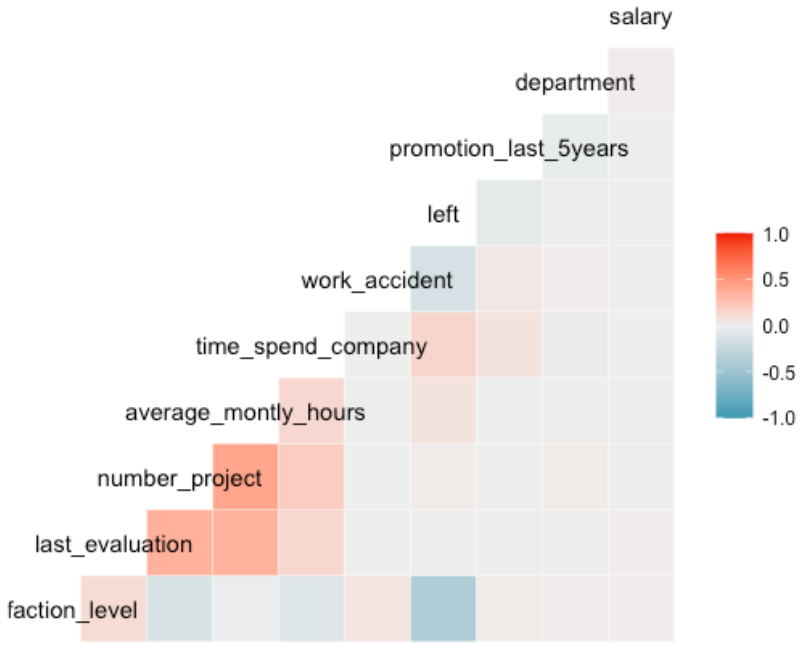

Figure 1-1. Correlation Heatmap of HR dataset.

1. 'left' & 'satisfaction_level'

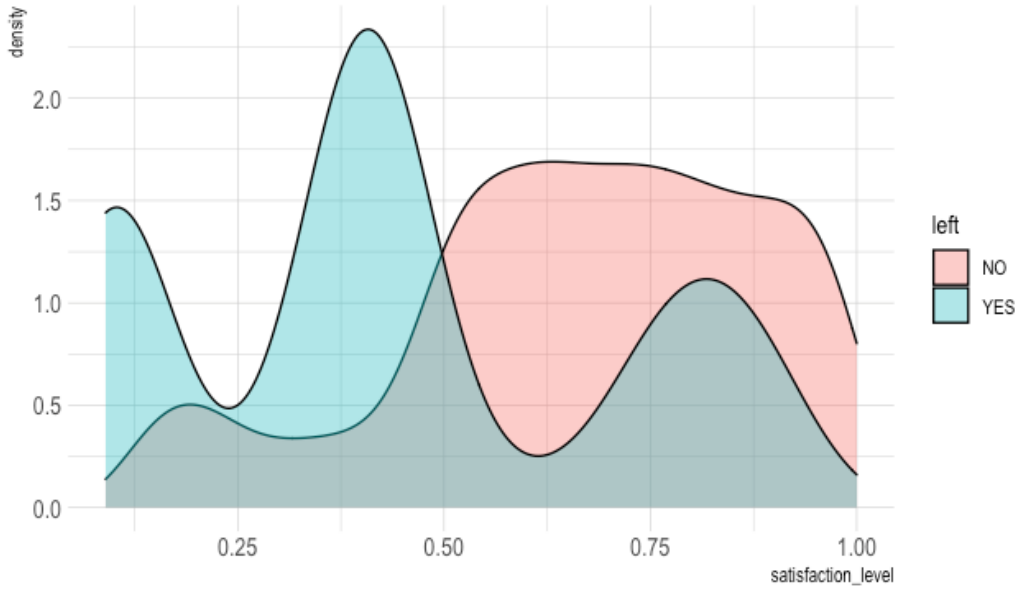

Figure 1-2. Density plot of 'left' and 'satisfaction_level'

The figure 1-2 is the density plot of 'left' and 'satisfaction_level'. 'satisfaction_level' is means how satisfied the employee is with their current job, possibly with the type of work or the number of hours worked. And the density plot is a representation of the distribution of a numeric variable. It uses a kernel density estimate to show the probability density function of the variable. This density plot shows the distribution of 'YES' and 'NO'. It can be seen, as the level of satisfaction increases, 'YES' gradually decreases and 'NO' gradually increases. When the degree of satisfaction is small, the distribution of 'YES' is more obvious and denser, which can show a strong negative correlation. There is a direct negative correlation between satisfaction and whether or not an employee leaves. The higher the satisfaction level, the less likely the employee is to leave.

2. 'left' & 'salary' & 'satisfaction_level'

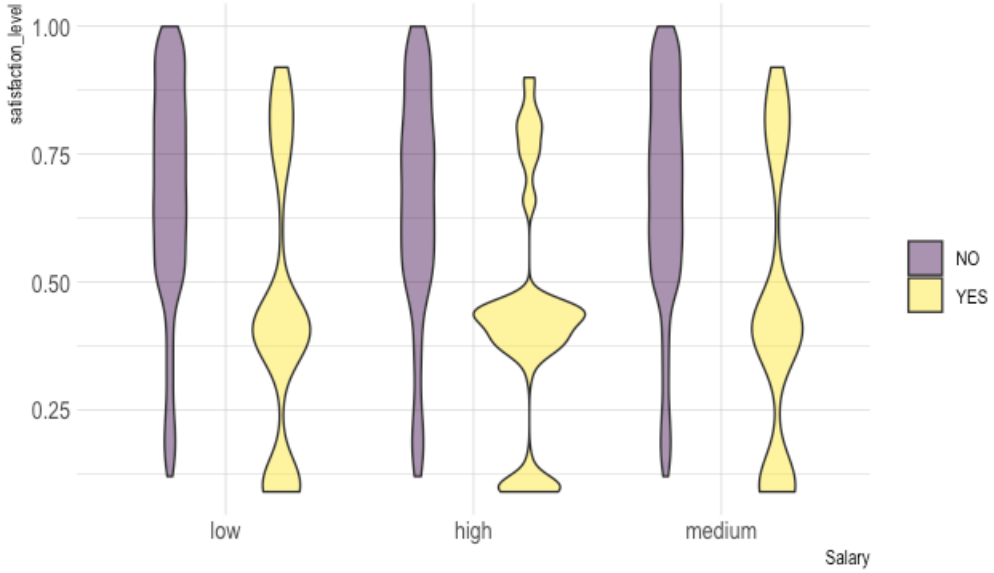

Figure 1-3. Grouped Violin chart of 'left', ''salary'' and 'satisfaction_level'

Each 'violin' represents a group or a variable. The shape represents the density estimate of the variable: the more data points in a specific range, the larger the violin is for that range. A grouped violin plot displays the distribution of a numeric variable for groups and subgroups. Here, groups are salary of the employee, and subgroups are 'YES' and 'NO' in variable 'left', and Purple is 'NO', yellow is 'YES'.

From figure 1-3, It can be seen, when satisfaction between 0.25 to 0.5, employees are more likely to choose to leave, and when satisfaction is greater than 0.5, employees are more likely to choose not to leave. This is the same trend as in Figure 1-2. But whether the employee's salary is high or low, the distribution between those who leave and those who do not leave is not very different at the same satisfaction value.

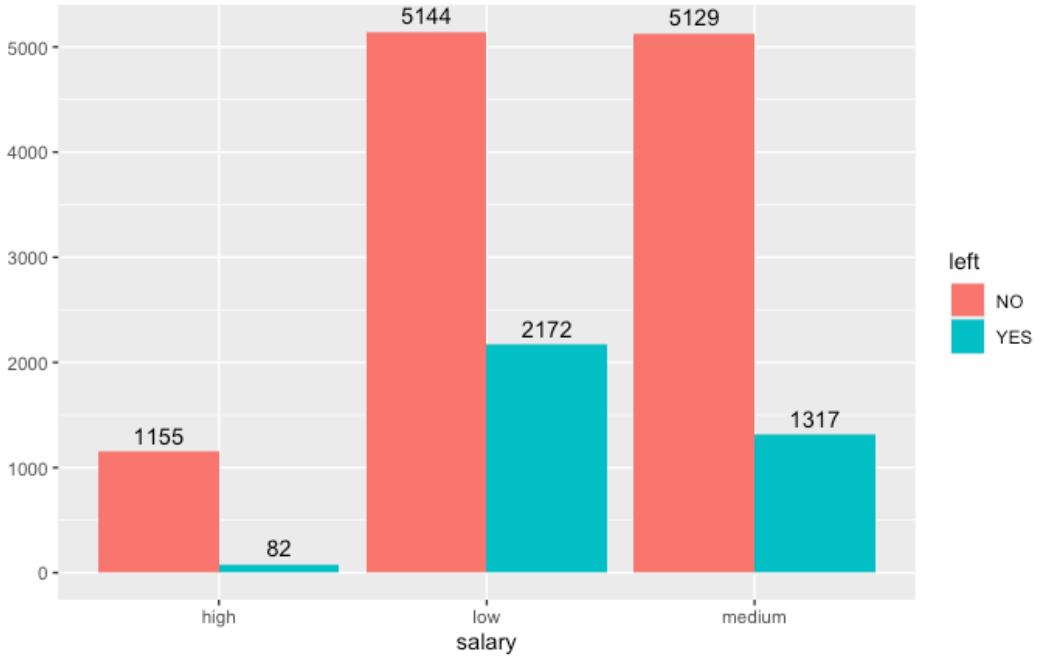

Figure 1-4. Barplot of 'left' and ''salary'.

The figure 1-4 is the Barplot of 'left' by ''salary'. This Barplot was created to clearly show the total number of employees who left at different salaries or not. Therefore, we can calculate the employees left rate.

'Salary'	high	low	medium
Left rate	6.63%	29.7%	20.4%

Table 1-4. Employees left rate by salary.

According to the table 1-4,the employees left rate on different salary. If satisfaction is not considered. It can be seen the higher of salary, the lower the left rate, and the lower the salary, the more likely of the employee to choose to leave. This suggests that there may be a relationship between whether an employee leaves or not with employee's salary, but there is no big direct correlation with satisfaction; the satisfaction is more important to the employee.

3. ‘left’ & ‘salary’ & ‘department’

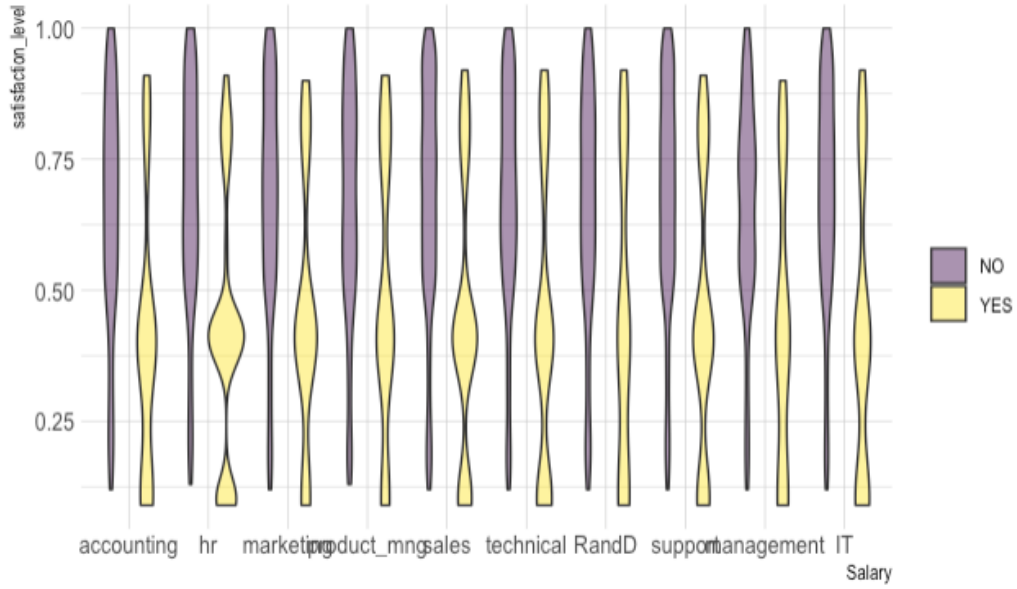

Figure 1-5. Grouped Violin chart of ‘left’, ‘‘department’ and ‘satisfaction_level’

This figure 1-5 Violin chart plot is quite similar with figure 1-3, only different is change the group from ‘salary’ to ‘department’. It same with before that when satisfaction between 0.25 to 0.5, employees are more likely to choose to leave, but we can see in ‘hr’, ‘marketing’ and ’sales’ department, the density is higher. It’s means employee are most likely to choose to leave than other department. This may be because in the company, employees in these three departments need to face a variety of customers on a regular basis, and may be more stressed and not well adjusted, so they are more likely to choose to leave when satisfaction is low.

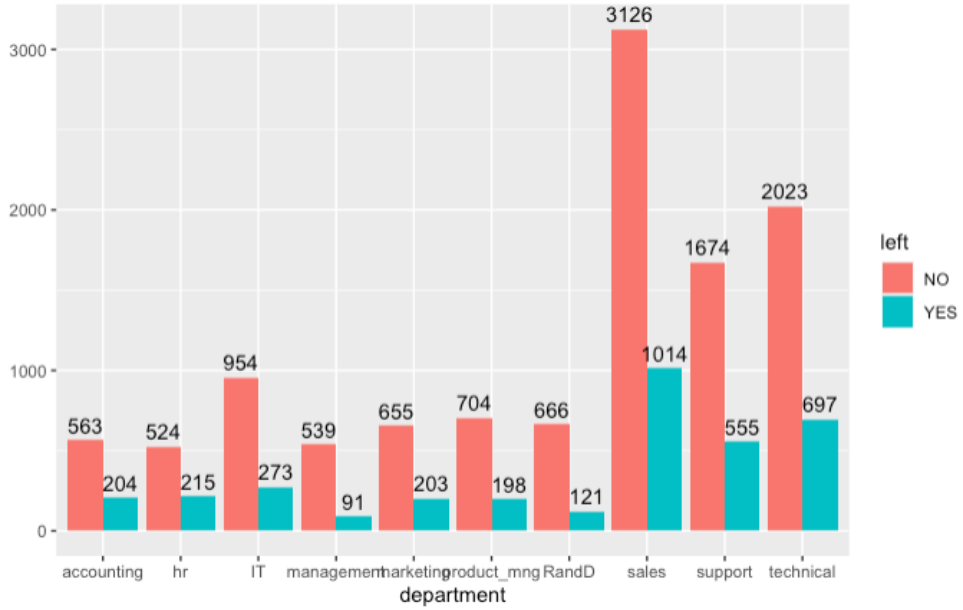

Figure 1-6. Barplot of ‘left’ and ‘‘department.

The figure 1-6 is the barplot of 'left' by ''department. This barplot was created to clearly show the total number of employees who left at different department or not. Therefore, we can calculate the employees left rate.

	accounting	hr	IT	management	marketing	Product_mmg	RandD	sales	support	technical
left rate	26.6%	29.1%	22.3%	14.4%	23.7%	22%	15.4%	24.5%	21.5%	25.6%

Table 1-5. Employees left rate by department.

According to the table 1-5,the employees left rate on different department. If satisfaction is not considered. It can be seen the department that most likely to choose 'YES' are 'hr', 'accounting', 'technical' 'sales' and so on. These departments are the core departments of some companies, which may have a high daily workload and high pressure, resulting in a high left rate.

4. 'left' & 'time_spend_company' & 'department'

Figure 1-7. Grouped Violin chart of 'left', ''department' and 'time_spend_company'.

In this Violin chart(figure 1-7), groups are department of the employee, and subgroups are 'YES' and 'NO' in variable 'left', and Purple is 'NO', yellow is 'YES'. This violin chart, consisting of 'left', 'time_spend_company' and 'department', is mainly for analyzing the relationship between 'left' and 'time_spend_company'. In their second year at the company, almost no employees choose to leave. From the third year to the sixth year, almost every year more employees want to leave the company than the people who choose to stay, no matter which department they work in. After the sixth year, there were almost no employees who wanted to leave the company.

This is a positive correlation, as the time spent in the company becomes longer, more people will choose to leave. Possible reasons for this are they feel they have low wages, no promotion space, or are dissatisfied with their position, or want to move to a bigger company, all of which may choose to leave the company in the first few years.

5. 'left' & 'number_project'

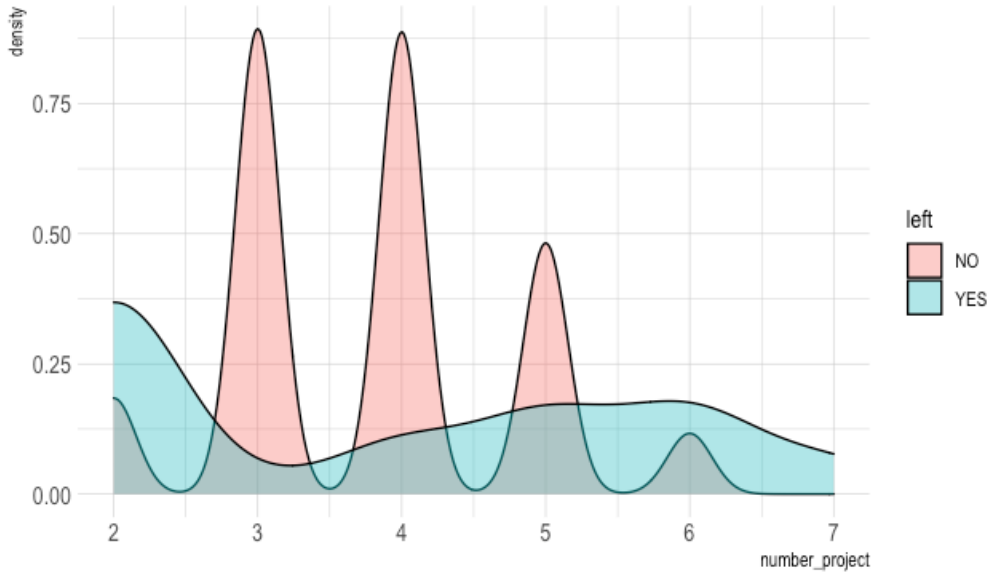

Figure 1-8. Density plot of 'left' and 'number_project'

The figure 1-8 is the density plot of 'left' and 'number_project'. 'number_project' is means how many project were completed in company by employees. It can be seen, as the number_project increases, 'NO' gradually decreases and 'YES' gradually increases. When the 'number_project' is 2, the distribution of 'YES' is also obvious and denser, so which can show a little positive correlation. There is a little positive correlation between number of project and whether or not an employee leaves. The more the project are complete or no project, the more likely the employee is to leave.

6. 'left' & 'promotion_last_5years'

The figure 1-4 is the barplot of 'left' by ''promotion_last_5years'. This barplot was created to clearly show the total number of employees whether left and promotion or not. Therefore, we can calculate the employees left rate. the table 1-6.

Table 1-6 shows that the left rate for employees who have not been promoted in the last five years is 24.2% and for those who have been promoted is 5.96%. There is clearly a negative correlation, with employees who have been promoted in the past generally choosing to stay, while those who have not been promoted have a 24.2% probability to leaving the position in company.

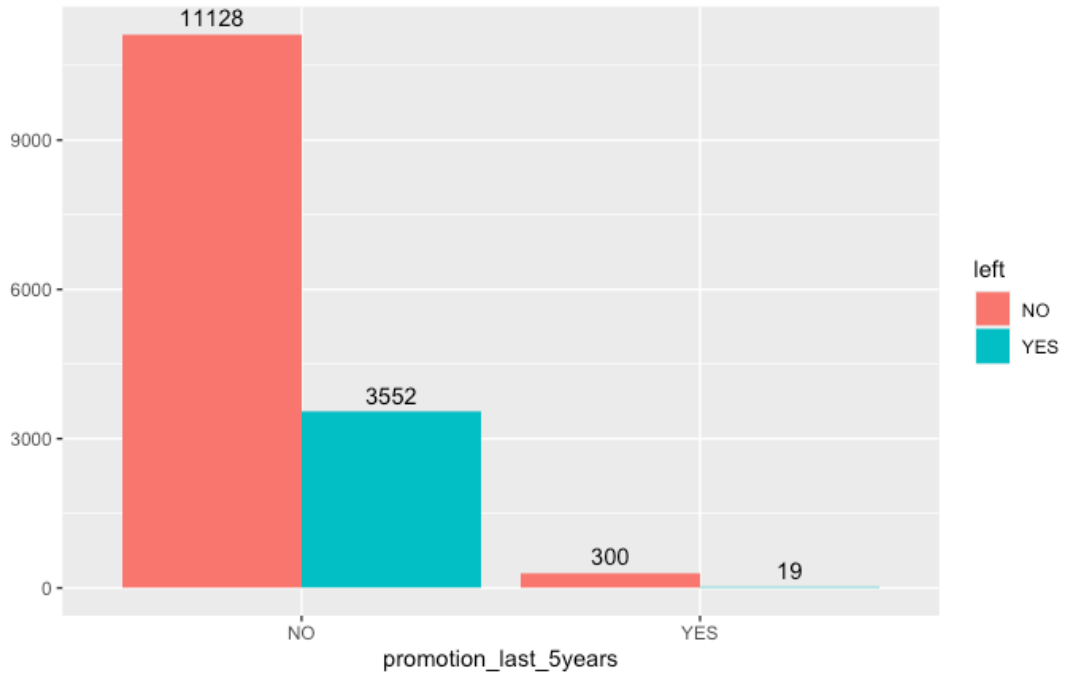

Figure 1-9. Barplot of 'left' and ''promotion_last_5years.

promotion in last 5 years	NO	YES
left rate	24.2%	5.96%

Table 1-6. Employees left rate by whether promotion in last 5 years.

4.0 Machine Learning Models

The three machine learning models in this project are created, analyzed and predicted using the CARET package in R Studio.

First, we need to partition the training sets from the testing sets. using set.seed() to create a starting point for randomly generated numbers, so that each time the code is run the same data sets. The "createDataPartition" in CARET does this by taking a stratified random sample of 0.75 of the data for training. Which means separate 75% as training data, remaining 25% as test data. We then create both the training and testing data sets which will be used to develop and evaluate the model.

4.1 Random Forest

The first method uses a random forest model that It builds and combines multiple decision trees, it can get more accurate predictions. It's a non-linear classification algorithm, because they choose predictors randomly at a time of training. Here we are creating the cross validation method that will be used by CARET to create the training sets. Cross validation means to randomly split the data into k-fold (in our case 5) data testing data sets and the repeated part just means to repeat this process k times (in our case 0, because too much time to run).

mtry	Accuracy	Kappa
2	0.9692444	0.9118442
10	0.9872000	0.9642881
18	0.9855111	0.9596727

Table 1-7. Random Forest model Accuracy.

Let's now inspect the results from table 1-7. The most important piece of information is the Accuracy, because that is what CARET uses to choose the final model. It is the overall agreement rate between the cross validation methods. The Kappa is another statistical method used for assessing models with categorical variables such as ours.

CARET chose the final value used for the model was mtry = 10., an accuracy of 98.7% and a Kappa of 96.4%.

	NO <dbl>	YES <dbl>
1	0.000	1.000
2	0.000	1.000
3	0.032	0.968
4	0.002	0.998
5	0.000	1.000
6	0.000	1.000

Table 1-8. Random Forest model predict probability of the result

Table 1-8. shows the probability of the predicted outcome of the random forest model, only the first six rows of the test dataset are shown, and the numbers indicate the probability of the employee choosing "YES" or "NO". That is, the probability of predicting the employee's leaving the job.

Actual \ Prediction	NO	YES
NO	2846	11
YES	28	864

Table 1-9. (Confusion Matrix) Validation of the predicted results

Confusion Matrix show, total 2874 'NO', there are 2846 are correctly classified as 'NO', 28 are wrong to classified as 'YES'; total 875 'YES', there are 864 are correctly classified as 'YES', 11 are wrong to classified as 'NO'.

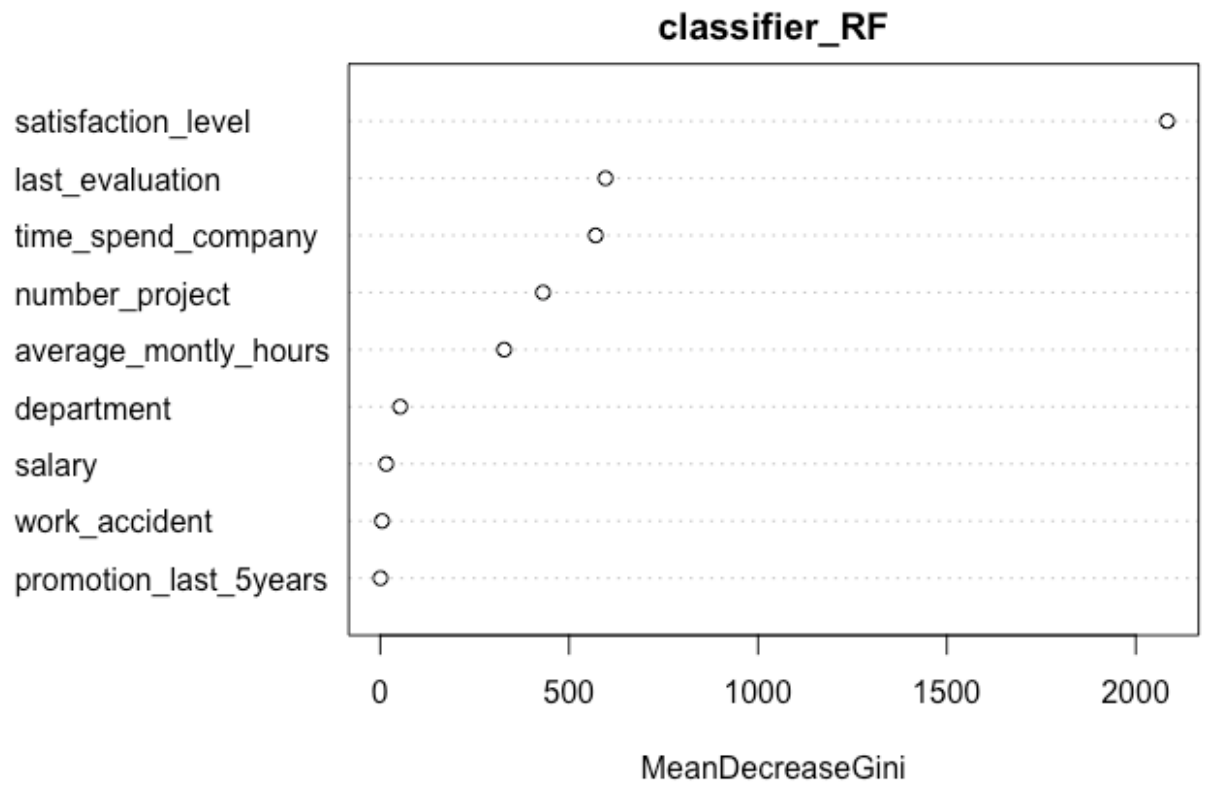

Figure 1-10. Variable importance plot

The Random Forest model also can created a Variable importance plot. The plot(Figure 1-10) clearly shows 'satisfaction_level' as the most important feature or variable followed by 'last_evaluation', 'time_spend_company' and 'number_project'.

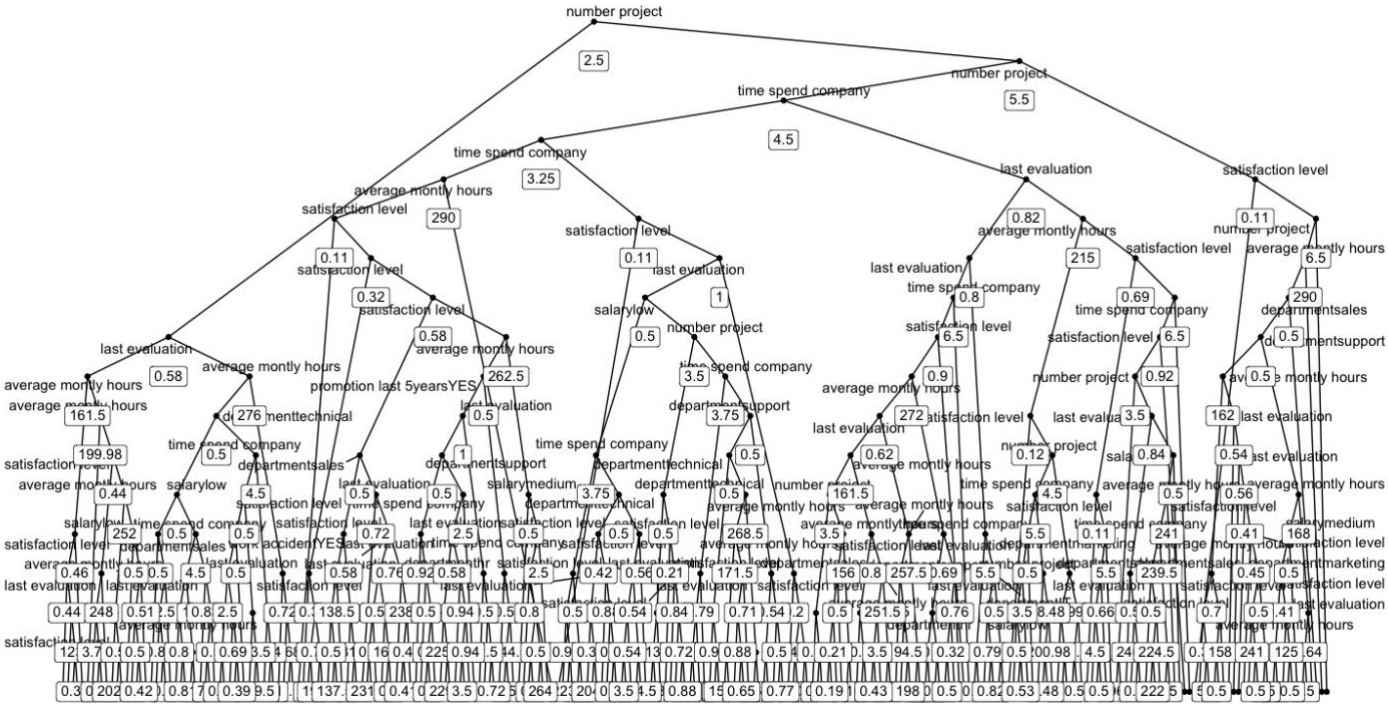

Figure 1-11. Random Forest tree plot

The above figure1-11 shows the decision tree diagram of the random forest model we used. The decision tree diagram can help us understand which decisions are the basis of the prediction, because we get a final model with ntree=500, and with a large amount of data, so Figure 1-11 is only a part of the complete decision tree diagram. At each split point, it means that this variable is less than or equal to this number, to the left if it is TRUE, and to the right if it is FALSE, until the final result . For example, the first one at the top, if the number of project $\leqslant$ 2.5 is TRUE, the data go to left split, if not $\leqslant$ 2.5, then the data go to left split.

4.2 k-Nearest Neighbors Algorithm (KNN)

K-Nearest Neighbor or K-NN is a Supervised Non-linear classification algorithm. K-NN is also a Non-parametric algorithm. It doesn't make any assumption about underlying data or its distribution. It is one of the simplest and widely used algorithm which depends on its k value(Neighbors). Here we are creating the cross validation method that will be used by CARET to create the training sets. we randomly split the data into k-fold (in our case 5) data testing data sets and the repeated part just means to repeat this process k times (in our case 5).

k	Accuracy	Kappa
5	0.9344536	0.8255241
7	0.9359114	0.8289188
9	0.9346670	0.8257535

Table 1-10. KNN model Accuracy.

CARET chose the final value used for the model was k = 7., an accuracy of 93.59% and a Kappa of 82.89%.

NO <dbl>	YES <dbl>
0.0000000	1.0000000
0.0000000	1.0000000
0.4285714	0.5714286
0.0000000	1.0000000
0.0000000	1.0000000
0.0000000	1.0000000

Table 1-11. Random Forest model predict probability of the result

Table 1-11. shows the probability of the predicted outcome of the KNN model which is the probability of predicting the employee's leaving the job. only the first six rows of the test dataset are shown.

Actual \ Prediction	NO	YES
NO	2678	179
YES	71	821

Table 1-12. (Confusion Matrix) Validation of the predicted results

Confusion Matrix show, total 2749 'NO', there are 2678 are correctly classified as 'NO', 71 are wrong to classified as 'YES'; total 1000 'YES', there are 821 are correctly classified as 'YES, 179 are wrong to classified as 'NO'.

4.3 Stochastic Gradient Boosting (GBM)

A Gradient Boosting Machine or GBM combines the predictions from multiple decision trees to generate the final predictions. we are creating the cross validation method that will be used by CARET to create the training sets. we randomly split the data into k-fold (in our case 5) data testing data sets and the repeated part just means to repeat this process k times (in our case 5).

interaction.depth	n.tree	Accuracy	Kappa
1	50	0.9053867	0.7167475
1	100	0.9103290	0.7355111
1	150	0.9147201	0.7486283
2	50	0.9424709	0.8353073
2	100	0.9648002	0.9012126
2	150	0.9698135	0.9153379
3	50	0.9677335	0.9094192
3	100	0.9720355	0.9214204
3	150	0.9739200	0.9267969

Table 1-13. GBM model Accuracy.

The final values used for the model were n.trees = 150, interaction.depth = 3, shrinkage = 0.1 and n.minobsinnode = 10. An accuracy of 97.39% and a Kappa of 92.68%.

NO <dbl>	YES <dbl>
0.040453534	0.9595465
0.219171829	0.7808282
0.112758601	0.8872414
0.025560015	0.9744400
0.007892455	0.9921075
0.040453534	0.9595465

Table 1-14. GBM model predict probability of the result

Table 1-14. shows the probability of the predicted outcome of the GBM model which is the probability of predicting the employee's leaving the job. only the first six rows of the test dataset are shown.

Actual \ Prediction	NO	YES
NO	2829	28
YES	80	812

Table 1-15. (Confusion Matrix) Validation of the predicted results

Confusion Matrix show, total 2909 'NO', there are 2829 are correctly classified as 'NO', 80 are wrong to classified as 'YES'; total 840 'YES', there are 812 are correctly classified as 'YES, 28 are wrong to classified as 'NO'.

4.4 Model Conclusion

	Accuracy
RF	98.72%
KNN	93.59%
GBM	97.39%

Table 1-16. Accuracy for each models.

Finally, comparing the accuracy of these three models, it is found that Random Forest has the highest accuracy, so in this case the Random Forest model fits best.

5.0 Conclude with Business Implications/Recommendations

From the above analysis we can find:

- 'satisfaction' is the most important indicator that affects the employee turnover rate. Other indicators such as 'work_accident', 'time_spend_company', and 'average_montly_hours' also have big impacts.
- If the satisfaction level can be improved to above 0.5, the turnover rate will decrease greatly.
- When employees have the same satisfaction level about the company, the salary levels don't have big influence on turnoverate, but hier salary levels will induce lower turnover rate in general.
- Some departments like sales, support, technical departments have higher turnover rate than other departments.

- Among the first six years employee's stay at the company, the longer they stay, the higher turnoverate can be seen. In this case, traning and career development can help them have a more clear future plan and more associated with the company.
- Maintaining the number of projects at around three can achieve low turnover rate.

To reduce the turnover rate. The key is to improve the satisfaction level. Here are some suggestions to decrease the turnover rate:

Work Design:

To improve employee engagement and sense of achievement by optimizing work design. Work design can include job rotation, work augmentation, and work enrichment. By doing this, giving employee more opportunities to improve themselves and earn more money. At the same time, keep employee have a balanced work and life.

Effective Communication:

Through communication to know emplyee's needs and complains. If their complains can be solved in time and proper needs can be satisfied, they are less likely to leave. This is far more important for deparments with higher turnover rates than other departments such as sales, technical, support departments.

Compensation and Incentives:

Provide fair, reasonable and competitive compensation, benefits and incentives. Firstly, to maintain employee, their salary level should be above the market average. When the salary maintains at a higher level, there are less people to leave. In addition, transparent performance evaluation system is needed to motivate employees.

Training and Career Development:

Training can enable employee with a sustainable career. Through training, employees can have the opportunity to get promoted and achieve self-improvement.

CASE 2 - Finance Practicum Data

6.0 Project Aims

In case 2,in order to understand the relationships between different financial metrics of companies and figure out ways to build a highly valued company, we analyze the financial performance of companies. By running machine learning models to explore the correlation of financial metrics, we summarize important factors that can help a company to improve its value.

7.0 Introduction of Dataset

This dataset was taken from professor Amba Sekhar which is used in BUSIE 702 course at NYIT. This dataset includes 504 observations on 22 variables with total 11,088 value. Table 2-1 shown below was the sample data we used for this case. The Response Variable is 'Current Market Cap. The remaining 21 variables are explanatory variables, include `Last Price', `Alpha`, `Beta`, `ROA`, `ROE', `Profit Margin`, `Asset Turnover` and so on. We had 14 character variables and 8 variables numerical variables.

	Current Market Cap	Last Price	Alpha	Beta	Revenue Growth Year over Year	ROA	ROE	…
A UN Equity	3.838E+10	128.49	0.127765	0.993224	18.3555	12.07887	25.36051	…
AAL UW Equity	1.054E+10	16.22	-0.22221	1.223408	72.35969	-6.11834	#N/A N/A	…
AAP UN Equity	1.157E+10	190.73	0.098263	0.928327	8.822874	6.332714	23.57857	…
AAPL UW Equity	2.353E+12	145.38	0.299396	1.129532	33.25938	29.62394	149.1901	…
ABBV UN Equity	2.601E+11	147.17	0.321832	0.723219	22.69016	11.38073	111.4479	…
ABC UN Equity	3.08E+10	147.03	0.217093	0.890334	12.68862	3.891975	#N/A N/A	…

Table 2-1. Preview of the Finance dataset

8.0 Data Cleaning

First, we first changed the name of each variable, for example from 'Current Market Cap' to 'Current.Market.Cap'. Removed the spaces to make the variable names more scientific. And assigned the name 'Equity' to the first column. Then, we converted 'Equity' and 'Industry.Group' to the factor variable, and converted all others variables to the numeric variable. First, the missing data ('#N/A N/A' value)of the whole dataset needs to be identified. According to table 2-2 below, there are 12 columns, total 238 NA values, those NA values all filled in with their variables means.

Equity	Current.Market.Cap	Last.Price
0	0	0
Alpha	Beta	Revenue.Growth.Year.over.Year
7	7	0
ROA	ROE	Profit.Margin
2	34	0
Asset.Turnover	Financial.Leverage	EBIT.Interest
3	27	87
Price.Earnings.Ratio	Price.to.Sales.Ratio	Short.and.Long.Term.Debt
24	6	0
Total.Equity	Operating.Return.on.Total.Invested.Capital	Weighted.Average.Cost.of.Cap
0	3	0
Free.Cash.Flow	Price.to.Book.Ratio	Industry.Group
0	36	0
WACC.Economic.Value.Added	Total.Compensation.Paid.to.Executives	
0	2	

Table 2-2. Number of missing data in Finance dataset.

9.0 Exploratory Data Analysis (EDA)

First we explored the Response variable (‘Current Market Cap’) for the prediction, and by creating histogram and box plots, we found that the data for this variable was not normally distributed and had a large number of outliers.

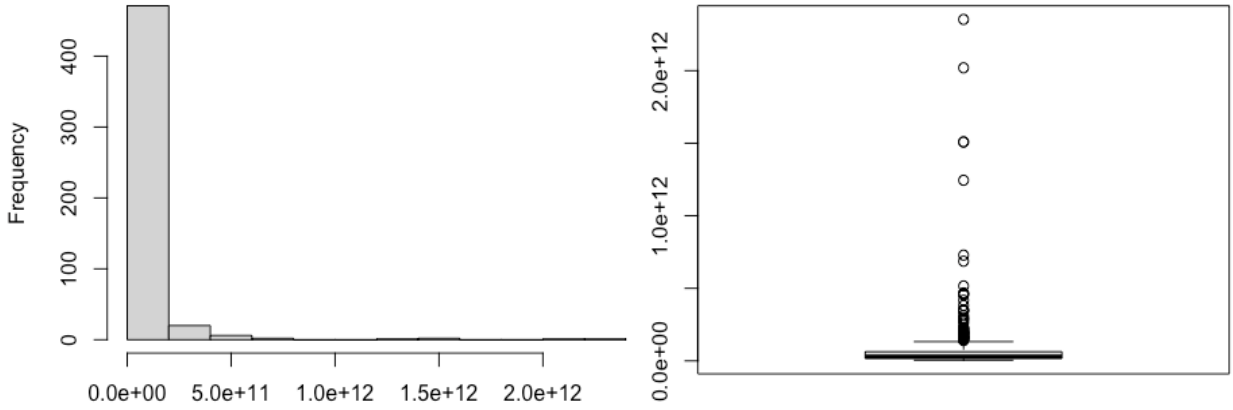

Figure 2-1. Histogram and box plots of ‘Current Market Cap’.

Using the box plot detection method, 57 outliers were detected. It was decided to keep them for the data integrity and accuracy.(table 2-3)

```
> boxplot.stats(finance_df$Current.Market.Cap)$out
 [1] 2.353002e+12 2.600656e+11 2.041073e+11 2.023020e+11 2.030616e+11 1.722600e+11 1.244839e+12 2.274786e+11 2.915785e+11 1.600418e+11
[11] 6.867916e+11 1.923270e+11 1.527823e+11 2.109611e+11 1.839854e+11 1.873786e+11 3.489509e+11 1.934388e+11 1.979406e+11 5.163124e+11
[21] 1.508659e+12 1.508659e+12 3.135475e+11 1.774217e+11 4.642319e+11 3.822865e+11 2.729768e+11 1.665892e+11 2.866156e+11 3.480317e+11
[31] 1.836739e+11 2.273649e+11 1.471498e+11 2.019489e+12 1.545865e+11 1.903472e+11 4.680000e+11 1.915470e+11 2.279354e+11 2.984997e+11
[41] 3.500334e+11 1.635676e+11 1.581664e+11 1.432634e+11 1.496231e+11 2.184749e+11 1.712522e+11 7.288848e+11 1.589113e+11 4.555855e+11
[51] 1.406776e+11 1.620749e+11 4.571641e+11 2.133839e+11 1.698836e+11 3.435209e+11 4.174209e+11
```

Table 2-2. Outliers in ‘Current Market Cap’

1. Correlation Heatmap

Then, exploring the relationships between variables. Since in this report we are mainly focused on company current market capitalization, the main focus on the variable 'Current.Market.Cap'. After converting the factor variable to numerical variable, we can using correlation function to check each variables correlation. The table 2-3 is the top 5 correlation value about the 'Current.Market.Cap'. Larger numbers indicate stronger relationships, with positive indicating positive correlation and negative numbers indicating negative correlation

(Variable Name)	Current.Market.Cap
Current.Market.Cap	1.000000000
Free.Cash.Flow	0.745445158
Total.Equity	0.545139962
Total.Compensation.Paid.to.Executives	0.478049011
WACC.Economic.Value.Added	0.445614694
Last.Price	0.328406549

Table 2-3. Correlation value of 'Current.Market.Cap'.

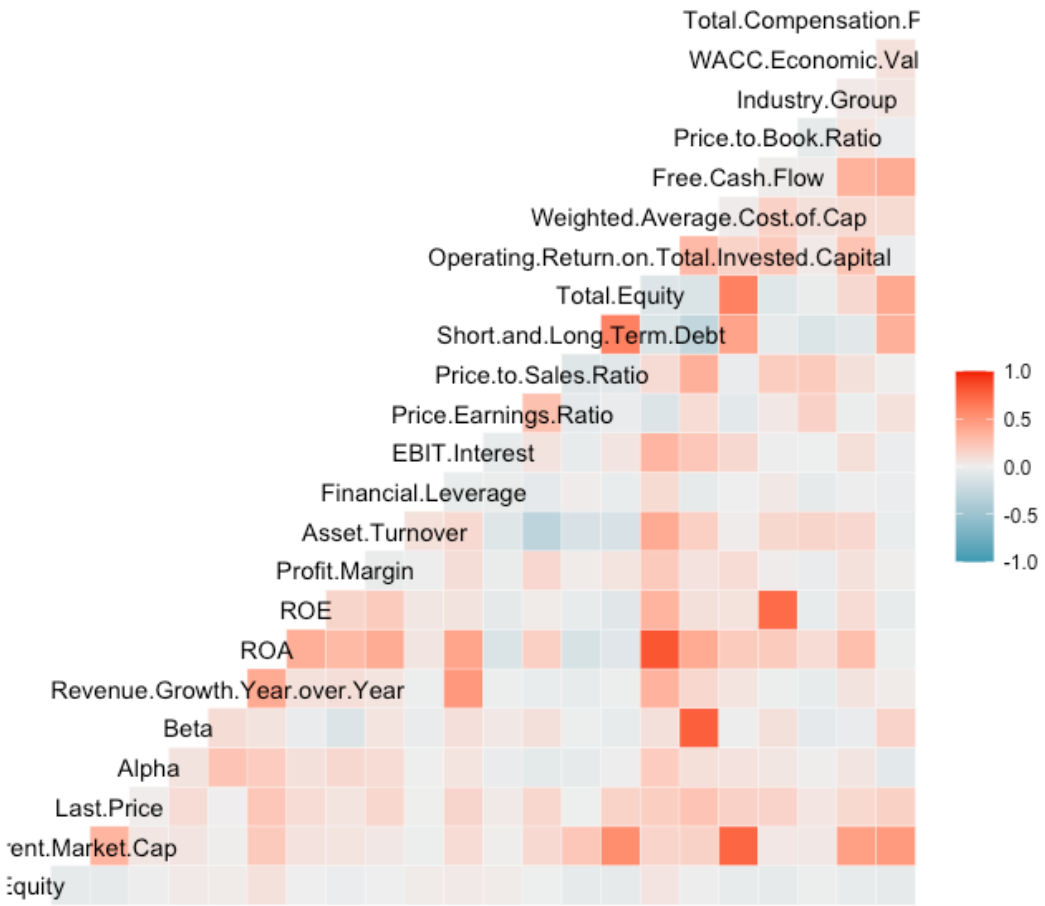

Figure 2-2. Correlation Heatmap of Finance dataset.

The figure 2-2 is the Correlation Heatmap of Finance dataset. The darker the color indicates a stronger relationship, with red indicating a positive correlation and blue a negative correlation. From table 2-3 and figure 2-2, It can be seen, the strongest relationship with 'Current.Market.Cap' is 'Free.Cash.Flow', followed by 'Total.Equity', then 'Total.Compensation.Paid.to.Executives', 'WACC.Economic.Value.Added' 'Last.Price' and so on. It is worth noting that all variables have a positive, or weak positive, correlation with ABC, with no direct negative correlation.

2. Show the data of 'Current.Market.Cap'

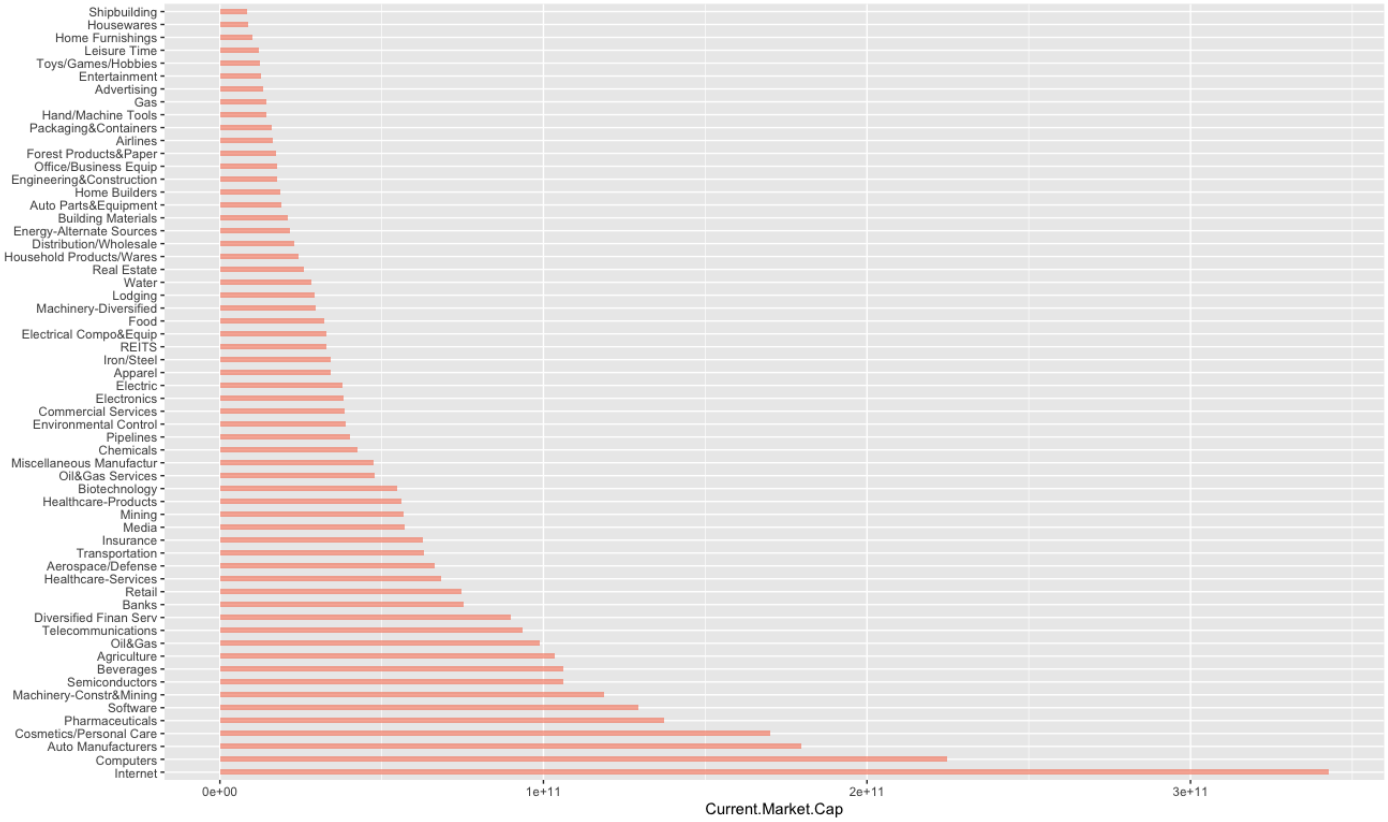

Figure 2-3. 'Current market cap' sorted by 'Industry.Group'.

First we need to check the 'Current market cap' status, so we are grouping data by 'Industry.Group', and other variables take their mean value. Then we can found the top 5 type high current market capitalization are Internet, Computers, Auto Manufacturers, Cosmetics/Personal Care and Pharmarceuticals. And found that the current market value of the Internet industry is extremely high. It shows that now is the information age, network and computer companies have better prospects and are more popular.

3. Correlation with 'Current.Market.Cap' & 'Free.Cash.Flow'

Relationship Hotspot We learned that the strongest relationship to current total market capitalization is free cash flow, so we created this linear scatter plot. From the figure, we can see that their has a strong linear relationship. With the increase of 'Free.Cash.Flow', the Current.Market.Cap' will also increase. Prove that a company's current market capitalization has a lot to do with how much free cash flowit has.

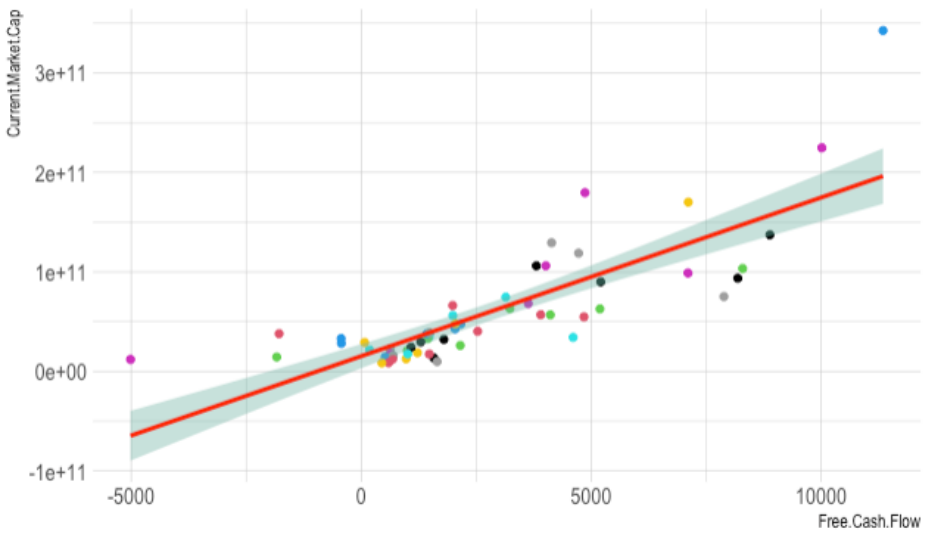

Figure 2-4. 'Linear Scatter plot ('Current.Market.Cap' & 'Free.Cash.Flow')

4. Circular barplot of 'Free.Cash.Flow'

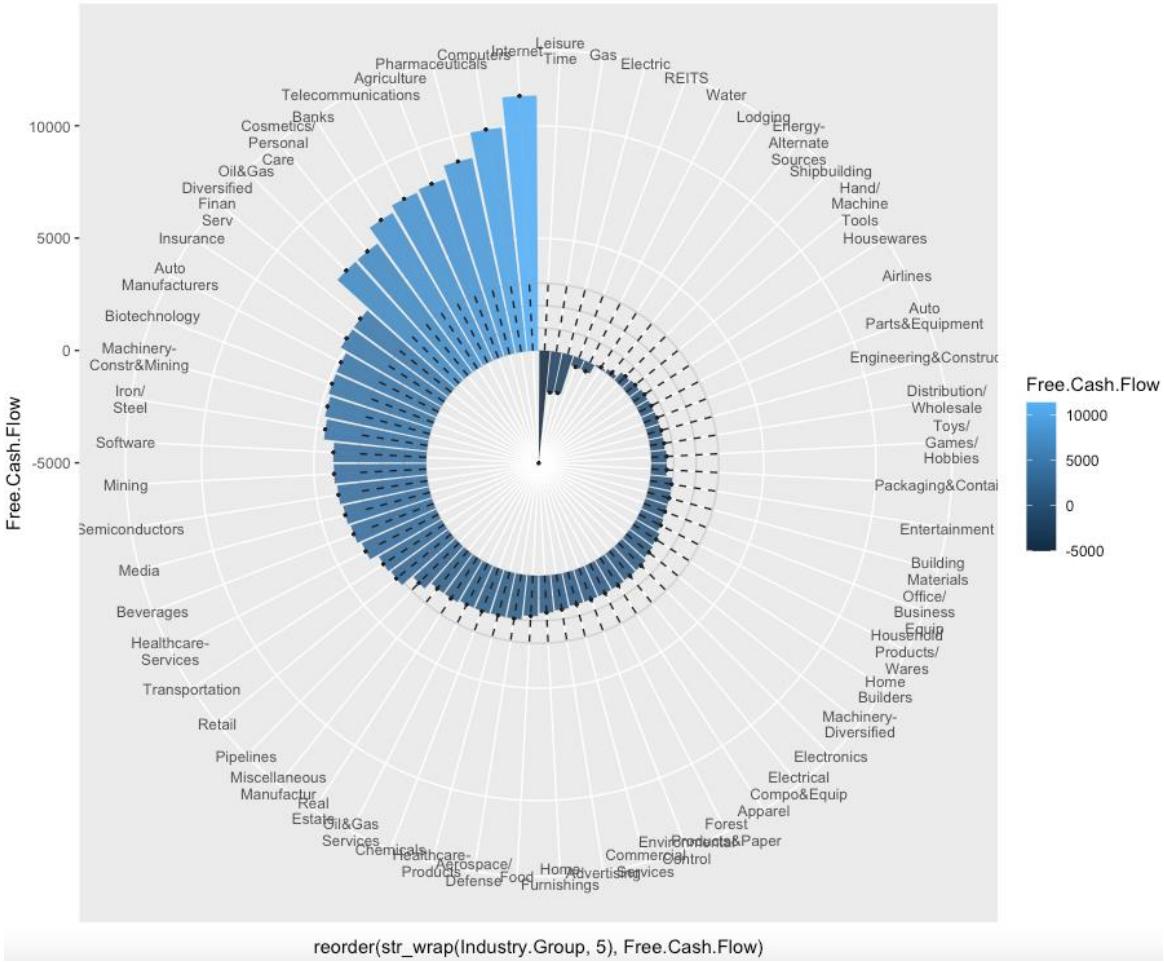

Figure 2-5. Circular barplot of 'Free.Cash.Flow'

Then we are going to check the 'Free.Cash.Flow' status, we are grouping data by 'Industry.Group', and other variables take their mean value. Then base on figure 2-5 we can found the top 5 type high Free.Cash.Flow are Internet, Computers, Pharmarceuticals, Agriculture and Telecommunications. It's quite fit with the correlation 'Current.Market.Cap'. And it's still the Internet technology companies on the top. We found one category of companies 'Leisure Time' has a huge negative free cash flow, and according to Figure 2-3, we can find that ' Leisure Time ' is not in the last place, but the fourth from the bottom. It can be found, Current market capitalization and free cash flow have a strong relationship, but not always.

5. Bubble chart with ('Current.Market.Cap' & 'Free.Cash.Flow' & 'Total.Equity')

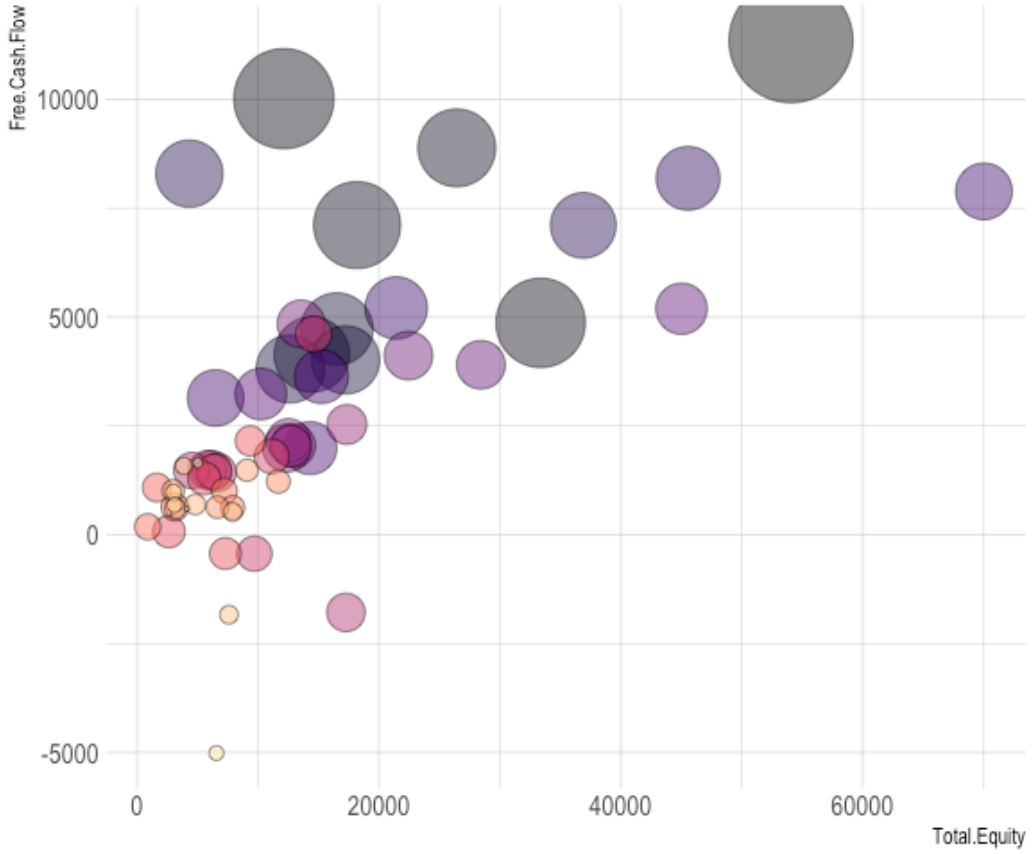

Figure 2-6. Bubble chart with ('Current.Market.Cap' & 'Free.Cash.Flow' & 'Total.Equity')

The figure 2-6 is the Bubble char show the relationship with 'Current.Market.Cap' & 'Free.Cash.Flow' & 'Total.Equity'. The larger the size of the bubble, the darker the color means the larger the 'current market cap', and vice versa. From this figure, we know 'Current.Market.Cap' & 'Free.Cash.Flow' & Total.Equity' they three all has positive correlation each other. This proves that as total equity grow, current market capitalization grows and free cash flow grows as well. But the rate of growth is not as fast as the rate of growth of free cash flow.

6. Bubble chart with ('Short.and.Long.Term.Debt' & 'Free.Cash.Flow' & 'Total.Equity')

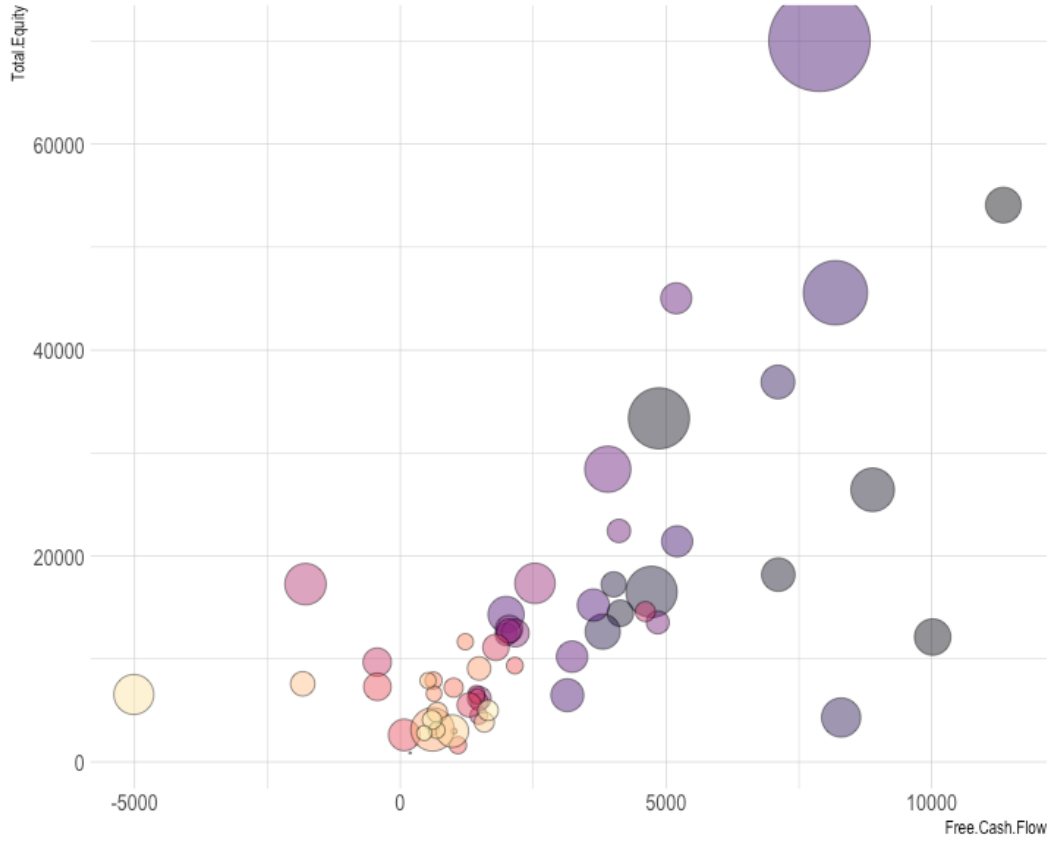

Figure 2-7. Bubble chart with ('Short.and.Long.Term.Debt' & 'Free.Cash.Flow' & 'Total.Equity')

This Bubble chart show the relationship with 'Short.and.Long.Term.Debt' & 'Free.Cash.Flow' & 'Total.Equity'. The larger the size of the bubble, the darker the color means the larger the 'Short.and.Long.Term.Debt', and vice versa. From this figure, we can found 'Short.and.Long.Term.Debt' & 'Free.Cash.Flow' & Total.Equity' they three all has positive correlation each other, but not really strong. This proves that as total equity grow, current market capitalization grows and free cash flow grows as well. But the rate of growth is not as fast as the rate of growth of free cash flow. This proves that the growth of cash flow may also mean the growth of debt, as the cash could be a bank loan. Total equity will also increase. But there is a huge risk of increasing debt in order to increase the current market capitalization, because debt is then accompanied by interest.

7. Bubble chart ('Current.Market.Cap', 'ROA' & 'Operating.Return.on.Total.Invested.Capital')

This is also a Bubble chart. Firstly, `ROA` refers to a financial ratio that indicates how profitable a company is in relation to its total assets. generally using ROA to determine how efficiently a company uses its assets to generate a profit. Return on Invested Capital (ROIC) is a return ratio that expresses recurring operating profits as a percentage of the company's net operational assets, bigger is good. In this plot (figure 2-8), the bubble is ROIC, the horizontal axis is the current market capitalization and the numerical axis is the ROA. As the current market capitalization and ROA increase, the ROIC is also gradually increasing.

Increasing a company's current market capitalization can be achieved by increasing ROA and ROIC. This may require the company to increase net income, or reduce debt.

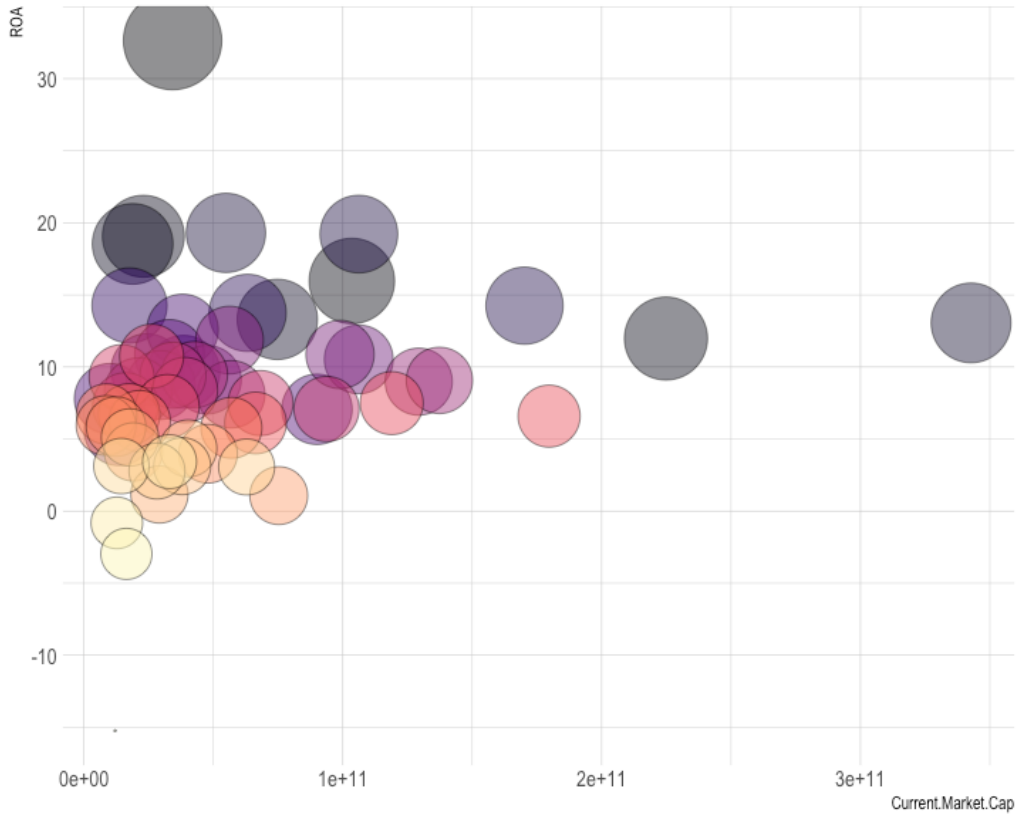

Figure 2-8.Bubble chart ('Current.Market.Cap', 'ROA' & 'Operating.Return.on.Total.Invested.Capital')

10.0 Machine Learning Models

The three machine learning models in this project are created, analyzed and predicted using by R Studio.

First, we need to partition the training sets from the testing sets. using set.seed() to create a starting point for randomly generated numbers, so that each time the code is run the same data sets. Taking a stratified random sample of 0.75 of the data for training. Which means separate 75% as training data, remaining 25% as testing data. We then create both the training and testing data sets which will be used to develop and evaluate the model.

10.1 Linear Regression (Stepwise)

In the first model, we use Linear Regression. The simplest form of regression is linear regression, which assumes that the predictors have a linear relationship with the target variable. The input variables are assumed to have a Gaussian distribution and are not correlated with each other (a problem called multi-collinearity).

lmFit <- lm(Current.Market.Cap ~., train)

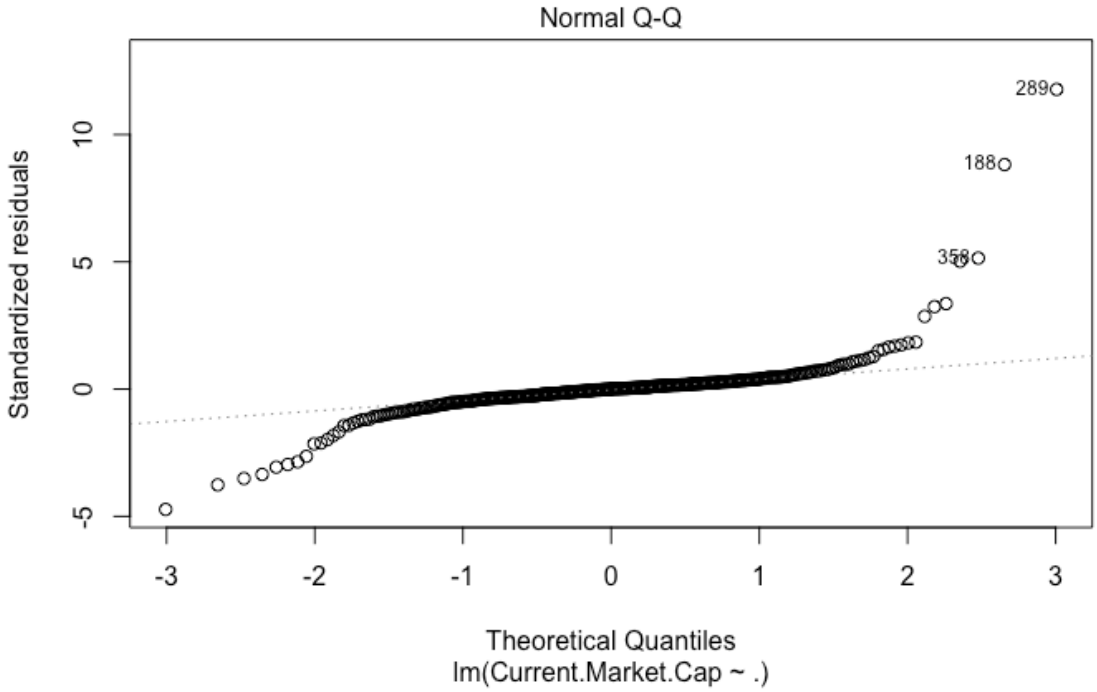

Figure 2-9. Normal Q-Q plot on linear regression model.

Using lm() to fit the linear regression model. After predict on test data, we will evaluate the result.

lm ()	Train Data	Test Data
R-square	0.8328	0.82

Table 2-4. lm() Model Evaluation

After that, we are perform variable selection by **Stepwise.** And then we got the importance variables and the best model. Here we are used the cross validation method to training the model. Cross validation will into k-fold (k=5).

Current.Market.Cap ~ Last.Price + Beta + Revenue.Growth.Year.over.Year + ROA + Asset.Turnover + Price.to.Sales.Ratio + Total.Equity + Weighted.Average.Cost.of.Cap + Free.Cash.Flow + Industry.Group + WACC.Economic.Value.Added + Total.Compensation.Paid.to.Executives

Then we can evaluate the **Stepwise** result.

Stepwise	Train Data	Test Data
R-square	0.8294	0.82

Table 2-4. Stepwise Model Evaluation

From the R-square above ,we can see the before and after improve are almost same, but we keep the highest value R-square from linear regression.

10.2 LASSO Regression

At the second model, we using LASSO Regression, Lasso regression is a method we can use to fit a regression model when multicollinearity is present in the data. To perform lasso regression, we setting alpha=1. Then in order to determine what value to use for lambda, we'll perform k-fold (k=10) cross-validation and identify the lambda value that produces the lowest test mean squared error (MSE).

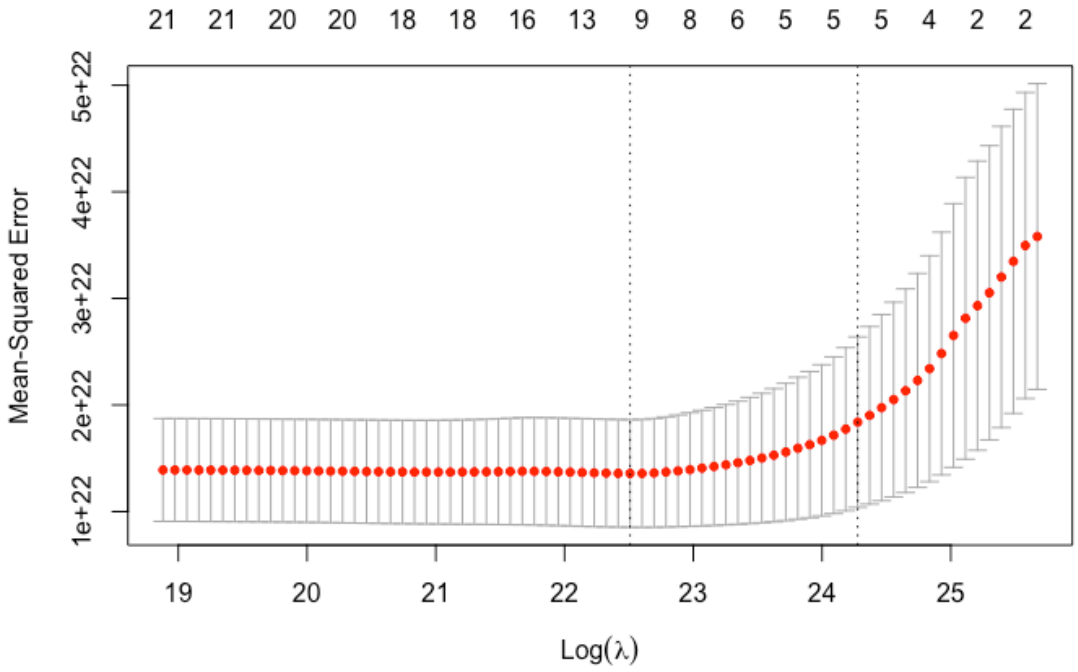

Figure 2-10. MSE vs Log(λ) on LASSO regression

The lambda value that minimizes the test MSE turns out to be 5963918151, log(λ) = 9.77, the first dash line on the Figure 2-10. Lastly, we can get the final model produced by the optimal lambda value. Need to notice is, it will no coefficient shown for the some predictor, because the lasso regression has the potential to remove predictors from the model by shrinking the coefficients completely to zero. This means it was completely dropped from the model because it wasn't influential enough.

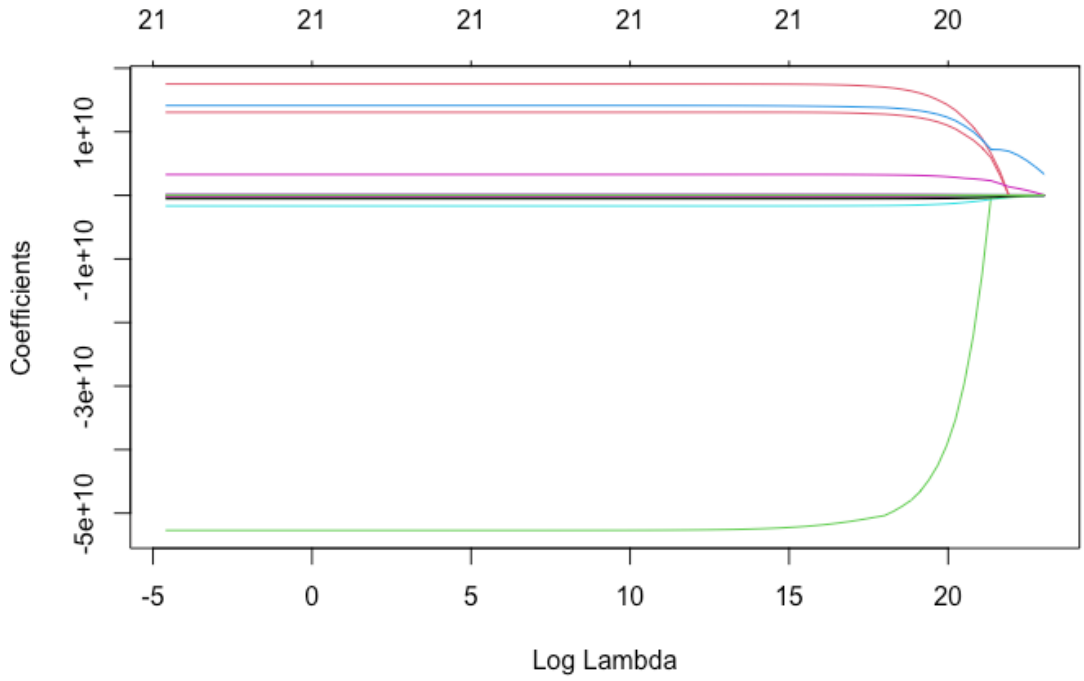

Figure 2-11. Trace plot on LASSO regression

Lastly, we can calculate the R-squared of the model on the training data and the testing data.

LASSO	Train Data	Test Data
R-square	0.8241	-1.55

Table 2-5. LASSO Model Evaluation

10.3 Ridge Regression

At the third model, we using Ridge Regression, Ridge regression is a method we can use to fit a regression model when multicollinearity is present in the data. Similar to LASSO. To perform Ridge regression, we setting alpha=0. Then in order to determine what value to use for lambda, we'll perform k-fold (k=10) cross-validation and identify the lambda value that produces the lowest test mean squared error (MSE).

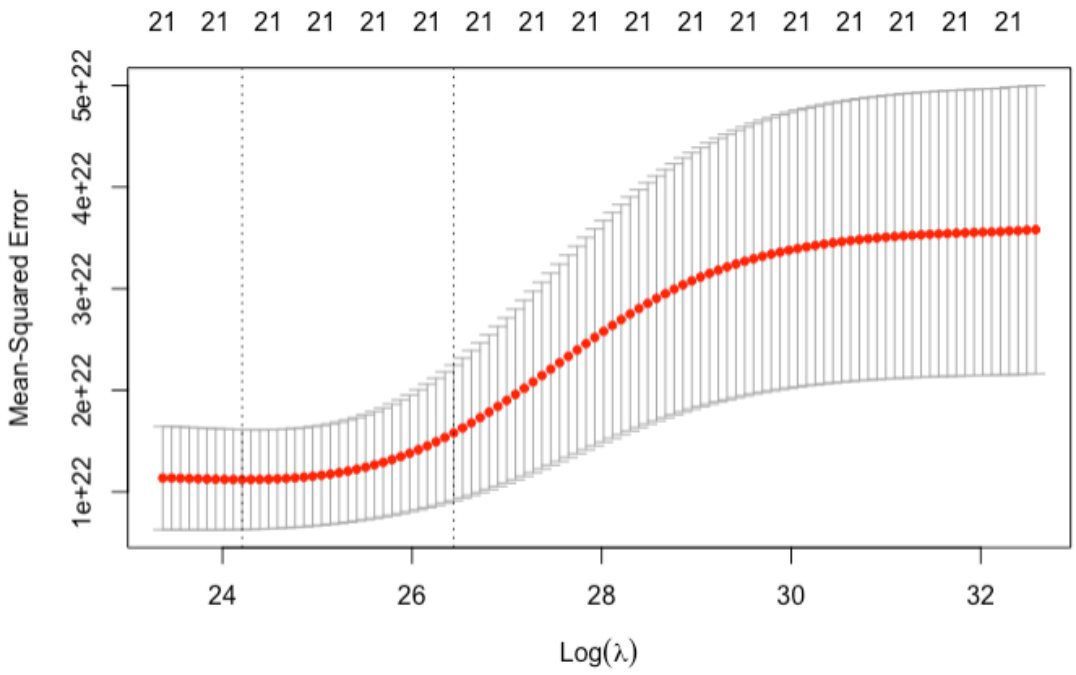

Figure 2-11. MSE vs Log(λ) on Ridge Regression

The lambda value that minimizes the test MSE turns out to be 32576574245, log(λ) = 24.21, the first dash line on the Figure 2-11. Lastly, we can get the final model produced by the optimal lambda value. Need to notice is, it will no coefficient shown for the some predictor, because Ridge regression shrinks all coefficients towards zero, but not remove the variable.

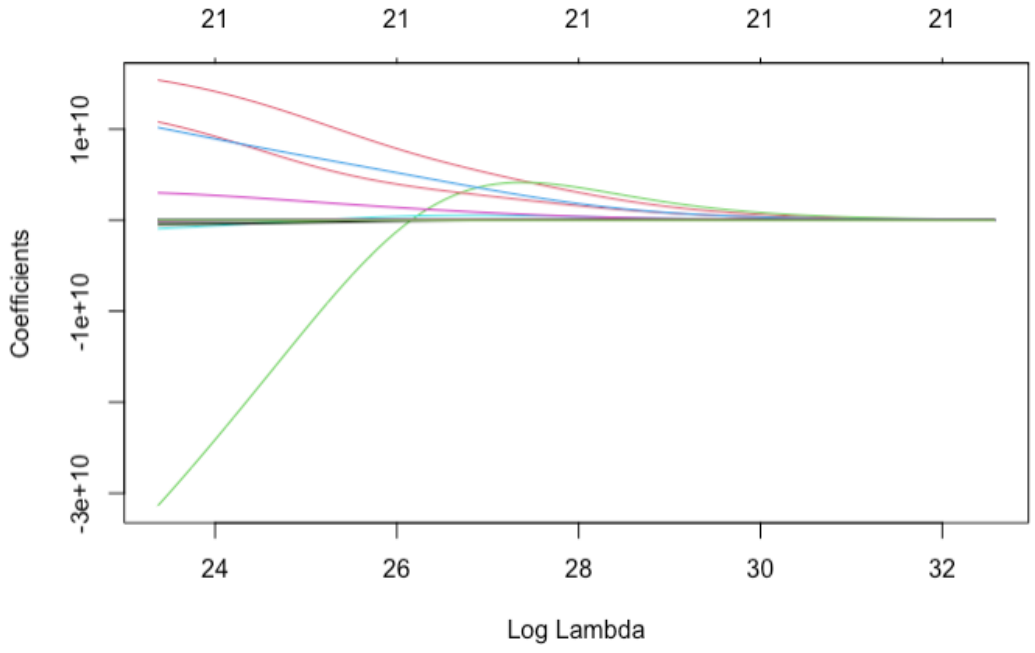

Figure 2-12. Trace plot on Ridge Regression

Lastly, we can calculate the R-squared of the model on the training data and the testing data.

Ridge	Train Data	Test Data
R-square	0.8216	-0.601

Table 2-6. Ridge Model Evaluation

10.4 Model Conclusion

	R-square
Linear Regression	0.8328
Stepwise	0.8294
LASSO Regression	0.8241
Ridge Regression	0.8216

Table 2-7. R-square for each models.

Finally, comparing the R-square of these each models, it is found that Linear Regression has the highest R-square, so in this case the Linear Regression model fits best.

11.0 Conclude with Business Implications/Recommendations

In conclusion, we can find that 'free cash flow' has the strongest positive correlation with equitys 'current market cap', followed by 'total equity', 'total compensation paid to executives', 'short and long term debt', 'ROA', 'price to sale ratio'. In other words, market supply and demand affect the current market cap of equities.

1) When there are more sufficient funds, the strength of funds to buy equities is strong, and the current market cap will rise. If the supply of funds in the equity market is tight and the demand for funds increases, the capital strength to buy equities becomes weak and the number of people selling stocks increases, then the current market cap of equities will fall.
2) Executive compensation is linked to the equity market value, which is conducive to the development of listed companies and safeguarding the interests of shareholders. When the company's performance increases, the salary of executives should naturally increase, and vice versa, the salary should decrease.
3) Equities with higher current market cap usually have higher short-term or long-term debts. If the debt costs of listed companies continue to increase, it will result in continuous reduction of profits and induce huge operating pressures for enterprises.
4) In general, those listed companies with higher ROA has better ability to earn money.
5) The price-to-sales ratio tends to decline as a company's revenue rises. This means that the investment value of the firm increases with the price-to-sales ratio; on the other hand, the higher the price-to-sales ratio, the higher the investment risk of the company, and the lower the investment value.

APPENDIX 1: Tabular Dashboard Support (Case 1)