# Understanding and Visualizing Data

#### **Part One – Draft a Data-Collection Plan**

Step 1: Identify a decision that requires, or would be enhanced by, data analysis.

What is the situation?

Is the startup a success or a failure?

Startups play a major role in economic growth. They bring new ideas, spur innovation, create employment thereby moving the economy. There has been an exponential growth in startups over the past few years. Predicting the success of a startup allows investors to find companies that have the potential for rapid growth, thereby allowing them to be one step ahead of the competition. Running a startup business is a key goal for me and I have been working towards achieving this goal. To start a business which would be successful requires research and good decision making. As such, I want to look at startups data and investigate which business startup(s) has been successful so far. What they have done to result into success or failure? This data analysis can help me decide on the type of business to invest in.

What are the parameters or options in the decision?

The parameters or options in the decision which I will consider are profit/loss revenue/expenses, regulation changes, policy changes and or environmental influences. To make a smart investment decision I need to consider all the risks in the business arena as well.

What other alternatives are available?

The available alternatives are analyzing revenue/expenses or profit/loss of a random sample of startups and then decide on investment. Another alternative can be coordinating with other investors, VCs and or entrepreneurs to assess the market and decide on investment. Some investors make small investments in a few startups and wait for returns or a preferrable situation and then decide on large investments.

Who are the key stakeholders?

Myself, other investors, venture capitalists and entrepreneurs.

How do you hope or expect data to help illuminate your decision?

The objective is to predict whether a startup which is currently operating turns into a success or a failure. Such analysis will help me understand the cases and variables in the outcome. There is data on investment, idea, duration, and other factors which can provide information about how and why they succeeded or failed. The data contains industry trends, investment insights and individual company information.

Step 2: Identify data that will help you better understand the situation.

What are the key performance indicators for your situation?

The KPIs:

* the total funding the startup has received since its start
* the rounds of funding the startup has received since its establishment
* has it been acquired?
* has it been categorized as the top 500?

What are the variables you will consider?

I am looking at different factors of success in startup businesses to help me decide. The type of business, the number of funding rounds which shows that investors are interested; the total amount of funding, whether the startup has been categorized as the top 500 and whether it was acquired.

Is each variable categorical or quantitative?

The data set has both categorical and quantitative variables. There are variables like funding total (USD) which is quantitative and there are categorical variables like “Is top 500”.

What purpose does each variable have in informing your decision?

Variables which show the funding amount or funding rounds help me decide which startup idea has received enough or considerable financial support to develop its product and services and become successful. In addition, categorical variables like “has VC, has Angel, is top 500” will also provide essential data in helping me decide the type of startups VCs have invested in, or the type of businesses that were later categorized in the top 500 category.

Step 3: Develop a data-gathering plan.

Where will the data come from?

I am going to use the dataset available in Kaggle on startup businesses.

Is it observational or experimental data?

The data is observational.

Who will collect it?

It is available in Kaggle which I will download and conduct the required analysis on.

How much data will you need (sample size)?

There are 923 rows of data in this dataset, I will need a random sample of 275 rows (30%) to help me analyze data and create a model to inform stakeholders.

How will you assure that it is representative of the population?

I will ensure a random sample of data is taken.

What steps will you take to mitigate potential bias?

A random sample of 30% from the dataset is a basic method to mitigate potential bias.

#### **Part Two – Identify Data Summaries and Visualizations**

What summary statistics will you use to inform your decision?

I am using mean, min, max and standard deviation.

Are you interested in statistics that are sensitive to or resistant to outliers, and why?

I am looking at statistics which are inclusive. I need to look at the overall picture including the outliers to better inform decision making.

What visualizations will you use to inform your decision?

Be sure to indicate what variables and what scale you are using for each visualization (e.g. histogram showing frequency of page hits per hour over a 24-hour period)

I will use a dashboard which showcases bar graphs, a line graph, a pie graph and a table of summary statistics. The dashboard includes the following visualizations:

* For the pie graph I am using two categorical variable -industries “category\_code” and “is top 500”.
* For the bar graph I am using a categorical variable -industries “category\_code” and a quantitative variable – Total funding “sum of funding (USD).
* For the second bar graph I am using three categorical variables – industries, has VC and has Angel.
* For the line graph I am using a categorical variable – industries and a quantitative variable – funding rounds.

#### **Part Three – Data and Your Decision**

Part a: Questions about your data-model-insight framework

What are you attempting to model with your data?

I am attempting to simplify the data to guide me towards making a decision to invest in a startup that will prosper and expand in the future.

What are the KPIs for the situation you are trying to understand?

The main question this analysis has to answer: Is the startup a success or a failure?

How much funding has the startup had so far?

The rounds of funding the startup has received.

Has it been categorized in the top 500?

Has it been acquired?

What is the relationship between your variables and the KPIs?

The variables in this data analysis model are measurements of KPIs. For example, the sum of funding or total funding is a measure of success of the startup in the total of funding it has received from investors. In addition, the funding rounds is another important measure of the KPIs which shows that the startup is a success/failure by the rounds of funding it has received. Another measure is the categorical variable- is top 500- which measures where the startup lies, if the startup is a success, it is categorized in the top 500. Lastly, has the startup been acquired or closed? Again this measures the ultimate goal of this analysis- is the startup a success or a failure?

What are the limitations of your model?

My model does not include individual startup details, or it doesn’t predict what would happen to startups in the future. A successful startup might end up a failure. It also doesn’t show who the VC or Angels are, which is a great drawback. In addition, it doesn’t include annual revenue or net income of the startups.

Do you feel your model, as defined, is “good enough” to inform your decision? Why or why not?

The model does provide primary data about startups, however additional data is required to help me decide. I need data about annual revenue/expenses or profit/loss. I need data about its VCs or angels and or stakeholders. I also need additional details about their services/products.