New_final_project

Haokun Zhang, Zhang Lu, Jonathan

2023/04/21

Contents

1 Introduction:

Companies listed on U.S. stock exchanges are required by law to have their financial statements audited by an independent auditor. This requirement is designed to provide assurance to stakeholders, such as investors, creditors, and regulators, that the company's financial statements are presented fairly and accurately in accordance with accounting standards and can help to build trust and confidence in the company. An audit fee then is necessary to be paid to the independent auditor for their services in conducting the audit of the company's financial statements. The audit fee is typically paid annually and covers the cost of the auditor's time, expertise and resources required to conduct the audit as well as potential legal and reputational risks associated with the audit.

To be specific, (1)Time: Auditors spend a significant amount of time reviewing a company's financial statements, internal controls, and other relevant information. The amount of time required depends on the size and complexity of the company, as well as the scope of the audit. Market capitalization, scale of assets, scale of revenue, scale of earnings etc. could be useful indicators of complexity of companies. (2) Expertise and resource required: Auditors are highly trained professionals with specialized knowledge in accounting, auditing, and financial reporting. They use this expertise to assess the accuracy and reliability of a company's financial statements, as well as to identify potential risks and areas for improvement. Among all auditors in the world, Big Four firms (Deloitte, PwC, EY, and KPMG) are usually considered to be more expertise

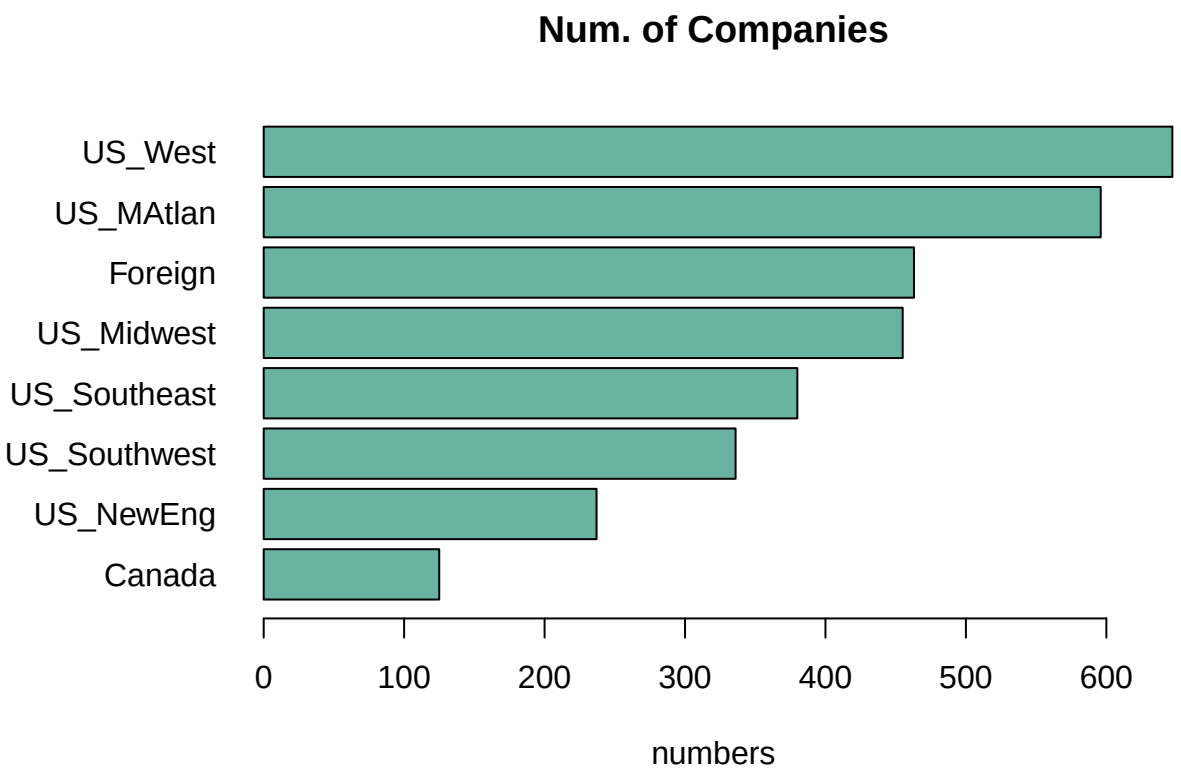
than non-Big 4 firms due to their extensive global networks, significant presence in multiple countries, heavy investment in training and development programs for their staff as well as great investment in technology and innovation. (3) Potential legal and reputational risk: If a material misstatement is later discovered in the financial statements, the auditors may be held liable for any losses suffered by investors or other stakeholders. Meanwhile, if an auditor's work is called into question or their reputation is damaged, it can be difficult for them to attract new clients or retain existing ones. An adverse opinion on internal control over financial reporting (ICFR) could be an indicator of potential legal and reputational higher risk for auditors.

In this project, we explore what are the key factors to determine audit fees because audit fees are important both for companies and auditors. For companies, predicting audit fees can help companies compare the cost of their audit with other companies in the same industry to assess their competitive position and identify opportunities for cost savings. For auditors, this research question could help them have a better audit fees negotiation process. A good reference of audit fees could on one hand help them ensure that their fees are reasonable and proportionate to the work performed, which can help to maintain the integrity of the audit profession, on the other hand, help auditors to assess their competitive position in the market.

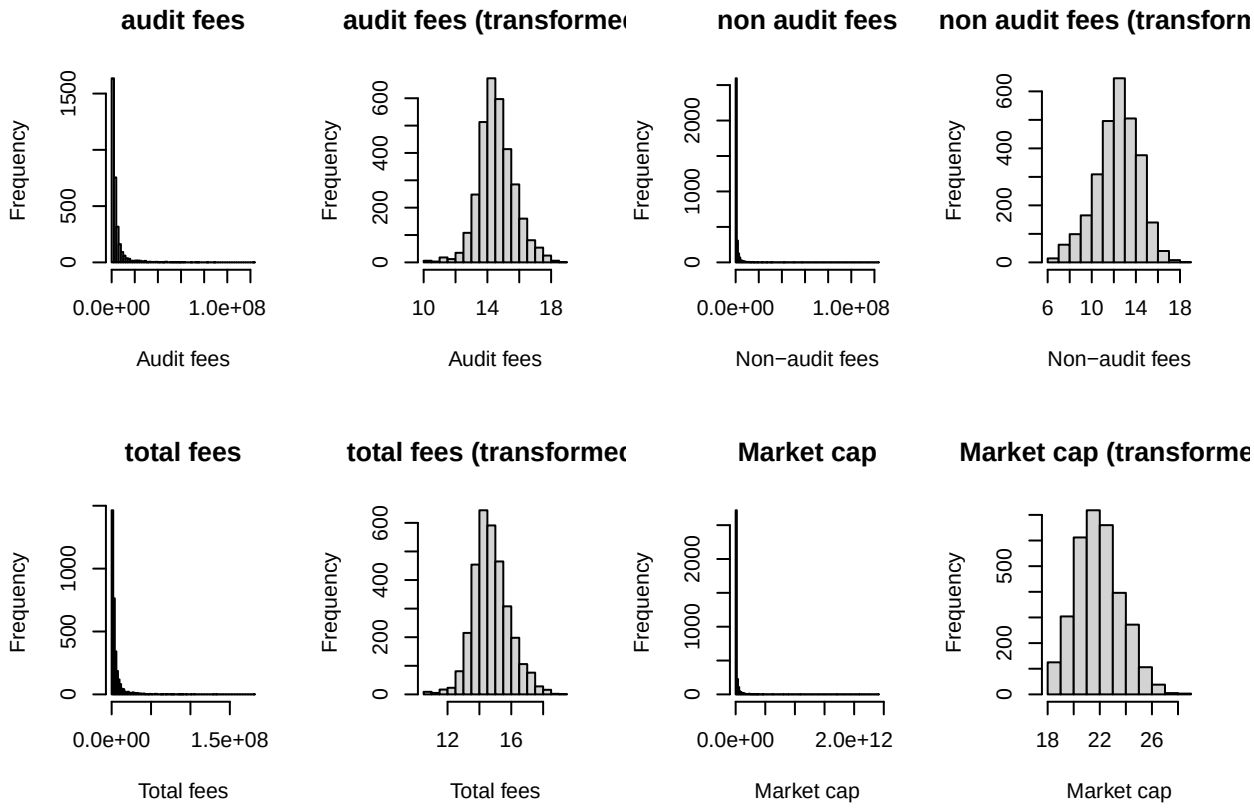
```
## [1] 0.9536894
```

2 Data visualization(EDA)

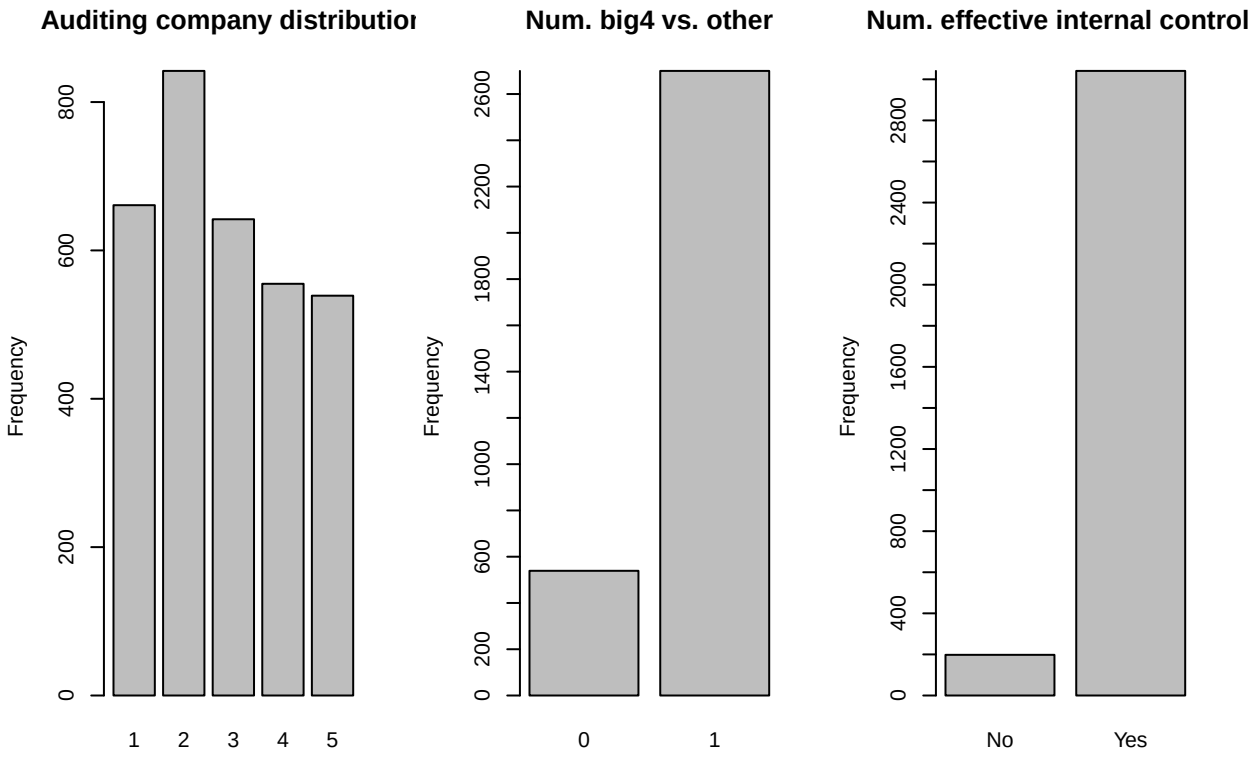
Plot 1: plot the number distribution of companies in different regions



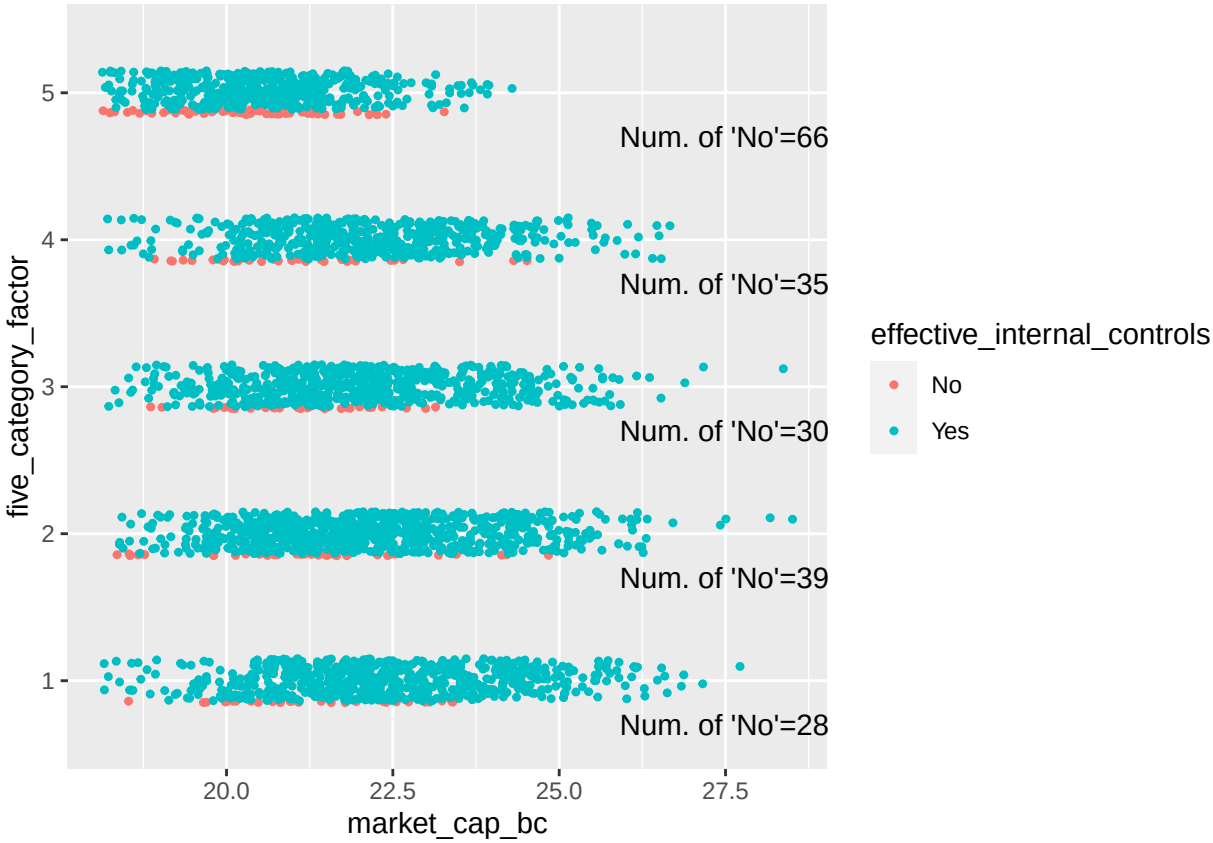
Plot 2: Use eight plots to display the effect of transformation on fee related variables



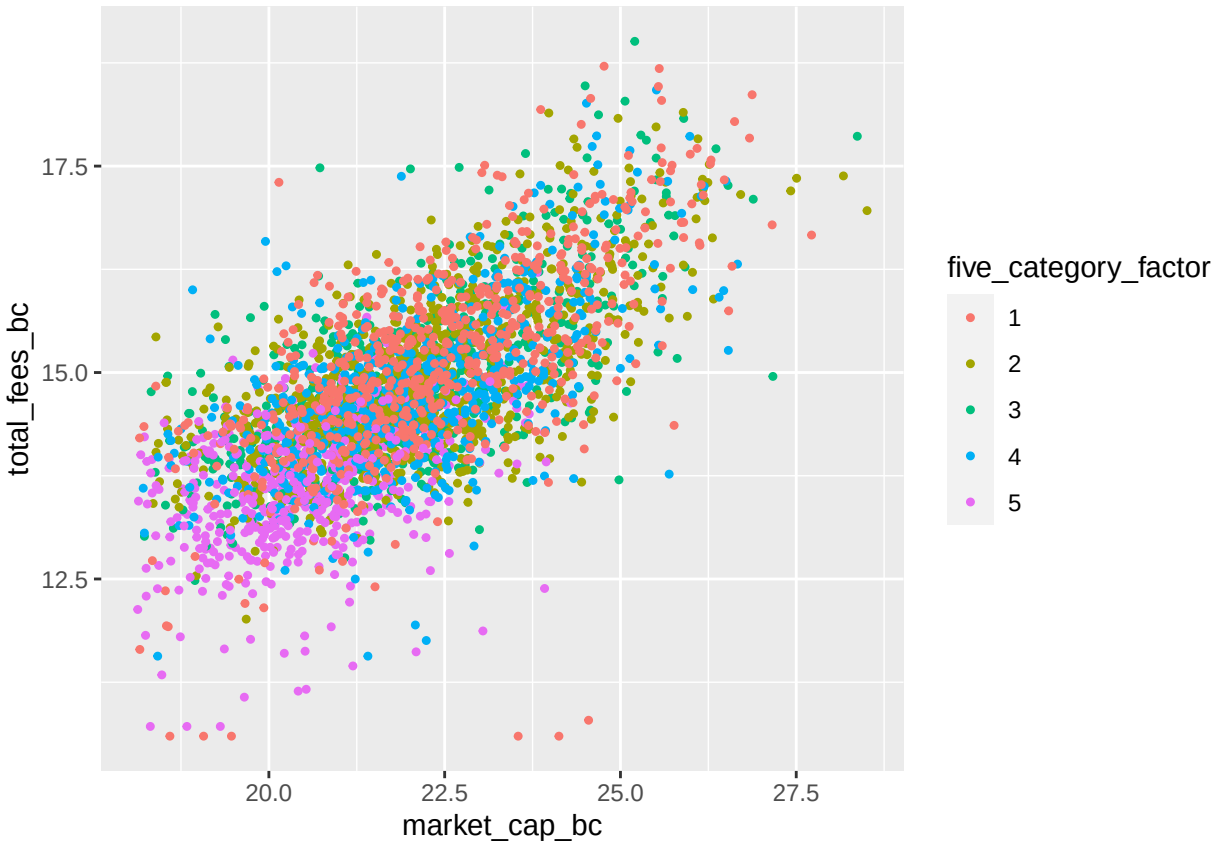
Plot 3: Use three plots to display the categorical data



Plot 4. Plot the transformed company market cap, total auditing fees, and effective internal control

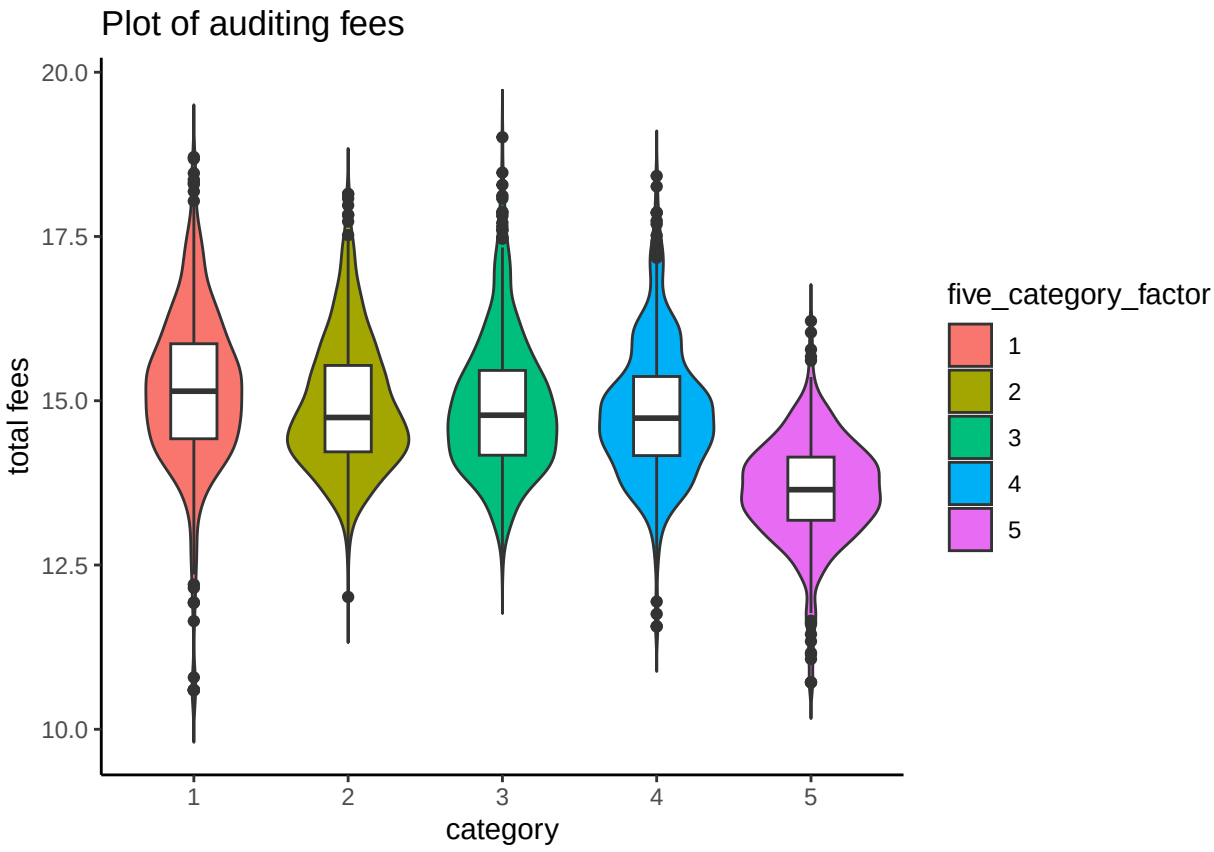


Plot 5: Plot the transformed company market cap vs. total auditing fees




```
## [1] 0.6904368
```

Plot 6: Plot the auditing fees

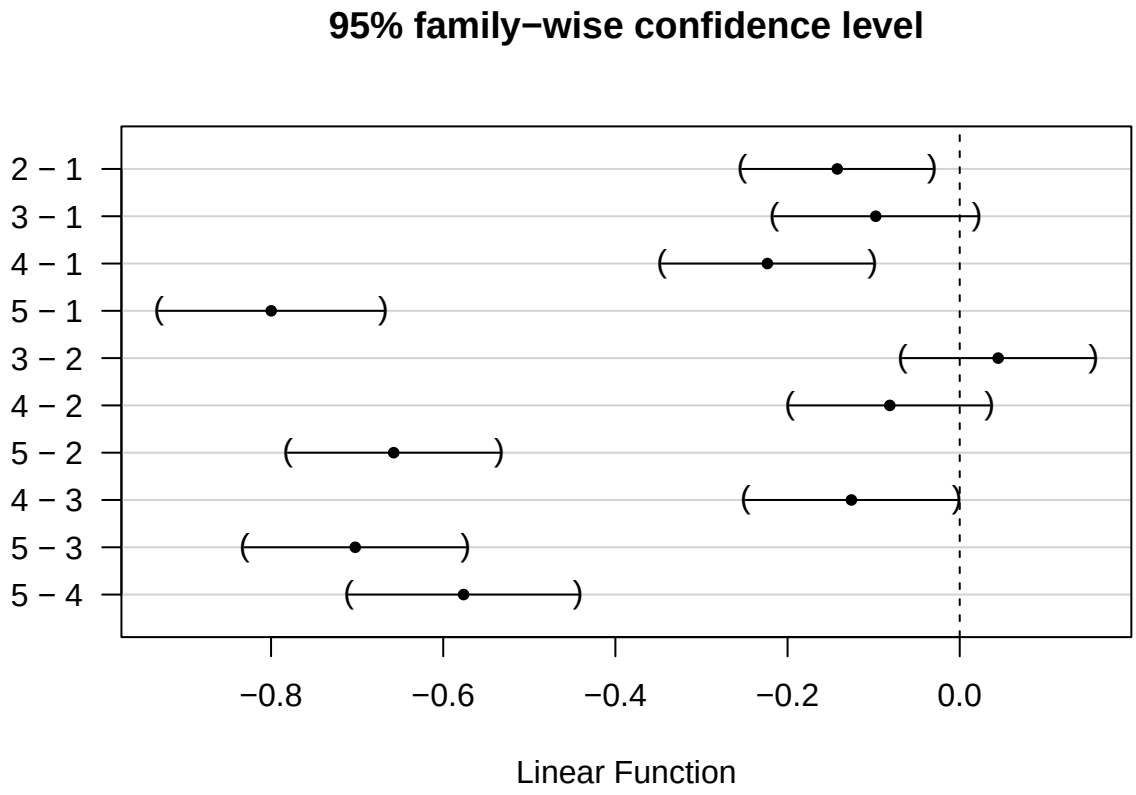


3 Perform a statistical test here to compare big4 vs non big4 when considering

```
## Anova Table (Type III tests)
##
## Response: total_fees_bc
##                                   Sum Sq   Df  F value    Pr(>F)
## (Intercept)                       101.36    1 168.8601 < 2.2e-16 ***
## five_category_factor               12.93    4   5.3853 0.0002547 ***
## market_cap_bc                     406.51    1 677.1899 < 2.2e-16 ***
## five_category_factor:market_cap_bc  20.26    4   8.4396 9.028e-07 ***
## Residuals                         1938.31 3229
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: lm(formula = total_fees_bc ~ five_category_factor + market_cap_bc,
##     data = df3)
##
```

```
## Linear Hypotheses:
##            Estimate Std. Error t value Pr(>|t|)
## 2 - 1 == 0 -0.14223    0.04051  -3.511  0.00404 **
## 3 - 1 == 0 -0.09759    0.04327  -2.255  0.15906
## 4 - 1 == 0 -0.22345    0.04490  -4.977  < 1e-04 ***
## 5 - 1 == 0 -0.79979    0.04787 -16.707  < 1e-04 ***
## 3 - 2 == 0  0.04465    0.04080   1.094  0.80866
## 4 - 2 == 0 -0.08122    0.04256  -1.908  0.31153
## 5 - 2 == 0 -0.65756    0.04504 -14.599  < 1e-04 ***
## 4 - 3 == 0 -0.12586    0.04512  -2.790  0.04190 *
## 5 - 3 == 0 -0.70220    0.04712 -14.904  < 1e-04 ***
## 5 - 4 == 0 -0.57634    0.04885 -11.798  < 1e-04 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```



4 Unsupervised Learning: Clustering

After exploring the data set, we continued to cluster with the companies based on the following values before and after previous transformation: audit fees($), total fees($), Market capitalization($), market-fee ratio, asset value($), revenue($) and earnings($). We used the Gap Statistic and K-means methods to determine the number of clusters.

4.1 Gap Statistic for Estimating the Number of Clusters

This method estimates a goodness of clustering measure, the "gap" statistic with a given range of number of clusters K. For each K, it compares the log value of dispersion of observations within a cluster to the estimated log value of dispersion. In this report, the maximum number of clusters to consider is 10.

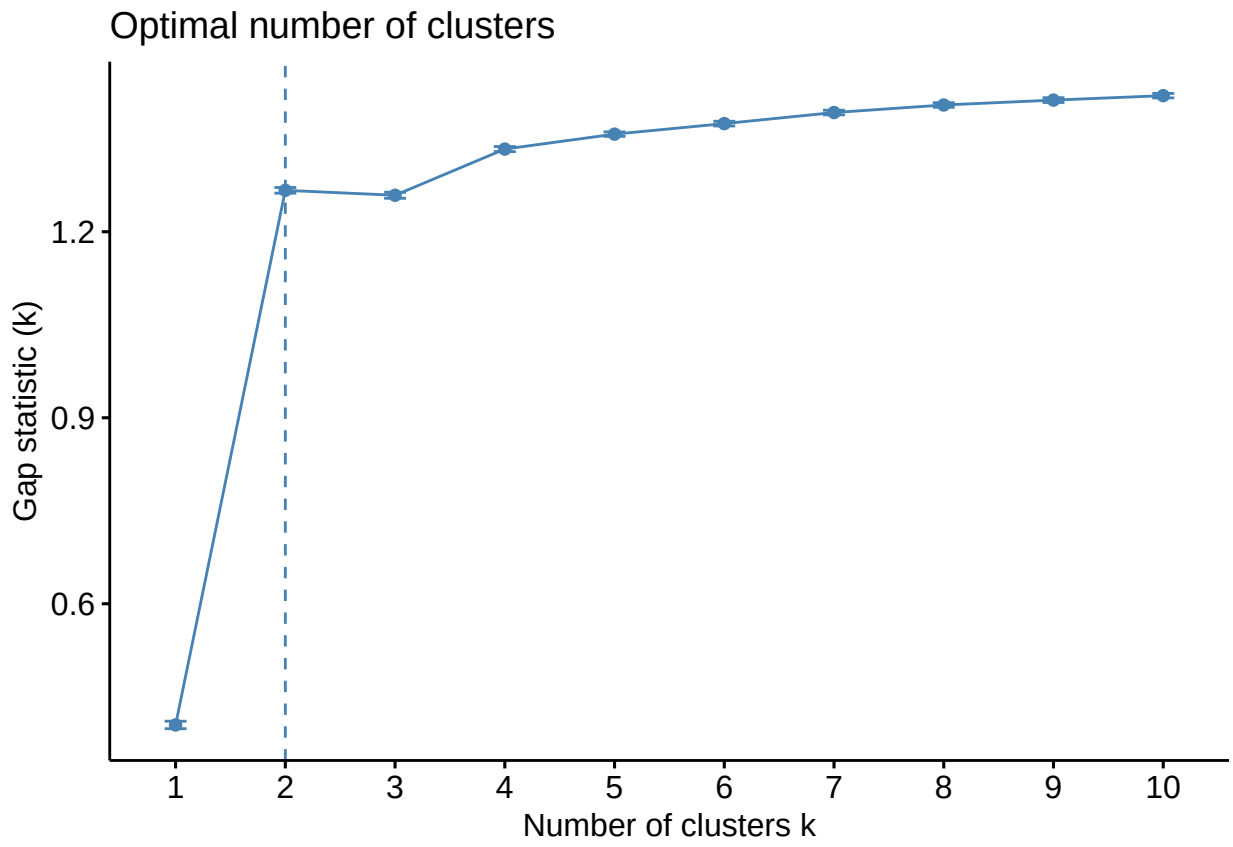


Figure 1: Gap Statistic Clustering Result

As The Figure **??**(fig: fig8) suggests, the Gap statistic estimates that the optimal number of clusters is 2.

4.2 K-means Clustering

The K-means clustering uses a principal component analysis to create clusters, and classify and partition objects into multiple groups. The objects are as similar as possible within, and as dissimilar to the objects in other groups as possible. K represents the number of groups. From the previous result, we select 2 sets of groups, with the same variables as Gap statistic used.

Table 1: Features of K-means Clustering Groups

cluster	audit_fees_bc	total_fees_bc	market_cap_bc	market_fee_ratio	assets_log	revenue_trans	earnings_trans
1	14.345	14.479	21.176	-6.697	21.047	17.665	-18.305
2	14.675	14.838	22.186	-7.348	22.386	21.174	19.244

The K-means clustering aggregated 2 clusters that have distinct features. The Figure **??**(fig:fig9) shows that these two groups overlap each other in large areas, which is expected because the box-cox and natural

Table 2: Untransformed Features of K-means Clustering Groups

cluster	audit_fees	total_fees	market_cap	market_fee_ratio	assets	revenue	earnings
1	3948548	4824005	17136157968	-7.174	2.111079e+10	7584571466	816700052
2	58787475	71092557	122335978468	-6.804	2.251131e+12	52604000000	15833555556

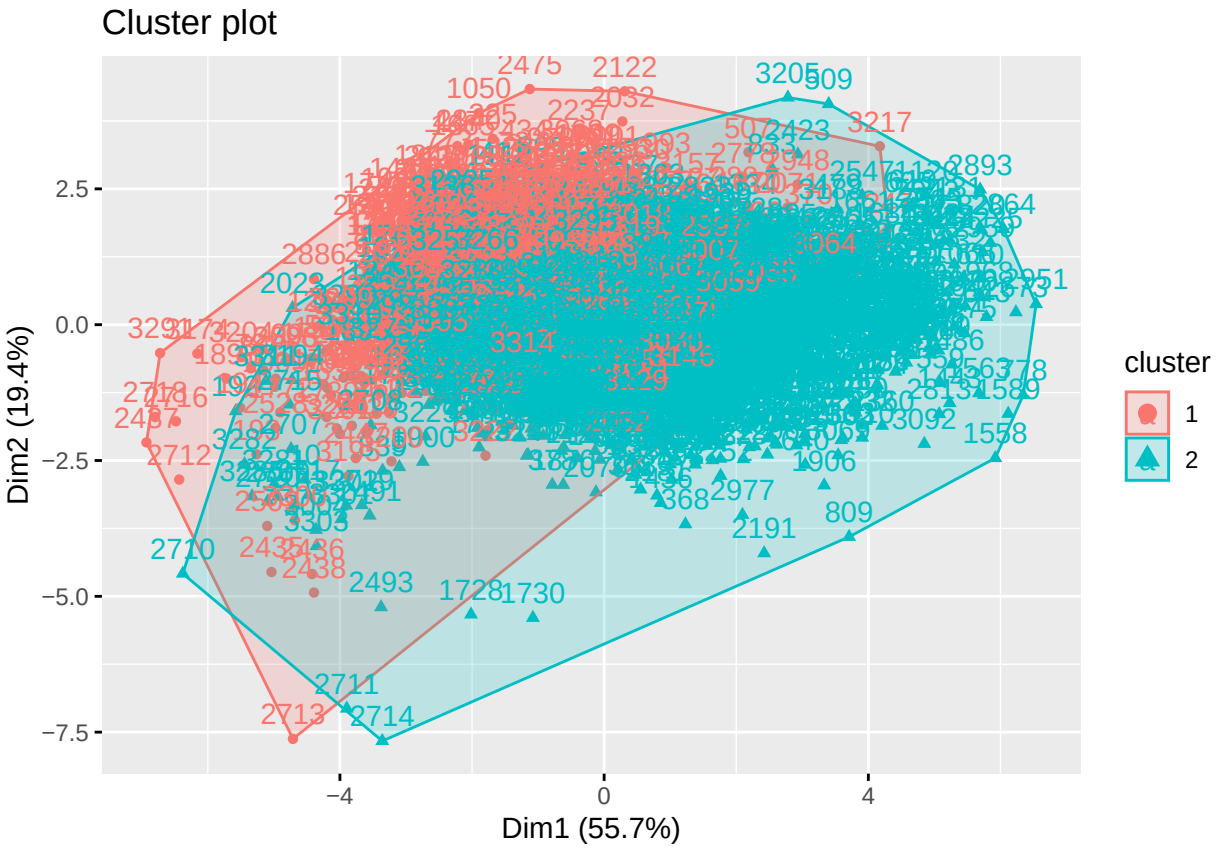


Figure 2: K-means Clustering Result of Untransformed Data

log transformation transformed value of the data to be closer to each other. Unsatisfied with the result, we conducted the same clustering with original data.

As shown in Table **??**(table:table2), companies in cluster 2 have smaller market caps and assets than companies in cluster 1. In cluster 1, the mean asset value is $2.2 trillion and mean market capitalization is 1.2 trillion. Cluster 2 has mean value of asset worth $0.2 trillion and 0.17 trillion market capitalization. As a result of less market capitalization, companies in cluster 2 have a mean market-fee ratio of 0.12% while companies in the second cluster has 0.06%. The cluster 2 also has less revenue and earnings than cluster 1. For instance, mean revenue of cluster 2 is approximately $0.78 trillion while the mean revenue of cluster 2 is $5.2 trillion.

In the cluster plot, The component of the first principal component accounts for 55.8% of the total variation, and the second principal component accounts for 19.4%. Together they explain 75.2% variation in the data set. The limited number of variables selected when clustering can explain this issue. Along with variables that have similar values to each other after transformation, the same reason could explain the large area that clusters overlap each other in Figure **??**(fig:fig9). The overlapped area is smaller in Fig **??**(fig:fig10) and the clusters are more distinctive.

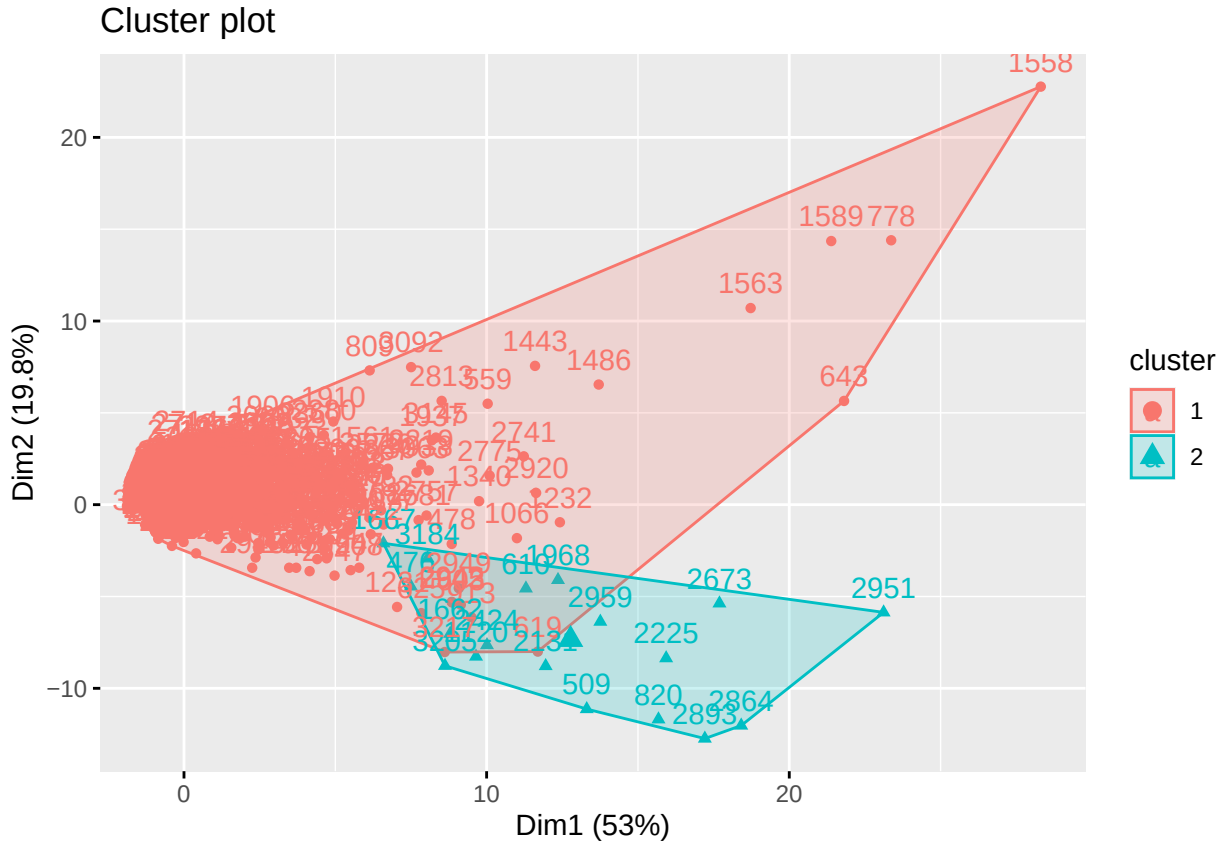


Figure 3: K-means Clustering Result of Untransformed Data

5 Supervised Learning

In the supervised learning part, we want to build several models and choose the best performance model to do the prediction of total audit fees. The workflow is shown below: Firstly, We want to try various models. We chose 5 models here: Linear model, KNN model,Random forest model, CART model and

Gradient-Boosting model. Secondly, we used K-fold cross validation to assess the model. Lastly, we would consider the MAE, RMSE and R-square to decide which model has the best performance. In the first step, we did the preliminary feature engineering, we abandon some features.

```
##
## Call:
## summary.resamples(object = res_total)
##
## Models: lm, knn, rf, cart, gbm
## Number of resamples: 5
##
## MAE
##           Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## lm   0.1247656 0.1248237 0.1275063 0.1307699 0.1366864 0.1400674    0
## knn  0.4545652 0.4557931 0.4575745 0.4602870 0.4583265 0.4751756    0
## rf   0.1248219 0.1307786 0.1325027 0.1334408 0.1386855 0.1404152    0
## cart 0.1939173 0.1974520 0.2020197 0.2021714 0.2075451 0.2099231    0
## gbm  0.1215159 0.1273792 0.1286229 0.1305952 0.1315960 0.1438621    0
##
## RMSE
##           Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## lm   0.1815099 0.1863813 0.1868630 0.2489354 0.2065249 0.4833980    0
## knn  0.5873146 0.6013985 0.6057226 0.6145919 0.6067716 0.6717523    0
## rf   0.1739837 0.1756560 0.1854713 0.1867223 0.1985352 0.1999654    0
## cart 0.2604609 0.2679060 0.2680983 0.2679842 0.2715478 0.2719078    0
## gbm  0.1688052 0.1771494 0.1792819 0.1885276 0.1824770 0.2349245    0
##
## Rsquared
##           Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## lm   0.8373952 0.9656563 0.9712942 0.9444037 0.9729593 0.9747132    0
## knn  0.6811461 0.7015815 0.7039611 0.7085261 0.7103870 0.7455549    0
## rf   0.9709799 0.9712032 0.9729899 0.9728555 0.9741428 0.9749617    0
## cart 0.9408363 0.9427923 0.9430377 0.9436327 0.9447049 0.9467924    0
## gbm  0.9564880 0.9743758 0.9747103 0.9718303 0.9752414 0.9783360    0
```

From the above figures, we. can see that linear model has the lowest MAE, Random Forest has the second lowest MAE. Gradient-Boosting model has the lowest RMSWE, and Random Forest also has the. second lowest RMSE.As a result, we decided to choose these 3 models initially and do the advanced feature engineering.

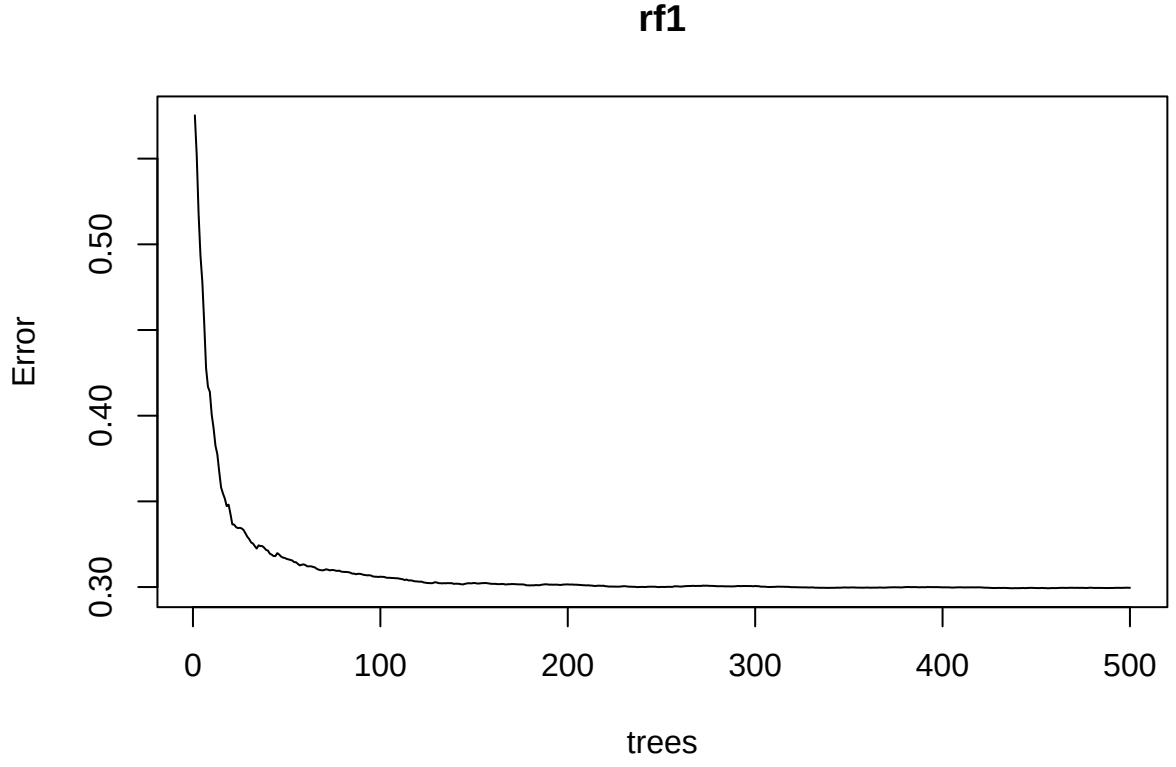
```
##
## Call:
## summary.resamples(object = res)
##
## Models: lm, rf, gbm
## Number of resamples: 5
##
## MAE
##          Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## lm  0.5018642 0.5072226 0.5117290 0.5136686 0.5196348 0.5278923    0
## rf  0.4221462 0.4297960 0.4309628 0.4328240 0.4352354 0.4459798    0
## gbm 0.4243101 0.4307670 0.4328323 0.4348067 0.4360918 0.4500325    0
##
```

```
## RMSE
##          Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## lm  0.6249813 0.6336866 0.6387257 0.6451767 0.6641223 0.6643678    0
## rf  0.5455421 0.5459103 0.5461181 0.5515208 0.5598408 0.5601929    0
## gbm 0.5502242 0.5527765 0.5538596 0.5580215 0.5595872 0.5736603    0
##
## Rsquared
##          Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## lm  0.6589044 0.6626211 0.6707226 0.6730335 0.6749222 0.6979971    0
## rf  0.7396508 0.7518743 0.7601270 0.7612184 0.7759304 0.7785097    0
## gbm 0.7411649 0.7526429 0.7529908 0.7553377 0.7636978 0.7661921    0
```

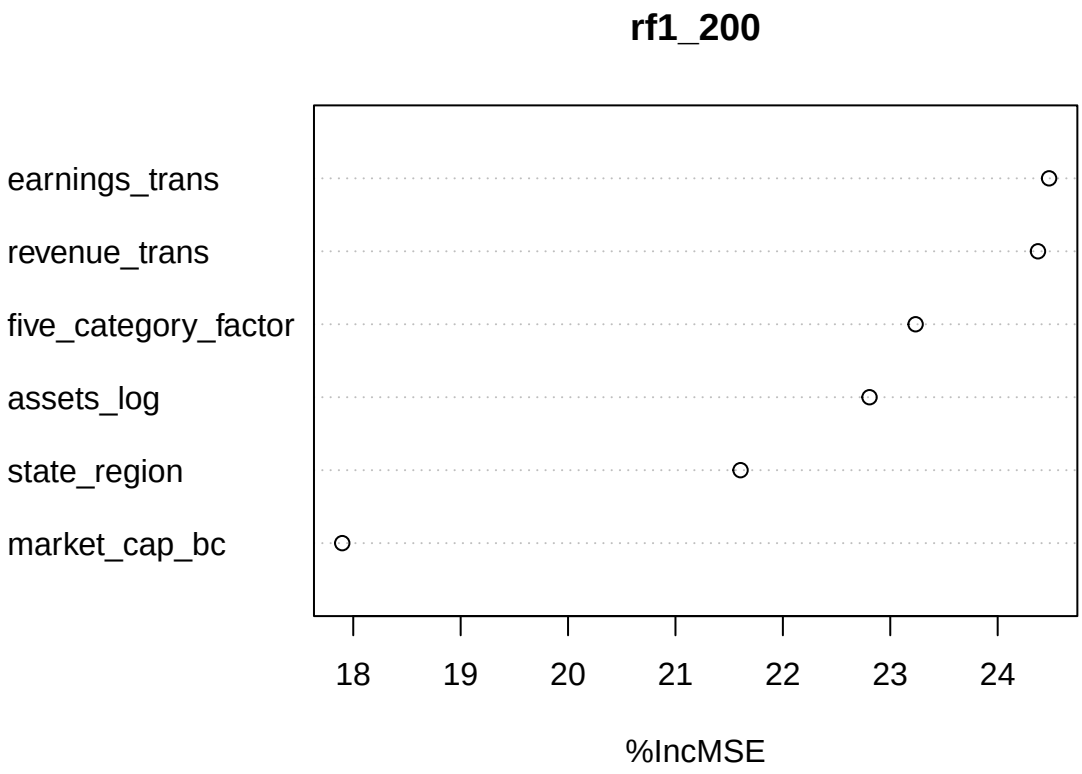
Depend on the above figures, after the advanced feature engineering, we saw that random forest model has the best performance, which has the lowest MAE, lowest RMSE and highest R-square.

So we chose Random Forest model to do the prediction of total auditing fee based on other variables.

```
##
## Call:
##  randomForest(formula = total_fees_bc ~ five_category_factor +      state_region + market_cap_bc + as
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 2
##
##           Mean of squared residuals: 0.2995484
##                     % Var explained: 76.11
```

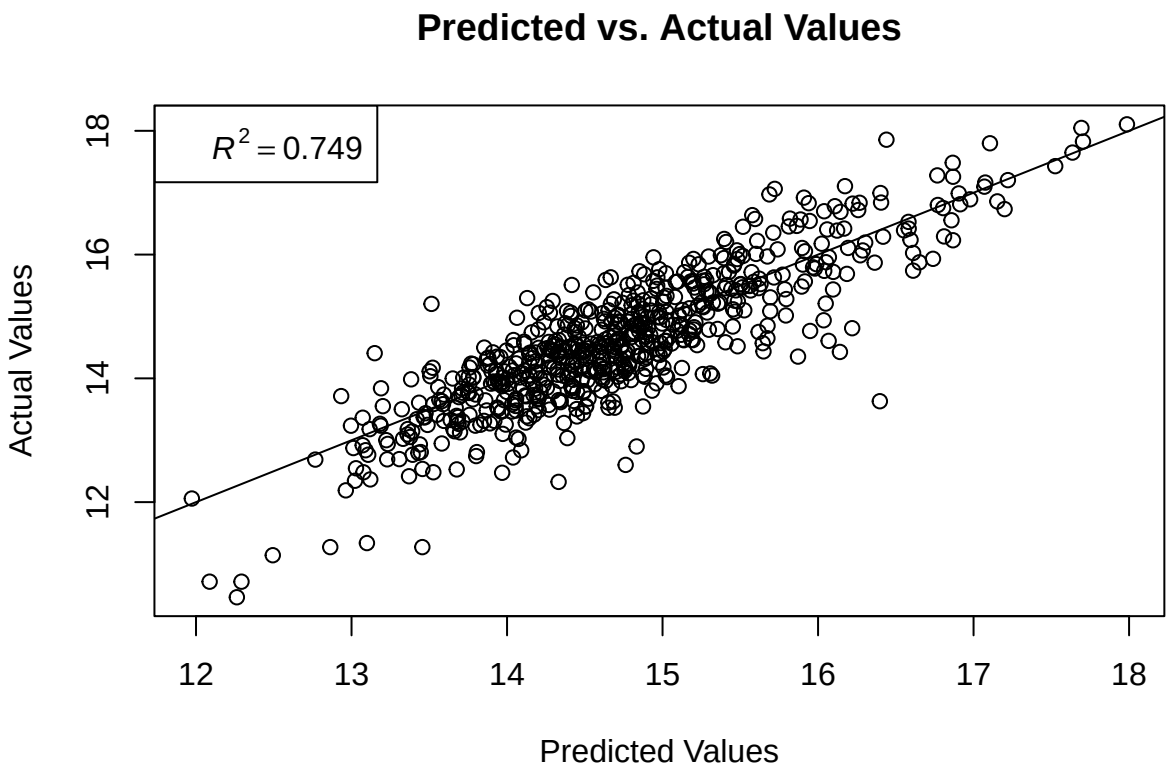


```
## [1] 0.5874249
```



The plot of rf1 shows out-of-bag MSE as a function of the number of trees used, We saw that when the number of tree is equal to about 200, the MSE is slowing down. So we choose number of tree = 200.

The explanation of RMSE: Then we saw the RMSE of random forest model is 0.5354485, which is lower than other models. The explanation of varImpPlot(rf1, type=1):These are the most influential variables of total audit fees. Revenues has the most influence on the total audit fees as we expected, if we delete the revenue variable, the accuracy of model will decrease about 45%. The other 5 variables: the region of the company, the market cap of the company, the asset of the company, the earnings of the company and whether the company is audited by the big 4, also have major influence on the model.



From the above figure of predict value v.s. actual value, we saw that our model has good performance. The R-square is 0.748, which means that about 75% of the variability observed in the total audit fee can be explained by the random forest model. However, we know that we can't just use R-square to assess the model performance. We also consider MAE and RMSE. Compare to other models, random forest model still has lowest MAE(the average absolute magnitude between the actual values and the predicted values) and RMSE(the average difference between values predicted by a model and the actual values). As a result, we can make sure that random forest model is the best performance model.

6 Conclusion

Companies listed on U.S. stock exchanges hire independent auditors to prepare their financial statements and present them annually. Usually, an audit fee is paid to the auditor and covers the cost of their services. In this project, we explored and determined key factors that have critical impact on audit fee decision made by auditors, as well as predicting audit fee with several parameters. We explored the internal control data of numerous companies with market capitalization above 75 million in 2021, used unsupervised learning method and found that these companies can be clustered to two groups with distinctive features. Supervised learning method allowed us to select Random Forest model that fits the best and predicts audit fee the most accurately.

Determining accurate audit fee is arguably one of the most important components in auditing process. This project contributes to predicting accurate audit fee, which gives companies a good reasonable reference during the negotiation with independent auditors, it also helps auditors to assess their competitiveness in the market.

Any predictive model is built based on the data given to learn, so is ours. Our result suggests that add more variables that could affect decision of audit fee, such as restatement history, previous auditing history

and companies' audit fees across a consecutive time frame will build more accurate and interesting predictive models. The data we used to train our models is limited, including companies with market capitalization below 75 million will potentially increase the model's predictive accuracy.

7 Appendix:

The auditor's report on internal controls over financial reporting (ICFR) is one of the few publicly observable ways for auditors to disclose unfavorable audit findings. Companies listed on U.S. stock exchanges with market capitalizations greater than $75 million must secure an independent audit opinion on ICFR under the Section 404(b) of the Sarbanes-Oxley Act of 2002, as amended by the Dodd-Frank Act of 2010. ICFR is the processes that ensure reliable financial reporting. Thus, while the auditor's opinion on the financial statements covers the product of financial reporting, the ICFR audit covers the processes that generate financial reports. The distinction is important because nearly every U.S. public company receives an unqualified (i.e., clean) audit opinion on its financial statements as the U.S. Securities and Exchange Commission (SEC) will not allow companies to trade on a stock exchange unless they correct any material errors the auditor detects in the financial statements. However, even if the audit opinion on the financial statements is clean, companies can still get an adverse audit opinion on ICFR if the auditor concludes that the company's internal controls over financial reporting are subject to one or more material weaknesses. An adverse opinion on ICFR indicates that the auditor believes that there are significant deficiencies in the company's internal controls that could potentially lead to material misstatements in the financial statements, which means that the auditor does not have confidence in the reliability of the company's financial reporting processes. Therefore, when auditors issue an adverse opinion on ICFR, auditors need to spend additional time, effort, and expertise required to mitigate the risk which could lead to material financial misstatements.