Group Project Report

Value Proposition

How can a venture capitalist improve the accuracy of their investments? Our aim is to predict the potential success of a startup founder, defined as raising the minimum of a series B funding.

Data Acquisition

Crunchbase provides access to Crunchbase Data for approved research and news organizations. With Crunchbase Data, one can incorporate the latest industry trends, investment insights, and rich company data right into their applications.

We acquired the data from GitHub and other online resources. We had access to the following Crunchbase datasets :

- Investments
- Founders
- Companies

We then summarised the above datasets and added features.

Data Manipulation

1. **Extraction of data on education**

The data on the education institute attended by a founder was extracted in two ways.
Firstly, directly by merging on founder name from the crunchbase website.
Secondly, we also extracted the data for whether or not the founder attended an ivy league college and their degree from the 'biography' of the founder by searching the biography text for keywords like 'Harvard University' or PhD.

For example, '*Sergey Sundukovskiy, Ph.D. has over 20 years of experience serving*…' would give a 1 value for Ph.D.

2. **Extraction of data on headquarter location**

The data on headquarter location was extremely drilled down in the raw format of *'City, State, Country'.* Moreover, for some startups only data on '*City Country'* was available. We extracted the data for different cities, states and countries and visualized it to find hotspots for investments

Feature Engineering

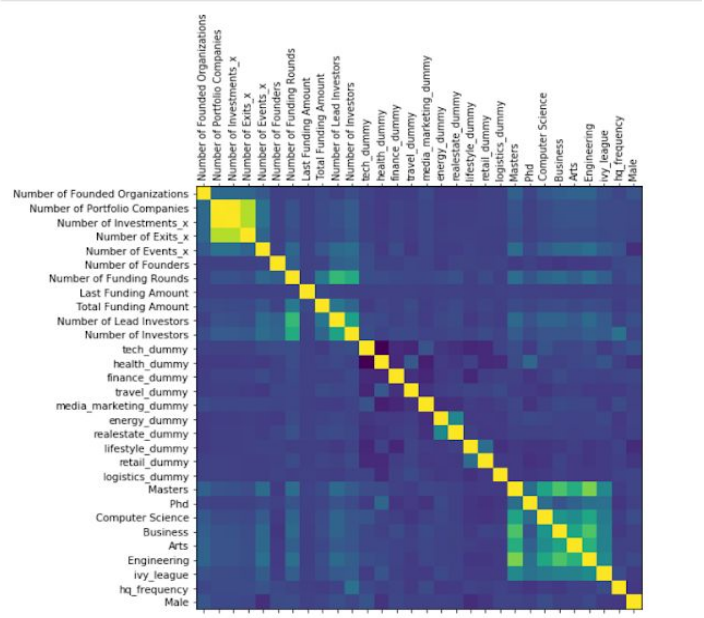
Creation of undergrad dummy, phd dummy, masters dummy: We created dummy variables that took on the value of one depending on the founders education level.

Creation of ivy league dummy: We created an *'ivy_league_dummy'* that took the value of one if the education of the start-up founder included any of the 8 ivy league colleges, MIT, Stanford or Caltech.

Creation of headquarters feature: We replaced the headquarter column which was in the format of *'City, State, Country'* with the frequency of occurrence of that location. For instance, if the headquarter location San Francisco, California, United States occurred 6348 times in our data, it was replaced by 6348.

Creation of industry dummy: The data on industry had several categories for each start up. We clubbed together the the different keywords from industry column and manually sorted them into 11 major categories namely and created dummies on those columns. The new categories were Tech, Finance, Healthcare, Logistics, Real Estate, Retail, Lifestyle and Fashion, Travel and Hospitality, Energy/Environment, Media and Marketing and Other. For example, industrys like '*Artificial Intelligence*' or '*Flash Storage*' were sorted into '*Tech*' and given 1 in the '*tech_dummy*' column.

Dropping of highly correlated features from correlation matrix:



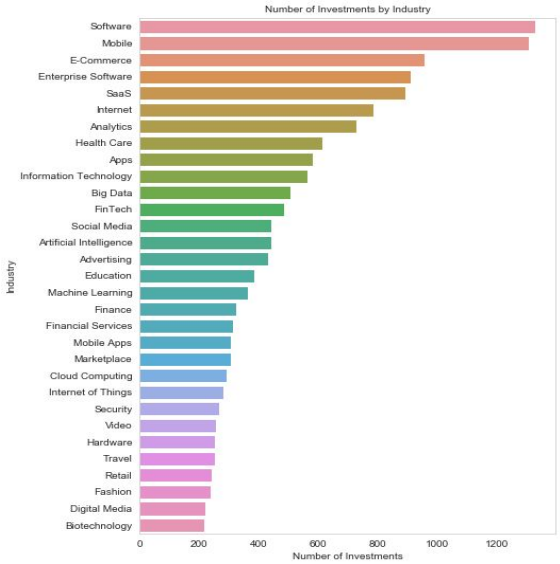
We dropped:

1. Number of lead investments

2. Number of partner investments
3. Last equity funding amount
4. Total equity funding amount
5. IPO status
6. other dummy
7. bachelor dummy

Data Visualisation

Fig 1. Number of Investments by Industry: The start-up space represents more than 520 different markets. However, tech industries like 'software', 'mobile' etc dominate all others.



The next two images compare the number of investments per round code with the amount invested in those rounds. The average value per investment increases at each stage (as we would imagine).

Fig2 : Total Investments per round type

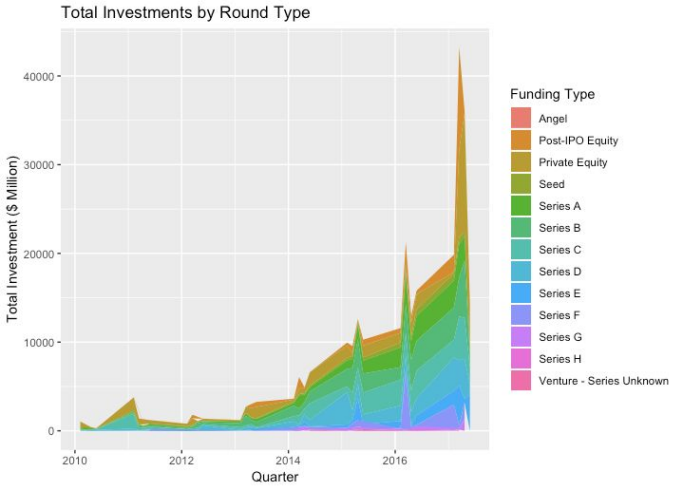


Fig3 : Number of investments per round type

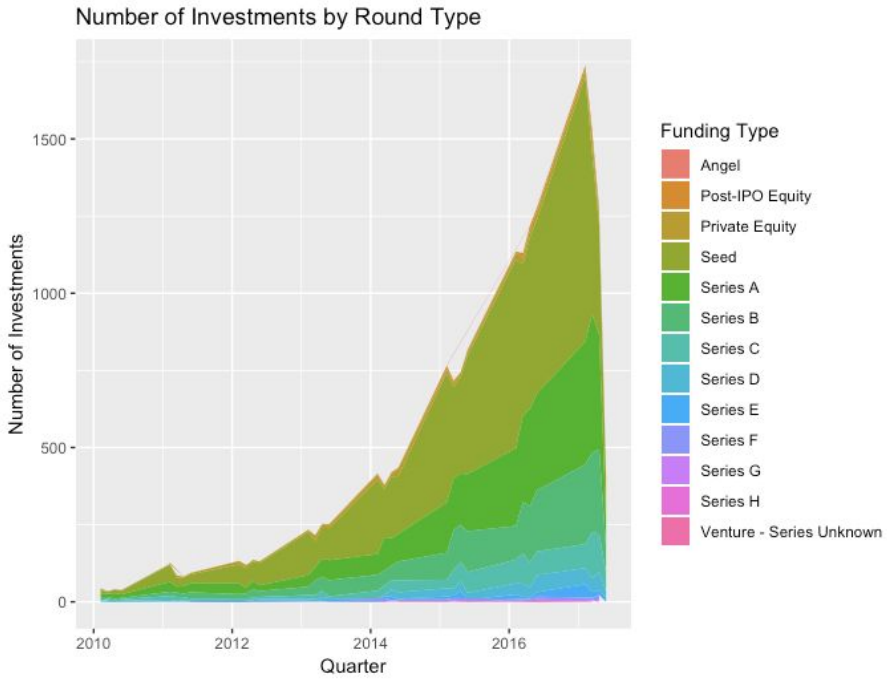


Fig 4: Frequency of investment by location.

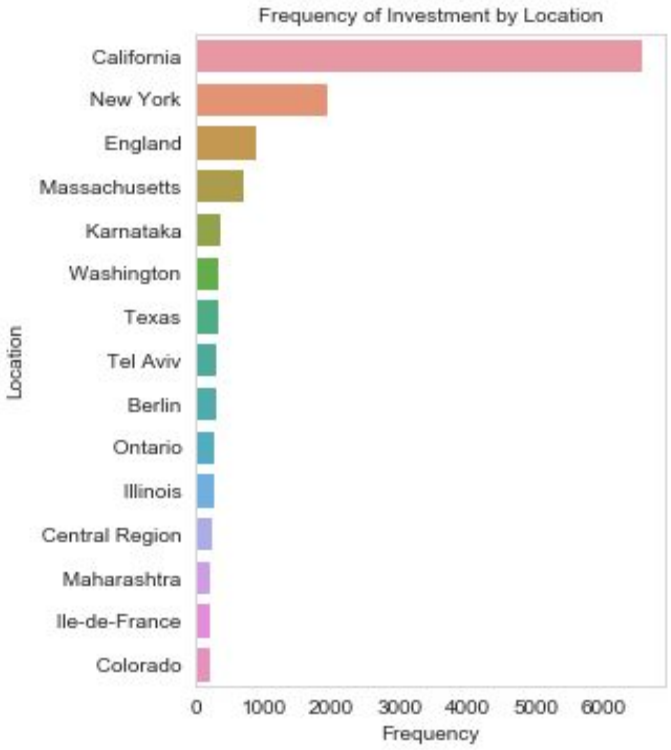
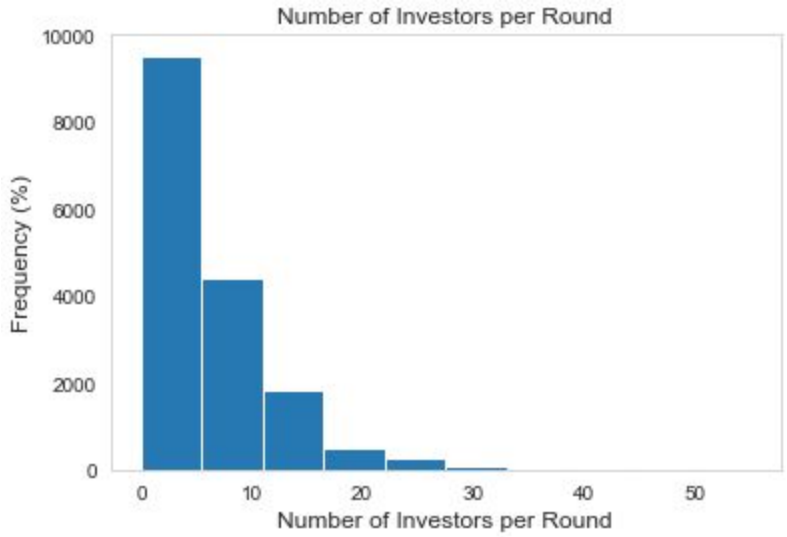


Fig 5: Number of investors per round

Syndication (or the act of co-investing in a company with other partners) plays an important role in the VC industry. In fact, almost 70% of the investments had at least two partners.

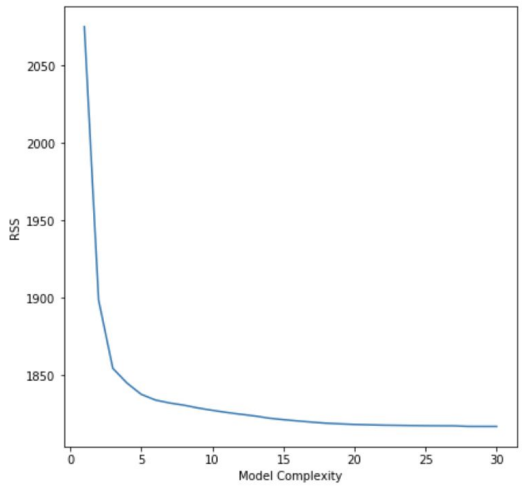


Modelling

1. **<u>Step 1:</u> perform forward stepwise best subset selection as baseline model:** We ran forward stepwise selection on all selected features, and plotted the minimum RSS of each model containing (1… 30) number of features. We chose n = 10, which has a reasonably low RSS. This way, we prevent selecting too many predictors, which can lead to overfitting. The final features chosen are:

```
-----------------
Foward Selection:
-----------------
Number of Lead Investors
Number of Funding Rounds
Masters
hq_frequency
ivy_league
Computer Science
Male
Number of Exits_x
Number of Founded Organizations
Business
```

   We graph the RSS on the y-axis against model complexity (the number of predictors of the model) on the x-axis.

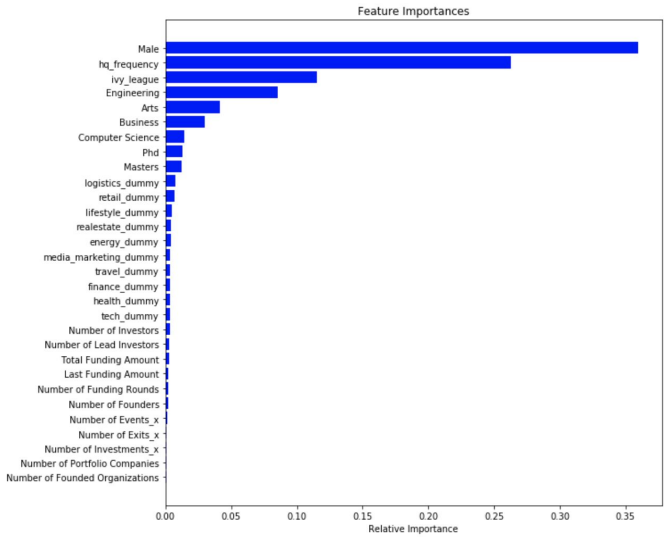


Using forward stepwise best subset selection, we selected the best linear model as our baseline linear model. This model yields an accuracy of 82.44% and MSE of 0.176. As a result, our objective is to produce a final model that beats these metrics.

2. **<u>Step 2:</u> Find the optimal hyperparameters for 4 candidate models.** We ran grid search and randomized grid search on the training dataset to find the optimal hyperparameters for our chosen candidate models.

   Random Forest We divided the model into 75-25% train test and ran random forest on it, using all the features. We performed a Grid Search Cross Validation on the model to find the parameter set that yields the highest accuracy score. The best parameters are as follows:

   - Max_depth = 15
   - Min_samples_leaf = 4
   - Min_samples_split = 8,
   - Number of estimators = 50



XGBoost: XGBoost is an implementation of gradient boosted decision trees designed for speed and performance.

- Colsample_bytree = 0.9063511098987536
- Gamma = 0.4738461983918303
- Learning rate = 0.29617277584129054
- Max_depth = 5
- Number of estimators = 143
- Subsample = 0.6781136557857271

SVC:

A Support Vector Classifier (SVC) is a discriminative classifier formally defined by a separating hyperplane. Given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples.

We used the cross validation grid search to find out the hyper parameters. The best kernel coefficient is 0.1 and best penalty parameter C is 10.

The accuracy of our SVC on training set is 84.96%

Logistic Regression: We also ran a logistic regression to predict success or failure. The most important features of the models are as follows.

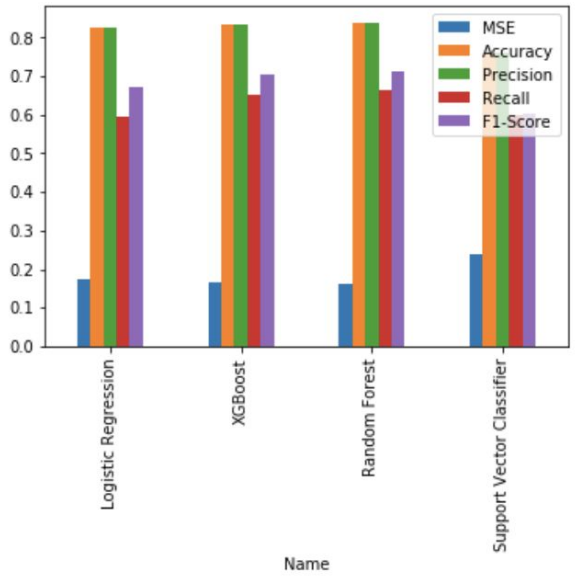
```
{'Arts': -0.044750991111363735,
 'Business': 0.193135911777947,
 'Computer Science': 0.22338590451682291,
 'Engineering': 0.1414911000626613,
 'Last Funding Amount': 0.47074045579130763,
 'Male': 0.28353125444912664,
 'Masters': 0.30308613441550253,
 'Number of Events_x': 1.2322971346269047,
 'Number of Exits_x': 1.1712374520918065,
 'Number of Founded Organizations': -1.0100408934357101,
 'Number of Founders': -0.26392842652377024,
 'Number of Funding Rounds': 9.748453889688799,
 'Number of Investments_x': 1.1299853714648207,
 'Number of Investors': -0.3119047183547719,
 'Number of Lead Investors': 12.814554101442019,
 'Number of Portfolio Companies': 1.282213141367116,
 'Phd': 0.20004151462538092,
 'Total Funding Amount': 3.8085185457044757,
 'energy_dummy': -0.06512719840479281,
 'finance_dummy': -0.22554820048061272,
 'health_dummy': -0.06084326547791157,
 'hq_frequency': -0.5924665568756124,
 'ivy_league': 0.3789985429342101,
 'lifestyle_dummy': -0.16866793423422588,
 'logistics_dummy': -0.13464686360957157,
 'media_marketing_dummy': 0.15813905725784352,
 'realestate_dummy': 0.032750652414070866,
 'retail_dummy': 0.19024572811330098,
 'tech_dummy': -0.028289165947819717,
 'travel_dummy': -0.0817171081792974}
```

Business, Computer Science and Engineering have a positive coefficient whereas Arts has a negative coefficient showing founders having CS, Eng, Business degrees succeed

All the results are intuitive, except for tech_dummy, number of investors and Headquarter frequency which have a negative coefficient.

3. **<u>Step 3:</u> K-Fold Cross Validation:** We calculated the performance metrics (including MSE, accuracy, precision, recall and F1-score) for all the models using 10-fold cross-validation with the optimal parameters that we got from Step 2. The table below shows the performance metrics of each of the 4 models used.

	Name	MSE	Accuracy	Precision	Recall	F1-Score
0	Logistic Regression	0.172907	0.827093	0.827093	0.597859	0.676666
1	XGBoost	0.075755	0.924245	0.924245	0.857003	0.872584
2	Random Forest	0.085875	0.914125	0.914125	0.822116	0.852860
3	Support Vector Classifier	0.169293	0.830707	0.830707	0.617855	0.688341



Evaluation of Models

Accuracy of the baseline model is 82.44% and MSE is 17.56%

Best choice: XGBoost and Random Forest

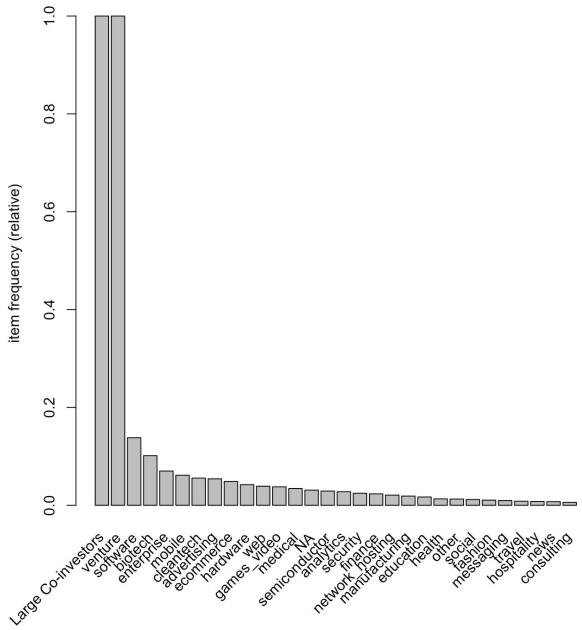
```
Accuracy of Test Set using XGBoost is 92.65%
MSE of Test Set using XGBoost is 7.35%
```

```
Accuracy of Test Set using Random Forest is 90.89%
MSE of Test Set using Random Forest is 8.85%
```

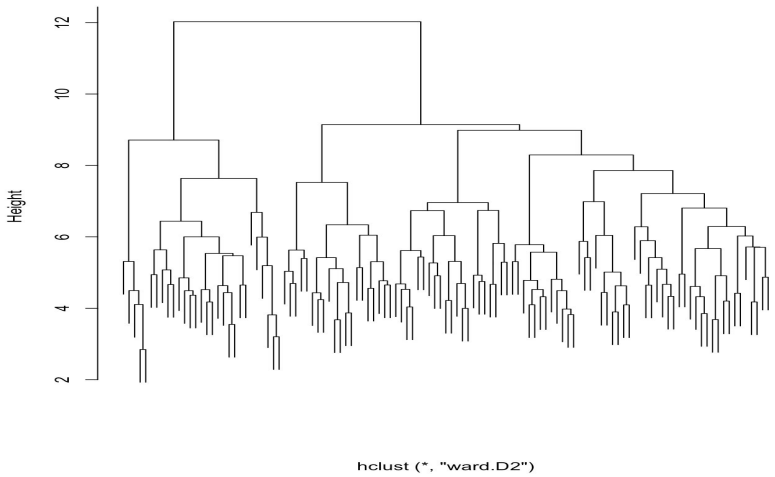
While Random Forest is easier to run and gives feature importance, XGBoost is more accurate

Market Basket Analysis

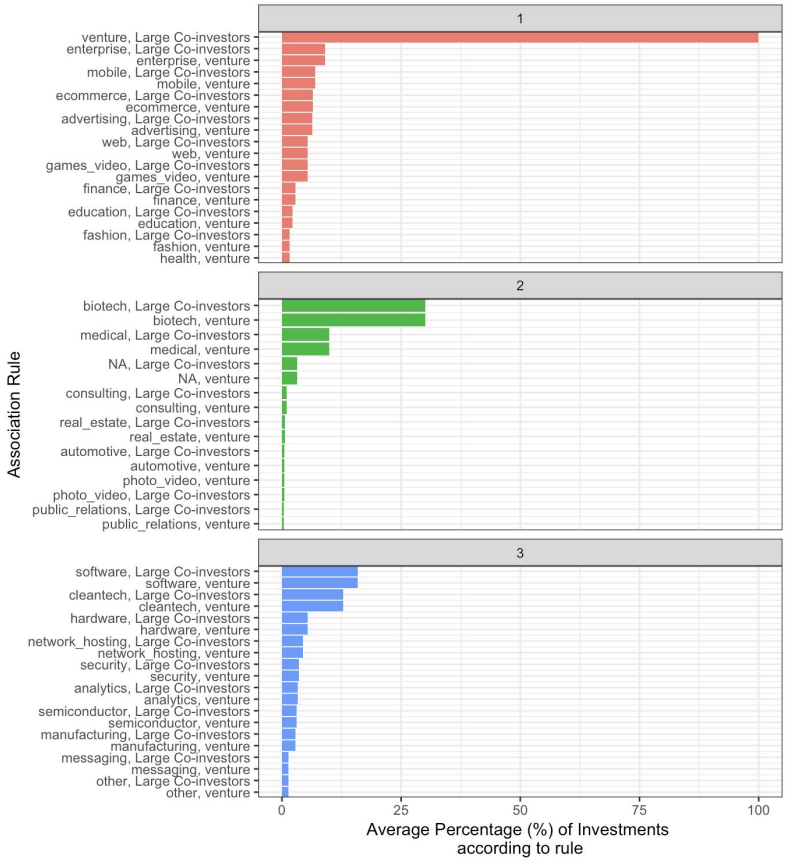
An algorithm was used to detect association rules (which features occurred together for a successful startup) and get feature likelihood, to understand to what extent the factors occur simultaneously for a VC portfolio. For this, 'arules' library was used and the model got 231 rules. Duplicates were removed and finally there were 81 rules.



Then visualisation of the rules used by VCs was conducted. Clustering on PCA components (30) was done, and three clusters were used. The clusters can be identified in the dendrogram below:

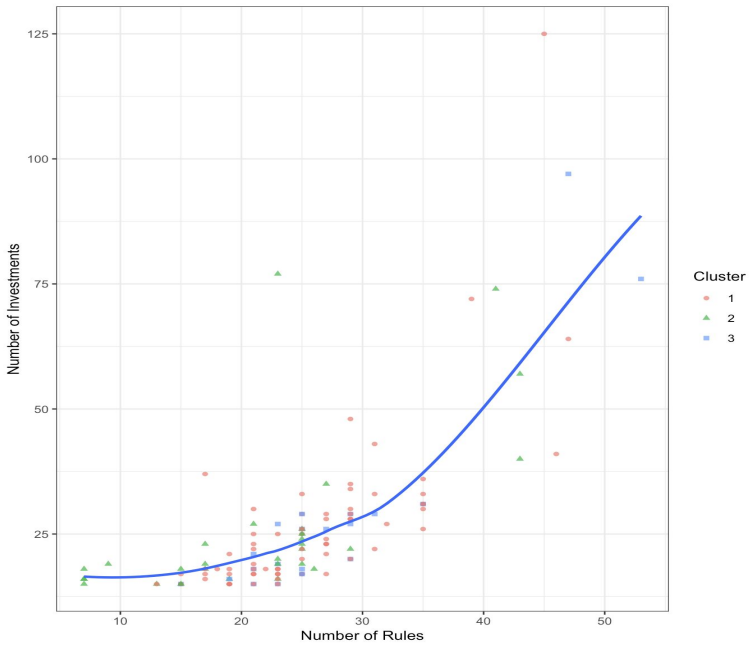


To understand how the clusters are unique, we checked for each rule which clusters have the highest average percentage of use (used two-sample test of proportions to filter only the rules that are significantly more preponderant in each cluster)



Interpretation

1. From market basket analysis, the largest VC firms in terms of the number of investments used the highest number of different rules which suggests portfolio diversification
2. Most of the largest VC firms were classified in cluster 3, which provides insight into the role of these firms in financing late stage rounds and software companies.




```
In [534]: df_main = df.drop(['Full_Name', 'Primary Job Title', 'Biography', 'Categories', 'Headquarters Location',
                            'Number of Lead Investments_x', 'Number of Partner Investments', 'Last Equity Funding Amount',
                             'Total Equity Funding Amount', 'IPO Status', 'other_dummy', 'Bachelors',
                           'Operating Status','Founded Date', 'Closed Date', 'Company Type', 'Funding Status', 'Last Funding Type',
                            'Last Equity Funding Type','Number of News Articles', 'Number of Employees', 'Acquisition Status',
                           'Last Funding Date', 'IPO Date', 'Number of Events_y'], axis = 1)
```

Following are the features we have removed due to high correlation or irrelevance to the target :

- 'Full_Name'
- 'Primary Job Title'
- 'Biography'
- 'Categories'
- 'Headquarters Location',
- 'Number of Lead Investments_x'
- 'Number of Partner Investments'
- 'Last Equity Funding Amount'
- 'Total Equity Funding Amount'
- 'IPO Status'
- 'Other_dummy'

- 'Bachelors'
- 'Operating Status'
- 'Founded Date'
- 'Closed Date'
- 'Company Type'
- 'Funding Status'
- 'Last Funding Type'
- 'Last Equity Funding Type'
- 'Number of News Articles'
- 'Number of Employees'
- 'Acquisition Status',
- 'Last Funding Date'
- 'IPO Date'
- 'Number of Events_y'

1. Even though the number of founder with business degrees is higher, we were able to deduce that the correlation between a founder succeeding if he has a computer science background is higher with good confidence. This can be due to the growing digital/software/data related startups

2. There is a good chance of a founder succeeding if his alma mater is an Ivy-League contrary to the popular belief of being a dropout. We hypothesize this can be true given the great alum network or maybe just the nature of the network a founder is exposed to in an Ivy league

3. Number of founded organisations is negatively correlated which can be attributed to distraction

4. More number of founders is negatively correlated to success. This is consistent with the idea: too many views (people) result into poor decisions

5. Startups with highest number of investments are focussed at portfolio diversification as observed in our market basket analysis

6. Most of the largest VC firms finance late stage rounds and tech companies. So more and more tech startups are blooming and most new start-ups are in the tech category