Financial Data Science II Final Project

Group PGH013: Dhruv Baid, Guolun Li, Shangyu Li, Uday Sharma, Yi Xin Xiang

December 5, 2022

1 Introduction

1.1 Data Description

The data are obtained from the SEC website, as a part of their Form D datasets. It contains information on the Form D submissions filed by companies in their EDGAR document submissions.

The file contains information on new securities offered, with each row corresponding to a unique EDGAR submission. There are 14187 rows, where each row corresponds to a Form D submission by a company, where the company is looking to offer new securities. The columns include accession number, which is a unique identifier for each submission, and other characteristics of that submission. There are 41 columns including the accession number.

1.2 Data Cleaning

Out of the 41 columns, 26 of them have one or more missing values. In tackling these missing values, we use the following philosophy: we make a distinction between not knowing something versus knowing that something is not applicable. NaNs should represent "lack of data", or cases where the actual value is unknown. If something is not applicable, then we fill it in with "Not Applicable" instead, since the fact that something is not applicable is additional information in itself, and should be included in the analysis.

After looking at the actual Form D, we were able to fill in the values of 22 out of the 26 columns – either with "False" or "Not Applicable" based on whether the variable is boolean or categorical, respectively. For 3 columns – sale date, authorized representative, and number of non-accredited investors – we leave the NaN values as is, as they represent "lack of data". The final column, over 100 recipient flag, contains all missing values. As imputing it either way doesn't add any value to the analysis, we drop the column.

1.3 Data Preprocessing

Finally, we convert the columns into the respective data types in Python for appropriate analysis in the next steps.

We end up with 9 numerical, 16 boolean, 6 categorical, 1 datetime, and 8 text columns.

2 Analysis 1

2.1 Research Question

Question: "Are there variables in the data set that could help predict the percentage of offering sold?"

The methods used to answer the question are Natural Language Processing and K-Means Clustering.

2.2 Methodology

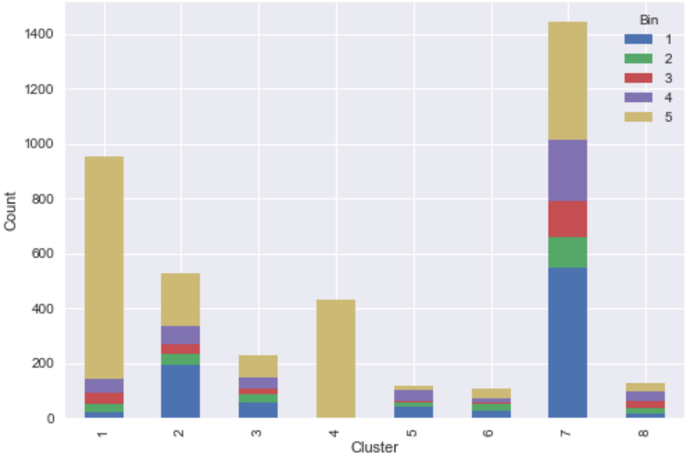
First, we set up the response variable y, equal to the percentage of offering sold, i.e., TOTALAMOUNTSOLD divided by TOTALOFFERINGAMOUNT. It is then divided into five bins – 0 to 0.2, 0.2 to 0.4, 0.4 to 0.6, and 0.8 to 1. Further analysis is done on these bins rather than the actual value of y.

For performing Natural Language Processing, we look at the following columns with text data: DESCRIPTIONOFOTHERTYPE, BUSCOMBCLARIFICATIONOFRESP, SALESAMTCLARIFICATIONOFRESP, FINDERFEECLARIFICATIONOFRESP, and GROSSPROCEEDSUSED_CLAROFRESP.

First, we tokenize each column to convert it to a form appropriate for NLP. Then, we use the term frequency-inverse document freqecy (TF-IDF) metric, which creates a matrix of how frequently each word shows up in each document, adjusted for its overall frequency. Now, the final step of the analysis is to use K-means clustering to group the data into 8 different clusters using this matrix. We are looking for clusters with members in y-bins we want. As we already know that the cluster contains members contain similar text, we can then identify certain texts which can help predict the percentage of offering that ends up getting sold.

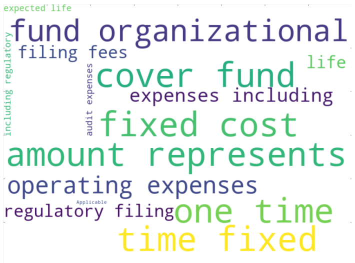
2.3 Results

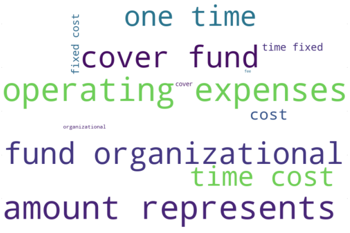
The plot for the 8 clusters we find is given below, where for each cluster, we also independently identify the y-bins the actual data belongs to.



2.3.1 Clusters 1 and 4

Here, percentage of offerings sold is high for a majority of the members, i.e., they belong to bin 5 (0.8 to 1). We find the rows have similar text in the GROSSPROCEEDSUSED_CLAROFRESP column. The word clouds generated for each cluster are given below (cluster 1 on the left, and cluster 4 on the right).





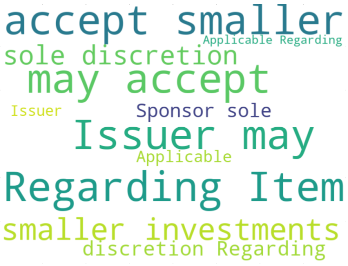
We notice a common feature to almost all the entries with 80-100% securities sold: the words “one-time cost” and “operating expenses”. As a proxy for the likelihood that the company does indeed sell most of its offerings, we look at the proportion of members of the two clusters that belong to bin 5, which is 89.63%.

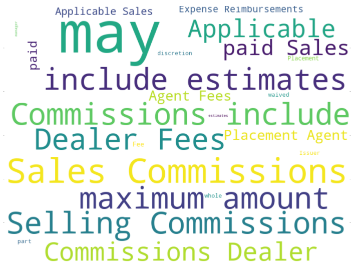
This can be a predictor for the percentage of securities sold – if a company makes a one-time payment to its promoters using the proceeds of its sales, or has plans to do so, and if it can be determined that it’s for the organizational and operating expenses, then it is 89.63% likely that the offering sells out completely. This could be attributed to the company being transparent and strong, as the payment is one-time and we can ascertain what it’s being put to use for.

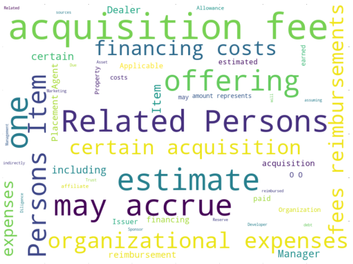
This way, we can use the text data we have to make a prediction about the percentage of total offerings that ends up getting sold. We employ a similar analysis for the following two clusters as well.

2.3.2 Cluster 5

Here, percentage of offerings sold is low for a majority of the members, i.e., they belong to bins 1 and 2 (0 to 0.4). We look at the word clouds for the SALESAMTCLARIFICATIONOFRESP (left), FINDERFEECLARIFICATIONOFRESP (middle), and GROSSPROCEEDSUSED_CLAROFRESP (right) columns.







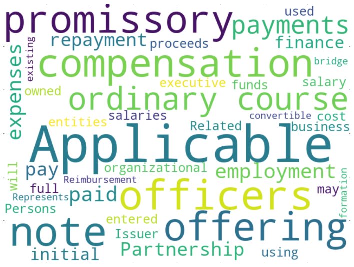
We make the following observations:

- For columns FINDERFEECLARIFICATIONOFRESP & GROSSPROCEEDSUSED_CLAROFRESP, the word “estimate” shows up frequently. If the company is vague about the amount of proceeds used to pay the promoters or the finders fee paid, and only provides as estimated value, then investors may worry that the money being raised isn’t being put to use in ways that are in their best interest.
- For column SALESAMTCLARIFICATIONOFRESP, words “accept smaller”, “discretion” and “smaller investments” show up frequently. If a company sometimes accepts investments below the minimum investment amount on the discretion of company management, it again brings into question the legitimacy of the offering and if there are some vested interests these investors have.

In both cases, it is 49.58% likely that the company only sells 0-20% of its offerings.

2.3.3 Cluster 8

Here, the percentage of offerings sold is medium for a majority of the members, i.e. they belong to bins 2, 3 or 4 (0.2 to 0.8). We look at the word cloud for column DESCRIPTIONOFOTHERTYPE.



The words “promissory”” and “note” show up frequently. We note that if the security being offered is a convertible promissory note or a SAFE (simple agreement for future equity), then it is 60.47% likely to sell between 20-80% of the amount being offered. This could be explained by the fact that start-ups, which usually offer these securities to early investors, can vary a lot in terms of quality and popularity.

3 Analysis 2

3.1 Research Question

Question: To what extent does `Investment Fund Type` influence the properties of the offering such as `Revenue Range`, `Total Offering Amount`?

Methods: we used Clustering (Agglomerative, K-Means, and Hierarchical) to answer this question.

3.2 Methodology

Our aim for this analysis was to come up with reasonable conclusions which could be supported by different clustering methods (i.e. would hold true independent of the methodology used). After the initial data pre-processing, we clustered the data using the three different methods mentioned above.

Agglomerative clustering involved computing the Gower distance matrix and passing this as the input to the function which would then organize the datapoints into clusters by minimizing the distances between each cluster's centroid and the points assigned to that cluster.

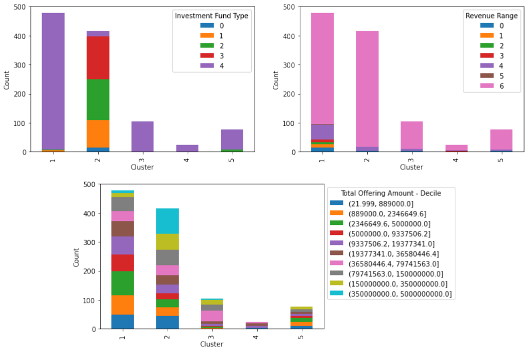
K-Means clustering involved passing 2 arguments to the function: a parameter $K = 10$, and the number of randomly chosen initial starting points the function would use to verify that it would output the set of K centroids/clusters which were closest in value to the ones which would theoretically minimize the sum of squared distances between each point and its corresponding cluster's centroid.

Finally, Hierarchical clustering involved grouping together sets of datapoints by their similarity, constructing a tree which would have fewer degrees of separation (in terms of the number of nodes/branches) between sets of points which were more similar to one another. After this, a level at which the tree would be 'cut' would be specified, separating the datapoints into clusters grouped together by similarity.

In all of the analysis (except for the initial Boxplot Analysis which we will discuss later), we discretized the `Total Offering Amount` by splitting it into quartiles, quintiles, and deciles. This was done so as to provide 3 levels of granularity when analyzing the possible effect the `Investment Fund Type` would have on `Total Offering Amount`. Note that `Revenue Range` was already discretized into levels, and so we did not need to change that.

3.3 Results

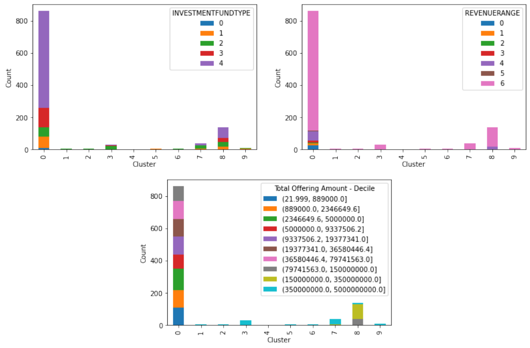
3.3.1 Agglomerative Clustering



Looking at the distribution of `Revenue Range` across clusters, we can find some interesting results. It is clear that Clusters 2, 3, and 5 are dominated by Revenue Ranges 4 and 6, whereas the lower end of the Revenue Ranges are found mainly in Cluster 1. From the distrbution of `Investment Fund Types`, we know that Cluster 2 contains most of `Investment Fund Types` 0, 1, 2, and 3, and the other Clusters are overwhelmingly comprised of `Investment Fund Type` 4. This seems to suggest that the first 4 `Investment Fund Types` lie at the higher end of the revenue range, whereas the last one is distributed across revenue ranges.

As for Total Offering Amount, Whether we divide it into quartiles, quintiles, or deciles, there does not appear to be any significant pattern across clusters. This could be taken as evidence that the `Investment Fund Type` - which has a clear distribution across clusters - does not influence the `Total Offering Amount` significantly, a conclusion which is supported by our preliminary boxplot analysis (see notebook).

3.3.2 K-Means Clustering

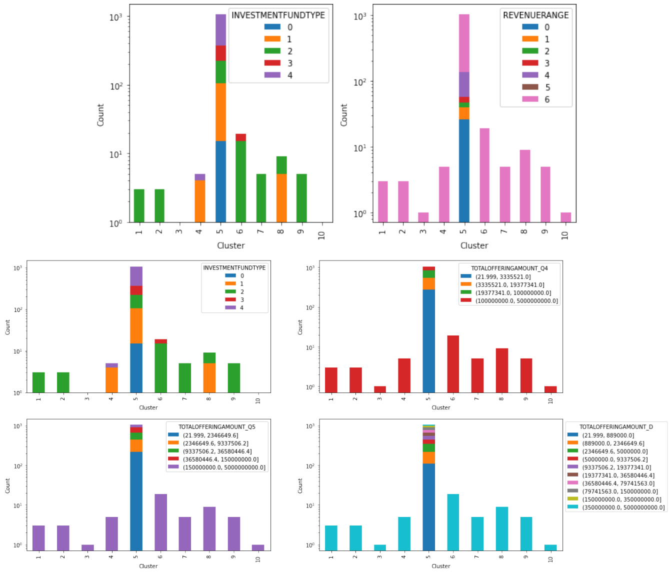


Clusters 3, 4, 6, 7, and 8 seem to be dominated by `Investment Fund Type` 2; Clusters 1 and 5 largely contain `Investment Fund Type` 4.

We can again see that Clusters which were comprised of a mix of Industry Group Types tended to also lie at the higher end of the Revenue Range. In particular, Clusters 0 and 5 support the analysis above - `Investment Fund Types` 0-4 lie at the higher end of the Revenue Range, whereas `Investment Fund Type` 5 is more evenly distributed.

Looking at the distribution of the deciles for Total Offering Amount, the highest decile is concentrated in Clusters 2, 3, and 8, and the second-highest decile is concentrated in Cluster 5. Comparing this to the distribution of `Investment Fund Types`, we can see a correspondence with Investment Group Type 2. Checking with the Agglomerate Clustering we performed earlier confirms this hypothesis as well. This suggests that Investment Group Type 2 lies at the highest end of `Total Offering Amount`. Aside from this, there is no obvious pattern which can be discerned.

3.3.3 Hierarchical Clustering



This section presents a very clear picture supporting the analysis performed in the previous two sections.

When we compare the distribution of `Investment Fund Types` across clusters to that of Revenue Ranges, it is immediately obvious that `Investment Fund Type` 2 accounts for the majority of `Revenue Range` 6 (as seen in Clusters 1, 2, 6, 7, and 9), and there is a mix across Revenue Ranges for the other Types (particularly as seen in Cluster 5).

When we look at the distribution of `Total Offering Amount` quintiles or deciles across clusters, a similar conclusion can be drawn: Clusters 1, 2, 6, 7, and 9 demonstrate that `Investment Fund Type` 2 accounts for a majority of the highest quintile/decile of `Total Offering Amount`, and the other Types are distributed more evenly.

3.3.4 Conclusion

From this segment, we can see how a seemingly simple tool like Clustering is able to quickly identify patterns in the data. This is useful, not only as a precursor to more refined analysis techniques but also as a way to summarize trends in the data. With regards to the question being asked, it appears that `Investment Fund Type` does play some role, albeit not a particularly significant one, in influencing such properties as `Revenue Range` and `Total Offering Amount`. Investment Fund Type 2 stands out from the other Types with respect to the two variables examined and is generally at the higher end of the range for both variables, potentially hinting at some underlying characteristic of this kind of Fund which might make it more attractive to certain client profiles.

4 Analysis 3

4.1 Research Question

Question: How does the relationship between the total amount sold and the total number of people who have already invested vary with the sale date?

The method involved is Nonparametric Regression.

4.2 Methodology

We perform log transformations on both `TOTALAMOUNTSOLD` and `TOTALNUMBERALREADYINVESTED` because both distributions are highly left skewed and ranges are large.

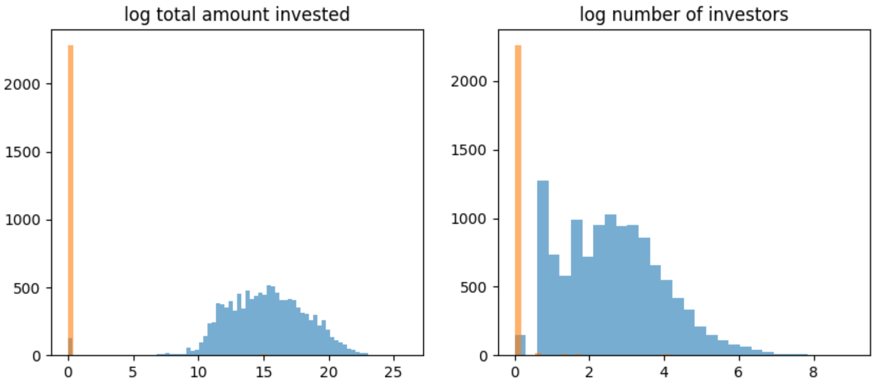
We divide `SALEDATE` into four bins - 1985 to 1994, 1995 to 2004, 2005 to 2014, 2015 to 2022 - to investigate how the relationship between `TOTALAMOUNTSOLD` and `TOTALNUMBERALREADYINVESTED` evolved over time. For each of the four time bins, we use locally weighted polynomial regression and tune the local span of data used for fitting.

In addition to an analysis of time series perspective, we also look at seasonal impacts on the relationship. We divide observations into four quarters according to the months of `SALEDATE`, and use nonparametric regression to fit the relationship of `TOTALAMOUNTSOLD` and `TOTALNUMBERALREADYINVESTED` in each quarter.

4.3 Analysis

4.3.1 Missing Value Consideration

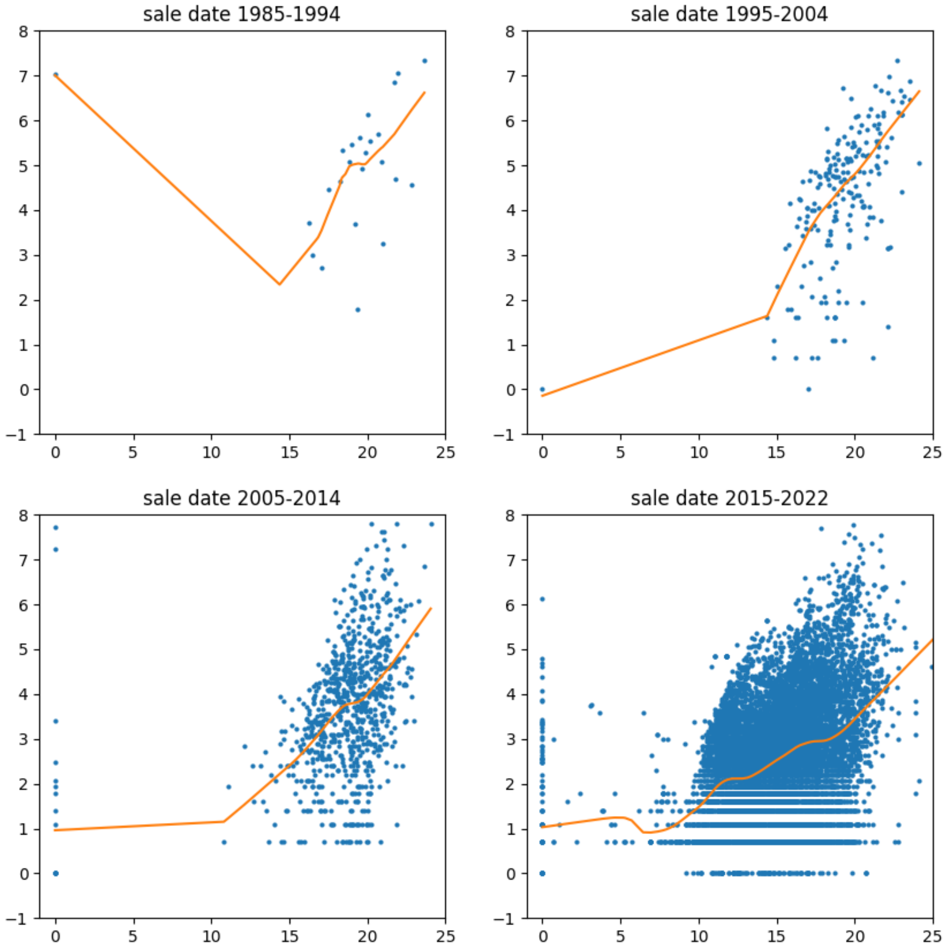
We create two sub-datasets differentiating on if sale dates are missings. The univariate distributions of two datasets look very different. Note in the missing sale date dataset, the majority of `TOTALAMOUNTSOLD` and `TOTALNUMBERALREADYINVESTED` are close to zero. We suggest these observations with variables values close to zero are special enough to be considered separately from the large population. Reasons are we could not find clear relationship between `TOTALAMOUNTSOLD` and `TOTALNUMBERALREADYINVESTED` when either value is fairly low. In the case of investor numbers are close to zero, the total amount sold can have large variation, and vice versa. Thus, missing values of sale dates will have less impact on the major focus of analysis.



4.3.2 Relationship Varying Over Time

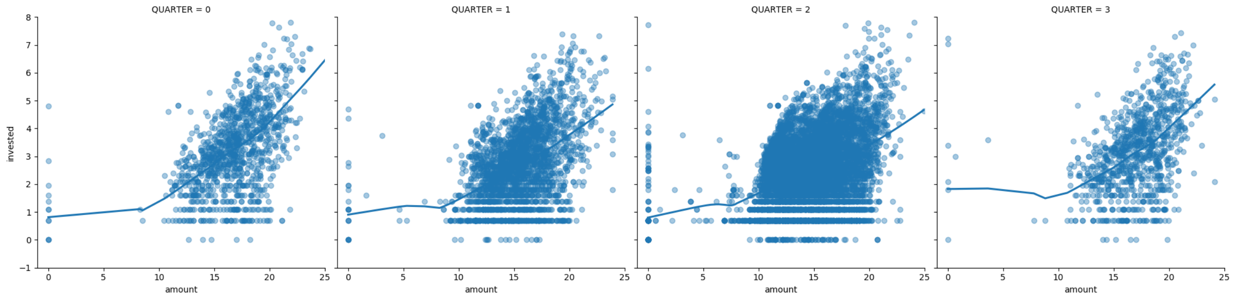
From the fitting line of nonparametric regression, there seems to be a positive correlation across time between the total amount sold and the total number of investors already invested.

The plots show increasing variance in more recent time periods. A greater number of small offering sales happened in 2015-2022. For a fixed number of investors, the range of total amount sold is growing wider from time period 2005-2014 to 2015-2022. Though the change is unclear for earlier times due to lack of data points.



Comparing the slope of regression curves across time, we do not see obvious differences in the first two decades. In 2005-2014, the curve is more flat compared to that of last ten years. The least steep curve is displayed in 2015-2022, suggesting an increase of amount sold of offerings may see less an increase in number of investors. Possible explanations may be the growing capital of investors over the time.

4.3.3 Seasonal Differences



Note the highest number of offerings set sale dates in the third quarter while the slope of the fitted curve is most flat. The first quarter has a noticeably steeper slope, which could mean less investors are willing to make a big amount of purchases in first quarter.

5 Analysis 4

5.1 Research Question

In this analysis, we conduct hypothesis testing to verify question 5: ”Is there a change in the minimum investment accepted over time, and does this vary with industry”?

5.2 Data Aggregation

We observed that the offering.csv data in HW1 refers only to the submissions in 2022 Quarter 3(2022 Q3). To get an accurate analysis result, we downloaded all data from 2016Q4 to 2022Q3 from SEC's website. We preserve only the observations with 'sale date' after 2017-01-01.

5.3 Data Preprocessing

The question concerns with only 3 variables; 'minimum investment accepted' ('min invest'), 'industry group type' and 'sale date'. Here 'min invest' is the response variable. The distribution of 'min invest' is too skewed and uneven, even after transformations, making any parametric distribution a bad fit for it. For the purpose of hypothesis testing, we discretized 'min invest' into 7 bins based on 6 thresholds: 0, 1000, 10000, 25000, 50000, and 150000. These numbers were selected based on the quantiles of 'min invest'.

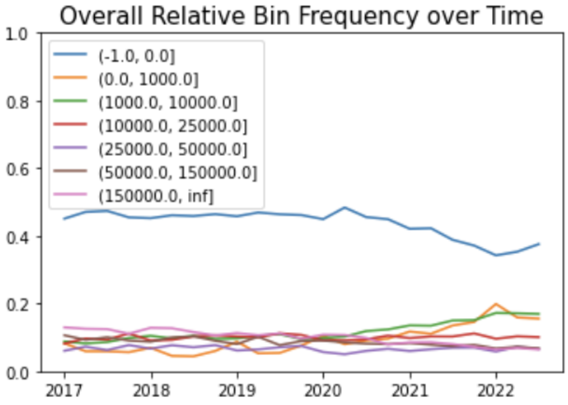
For a given bin B and a given industry I, each time period t (here t is measured in quarters, not days), we define

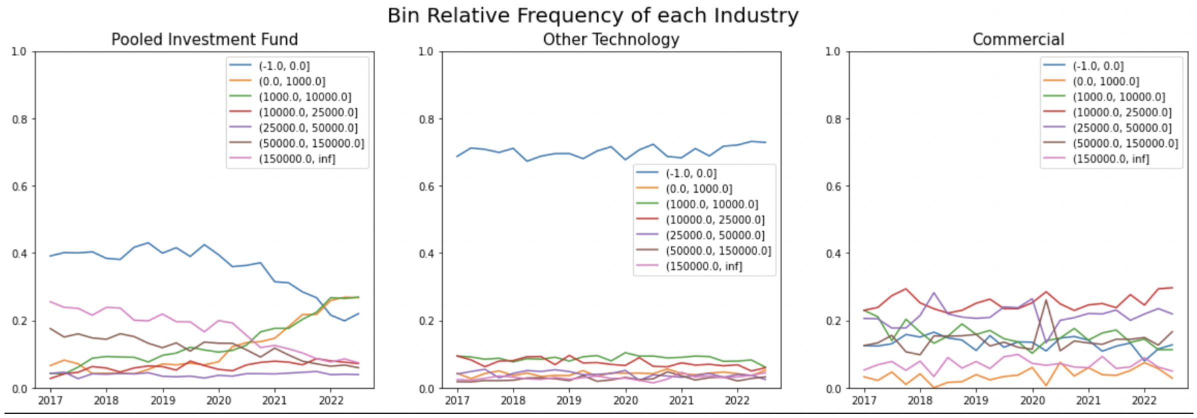
$$\begin{cases} p_B^t = \frac{\text{total number of observations within B in this quarter}}{\text{total number of observations in this quarter } t} \\ p_{B,I}^t = \frac{\text{total number of observations within bin B and industry I in this quarter}}{\text{total number of observations within industry I in this quarter} t} \end{cases}$$

These two variables respectively represent the overall and per-industry relative frequency of bin B. For example, if I indicates the Commercial industry, and $B = (0, 1000]$, then $p_{B,I}^3 = 0.125$ means that for all the observations in the commercial industry whose sale date lies in Q3 2017, 12.5% of them have a minimum accepted investment amount between 0 and 1000. Notice $t = 3$ corresponds to Q3 2017 because we started at Q1 2017. To address question 5, we will analyze how p_B^t and $p_{B,I}^t$ change over time. Note that we analyzed the trend of each bin B independently.

5.4 Trend Plots

We plot the trend of each p_B^t and $p_{B,I}^t$ over time. The time series trend vary greatly among bins and industries.





5.5 Hypothesis Testing

We conducted the likelihood ratio (LLR) test to test if the relative bin frequency in any quarter is different from the other quarters', i.e. if the relative bin frequency changes over time. We also conducted the Mann-Kendall(MK) test to determine if there is an increasing or decreasing trend in the bin ratio series. For the LLR test, the test result will either be 'changing trend', or 'no trend' (not changing). For MK test, the test result will be one of the following; 'increasing trend', 'decreasing trend', 'no trend', or 'inconclusive'. Notice that the MK test might be inconclusive because the independence assumption might be violated.

For the technical details of these two tests and mathematical derivation of LLR test, see the notebook.

5.6 Test Result - Overall

Below we list the test results of p_B^t for each bin for LLR and MK tests. It can be seen that, for 5 out of 7 bins, the test results of LLR and MK match. That is, when LLR test indicates a changing trend, MK test indicates either increasing or decreasing trend. However, for the middle two bins ($(10000, 25000]$ and $(25000, 50000]$), the test results differ.

Test	$(-1, 0]$	$(0, 1000]$	$(1000, 10000]$	$(10000, 25000]$	$(25000, 50000]$	$(50000, 150000]$	$(150000, \infty)$
LLR test	changing	changing	changing	changing	changing	changing	changing
MK test	decreasing	increasing	increasing	no trend	no trend	decreasing	decreasing

5.7 Test Result - Industry Effect

In total, there are 35 industries. Due to the 2 page limit here, we choose to analyze the 3 biggest industries; Pooled Investment Fund, Other Technology and Commercial. Moreover, we will be looking only at the relative frequency of the first bin i.e. observations with 'min invest amount' of 0. This correspond to the blue lines in above plot.

Test	Pooled Investment Fund	Other Technology	Commercial
LLR	Changing	No Trend	No Trend
MK	Decreasing	No Trend	No Trend

5.8 Conclusion

In conclusion, we represent the overall distribution of 'minimum accepted investment amount' using the relative frequency distribution within each bin. We analyze the change in relative frequency from 3 aspects; plots, LLR test, and MK test. The test results of the MK and LLR tests agree most of the time. They also align with the plots.

Based on the LLR and MK tests, we can say that the overall relative frequency changes over time for 5 out of 7 bins. We then conduct the same analysis separately for 3 industries, but only on the first and largest bin. Based on LLR and MK test, we see that the 'Pooled Investment Fund' industry exhibits a decreasing trend, whereas 'Other Technology' and 'Commercial' industries stays constant over time. This means that the proportion of 'zero minimum accepted investment' offerings decreased over time in the 'Pooled Investment Fund' industry, but remained constant for 'Other Technology' and 'Commercial' industries.