Untitled

Section 1: Business Objective & Data Structure

Business Objective

Investment strategies for investing in start-up companies are widely based on intuition or past experience. As a result, investors rely primarily on the need being addressed, background of the founders, size of the market being addressed and the ability of the company to scale after tasting early success. The question we pose here is, "can we perform some rigorous analysis that can be used to identify relevant factors and score prospective start- ups based on their potential to be successful". This model/ analysis will then allow investors to make more informed decisions and rely less on their intuitions.

Data Structure

The data has already been split into train, 'CAX_Startup_Train', and test data, 'CAX_Startup_Test'. Data dictionary was also provided that describes all the 51 variables included. The Target variable is a binary class with 234 known observations in the train set and 80 observations to be scored. I combined the data set to wholisticall verify the quality and pre-process all the independent variables for modeling and scoring. I ran a missing values check to and found out there were 80 observations that contains missing values.

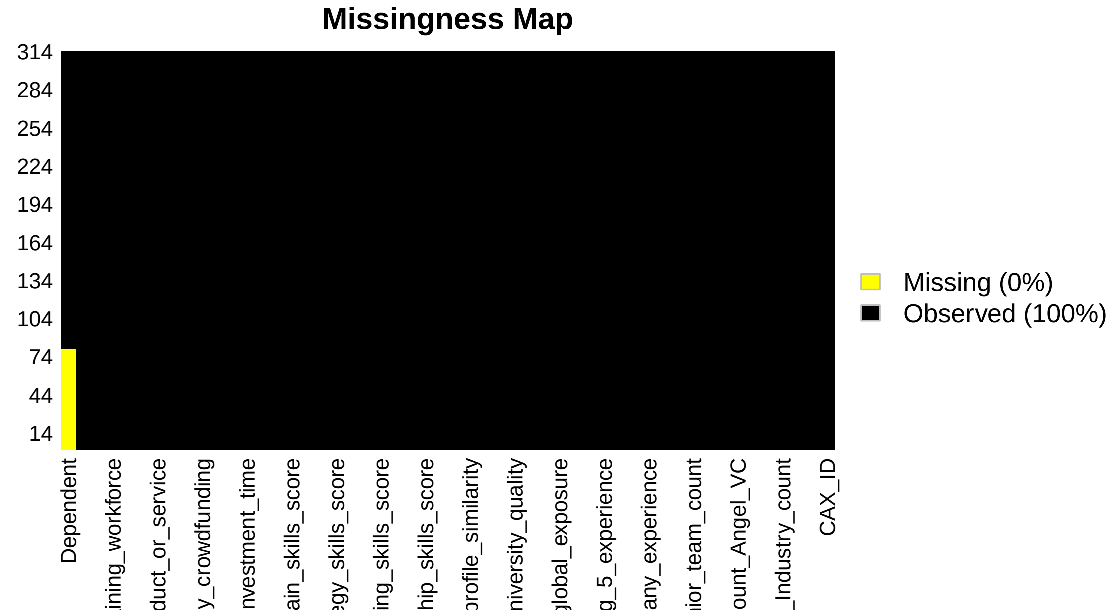


Fig 1 Missing Values

The plot of the missing values showed that they only occurred within the target variable 'Dependent' representing the observations to be predicted. The remainder of the data has no missing value. A glimpse of the data frame showed that numeric and character data types. I converted some of the all of the character data type to categorical data and some coded as integer to categorical data.

Section 2: Exploratory Data Analysis

To effectively explore the data and get insights from the visualizations, I split the data back to train and test and then split the traininto two sub-data frames, continuous and categorical data frames.

I started by plotting the distribution of the continuous variables before moving on the the categorical variables. I grouped some of the plots for comparison and the first set of plots were to understand the distribution of investors to the startups.

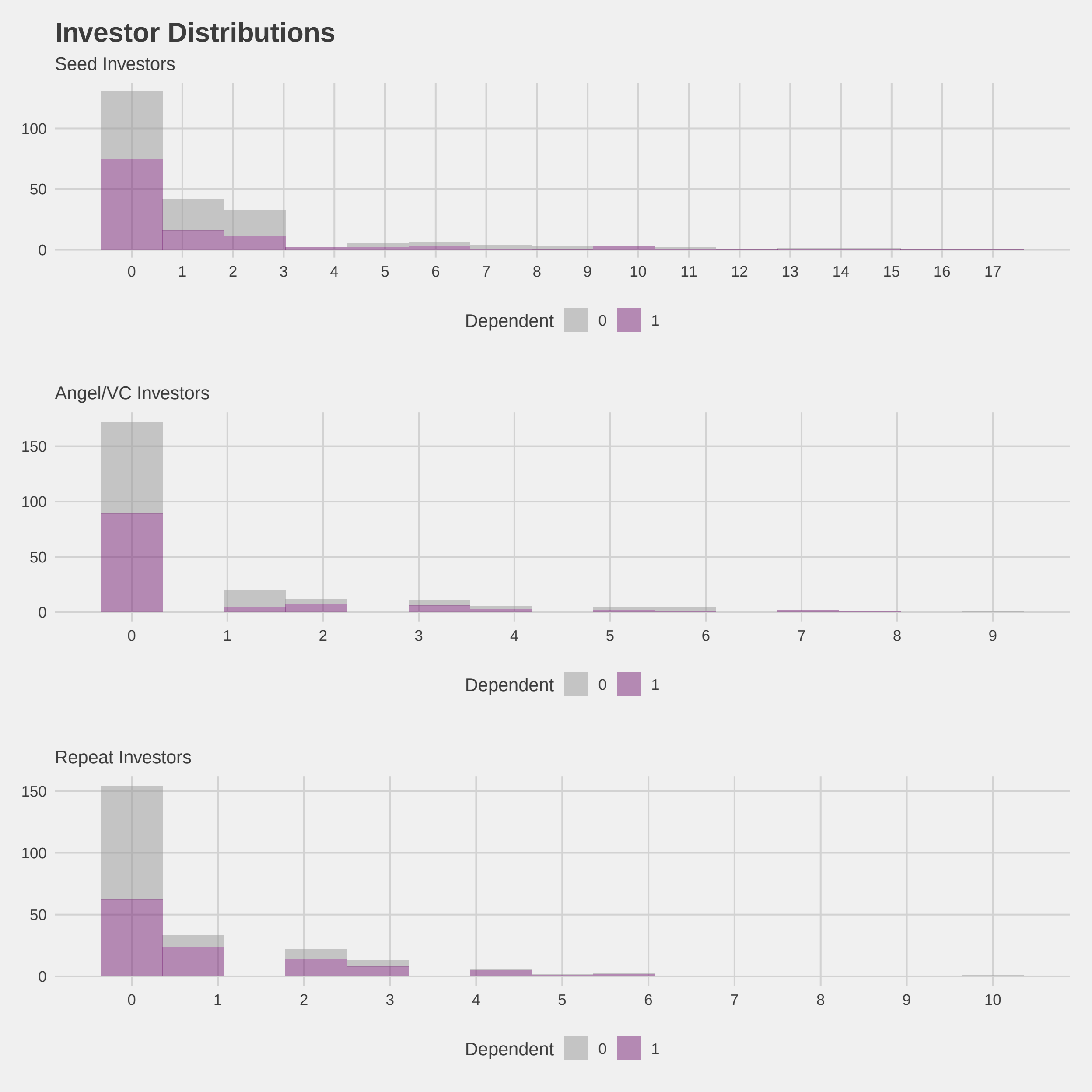


Fig. 2.1 Investor Distributions

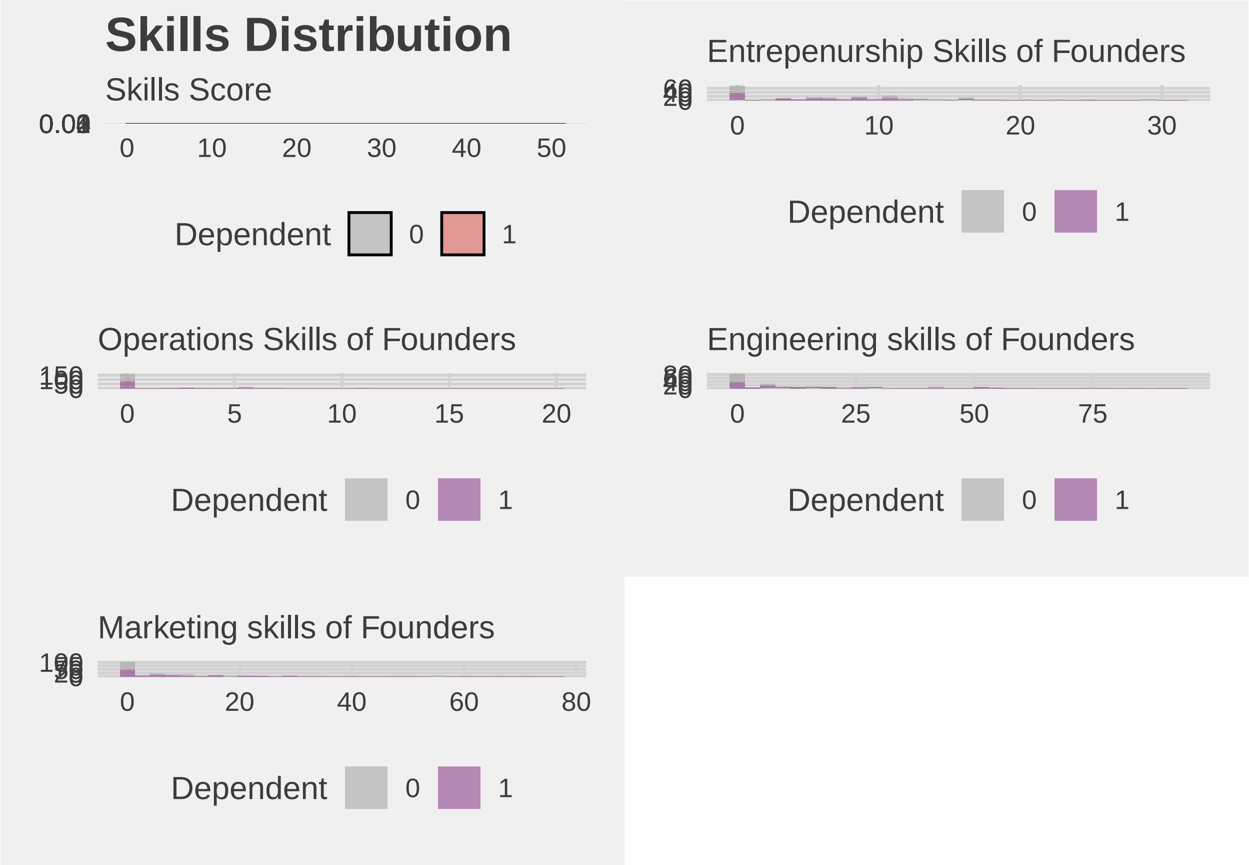
We find that all distributions are right skewed with evidence of outliers. In addition, over 75% of the startups had no investors in seed or as Angel/VC.

```
# Skills distribution
```

```
skill.score.plt1 <-cnt_train%>%
  ggplot(aes(Founders_skills_score, fill=Dependent)) +
  geom_density(alpha=0.5)+
  theme_fivethirtyeight()+
  scale_fill_manual(values = c("#999999", "#d84242")) +
  labs(title= ’Skills Distribution’,subtitle = ’Skills Score’)

skill.score.plt2 <-cnt_train%>%
  ggplot(aes(Founders_Entrepreneurship_skills_score, fill=Dependent)) +
  geom_histogram(alpha=0.5)+
  theme_fivethirtyeight()+
  scale_fill_manual(values = c("#999999", "#7a2477")) +
  labs(subtitle = ’Entrepenurship Skills of Founders’)

skill.score.plt3 <-cnt_train%>%
  ggplot(aes(Founders_Operations_skills_score, fill=Dependent)) +
  geom_histogram(alpha=0.5)+
  theme_fivethirtyeight()+
  scale_fill_manual(values = c("#999999", "#7a2477")) +
  labs(subtitle = ’Operations Skills of Founders’)

skill.score.plt4 <- cnt_train%>%
  ggplot(aes(Founders_Engineering_skills_score, fill=Dependent)) +
  geom_histogram(alpha=0.5)+
  theme_fivethirtyeight()+
  scale_fill_manual(values = c("#999999", "#7a2477")) +
  labs(subtitle = ’Engineering skills of Founders’)

skill.score.plt5 <-cnt_train%>%
  ggplot(aes(Founders_Marketing_skills_score, fill=Dependent)) +
  geom_histogram(alpha=0.5)+
  theme_fivethirtyeight()+
  scale_fill_manual(values = c("#999999", "#7a2477")) +
  labs(subtitle = ’Marketing skills of Founders’)

skill.score.plt6 <-cnt_train%>%
  ggplot(aes(Founders_Leadership_skills_score, fill=Dependent)) +
  geom_histogram(alpha=0.5)+
  theme_fivethirtyeight()+
  scale_fill_manual(values = c("#999999", "#7a2477")) +
  labs(subtitle = ’Leadership Skill of Founders’)

skill.score.plt7 <- cnt_train%>%
  ggplot(aes(Founders_Data_Science_skills_score, fill=Dependent)) +
  geom_histogram(alpha=0.5)+
  theme_fivethirtyeight()+
  scale_fill_manual(values = c("#999999", "#7a2477")) +
  labs(subtitle = ’Data Science skill of Founders’)

skill.score.plt8 <-cnt_train%>%
  ggplot(aes(Founders_Business_Strategy_skills_score, fill=Dependent)) +
  geom_histogram(alpha=0.5)+
  theme_fivethirtyeight()+
```

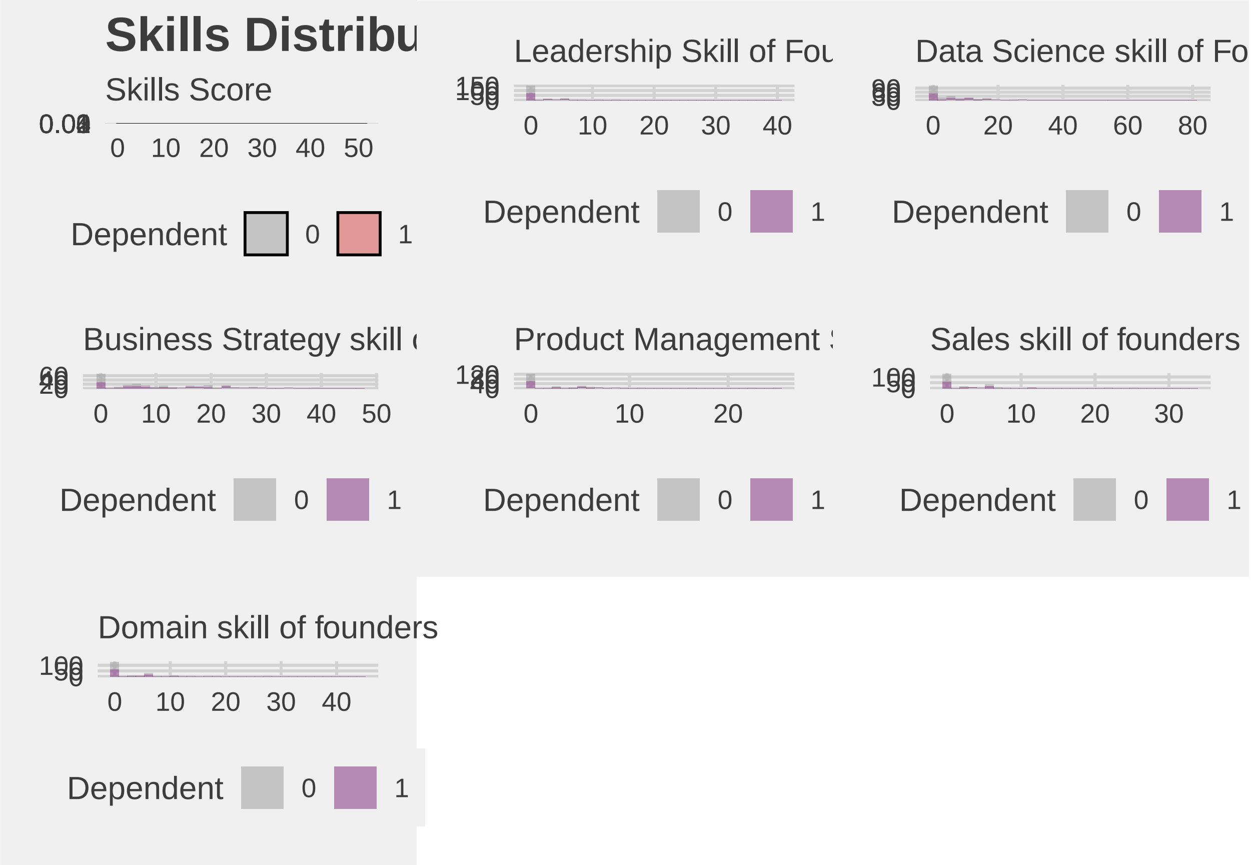
```r
  scale_fill_manual(values = c("#999999", "#7a2477")) +
  labs(subtitle = ’Business Strategy skill of Founders’)

skill.score.plt9 <- cnt_train%>%
  ggplot(aes(Founders_Product_Management_skills_score, fill=Dependent)) +
  geom_histogram(alpha=0.5)+
  theme_fivethirtyeight()+
  scale_fill_manual(values = c("#999999", "#7a2477")) +
  labs(subtitle = ’Product Management Skill of Founders’)

skill.score.plt10 <- cnt_train%>%
  ggplot(aes(Founders_Sales_skills_score, fill=Dependent)) +
  geom_histogram(alpha=0.5)+
  theme_fivethirtyeight()+
  scale_fill_manual(values = c("#999999", "#7a2477")) +
  labs(subtitle = ’Sales skill of founders’)

skill.score.plt11 <- cnt_train%>%
  ggplot(aes(Founders_Domain_skills_score, fill=Dependent)) +
  geom_histogram(alpha=0.5)+
  theme_fivethirtyeight()+
  scale_fill_manual(values = c("#999999", "#7a2477")) +
  labs(subtitle = ’Domain skill of founders’)


grid.arrange(skill.score.plt1,skill.score.plt2,skill.score.plt3,skill.score.plt4,skill.score.plt5)
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



```
grid.arrange(skill.score.plt1,skill.score.plt6,skill.score.plt7,skill.score.plt8,skill.score.plt9,skill
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```


Skills Distribu
Skills Score
0.04
0 10 20 30 40 50
Dependent 0 1
Leadership Skill of Fou
150
0 10 20 30 40
Dependent 0 1
Data Science skill of Fo
0 20 40 60 80
Dependent 0 1
Business Strategy skill
60
20
0 10 20 30 40 50
Dependent 0 1
Product Management
120
40
0 10 20
Dependent 0 1
Sales skill of founders
100
50
0 10 20 30
Dependent 0 1
Domain skill of founders
100
50
0 10 20 30 40
Dependent 0 1