APPLIED ANALYTICS FRAMEWORKS AND METHODS - II

PROJECT

BUILT TO LAST

Guided By:

Vishal Lala

Group Members:

Kahsavyah Buckley (kbb2129)
Hiral Raval (hr2483)
Aimee Tran(tdt2118)
Romaine Knight (rk3055)

STATEMENT OF PROBLEM

As mentioned earlier, the dataset is about the investment series for startups with some demographical information about the startups. Following are the few questions we are seeking answers/ explanations for:

i.   Which are the top markets gaining maximum investments?

ii.   Startups of which origin receive maximum funding?

iii.   Which are the states of the top country with maximum funding to startups?

iv.   For operating companies, what are the categories that have raised the most funding on average?

v.   Since the dataset contains a lot of variables, can the variables be factored together to ease out further analysis?

ANALYSIS TECHNIQUE

It is believed that predictions from smaller and simpler models are more stable across the samples. Also, in a business set up, simpler models with fewer variables are easier to interpret and communicate to stakeholders.

With increase in number of predictors, the chance of finding correlations among a set of predictors increase. This multicollinearity inflates standard errors of coefficient, which leads to fallacious conclusions.

Our dataset has 39 variables and we intend to find out multicollinearity in this dataset.

Thereby, our purpose was to capture the variance in these large number of variables in a smaller set to ease out the future analysis process. Hence, we chose *'dimension reduction technique'* to achieve this.

The steps in performing dimension reduction are:

    i.   Prepare the data

    ii.   Check suitability for performing this analysis

    iii.   Determine number of factors/components

    iv.   Map variables onto factors/ Describing Components

There are two dimension reduction techniques:

    i.   Factor Analysis

    ii.   Principal Component Analysis

We started with the implementation of *Exploratory Factor Analysis*. Factor Analysis makes use of *latent variable* – which is difficult to measure in real life with a single variable. Instead, it is seen through the relationships it causes in a set of variables. Furthermore, it decomposed *shared variance* in the dataset.

However, on implementing various usages of exploratory factor analysis using minimum residual, a warning was obtained. The warning read as

"*The estimated weights for the factor scores are probably incorrect. Try a different factor score estimation method.*"

Hence, we proceeded with *Principal Component Analysis*. Principal Components Analysis' approach is to create one or more index variables from a larger set of measured variables. It does this using linear combination (basically a weighted average) of a set of variables. These created index variables are called *components*. This approach decomposes *total variance* of the variables.
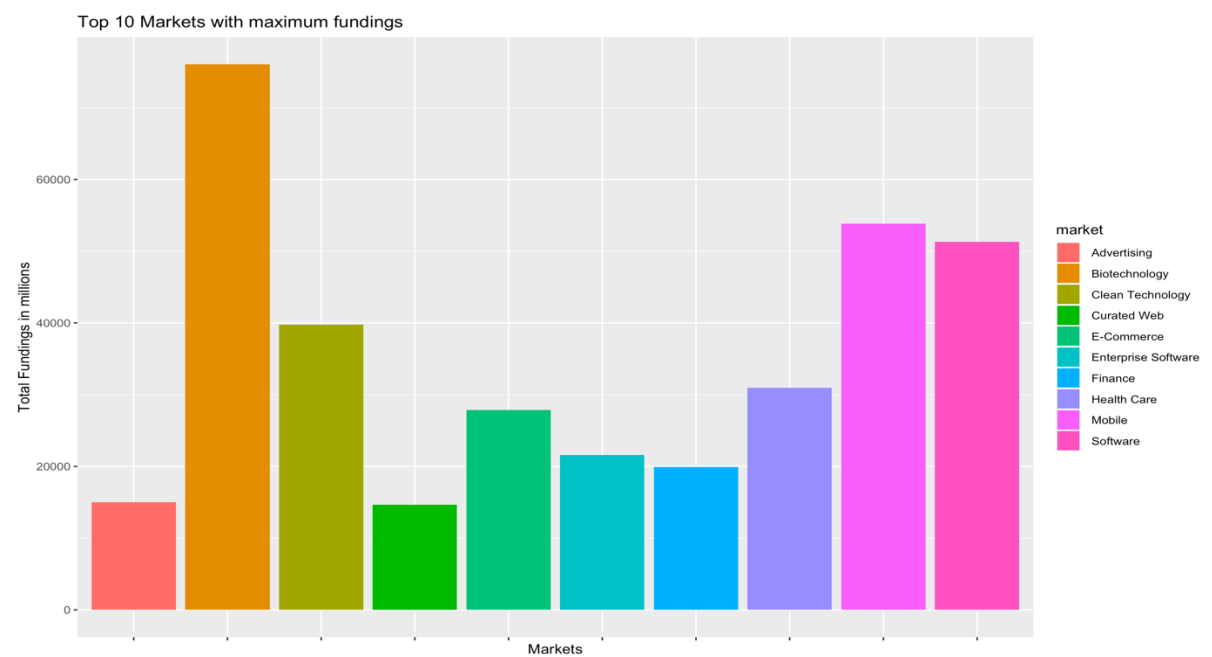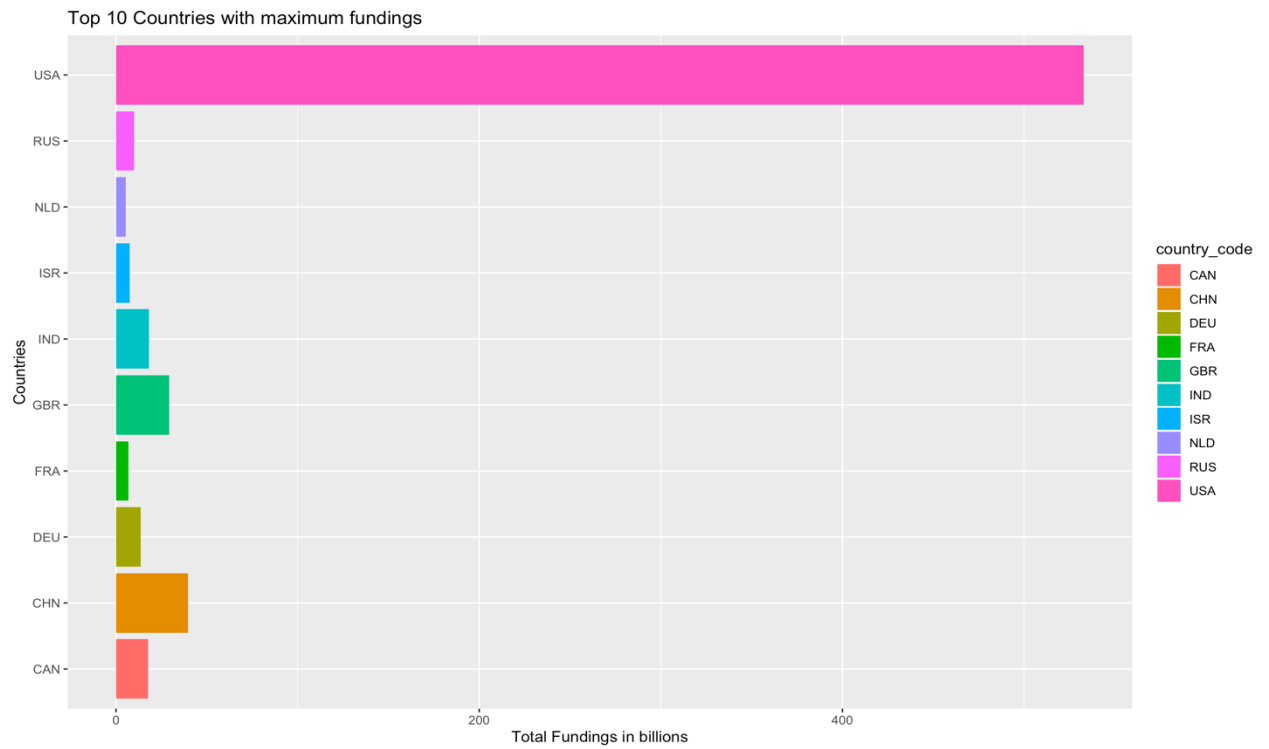
RESULTS OF ANALYSIS

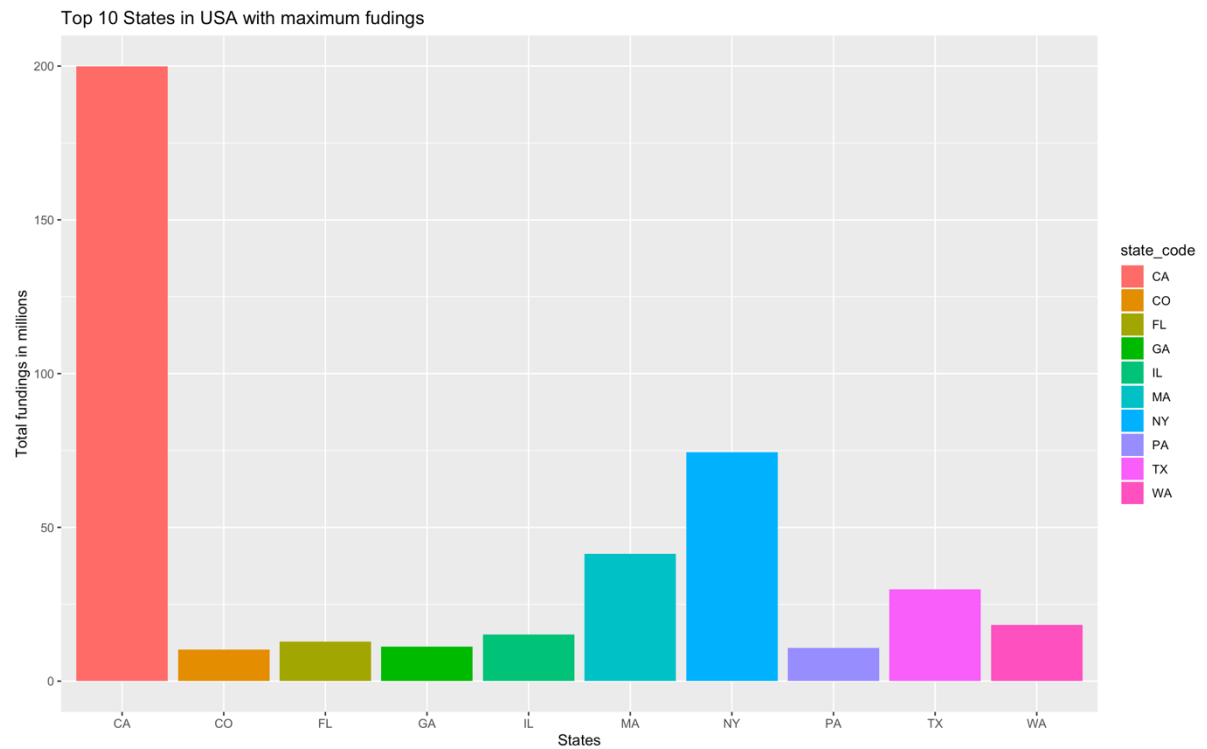## 1. Preliminary Analysis

I. Founded Year Analysis

Year Founded
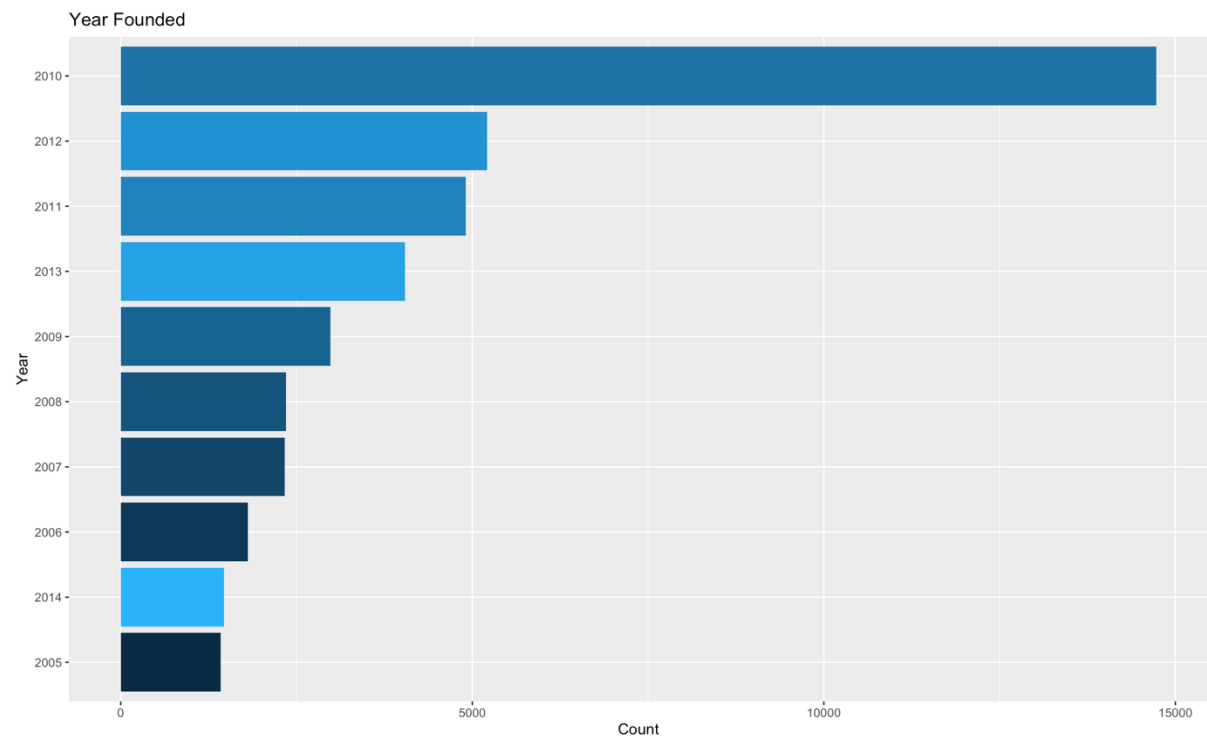
II. Market Leaders with Maximum Funding

Top 10 Markets with maximum fundings

III.   Countries that observed maximum funding

Top 10 Countries with maximum fundings

IV.   Top 10 USA states that observed maximum funding

Top 10 States in USA with maximum fudings

V.   Top markets of '*operating*' companies

Top 10 markets of operating startups
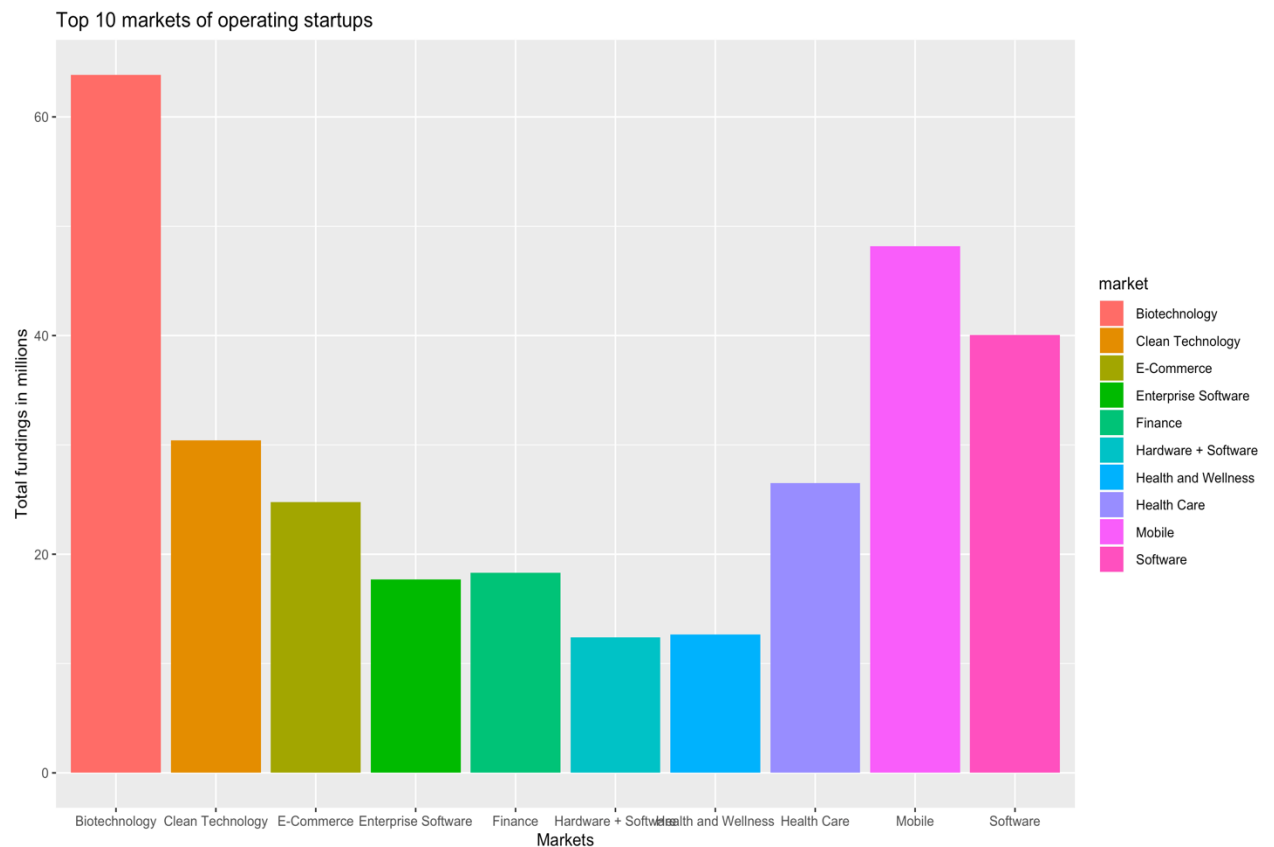
2. FACTOR ANALYSIS

i.   Checking Suitability for Factor Analysis

Barlett's Test of Sphericity was used.

p-value obtained was 0.005681558. This is a good indication to continue factor analysis.

Correlation Plot

ii.   Determining Numbers of Factors

Following methods were used and their results were:

    a.   Eigen Value: 8

    b.   Parallel Analysis: 7

As above-mentioned, while implementing parallel analysis, a warning was obtained that the estimated weights for factor scores might be incorrect, and suggested that we use a different a different strategy. Hence, Principal Component Analysis was chosen.
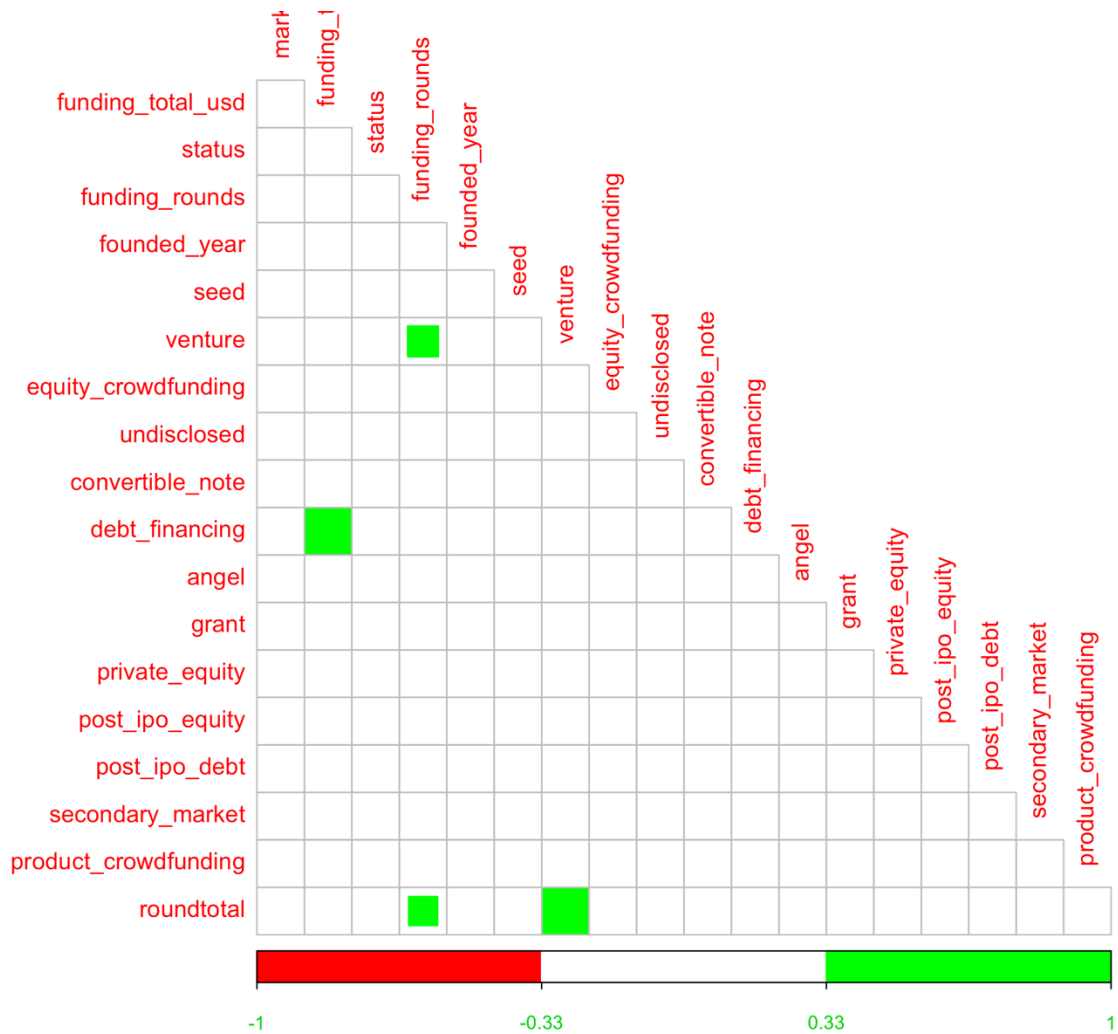
3. PRINCIPAL COMPONENT ANALYSIS

i. Checking Suitability for Factor Analysis

Barlett's Test of Sphericity was used.

p-value obtained was 0.005681558. This is a good indication to continue

factor analysis.

The data was then split into test and train.

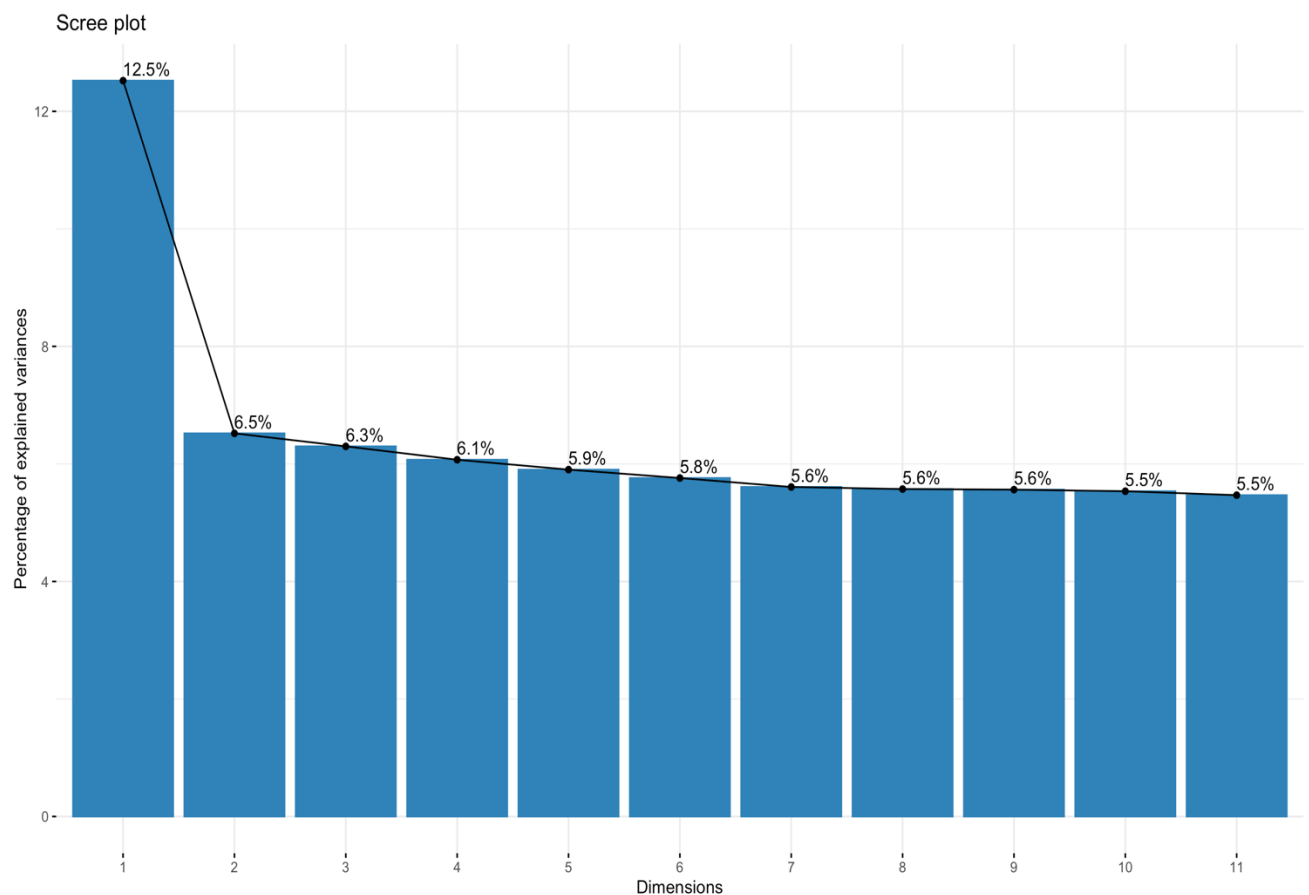### ii. Determining Number of Components

Following methods were used and their results were:

a. **Scree Plot:** 6

b. **Eigen Value**: 9

c. **Parallel Analysis:** 5

Based on the above analysis, we proceed with **number of factors = 6**

Scree plot

iii.   Describing Components

Size of the loadings/contribution indicates the importance of the variable in describing a component. Following displays contribution of every element to the components:

```
                       Dim.1 Dim.2 Dim.3 Dim.4 Dim.5 Dim.6
market                  1.25  0.50 25.34 11.00  1.82  2.78
status                  0.87  4.73 12.74  2.18  5.22 10.24
funding_rounds         18.77  0.32  0.96  0.11  0.14  3.83
founded_year            1.96 20.99 15.87  1.29  0.18  0.40
seed                    0.11  7.05  5.62  3.63  2.77 38.60
venture                38.04  0.05  0.11  1.77  0.09  0.80
equity_crowdfunding     0.02  0.00  0.18  0.15  0.01  4.42
undisclosed             0.05  1.53  0.10  0.86 43.68  0.15
convertible_note        0.05  0.31  2.05 13.29  2.38 11.35
debt_financing          0.35  2.07  0.17  7.00  5.30  0.98
angel                   0.10 24.59 15.00 11.79  0.21  0.05
grant                   0.04  0.18  0.70  1.81  0.13  3.74
private_equity          0.36  4.95  1.85 30.49  3.59  2.85
post_ipo_equity         0.54  7.16  3.51  1.10 31.54  0.76
post_ipo_debt           0.05  1.45  0.01  2.89  2.16  0.37
secondary_market        1.12 23.55 14.97  7.84  0.13  0.05
product_crowdfunding    0.01  0.32  0.55  0.48  0.60 17.91
roundtotal             36.31  0.25  0.27  2.32  0.07  0.73
```

CONCLUSIONS

1. From the charts in the analysis section, we can conclude:

   a. It is interesting to know that most of the companies across the globe were founded immediately after the great recession of 2008.

   b. Top markets that gathered maximum funding were Biotechnology, Mobile, Software, and Clean Technology.

   c. Startups of a few countries seem to be receiving enormous funding, indicating these countries boost and encourage growth of startups. Such countries are USA, Russia, Great Britain, Canada.

   d. USA being the most favorable country, the states California, Massachusetts, New York seem to promote startups better than the other states in terms of total funding.

e. Of all the currently operating companies, investors seem to be more interested in those who belong to the markets of Biotechnology, Mobile, Software.

2. The 39 variables of this dataset can be reduced to 6 components without significant loss in information for easier analysis and predictions.
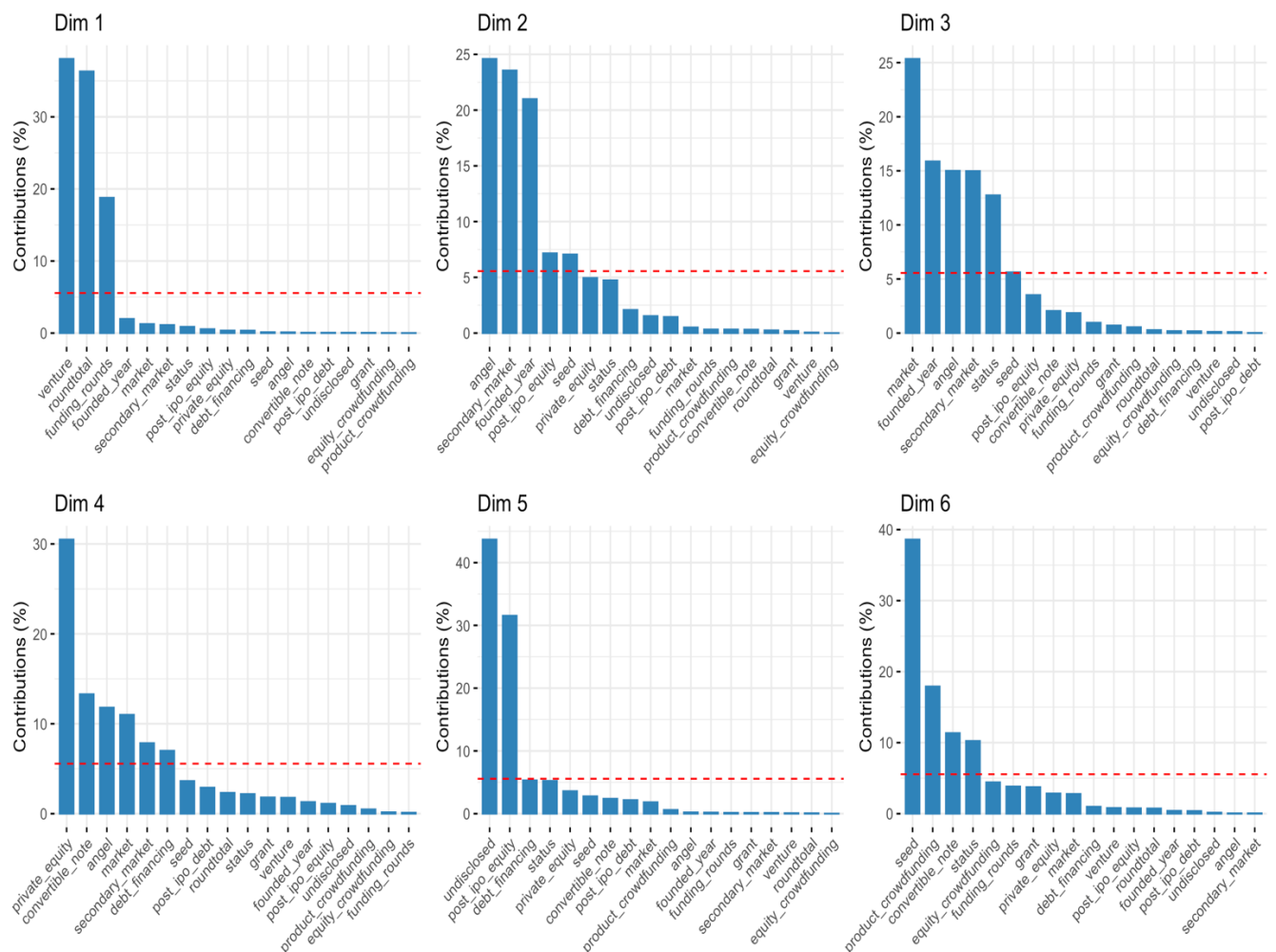
Following are the variables in each component:

Components	Variables
1	funding_rounds venture roundtotal
2	founded_year angel secondary_market
3	market status
4	convertible_note debt_financing private_equity post_ipo_debt
5	undisclosed post_ipo_equity
6	seed equity_crowdfunding grant product_crowdfunding

RECOMMENDATIONS

1. For aspiring entrepreneurs, it is advisable that they build their company either on Biotechnology, Mobile, Software, Clean Technology as these are the markets segments that assure maximum funding.

2. Entrepreneurs are highly likely to obtain more funding if they register their companies in countries such as USA, Great Britain, Canada or Russia.

3. For any further analysis based on the total funding received, it is advisable that the dataset is reduced in dimensions. The dataset can be reduced to above mentioned principal components without significant loss in information.