Business Analytics for Beginners Using R - Part I

Chinedu Okechukwu

10/26/2020

Problem Statement

Investors need to move beyond the value-proposition of start-up companies in determining their success or failure to inform their investing strategies. As more information and data on start-ups is now available, investors want to discover factors from this information that affect the success of these start-up companies.

The primary objective of this data analytics project is to identify these factors or attributes of successful start-up companies by performing some analytical procedures on many informational attributes acquired on 472 start-up companies. The aim is to get the data into a more useable format to perform some form of exploration and statistical tests.

Section 1: How the data was treated including missing value?

The csv file read to R contained 472 observation of 116 variables on start-up companies. Majority of the variables were classified as characters with a few as integers.

Data frame had 264 rows of incomplete cases, i.e rows with NA values. Data also had values with "No Info" or blanks that needs to be changed to NA values. These values where replaced with NAs which increased the number of rows with incomplete cases to 456.The missing value plot is shown in Figure below.

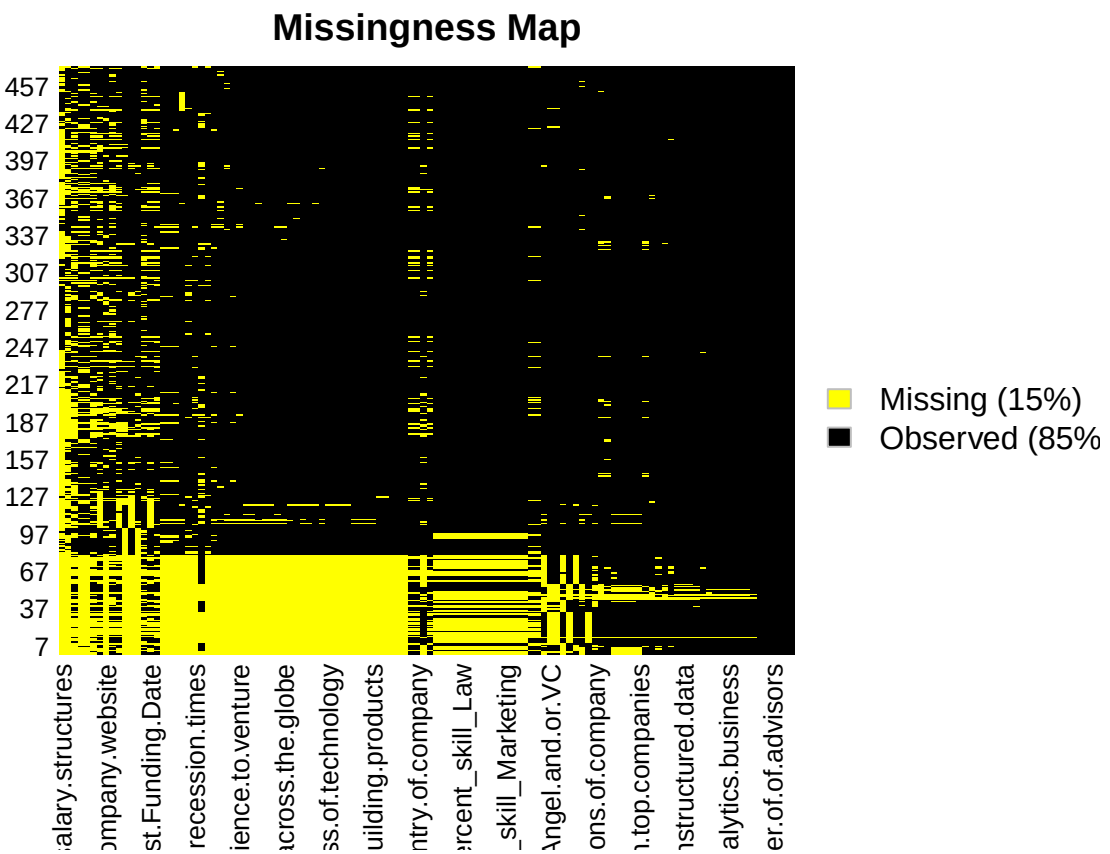

Fig 1 Missing Value Plot

I handled the 15% missing value as follows:

1. I removed variables that had more than 40% missing values which resulted in the removal of the following 3 variables I calculated the proportion of missing values for each variable and set a rule to keep variables that have 40% or less missing values. The following columns(variables were removed as a result)

```
##                                  variable Percent.Missing
## 1              Employees.count.MoM.change            43.4
## 2 Employee.benefits.and.salary.structures            74.4
## 3                        Client.Reputation            58.1
```

2. I dropped observation with more than 40% of the variables missing

3. I imputed the remaining missing values with the median for numeric variables and the mode for categorical variables

Section 2: Details of additional features created

I created two additional variables- Investor.count from as a sum of the total number investors for each startup and "Industry. Count" as the total number of industries a startup belongs to.I then dropped their originating variables along with the following:
. Company_Name
. Short.Description.of.company.profile

. Est..Founding.Date & Last.Funding.Date (Age of company is already in the data)
. Specialization.of.highest.education- (multiple values but highest education will suffice)
. Time.to.maturity.of.technology..in.years (this had only one value, hence low variability)