Financial Literacy

Self-Evaluation and Reality

Yangsijia Wang
Dr. Kristina Sendova
August 15, 2022

Contents

1.Abstract

This study is on the topic of financial literacy, with the data source containing information on clients' demographic information and self-evaluation, change in account value, and trade record, three major problems were investigated: first, whether a client's demographic traits are related to his/her self-evaluation of financial knowledge level; second, does the trading behaviour differ for clients who self-identified as in different financial knowledge groups; and third, do people who self-identified as financially knowledgeable have better investment result. Data manipulation was done using SQL and R. Exploratory analysis including multiple types of plots and proportion tables was used to derive the hypothesis of the potential relationship between variables, and hypothesis tests, contingency tables were used to prove the significance of the relationship. It is observed that demographic traits including marital status, gender, maximum age attained, the residence of living, income and net worth are related to a client's evaluation of his/her financial knowledge level, but the relationship was not strong enough to make predictions based on available information. Investors who think of themselves as more financially knowledgeable tend to choose a more risky portfolio, and they trade more frequently with a clear preference for agent operations – this type of trade is also commonly used for the group of investors who believe that they lack enough knowledge in this field; while investors in between prefer principal operations. For the actual investment performance, investors who self-identified as equipped with more financial knowledge do have better performance on average, with the internal rate of return used as the standard of evaluation, it is also observed that the investors who see themselves as having an excellent amount of financial knowledge also have the largest diversity in terms of their performance.

2.Introduction

What is financial literacy? It is a term used to describe an investor's financial knowledge base as well as his/her ability to make use of this knowledge to make wise investment decisions. Financial literacy is a topic that is drawing increasing attention from society, both the Canadian government and other large financial institutions have programs and resources helping people increase their financial knowledge, and the government also has a national financial literacy strategy from 2021-2026 (not sure whether to site). As part of the financial wellness lab research, this project is centred on investors' self-evaluation of their financial literacy level. Several questions were raised surrounding this topic, including the demographic features that potentially lead to one's evaluation, the risk preference and trading behaviour of the investors in different financial literacy groups, and the accuracy of an investor's self-evaluation.

The anonymous data used for the project was provided by the cooperating industry partner, including the Know Your Client (KYC) information, trade record and position of accounts, reference of risk rating for different securities (ref_risk_rating), recorded in table format. The information recorded in the KYC table includes the clients' demographic traits and risk preferences, with each row corresponding to a record of an investor at a certain age. Apart from age, net worth, annual income, marital status, gender, retirement condition and residence, which are demographic predictors used in this project, the most important classification is the client's investment knowledge level, which is evaluated by the clients themselves and is used as the basis for group division in this project. This classification has four levels: excellent, good, fair, and nil (negligible), most of the analyses in the project were investigations of the potential difference in terms of behaviours

and results among these four groups. Apart from the fields above, the risk tolerance (voluntarily reported) of different accounts was recorded using a five-level factor, ranging from low to high. The KYC table is interconnected with the trade table and position table by client id and/or account id, this connection ensured the de-identification and data anonymization, while enabling the connection between the traits of each account and their respective investment performance. The information used in the position table was the market value of the securities purchased by each account at the start and end date of record, and the data used in the trade table was the transaction date and amount each time.

It was discovered that the demographic information of an investor is not closely related to one's investment knowledge level, according to the investor's self-identified knowledge level group, and one's self-determined risk tolerance is positively aligned with his/her financial knowledge level base; besides, people who self-identified as more financially knowledgeable tend to trade more frequently and have a higher return.

3.Methodology and Results

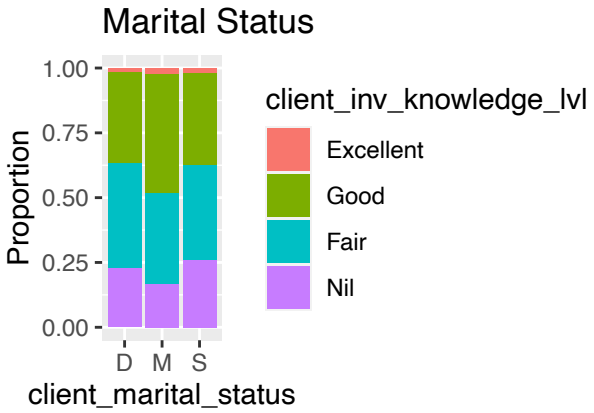
3.1 Demographic Information and Self-evaluation

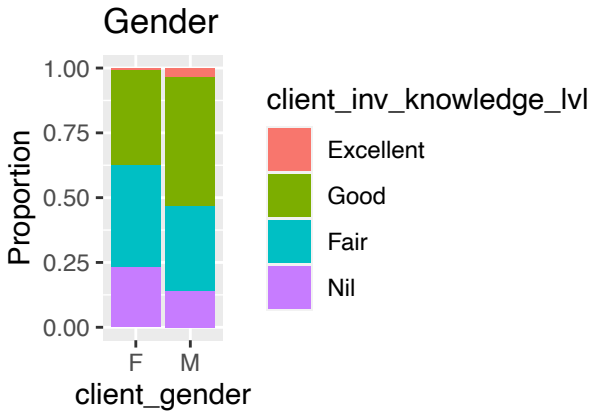
For the exploration of the potential relationship between the demographic traits and an investor's self-evaluation of financial knowledge level, the KYC table is used, which includes 29128950 rows of observations. The observations include duplicate records for different accounts of the same client, as well as the same demographic information of the same client at different ages. For our target of study, one client was limited to one account record at the maximum age recorded due to the repeating nature of other information. After the first-step process in MySQL and the elimination of the rows including incomplete information, we are left with 25,836 clients for this section of the study. Among all the clients, 43% identify their investment knowledge as "good", 36% as "fair", 19% as "nil" (negligible), and 2% of clients self-identified their knowledge level as "excellent", this result is derived using a proportion table (appendix 7.3.1).

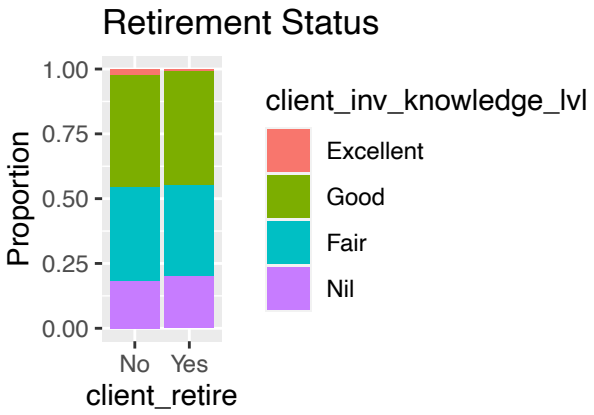
The relationship between the self-identified investment knowledge groups and different demographic variables was investigated using proportion tables as well as proportional stacked bar graphs. The first-step hypothesis of potential linkage is derived after observing the graphs, then proved using contingency tables and hypothesis tests. Specifically, for categorical variables (marital status, gender, residence, retired or not), chi-squared tests are used to examine the significance of the relationship between the demographic trait and the financial knowledge group by seeing the small p-value as the signal of connection; while for continuous variables (maximum age attained, net worth, annual income), nested multinomial models are built and compared using the deviance test, and the large p-value suggests insignificant linear relationship.

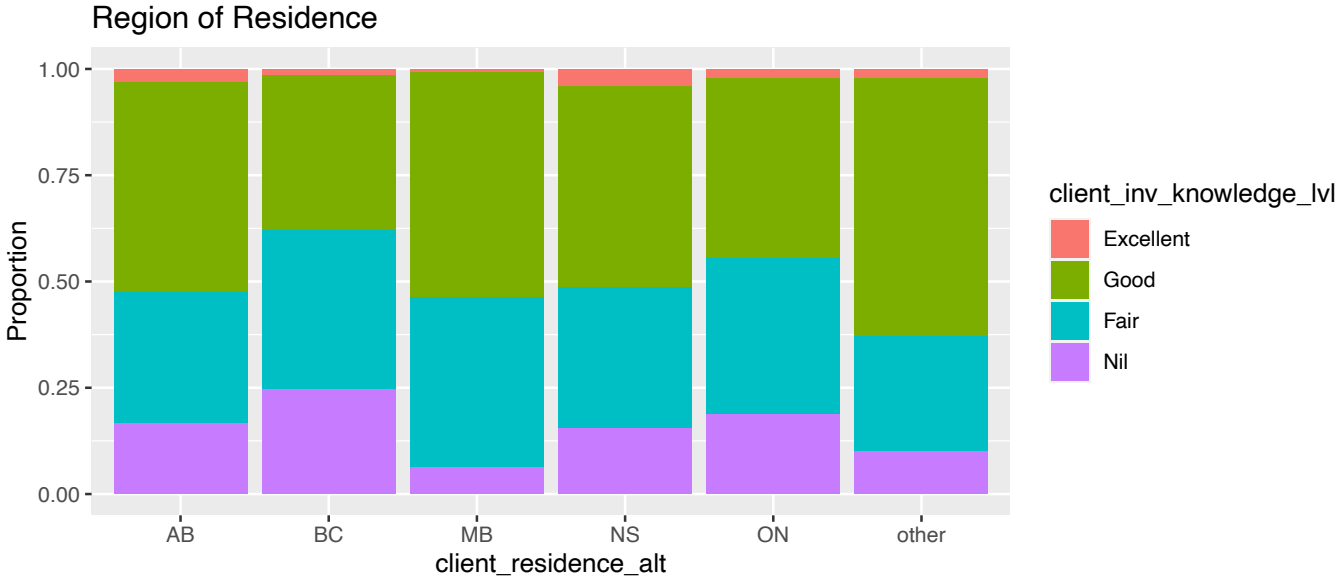
The data used for the evaluation of relationship was modified for three variables: for the categorical variable of marital status, one of the statuses was labelled as "*", with no further definition in the data dictionary for this dataset, therefore all records with this marital status were excluded from the analysis of this section (2,168 in total). For the continuous variables, it was observed in the boxplot of income/net worth distribution for different self-identified financial knowledge groups with all observations, that there is a point of extreme large value for each case. These two points were identified as abnormal and thus excluded from the rest of the analysis. After the exclusion of

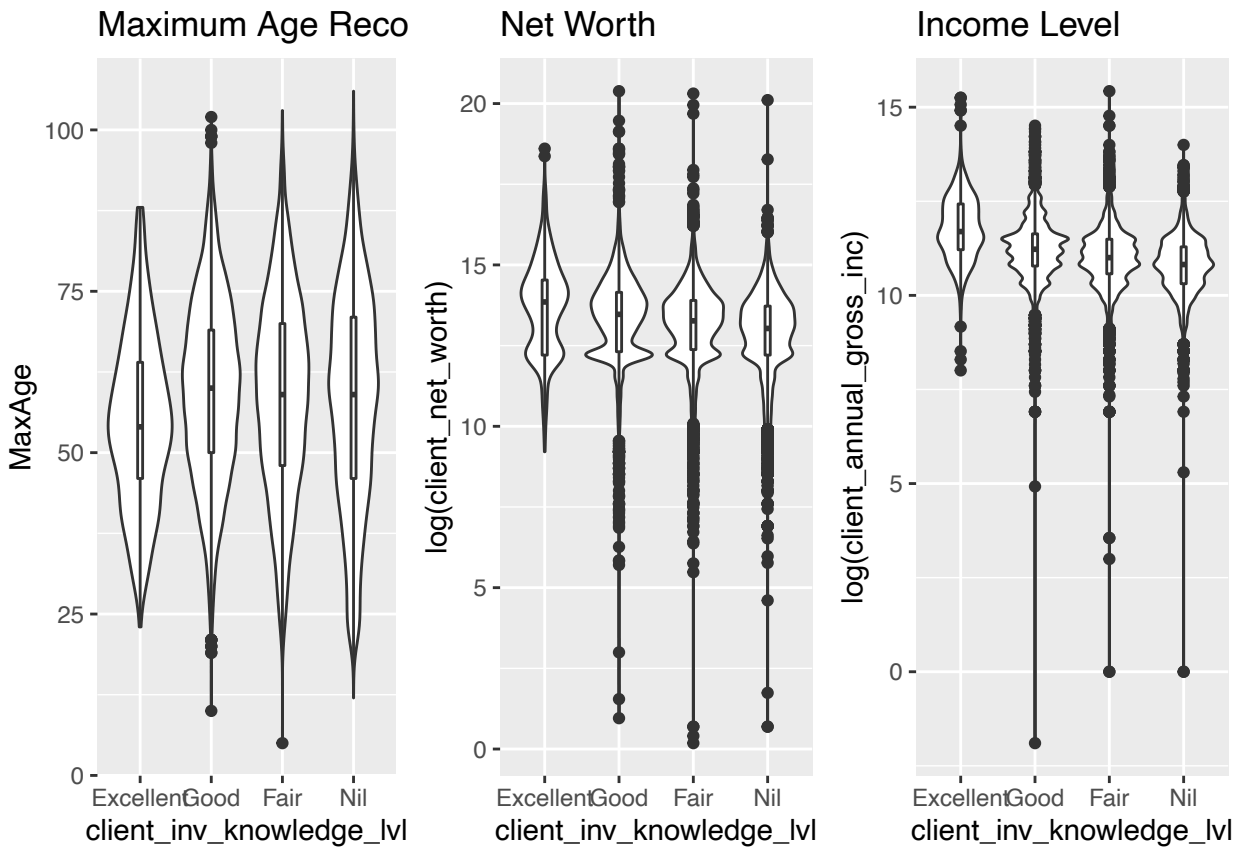
outliers, the trend was still unclear for these two variables due to the large difference in the wealth of the investors, therefore, it was the log of net worth & annual income that was finally included in the plots for the relationship study (see appendix for the details of plots).











From the plots, it was argued that all demographic traits are related to people's self-evaluation of financial knowledge, so hypothesis tests were performed to prove this argument. The dataset used for the hypothesis tests was the data excluding all outliers and unidentifiable records stated above. First, a full model is fit using all seven demographic predictors, and AIC selection was done to see whether there are predictors that can be excluded. The resulting model did not exclude any of the predictors. Then, the significance of individual predictors were tested: for the four categorical variables, contingency tables were created, and chi-squared tests were performed on the null hypothesis of the individuality of predictors (the client's self-classified investment knowledge level is independent of his/her marital status/gender/region of residence/retirement status). The p-values for the chi-squared tests were all small in the four cases, therefore the null hypothesis was rejected, and all categorical predictors were significant. Nested models were compared using the deviance test for the three continuous variables (maximum age recorded, net worth, annual income): a smaller model that only exclude the variable that we want to test its significance was built, and a small value of the p-value from the deviance test between the full model and the smaller model suggests the significance of the variable. A large p-value for the deviance test of the variable net worth suggested the lack of enough connection between it and financial knowledge level, this may be due to the correlation between annual income and net worth, so the inclusion of one of them is enough. The results can be concluded, that all demographic traits included in the available dataset except for net worth are significant predictors, they are all related to an investor's classification of his/her financial knowledge. However, this connection was not strong enough for the prediction of one's financial knowledge group based on the available demographic traits. If the dataset is randomly divided into a train and test set, then build a multinomial model using all seven demographic predictors based on the training set, the fit of this model on the test set is 45.8%. The balanced accuracy of 53.5% was also derived, which was not a large improvement. As

an alternative method, the random forest was used for the prediction of financial knowledge level classification as well, the prediction accuracy of 48.65% did not differ much from the model built by multinomial regression.

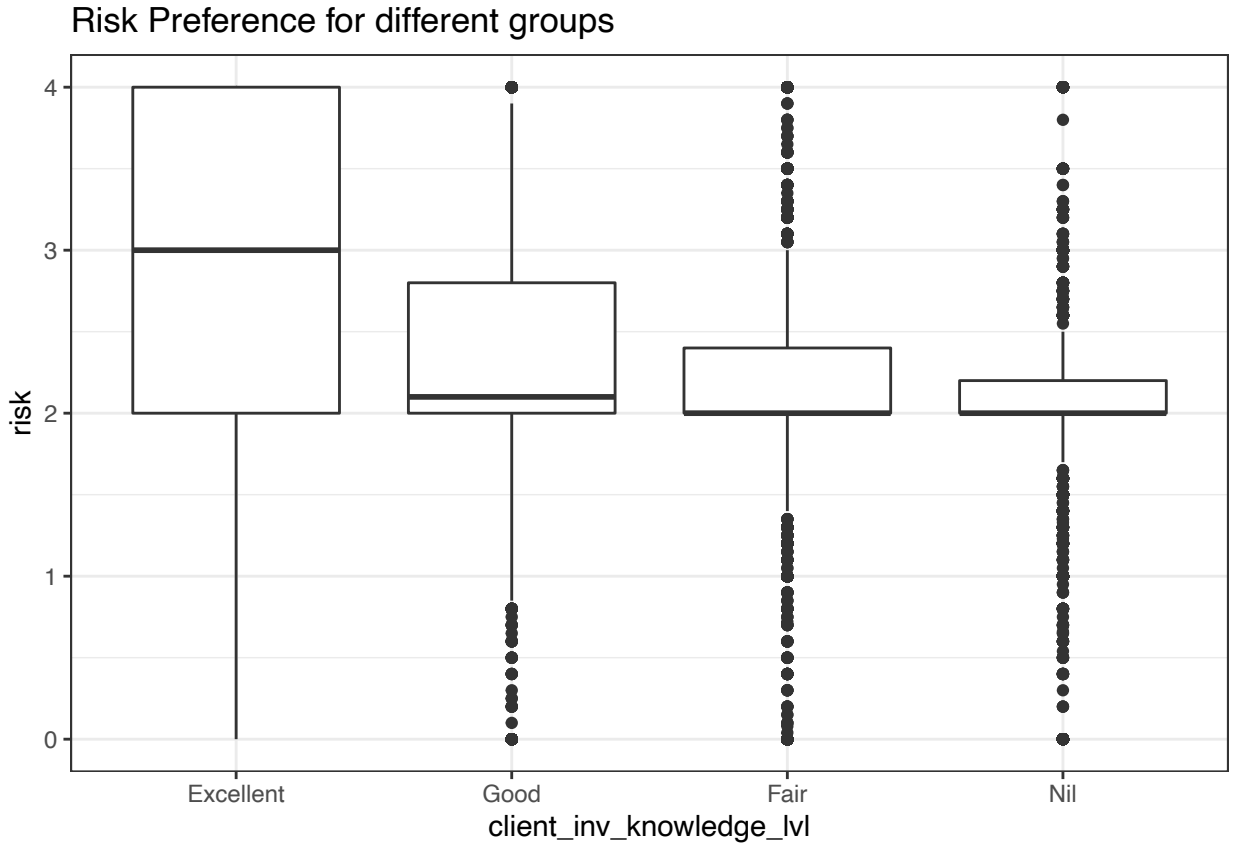
Therefore, it could be concluded that although demographic traits including gender, marital status, residence, retirement condition, maximum age recorded, and annual income were related to an investor's grouping of his/her investing knowledge, this relationship was not strong and decisive (detail in appendix). It was also worth noticing, that net worth was shown to be a significant predictor for the random forest model, which was different from the conclusion gained by the multinomial model. After testing the correlation between the predictor annual income and net worth, it was concluded that they did not have a noticeable linear correlation, thus the reason for the random forest model attaching importance to net worth was worth further investigation.

3.2 Risk Prefrence for Different Self-identified Fianncial Knowledge Groups

The analysis of the relationship between risk tolerance and the clients' financial literacy level is done in two sections: first, according to clients' classification of knowledge group, what is the risk tolerance they think they have; second, what is the actual risk tolerance of people in different self-identified financial knowledge groups.

For the first section, investors' ideal risk tolerance: it was recorded in the KYC table with a factor of five levels: low, low-medium, medium, medium-high, and high. Clients' can assign a percentage to each level, which finally adds up to 100%. The analysis begins with quantifying this levelled factor into a continuous one, by assigning the value of 0 to the level low, 1 to the level low-medium, 2 to the level medium, 3 to the level medium-high and 4 to the level high, then multiply the number by the percentage a client assigned to this level, adding up the resulting value of each level, that was the resulting score of ideal risk preference for this client.

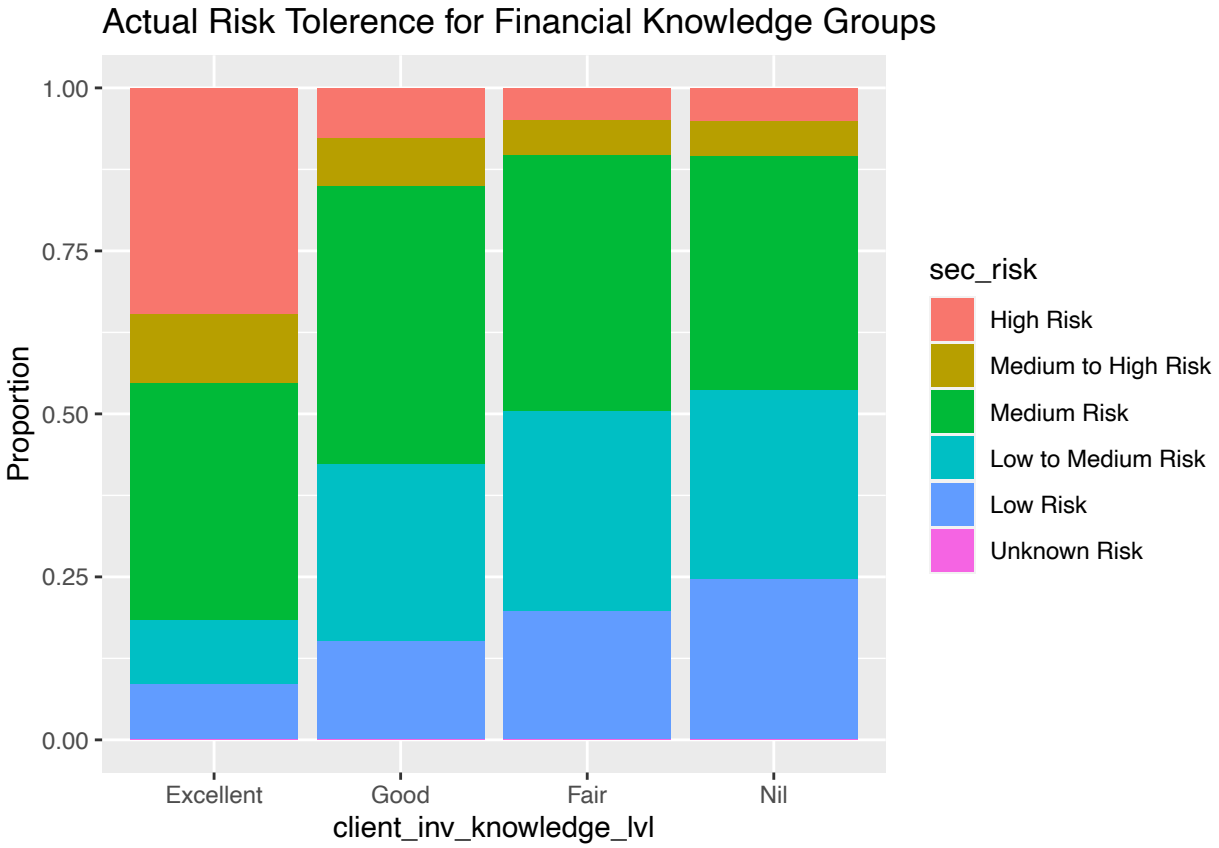
A combined boxplot was created to see the ideal risk preference score distribution of the four self-classified investment knowledge group, and it can be seen that the better the investment knowledge an investor think he/she has, the more risk they tend to afford. The mean risk preference score increased from the "nil" group to the "excellent" group, and the excellent group also had the largest variation in terms of the risk preference score among the four groups. Then, a linear model was built to test the significance of the relationship between the financial knowledge group and risk preference, the small p-value of the model support the argument that the relationship was significant.



The risk preference difference for different demographic groups was also analyzed, with linear models built individually between each demographic variable in the section above and risk preference as the response variable. It is observed that among all seven models, all had a small p-value except for the model with net worth, suggesting that the predictors of marital status, gender, retirement status, region of residence, annual income, and maximum age recorded were all related with a person's ideal risk preference (details in appendix 7.3.2).

For the analysis of the actual risk tolerance of investors, the risk rating of the securities they purchase was tracked, and the information was gathered to see the percentage distribution of securities of each risk level (five risk levels mentioned above plus unknown risk category) on a group-wide basis. To achieve this purpose, the trade table, ref_risk_rating table and KYC table were connected: the trade table includes the record of the security each client purchased, and their risk level was mentioned in the ref_risk_rating table, so these two tables were connected using security id; besides, the account id was used to link the KYC table, which includes the investment knowledge grouping, and the trade table, which contains the actual investment choices of investors. Proportional bar graphs for the population and regular bar plots of each financial knowledge were then drawn based on this new table to derive the result. From the proportional bar graph by financial knowledge level (appendix 7.3.2), it could be seen that investors who self-identified as having more investment knowledge tend to choose higher-risk securities, it was most obviously shown in the percentage distribution of each risk level of securities for group "excellent", where 34.8% of the securities purchased were of high risk (appendix 7.3.2). Investors who think they have less financial knowledge tend to be more conservative and preferred to choose lower-risk financial instruments. A hypothesis test with the null hypothesis "one's actual risk tolerance is the same as he/her assumed" was also performed, with a chi-squared test giving the small p-value, this null hypothesis was rejected.

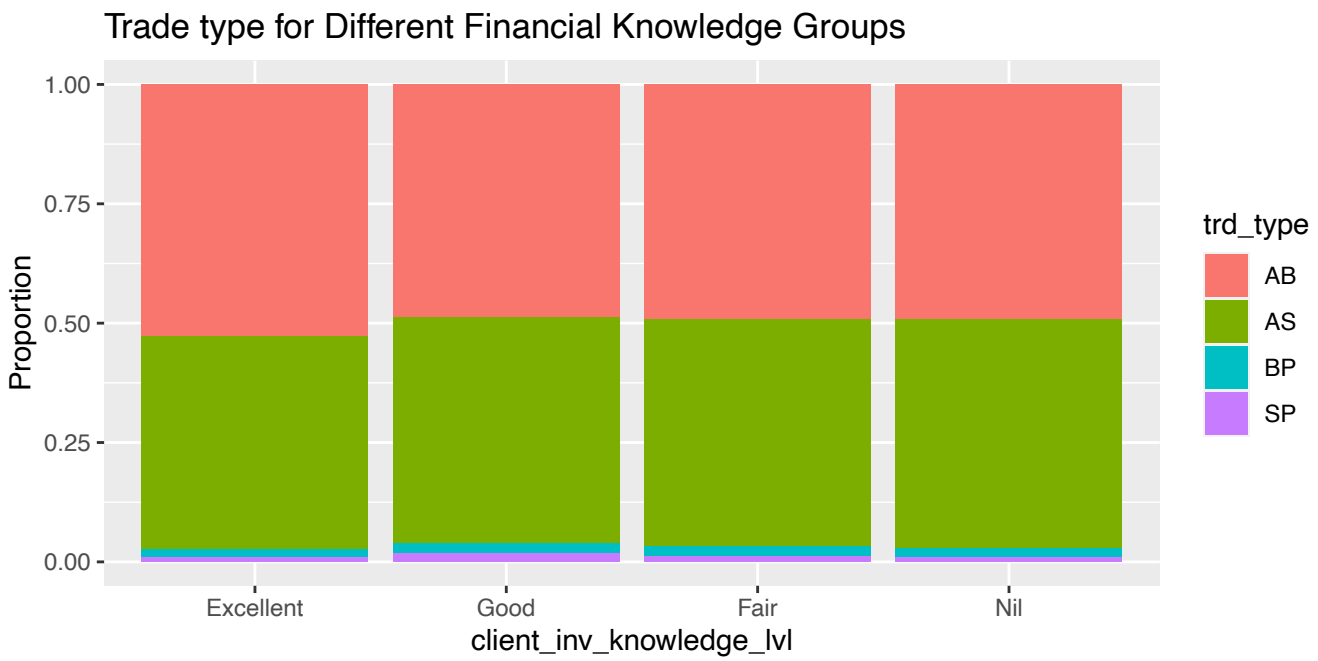
Therefore, the conclusion drawn from the actual practice of risk tolerance for each financial knowledge group matched the result of ideal risk preference that investors evaluated themselves, indicating that investors generally followed their desired risk preference during investment activities.



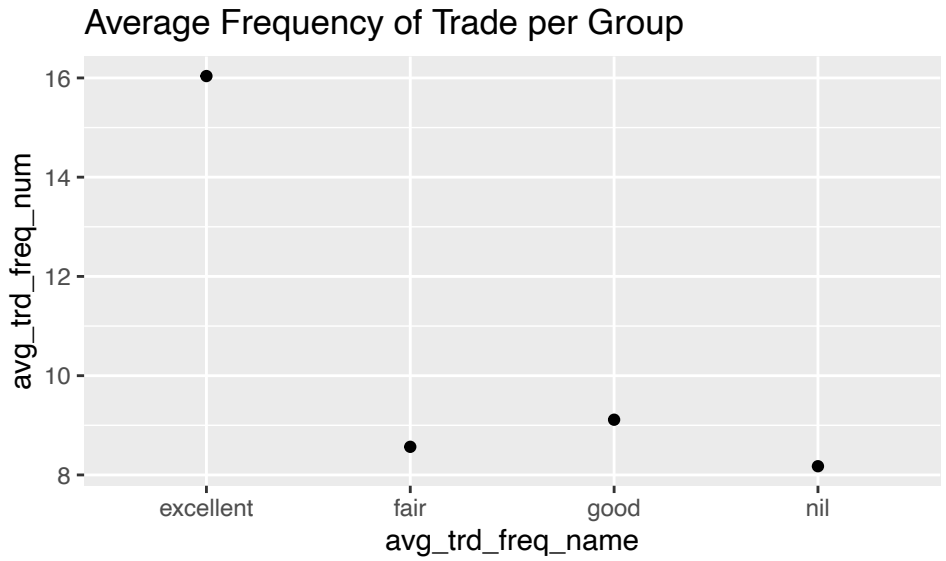
3.3 Trading Behaviour for Different Self-identified Fianncial Knowledge Groups

It was also worth investigating if investors who believe they have different levels of investment knowledge would trade differently. Information provided in the dataset includes the trading type, trading frequency and financial instrument type of each transaction for each account, so our analysis was done in these three aspects.

For the trading type (agency or principle buy/sell), the table created for the actual risk tolerance section was used again, since it contains the trading type (originally from the trade table) and financial knowledge group division result from the KYC table. A proportional bar graph was drawn directly for the trade type distribution for each group (appendix 7.3.3), and it could be seen that most investors, regardless of which group they are in, prefer agency operations to principle operation, though groups "good" and "fair" contain a larger percentage of principle operations, the difference was not significant generally. It was also observed from the graph, that the type of operations for group fair and nil were similar at first glance. The result of a chi-squared test with small p-value supported this hypothesis, that the trade type for these two groups did have similarities.

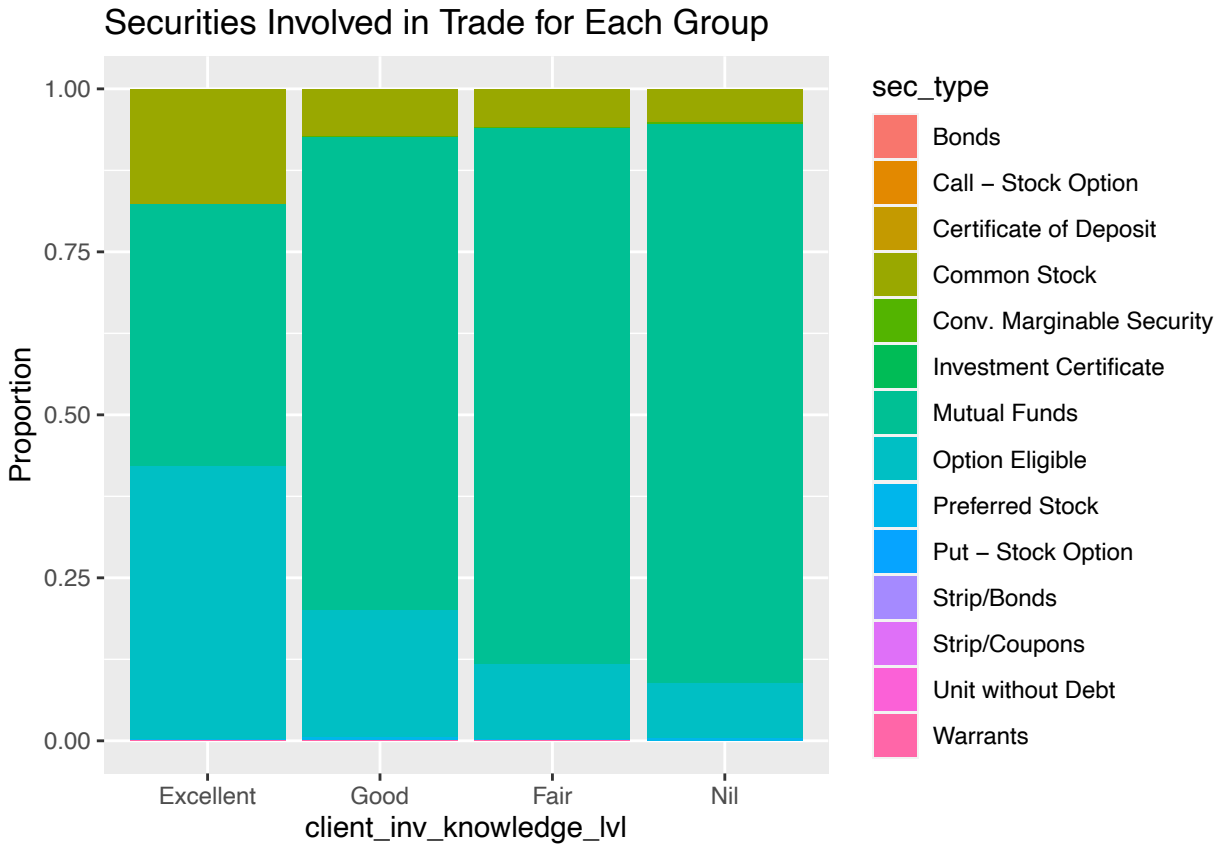


The study of trading frequency for each group led to a different conclusion. By counting the trade frequency of each account and averaging the frequency of accounts in each financial knowledge group (done by looping through the table containing trade information), it was discovered that the trading frequency was significantly higher for group "excellent", with the average of 16 times/account. The frequency of trade decreased as the financial knowledge level decreased, but the difference among the three other groups was not as significant (appendix 7.3.3).



The security type involved in the transactions of each group also differs. The table derived from the actual risk tolerance section can be directly applied here, as information regarding security type was contained inside. By drawing the proportional bar graph and the table (appendix 7.3.3), it can be concluded that although investors had a wide range of financial instruments to choose from (14 types included in our study), the choice of the majority was limited to three types for all groups: common stock, mutual fund, and option eligible. Among all the groups, groups "fair" and "nil" performed similarly, while an overall trend of increase in the proportion of common stock and option eligible security in the portfolio can be seen as people being more confident with their

financial knowledge.



3.4 Accracy of Financial Knowledge Self-evaluation

The last question to investigate is whether the classification of financial knowledge level done by the investor themselves actually matches their investment performance – whether people who think they have more investment knowledge has a higher return. The investment performance is measured by internal rate of return(IRR) in this project: using the book value at the start and end of the recorded time period (in position table), and the transaction record given in the trade table, the cashflows in the recorded period can be arranged as a table with two columns, date and the value of the account. It was noticed that the transactions recorded in the trade table were based on a longer time period than the records in the position table, so only the transactions in the time frame determined by the recorded date in the position table were used for the IRR calculation. Then, using a for loop, each account's timeline was first arranged in the unit of year, and the built-in IRR function in the jrvFinance package was used to calculate the internal rate of return for each account. Due to the large volume of work for 45,790 accounts, this calculation was separated into five groups to be performed, and it is suggested that the result for each group be saved as an individual CSV file. The files were then combined to compose the complete IRR file for all the accounts.

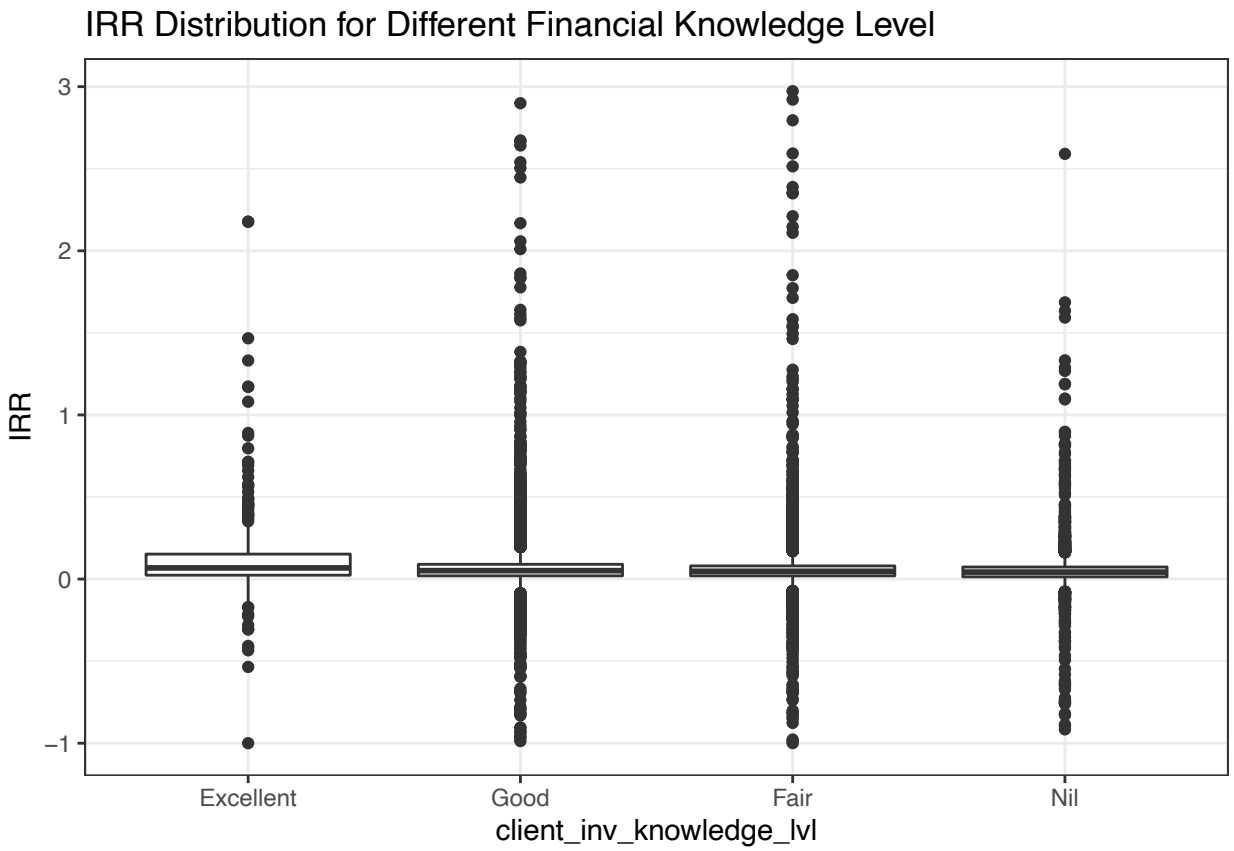
It was observed that there are three types of abnormal results for the IRR calculation: zero, not applicable, and infinity. First, for the accounts with IRR not applicable, it is observed that this type of situation happened because of the unexpectedly large amount of cash outflow which cannot be balanced by the cash inflow in the observation period. Second, for the accounts with IRR =

infinity, this type of situation was due to an unexpected cash inflow than outflow. The first two situations were caused by similar reasons, and the root of the problem may lie in the data source itself. There may be unrecorded transactions from the bank side that were not included in the trade table, thus leading to the occurrence of unbalanced money inflow or outflow at some time point. The problem was then fixed by the reconstruction of the cash flow table by using the adjusted book value and unrecorded cash flow data from the bank, and the following analysis was on the basis of the revised data. Third, for the accounts with IRR = 0, these only composed a very small proportion of the whole dataset(32 in the nil group, 24 in the fair group, 32 in the good group, 3 in the excellent group) and there wasn't abnormality in terms of cash flow, so those can be included.

The revised dataset of IRR includes 30,827 entries, and it was connected with the KYC table and trade table to link the performance of each deletion of accounts with unreasonable records in the revised cashflow table composition stage. From the revised table, another round of filtering with IRR<3 was done, keeping 99.71% of records while keeping out the extremely large IRR. This operation is for tracking the trend of the majority, making the data influenced less by the extreme cases.

The distribution of the internal rate of return for each group is shown in the graph below. It can be seen that people who self-identified as having more financial knowledge did perform better according to the available data, with the mean internal rate of return increasing as the evaluation of the investing ability of an individual increase. It was argued that this performance might be related to the type of trade an investor chose: agent operations may or may not outperform principal transactions, and this factor may affect the IRR as well. However, the dataset suggested that one account may use multiple types of trade for different transactions, so it was difficult to link a specific trade type to an account, thus the effect of the types of trade on the return of the investment was hard to identify.

Since the distribution of each group's IRR all have long tails and spread out, while the standard deviation was not high, indicating the extreme cases to have large numerical differences but relatively small counts, a closer look at the mean and median of each group was done to identify the performance at an average level. It was discovered that the difference between the mean and median in group "nil" was relatively small, and it increases as the self-evaluation of the financial literacy of an investor increases. The difference became 1.6% for the level of good, and 4.88% for the group "excellent". This difference between mean and median, with mean always higher than median, suggested that there are a smaller proportion of people in each group having comparatively higher return than average. Although the return of investment increases as the self-identified investment knowledge level increases, there wasn't a very large difference in the median performance of each group, varying from 4.2% to 6.8%. However, the return of investment had a larger variation as the financial knowledge level group increased, indicating a positive relationship between the self-evaluation of investment knowledge and the possible difference in return.



Apart from the exploratory analysis, a nested model test for investigating the potential relationship between IRR and client investment knowledge was also performed, the multinomial model between investment knowledge level and IRR and a null model were built, and the difference of deviance between the two models was used to evaluate the importance of the predictor client investment knowledge level, the small result p-value leads to the conclusion that these two variables are related. However, IRR was not significant enough to predict an investor's investment knowledge level, according to the data we have, by splitting the trait-test set and predicting knowledge level using the training model, the accuracy was 43%, and balanced accuracy was slightly improved to be 50%, but either of these two data was significant. Therefore, it can be concluded that IRR was related to the financial knowledge level of an investor, but this predictor solely was not sufficient to be used to predict the self-evaluation of the financial knowledge level of an investor.

4. Conclusions

From the analysis above, the conclusion of this project can be drawn as below:

Demographic information including marital status, gender, region of residence, retirement status, maximum age recorded, and annual income were related to an investor's classification of his/her investment knowledge level, while the connection between the predictor net worth and the knowledge level was proved to not have a significant linear relationship. Although it can be demonstrated that the variables were connected to one's evaluation in the financial aspect, the connection was not close and practical for the prediction of financial knowledge level based on these demographic traits.

The risk preference expressed by the investors aligned with their actual risk tolerance, that clients who think they know more about investment leant toward higher risk, which was reflected in the risk level of the securities they choose. The difference between the group "excellent" and other groups was significant.

All four financial knowledge groups preferred agency operations to principle operations as the type of trade, and the trading frequency was higher for investors who self-identified as having more investment knowledge. The trading frequency for the group "excellent" was much higher, while the difference among the three other groups was not large. The three most commonly purchased security types were common stock, mutual funds, and option-eligible securities.

The actual performance of the investors was measured by the internal rate of return, and the IRR of different financial knowledge groups was compared. The problems in calculated IRR for the accounts revealed the problem of the potential incompleteness of the dataset: the lack of some large-amount transactions in the trade table. After the reconstruction of the IRR table, it was observed that the more investment knowledge an investor thinks he/she has, generally the higher the return of investment was. The variation of the return on investment generally increases as the self-evaluated financial knowledge level increases, and a small proportion of investors in each group had much higher returns than the median level, making the mean return for each group higher. The median return for the four groups did not differ as much as the mean. It was concluded that the internal rate of return was related to the grouping of financial knowledge level, but it was not sufficient by itself to be used to make predictions on the grouping by investment knowledge.

5.General Discussion

It was suggested in the Executive Summary For the Financial Consumer Agency of Canada in 2016, that although financial knowledge was important for the success of investment decisions, it was not the only factor of influence. The confidence of the investor also affects the result of the investment — "Confidence seems to direct seniors and near-seniors with low knowledge toward financially desirable behaviours in several key domains" (Hui, Nguyen, Palameta, Gyarmati, 3). This conclusion can be reflected in this project: the investment knowledge level we used as the basis for group division was a voluntary classification done by the investor themselves, and the group division itself was influenced by the confidence level of the investor. Therefore, the success of the investors in the group "excellent" may not only be resulted from financial literacy, but also because of their confidence in investment ability, and this applies to the three other groups as well. If the effect of solely financial literacy on investment decisions is to be studied, it should be based on a population that is similarly confident with their investment ability, according to the study referred to above, but this condition is too hard to be met. At the current stage, it is personally stated that the psychological effect on investors is hard to be isolated.

For the second stage of this project, studies surrounding the calculated IRR dataset will be performed. The study is divided into several sections: the difference in return of investment for different gender; the relationship between IRR and portfolio volume; the influence of trading fees and the use of agents; and traits for different levels of internal rate of return.

6.Acknowledgement

Thank you to Dr.Kristina Sendova, Dr.Shu Li, and Dr.Miao Yang. This project would not be accomplished without your suggestions and guidance, it was a great pleasure working with you.

Thank you to Financial Wellness Lab, the USRI program and the department of Statistical and Actuarial Science, Faculty of Science for giving me the opportunity to participate in the research.

7.References

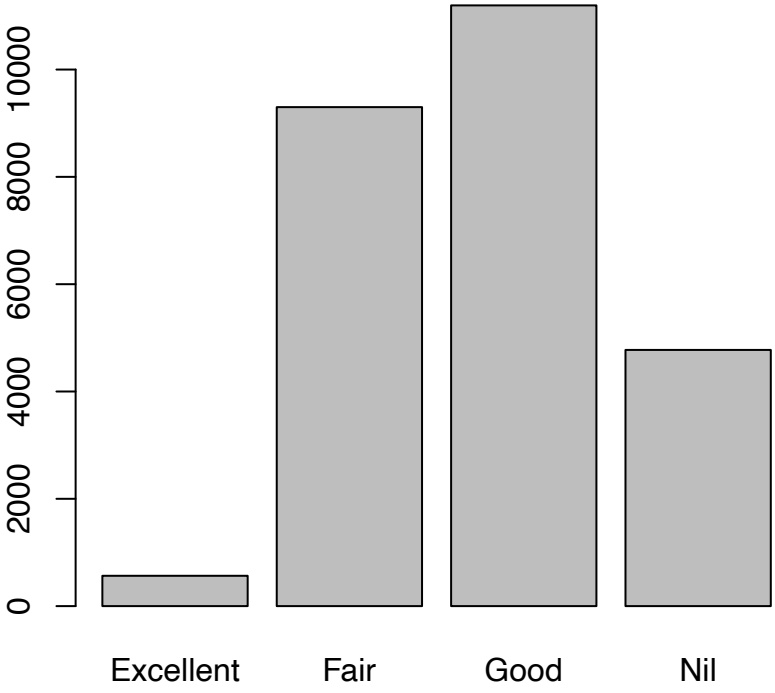
Hui, T. S.-wai, Nguyen, C., Palameta, B., & Gyarmati , D. (2016, May). The role of financial literacy in financial decisions and retirement preparedness among seniors and near-seniors. https://www.canada.ca. Retrieved July 18th, 2022, from https://www.canada.ca/content/dam/canada/financial-consumer-agency/migration/eng/resources/researchsurveys/documents/financial-decisions-retirement-preparedness.pdf

8.Appendix

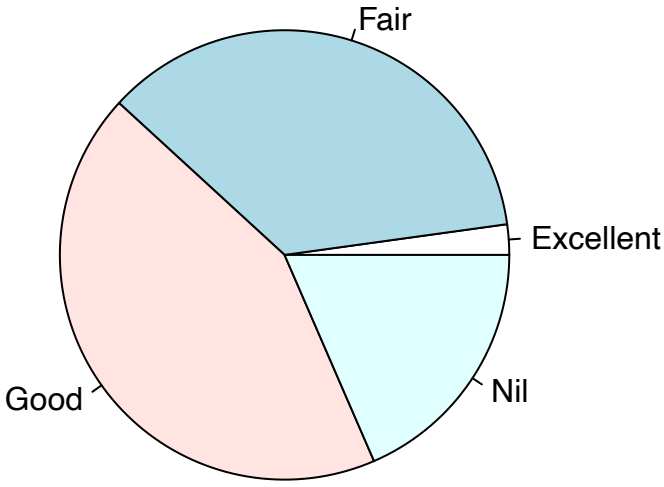
For Section 3.1

Distribution of investor in the four investment knowledge groups

```
##
## Excellent      Fair      Good       Nil
##       565      9300     11196      4775
```




```
##
##  Excellent       Fair       Good        Nil
## 0.02186871 0.35996284 0.43334882 0.18481963
```

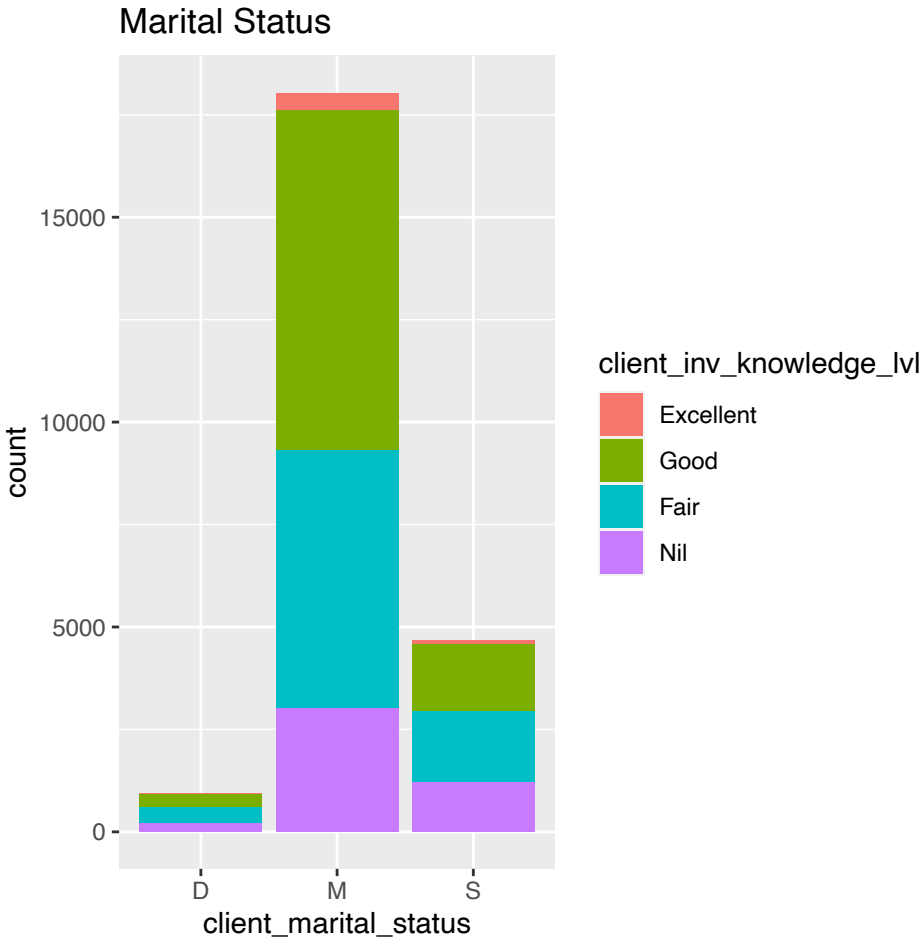


Relationship between each demographic variables and financial knowledge group

Marital Status

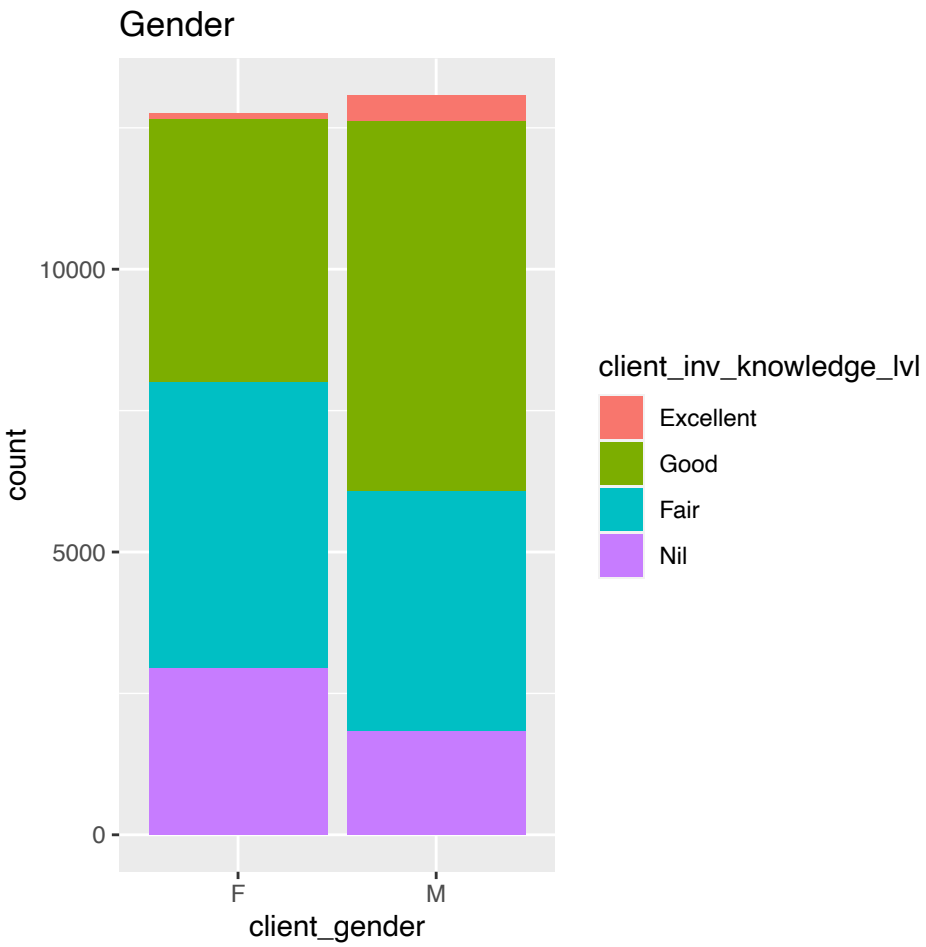
Since records with label star in this column are not in any defined marital status category, it is decided to be excluded from the analysis on this aspect.

```
##
##                          D           M           S
##   Excellent 0.0006760183 0.0183369951 0.0039716072
##   Good      0.0138583742 0.3502619571 0.0700101403
##   Fair      0.0163089403 0.2667737029 0.0724607064
##   Nil       0.0090417441 0.1268801758 0.0514196383
```



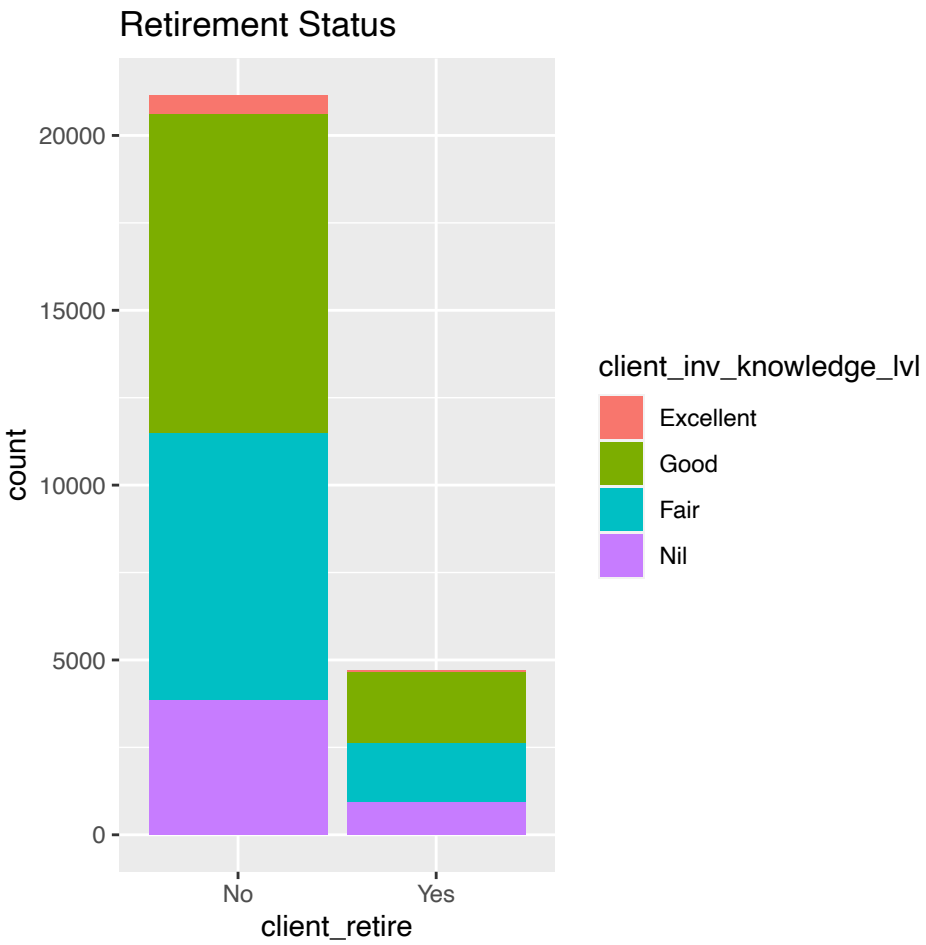
Gender

```
##
##                        F          M
##   Excellent 0.00394798 0.01792073
##   Good      0.18063942 0.25270940
##   Fair      0.19527017 0.16469268
##   Nil       0.11422047 0.07059916
```



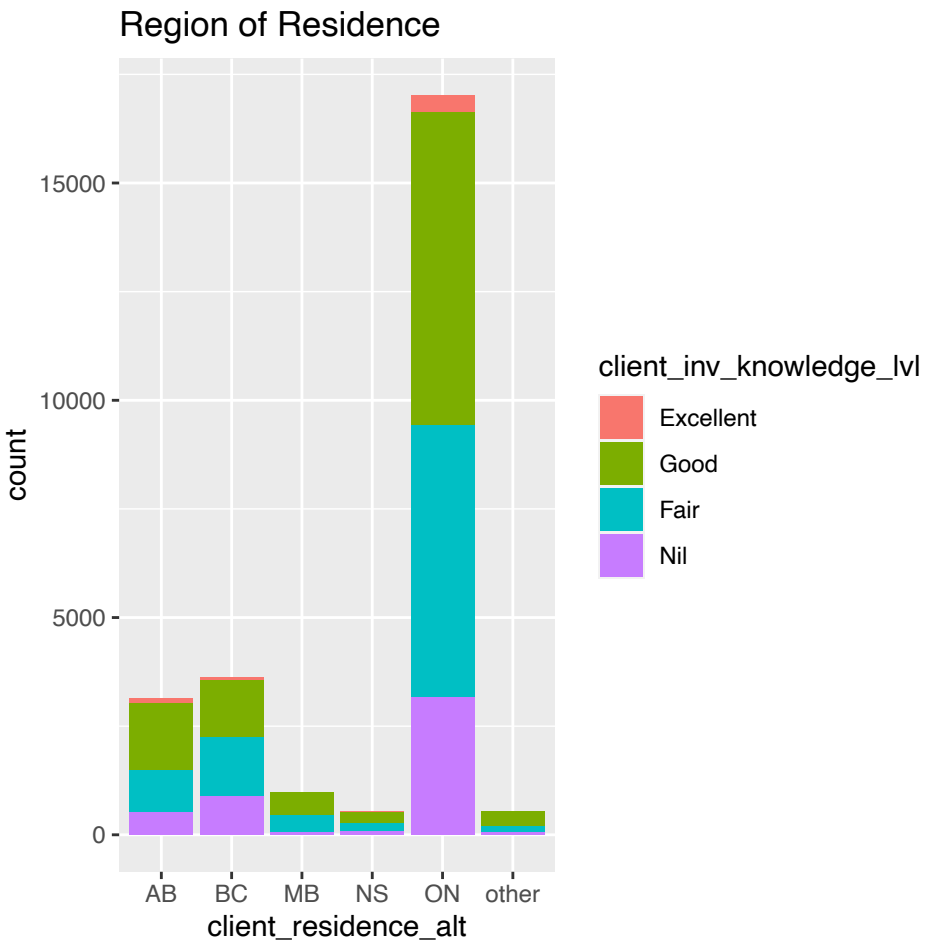
Retirement Status

```
##
##                      No         Yes
##   Excellent 0.020088249 0.001780461
##   Good      0.353963462 0.079385354
##   Fair      0.295711410 0.064251432
##   Nil       0.148513702 0.036305930
```

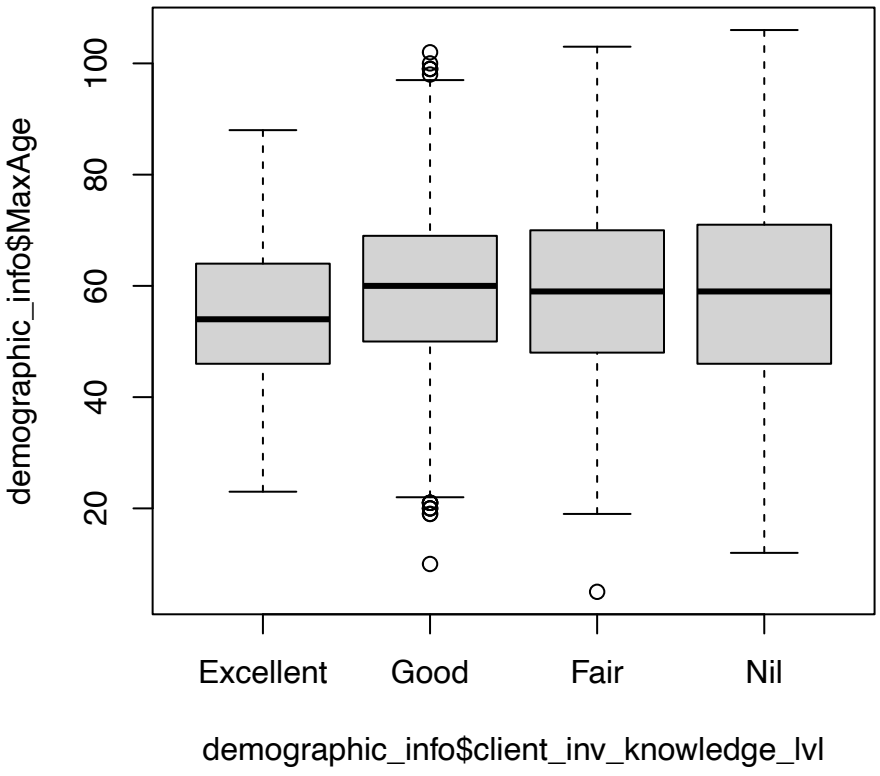


Region of Residence

```
##
##                       F          M
##   Excellent 0.00394798 0.01792073
##   Good      0.18063942 0.25270940
##   Fair      0.19527017 0.16469268
##   Nil       0.11422047 0.07059916
```

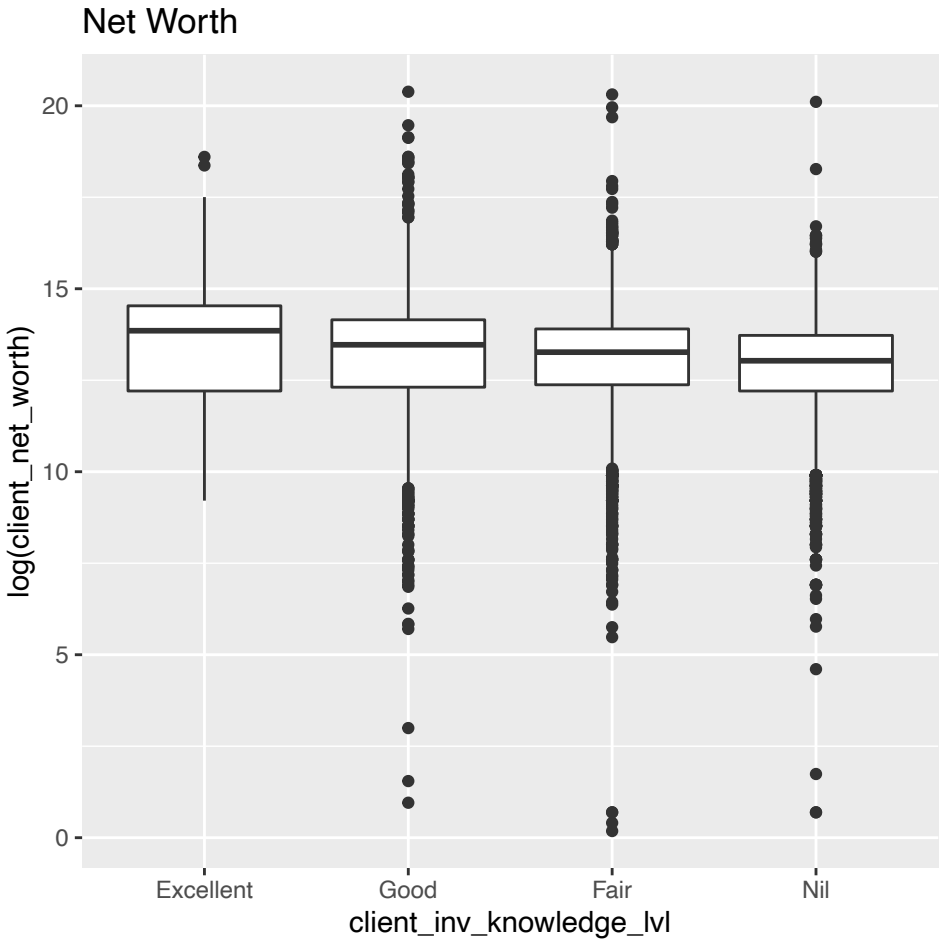


Maximum Age attained



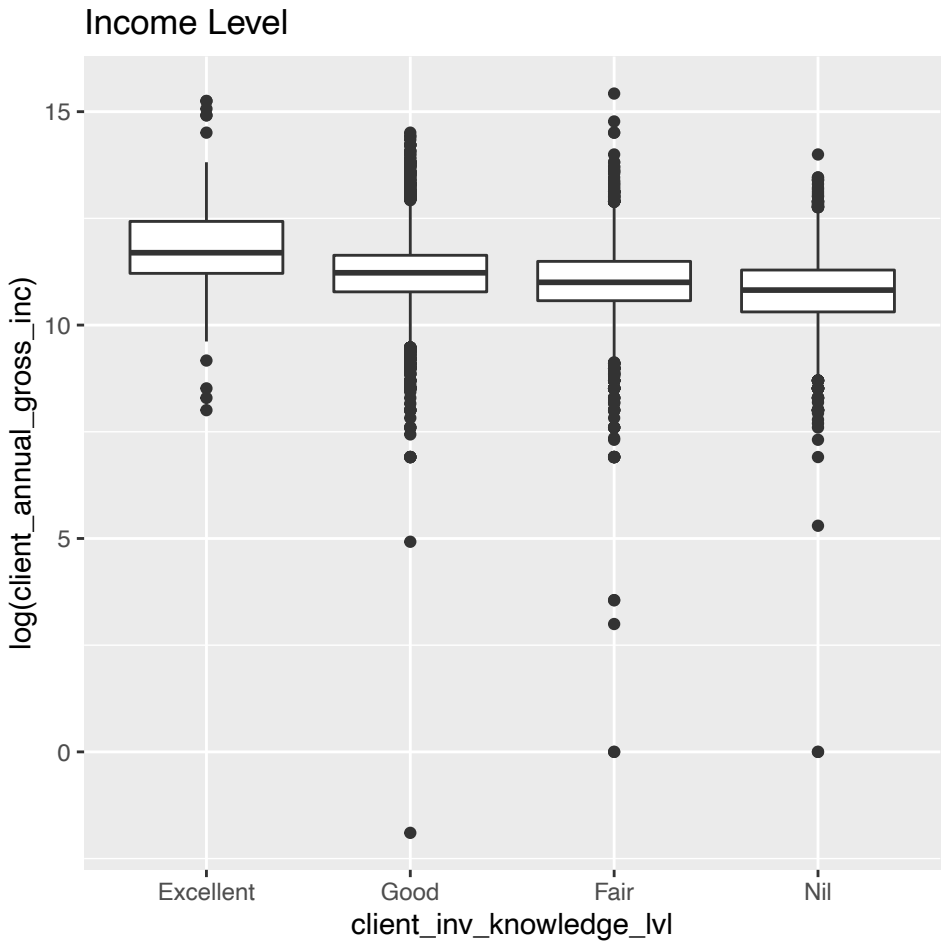
Net worth

During the first plot, it was discovered that there was an outlier of extremely large, or unattainable value, it was then excluded from the following analysis. The final plot was with net worth taken log, due to its large value, making the trend hard to identify.



Annual Income

Similar situation as above happened, the only outlier was excluded.



AIC Test Result

```
summary(full_mod_aic)
```

```
## Call:
## multinom(formula = client_inv_knowledge_lvl ~ client_gender +
##     client_marital_status + MaxAge + client_retire + client_annual_gross_inc +
##     client_residence_alt, data = demographic_info1)
##
## Coefficients:
##           (Intercept) client_genderM client_marital_statusD
## Excellent -2.66005149      1.6624761            -1.03123066
## Good      -0.38988666      0.5584722            -0.28668698
## Fair      -0.08598131      0.1770925            -0.01131091
##           client_marital_statusM client_marital_statusS       MaxAge
## Excellent             -0.7010202             -0.9278007 -0.003897381
## Good                   0.1939368             -0.2971365  0.008890329
## Fair                   0.1195886             -0.1942590  0.004967126
##           client_retireYes client_annual_gross_inc client_residence_altBC
## Excellent      -0.49834413           8.924807e-06             -0.6348350
## Good            0.03410959           7.227198e-06             -0.5478183
## Fair           -0.06561711           4.036331e-06             -0.1796257
```

```
##           client_residence_altMB client_residence_altNS client_residence_altON
## Excellent               0.076190              0.8089086            -0.23037802
## Good                    1.200774              0.1646978            -0.22772015
## Fair                    1.328002              0.2162700             0.07737054
##           client_residence_altother
## Excellent                 0.2042781
## Good                      0.7814623
## Fair                      0.4395113
##
## Std. Errors:
##             (Intercept) client_genderM client_marital_statusD
## Excellent 2.513564e-06   1.988502e-06           7.259240e-08
## Good      5.726849e-06   2.656826e-06           2.105368e-07
## Fair      5.870321e-06   1.877727e-06           3.003051e-07
##           client_marital_statusM client_marital_statusS       MaxAge
## Excellent           2.004676e-06           4.363122e-07 0.0001511012
## Good                4.248828e-06           1.268451e-06 0.0003872858
## Fair                4.000135e-06           1.579757e-06 0.0004075249
##           client_retireYes client_annual_gross_inc client_residence_altBC
## Excellent     3.682315e-07            2.983463e-07           3.172518e-07
## Good          2.026996e-06            2.903777e-07           1.017531e-06
## Fair          2.299528e-06            3.099405e-07           1.196978e-06
##           client_residence_altMB client_residence_altNS client_residence_altON
## Excellent           4.921273e-08           1.282345e-07           1.634851e-06
## Good                1.859598e-07           1.806234e-07           3.622172e-06
## Fair                1.951566e-07           1.497330e-07           3.793380e-06
##           client_residence_altother
## Excellent              3.227329e-08
## Good                   8.990866e-08
## Fair                   6.031045e-08
##
## Residual Deviance: 51228.78
## AIC: 51300.78
```

Detail breakdown of hypothesis tests for different demographic variables, model accuracy

H0– investment knowledge of a client is independent of her/his marital status.

```
know_marital_revised = table(demographic_info_marital$client_inv_knowledge_lvl,
                             demographic_info_marital$client_marital_status)
summary(know_marital_revised)
```

```
## Number of cases in table: 23668
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 308.74, df = 6, p-value = 1.095e-63
```

```
know_marital_revised
```

```
##
##                D    M    S
##   Excellent   16  434   94
##   Good       328 8290 1657
##   Fair       386 6314 1715
##   Nil        214 3003 1217
```

```
#For tables larger than 2x2, the chi-square approximation can be good
#even if some expected counts are less than 5.
chisq.test(know_marital_revised)
```

```
##
##  Pearson's Chi-squared test
##
## data:  know_marital_revised
## X-squared = 308.74, df = 6, p-value < 2.2e-16
```

H0– investment knowledge of a client is independent of her/his gender.

```
know_gender = table(demographic_info$client_inv_knowledge_lvl,
                    demographic_info$client_gender)
chisq.test(know_gender)
```

```
##
##  Pearson's Chi-squared test
##
## data:  know_gender
## X-squared = 869.93, df = 3, p-value < 2.2e-16
```

H0– investment knowledge of a client is independent of her/his retirement status.

```
know_retire = table(demographic_info$client_inv_knowledge_lvl,
                    demographic_info$client_retire)
chisq.test(know_retire)
```

```
##
##  Pearson's Chi-squared test
##
## data:  know_retire
## X-squared = 45.999, df = 3, p-value = 5.676e-10
```

H0– investment knowledge of a client is independent of her/his Region of Residence.

```
know_region = table(demographic_info$client_inv_knowledge_lvl,
                    demographic_info$client_residence_alt)
chisq.test(know_region)
```

```
##
##  Pearson's Chi-squared test
##
## data:  know_region
## X-squared = 395.93, df = 15, p-value < 2.2e-16
```

For the three predictor below, a full model with all predictors and models with the predictor evaluated excluded were constructed, and chi-squared tests were performed. H0– investment knowledge of a client is independent of her/his age.

```
pchisq(52.29755,full_mod$edf-mod_ex_MaxAge$edf,lower=F)
```

```
## [1] 2.588155e-11
```

H0– investment knowledge of a client is independent of her/his net worth.

```
pchisq(6.052035,full_mod$edf-mod_ex_net_worth$edf,lower=F)
```

```
## [1] 0.1091059
```

H0– investment knowledge of a client is independent of her/his annual income.

```
pchisq(713.1236,full_mod$edf-mod_ex_client_annual_gross_inc$edf,lower=F)
```

```
## [1] 2.994387e-154
```

Fit of the Full Model

```
#Table result comparison:
xtabs( ~ pred_result + test$client_inv_knowledge_lvl)
```

```
##             test$client_inv_knowledge_lvl
## pred_result  Nil Excellent Good Fair
##   Nil         52         1   16   34
##   Excellent    0         1    2    0
##   Good       603       140 2200 1405
##   Fair       663        16  880 1087
```

```r
#We can compute the proportion correctly classified as:
(81+2+961+2208)/nrow(test)
```

```
## [1] 0.4580282
```

```r
#Note: the data used for the accuracy of the prediction
#is based on one random split.

#Balanced accuracy
library(yardstick)
bal_accuracy_vec(test$client_inv_knowledge_lvl,pred_result) #53.5%
```

```
## [1] 0.5374454
```

Random Forest Model

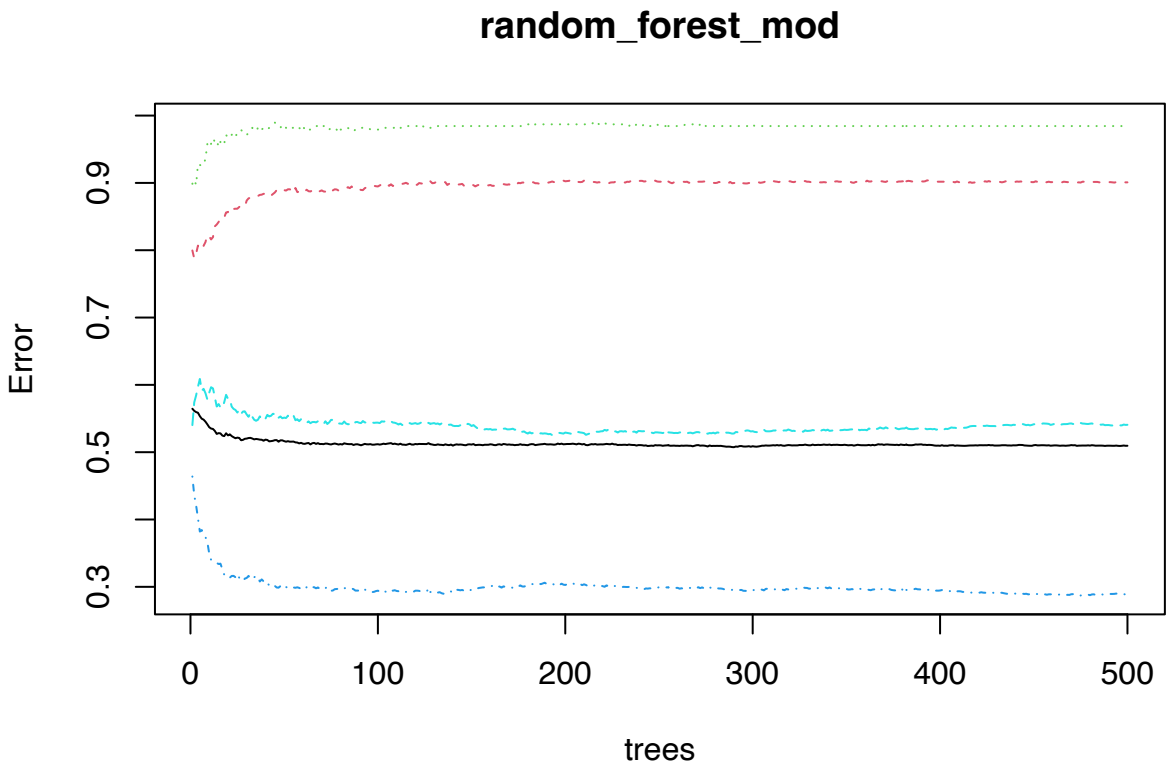
```r
confusionMatrix(random_pred, test$client_inv_knowledge_lvl) #The accuracy is 48.65%
```

```
## Confusion Matrix and Statistics
##
##            Reference
## Prediction   Nil Excellent Good Fair
##   Nil        123         1   56   99
##   Excellent    0         2    1    0
##   Good       549       142 2212 1266
##   Fair       646        13  829 1161
##
## Overall Statistics
##
##                Accuracy : 0.4927
##                  95% CI : (0.481, 0.5044)
##     No Information Rate : 0.4363
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.1597
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
## Statistics by Class:
##
##                      Class: Nil Class: Excellent Class: Good Class: Fair
## Sensitivity             0.09332        0.0126582      0.7140      0.4596
## Specificity             0.97302        0.9998559      0.5110      0.6747
## Pos Pred Value          0.44086        0.6666667      0.5306      0.4383
```

```
## Neg Pred Value          0.82481          0.9780189      0.6977      0.6933
## Prevalence              0.18563          0.0222535      0.4363      0.3558
## Detection Rate          0.01732          0.0002817      0.3115      0.1635
## Detection Prevalence    0.03930          0.0004225      0.5872      0.3731
## Balanced Accuracy       0.53317          0.5062571      0.6125      0.5672
```

```
#There isn't much difference compared to linear model

plot(random_forest_mod) #The error rate hasn't decreased much
```

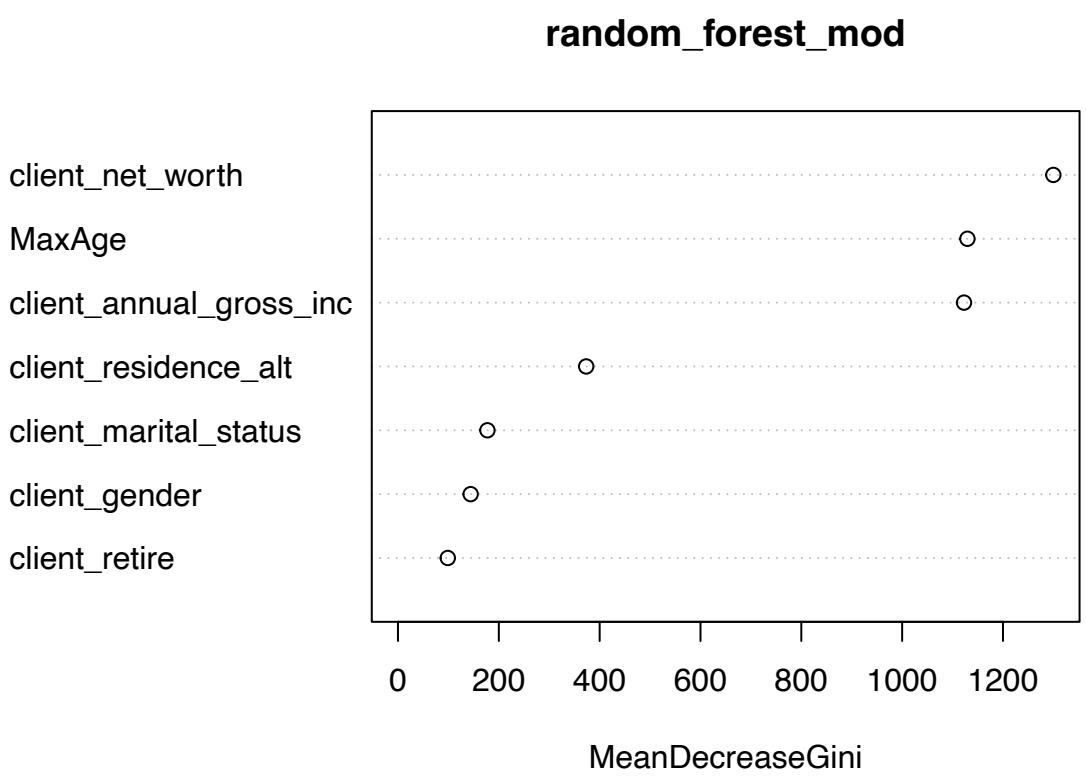



```
#since the number of trees goes beyond 50.

# Importance & Variable importance plot
importance(random_forest_mod)
```

```
##                         MeanDecreaseGini
## client_gender                  144.13766
## client_marital_status          177.40766
## client_net_worth              1299.91343
## MaxAge                        1129.46562
## client_retire                   98.77586
## client_annual_gross_inc       1122.72264
## client_residence_alt           373.32446
```

```
varImpPlot(random_forest_mod)
```




```
#It can be seen that net worth becomes the most important predictor,
#which is quite different from the result given by linear model.
```

```
#Correlation between annual income and net worth is not high.
cor(demographic_info1$client_annual_gross_inc, demographic_info1$client_net_worth)
```

```
## [1] 0.09754103
```

For Section 3.2

Linear model– risk and investment knowledge relationship

Linear model– risk and demographic variables relationship risk tolerance for different group of investors

Retirement

```
##
## Call:
## lm(formula = risk ~ client_retire, data = risk_info)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -2.2411 -0.2411 -0.1494  0.3589  2.0506
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.241088   0.004885  458.76   <2e-16 ***
## client_retire1 -0.291695   0.011460  -25.45   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7103 on 25834 degrees of freedom
## Multiple R-squared:  0.02447,    Adjusted R-squared:  0.02443
## F-statistic: 647.9 on 1 and 25834 DF,  p-value: < 2.2e-16
```

Residence

```
##
## Call:
## lm(formula = risk ~ client_residence_alt, data = risk_info)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3218 -0.2204 -0.2204  0.3796  2.0845
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 2.25541    0.01270 177.588  < 2e-16 ***
## client_residence_altBC     -0.30735    0.01734 -17.723  < 2e-16 ***
## client_residence_altMB      0.11820    0.02594   4.557  5.2e-06 ***
## client_residence_altNS     -0.33991    0.03305 -10.284  < 2e-16 ***
## client_residence_altON     -0.03502    0.01382  -2.534   0.0113 *
## client_residence_altother  0.06635    0.03295   2.014   0.0440 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7104 on 25830 degrees of freedom
## Multiple R-squared:  0.02428,    Adjusted R-squared:  0.02409
## F-statistic: 128.6 on 5 and 25830 DF,  p-value: < 2.2e-16
```

Marital Status

```
##
## Call:
## lm(formula = risk ~ client_marital_status, data = risk_info)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -2.1951 -0.1934 -0.1934  0.4049  1.8590
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              2.149451   0.015442 139.194  < 2e-16 ***
## client_marital_statusD -0.008455   0.028038  -0.302  0.76298
## client_marital_statusM  0.043909   0.016344   2.687  0.00722 **
## client_marital_statusS  0.045665   0.018678   2.445  0.01450 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.719 on 25832 degrees of freedom
## Multiple R-squared:  0.0004538,  Adjusted R-squared:  0.0003377
## F-statistic: 3.909 on 3 and 25832 DF,  p-value: 0.008388
```

Gender

```
##
## Call:
## lm(formula = risk ~ client_gender, data = risk_info)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2537 -0.2537 -0.1208  0.3791  1.8792
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.120847   0.006338  334.63   <2e-16 ***
## client_genderM 0.132893   0.008911   14.91   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7161 on 25834 degrees of freedom
## Multiple R-squared:  0.008537,   Adjusted R-squared:  0.008498
## F-statistic: 222.4 on 1 and 25834 DF,  p-value: < 2.2e-16
```

Net worth

```
##
## Call:
## lm(formula = risk ~ client_net_worth, data = risk_info)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1881 -0.1881 -0.1881  0.4119  1.8119
##
```

```
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.188e+00  4.474e-03 489.048   <2e-16 ***
## client_net_worth -1.254e-61  4.794e-61  -0.262    0.794
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7191 on 25834 degrees of freedom
## Multiple R-squared:  2.648e-06,  Adjusted R-squared:  -3.606e-05
## F-statistic: 0.0684 on 1 and 25834 DF,  p-value: 0.7937
```

Annual gross income

```
##
## Call:
## lm(formula = risk ~ client_annual_gross_inc, data = risk_info)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1923 -0.1882 -0.1874  0.4118  1.8129
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             2.187e+00  4.489e-03 487.195   <2e-16 ***
## client_annual_gross_inc 1.046e-08  4.090e-09   2.558   0.0105 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7191 on 25834 degrees of freedom
## Multiple R-squared:  0.0002532,  Adjusted R-squared:  0.0002145
## F-statistic: 6.543 on 1 and 25834 DF,  p-value: 0.01053
```

Maximum age recorded

```
##
## Call:
## lm(formula = risk ~ MaxAge, data = risk_info)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.53173 -0.29893 -0.07625  0.40290  2.22740
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.7847674  0.0179190  155.41   <2e-16 ***
## MaxAge      -0.0101217  0.0002948  -34.34   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7033 on 25834 degrees of freedom
## Multiple R-squared:  0.04365,    Adjusted R-squared:  0.04361
## F-statistic:  1179 on 1 and 25834 DF,  p-value: < 2.2e-16
```

Actual risk preference

Table exhibition of the actual risk tolerance for different groups

```
##
##             High Risk Medium to High Risk Medium Risk Low to Medium Risk
##   Excellent      5163                1557        5412                1456
##   Good          12620               11980       68943               44268
##   Fair           6384                6672       49907               39092
##   Nil            2845                3109       20214               16354
##
##             Low Risk Unknown Risk
##   Excellent     1262            1
##   Good         24418            7
##   Fair         25072           13
##   Nil          13938           10
```

Proportional Table

```
##
##                High Risk Medium to High Risk  Medium Risk Low to Medium Risk
##   Excellent 3.476534e-01        1.048414e-01 3.644199e-01       9.804054e-02
##   Good      7.778791e-02        7.384304e-02 4.249550e-01       2.728618e-01
##   Fair      5.021236e-02        5.247758e-02 3.925358e-01       3.074721e-01
##   Nil       5.038073e-02        5.505578e-02 3.579600e-01       2.896051e-01
##
##                 Low Risk Unknown Risk
##   Excellent 8.497744e-02 6.733553e-05
##   Good      1.505091e-01 4.314702e-05
##   Fair      1.971999e-01 1.022495e-04
##   Nil       2.468213e-01 1.770852e-04
```

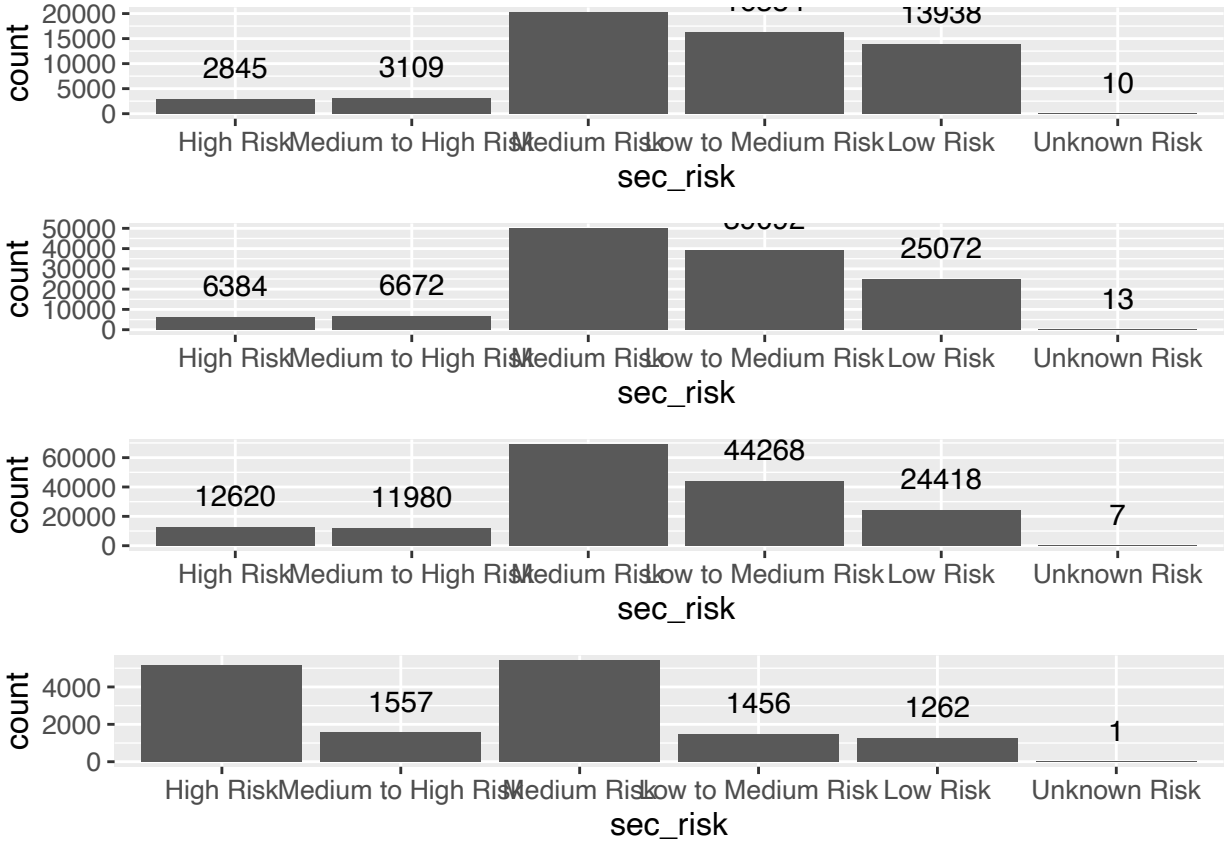
Hypothesis test

```
#H0: Actual risk preference OF different groups are the same.
chisq.test(sec_risk_distr)
```

```
## Warning in chisq.test(sec_risk_distr): Chi-squared approximation may be
## incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  sec_risk_distr
## X-squared = 23081, df = 15, p-value < 2.2e-16
```

Group-wise analysis



For Section 3.3

Test for similarity between the trading behaviours of group nil and fair

```
##
##  Pearson's Chi-squared test
##
## data:  table_trade_fair_nil
## X-squared = 421.52, df = 9, p-value < 2.2e-16
```

```
#Code for the plot "Trade type for Different Financial Knowledge Groups"
ggplot(trade_type_info,
       aes(x = client_inv_knowledge_lvl,
           fill = trd_type)) +
  geom_bar(position = "fill") +
```

```
  labs(y = "Proportion") +
  ggtitle("Trade type for Different Financial Knowledge Groups")
```

Proportion of the trade type in each group

```
##
##                      AB          AS          BP          SP
##   Excellent 0.526579739 0.447453472 0.016326758 0.009640031
##   Good      0.488126304 0.473174032 0.021291353 0.017408311
##   Fair      0.491153954 0.475732150 0.021336311 0.011777585
##   Nil       0.491766786 0.480750974 0.017317576 0.010164664
```

Average trade frequency in each group

Nil

```
## [1] 8.175298
```

Fair

```
## [1] 8.566591
```

Good

```
## [1] 9.112683
```

Excellent

```
## [1] 16.03753
```

Table representation of financial instrument type and choice by group

```
##
##              Bonds Call - Stock Option Certificate of Deposit Common Stock
##   Excellent      6                   9                     0         2610
##   Good          90                   5                     0        11718
##   Fair         140                   2                     1         7310
##   Nil           34                   0                     0         2858
##
##             Conv. Marginable Security Investment Certificate Mutual Funds
##   Excellent                         6                     0         5969
##   Good                            329                    17       117464
##   Fair                            186                     9       104548
##   Nil                             151                    10        48438
```

```
##
##             Option Eligible Preferred Stock Put - Stock Option Strip/Bonds
##   Excellent            6211               21                  0           0
##   Good                31782              818                  0           2
##   Fair                14576              356                  5           0
##   Nil                  4762              217                  0           0
##
##             Strip/Coupons Unit without Debt Warrants
##   Excellent             0                 6       13
##   Good                  2                 0        9
##   Fair                  3                 1        3
##   Nil                   0                 0        0
```

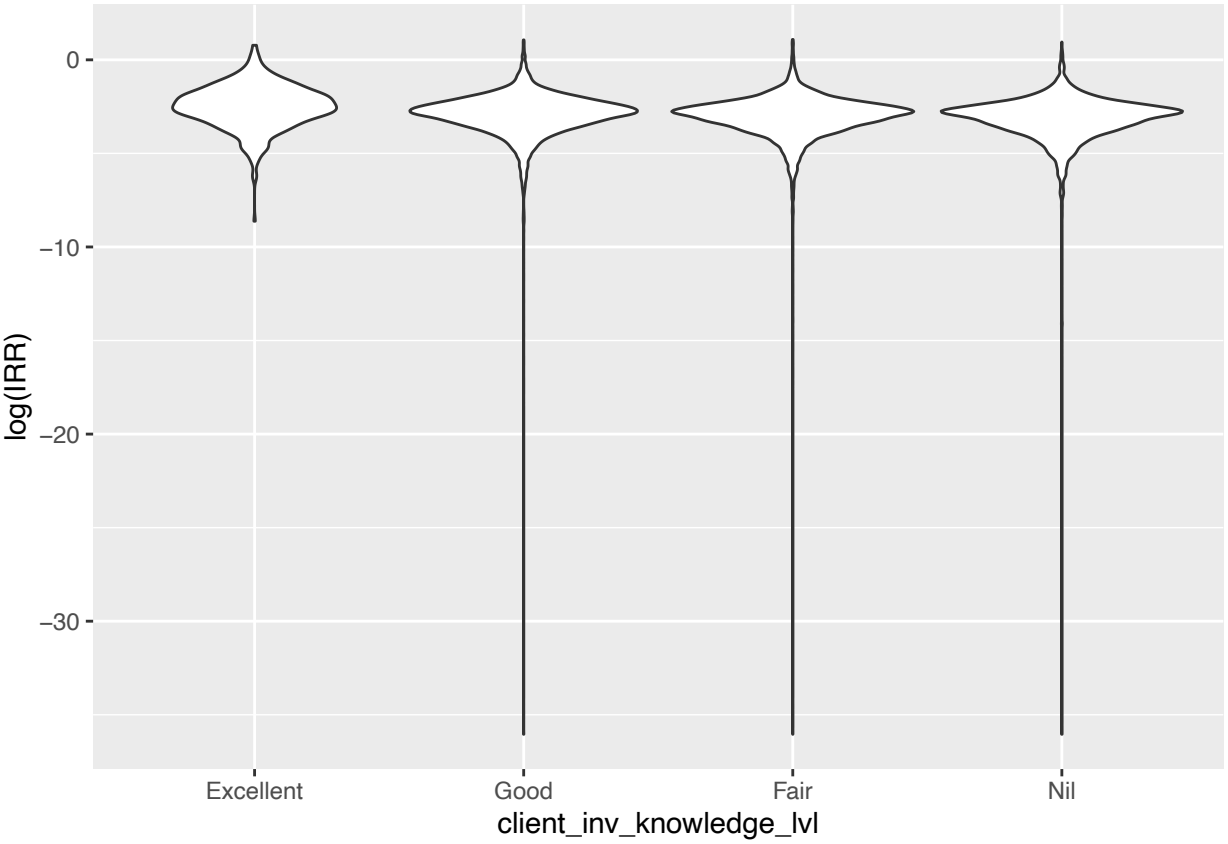
Proportion table

```
##
##                    Bonds Call - Stock Option Certificate of Deposit
##   Excellent 4.040132e-04        6.060198e-04           0.000000e+00
##   Good      5.547474e-04        3.081930e-05           0.000000e+00
##   Fair      1.101148e-03        1.573069e-05           7.865345e-06
##   Nil       6.020896e-04        0.000000e+00           0.000000e+00
##
##             Common Stock Conv. Marginable Security Investment Certificate
##   Excellent 1.757457e-01              4.040132e-04           0.000000e+00
##   Good      7.222811e-02              2.027910e-03           1.047856e-04
##   Fair      5.749567e-02              1.462954e-03           7.078811e-05
##   Nil       5.061094e-02              2.673986e-03           1.770852e-04
##
##             Mutual Funds Option Eligible Preferred Stock Put - Stock Option
##   Excellent 4.019258e-01    4.182210e-01    1.414046e-03       0.000000e+00
##   Good      7.240317e-01    1.958998e-01    5.042038e-03       0.000000e+00
##   Fair      8.223061e-01    1.146453e-01    2.800063e-03       3.932673e-05
##   Nil       8.577652e-01    8.432796e-02    3.842748e-03       0.000000e+00
##
##              Strip/Bonds Strip/Coupons Unit without Debt     Warrants
##   Excellent 0.000000e+00  0.000000e+00      4.040132e-04 8.753619e-04
##   Good      1.232772e-05  1.232772e-05      0.000000e+00 5.547474e-05
##   Fair      0.000000e+00  2.359604e-05      7.865345e-06 2.359604e-05
##   Nil       0.000000e+00  0.000000e+00      0.000000e+00 0.000000e+00
```

Divide into groups

For Section 3.4

Additional plot for analyzing the relationship between IRR and financial Knowledge level, violin plot with IRR take log



Details of the IRR mean and median for each fianncial knowledge group

```
#Nil
#there are 4776 clients in this group.
length(unique(nil_perform$client_id))
```

```
## [1] 4776
```

```
mean(nil_perform$IRR) #0.05100785
```

```
## [1] 0.05100785
```

```
median(nil_perform$IRR) #0.04243237
```

```
## [1] 0.04243237
```

```
sd(nil_perform$IRR)
```

```
## [1] 0.1204877
```

```
#Small difference between the mean and median.

#Fair
#there are 9479 clients in this group.
length(unique(fair_perform$client_id))
```

```
## [1] 9479
```

```
mean(fair_perform$IRR) #0.0605187
```

```
## [1] 0.0605187
```

```
median(fair_perform$IRR) #0.04722265
```

```
## [1] 0.04722265
```

```
sd(fair_perform$IRR)
```

```
## [1] 0.1393596
```

```
#larger difference between the mean and median.

#Good
#there are 11428 clients in this group.
length(unique(good_perform$client_id))
```

```
## [1] 11428
```

```
mean(good_perform$IRR) #0.06736891
```

```
## [1] 0.06736891
```

```
median(good_perform$IRR) #0.05120008
```

```
## [1] 0.05120008
```

```
sd(good_perform$IRR)
```

```
## [1] 0.1439304
```

```
#larger difference between the mean and median.

#Excellent
#there are 538 clients in this group.
length(unique(excellent_perform$client_id))
```

```
## [1] 538
```

```
mean(excellent_perform$IRR) #0.1167546
```

```
## [1] 0.1167546
```

```
median(excellent_perform$IRR) #0.06793897
```

```
## [1] 0.06793897
```

```
sd(excellent_perform$IRR)
```

```
## [1] 0.2310799
```

```
#very large difference between the mean and median.

#The mean return of investment generally increases as financial knowledge increases, but
```

Hypothesis test of the relationship between financial knowledge level and IRR

```
#Nest model
pchisq(100.5167,actual_return_mod$edf-actual_return_compare_mod$edf,lower=F)
```

```
## [1] 1.203353e-21
```

```
#Nest model test
#A small p-value can be seen,
#which leads to the conclusion that the predictor IRR is significant.
```

Prediction of grouping using IRR

```
## # weights:  12 (6 variable)
## initial  value 25444.046704
## iter  10 value 20518.401285
## final  value 20511.327252
## converged
```

```
##              test1$client_inv_knowledge_lvl
## pred_result_1 Excellent Good Fair  Nil
##     Excellent         0    0    0    0
##     Good            165 3431 2842 1412
##     Fair              0    4    0    0
##     Nil               0    5    5    3
```

Accuracy of prediction

```
## [1] 0.432058
```

Notice that it may be different from the table above, since the table was a random result of prediction, and it differs from time to time as the train and test set selected differs.

Balanced Accuracy

```
## [1] 0.4998705
```