Analysis of 2015 Top 100 Analytics Startups

Data Overview

This dataset describes information about the top 100 analytics startups in 2015 (i.e., 100 records in total in the sample), which was offered by Mattermark[1]. Operating a business-to-business (B2B) business model, these startups mainly develop and sell analytics software to enterprises (Columbus 2015). This dataset was selected due to an interest in startups and what factors may correlate with a high growth score, the ability to get more funding as well as to grow sustainably.

Looking at the dataset, some details are included regarding startup name, its growth score, funding stage, total funding (in USD) a startup received, date of last funding and its location. Specifically, the variable of *growth score* measures how quickly a company gains traction at a given point in time. It is calculated by Mattermark through the combination of the Mindshare Score (web traffic, social traction) as well as business growth metrics (e.g. employee count over time, funding). According to the illustration of Mattermark, its underlying assumption is that companies who see growth across these signals are shipping products and talking to customers, and are more likely to continue to grow as a result (Columbus 2015).

Hypotheses

Before starting this analysis, we hoped to discover what factors were correlated to growth score and stage of startups in the dataset. Our hypotheses and reasoning were:

1. Total funding is positively related to funding stage
   Our reasoning is that by the time a startup progresses to a late stage, it would have accumulated a lot of total funding.

2. Total funding is positively related to growth score
   We expect that funding is positively related to growth score, as funding is one of the factors that is considered when calculating the growth score. Other factors that are used to calculate this score, such as employee headcount and web traffic are not shown in this dataset.

3. Location is related to funding stage

[1] The dataset was downloaded from https://raw.githubusercontent.com/curran/data/gh-pages/mattermark/2015-top-100-analytics-startups.csv.

We expect that a higher proportion of late-stage startups would be found in the Bay Area compared with outside the Bay Area as the Bay Area is a magnet for startups due to a high concentration of skilled labour and access to resources required to start a company.

4. Location and funding are related
   We expect that higher startup funding is found in the Bay Area compared with outside the Bay Area. We expect that the median funding and maximum funding is higher in the Bay Area compared with outside the Bay Area.

Data Wrangling

After loading this dataset into Python, we observed it in detail and then conducted dataset wrangling as follows.

First of all, we read the dataset and found that the column names are capitalized and contain spaces between words, which are not usable to code. As a result, we lowered them and excluded the spaces to facilitate future analysis.

Secondly, the datatype of the whole dataset was checked, which showed that all values were objects except the growth score. To conduct statistical analysis on the total funding, we need the values to be numerical, so we converted the datatype of *'t_funding*' from object to floating.

In addition, the '*date_l_funding*' indicates that most startups received last funding during the period of 2012-2015. In order to obtain accurate information about this, we changed the '*date_l_funding*' from the datatype of object to datetime first. Then, according to this, a new column named '*year_l_funding*' was created, which displays the specific year of startups accepting their last funding.

Lastly, missing values of the dataset were checked thoroughly through creating a function (i.e., "*check_missing_v*"). The result suggested there are not any missing values in this dataset.

Once the basic dataset checking and wrangling was completed, we were left with a dataset which could be used for the following analytics.

Data Analysis

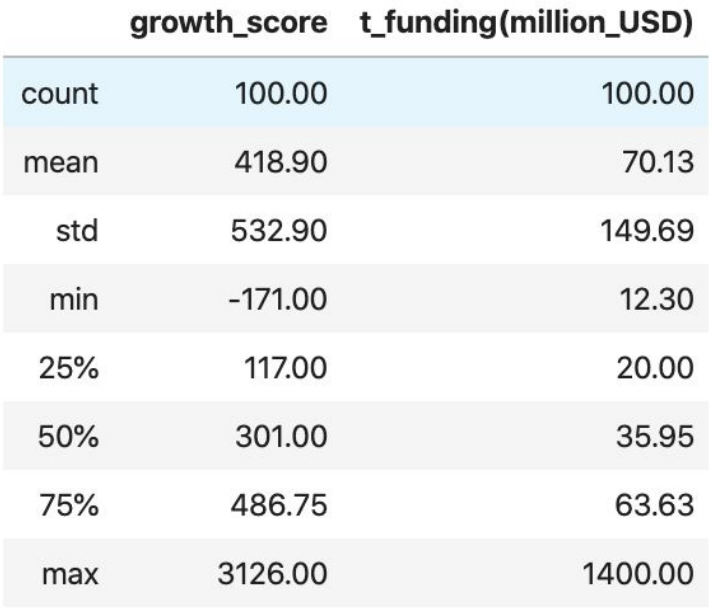
Our overall approach to the data analysis was to examine how the total funding, location, and year of last funding received were related to the funding stage and growth score. Specifically, we examined the relationship between the following:

1. Total funding and the growth score by plotting a scatter plot and examining the correlation

2. Total funding and funding stage by plotting a series of box plots, pie charts and bar charts
3. Location and total funding received, by dividing the startup data into “Bay Area” and “Outside the Bay Area” and plotting the percentage of funding received by startups in each location category
4. Location and funding stage, by plotting the number of startups in each funding stage in the Bay Area and Outside the Bay Area, and comparing the data

Descriptive analysis

First, we focused on how many startups there are in different ranges of the growth score and total funding. Their growth score ranges from -171 to 3126 and the total funding is between 0 and 1400 (Figure 1). However, more than 90% of 100 startups have a growth score of less than 800 and their received total funding is between 0-180 million USD (Figure 2).



	growth_score	t_funding(million_USD)
count	100.00	100.00
mean	418.90	70.13
std	532.90	149.69
min	-171.00	12.30
25%	117.00	20.00
50%	301.00	35.95
75%	486.75	63.63
max	3126.00	1400.00

Figure 1 Descriptive Statistics of Growth Score and Total Funding

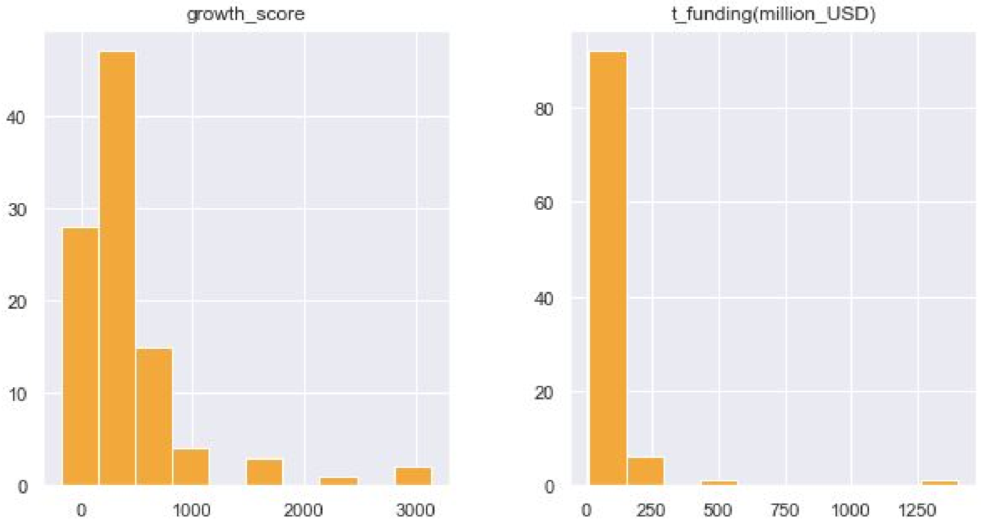


Figure 2 Distribution of Growth Score and Total Funding of 100 Startups

Secondly, we explored the relationship between startups' growth score and their total funding because the growth score is calculated partially based on funding (Columbus 2015). Figure 3 shows that total funding and growth scores are not strongly correlated. We note there are some outliers in Figure 3 which may influence this relationship. Therefore, we decided to see if there is any difference in the relationship if we exclude outliers. After adjusting the range of total funding (less than 500) and growth score (less than 1500), we deleted 7 outliers. Then we draw a scatter plot using the subsample of 93 startups (Figure 4). It indicates a slightly positive relationship between total funding and growth score.

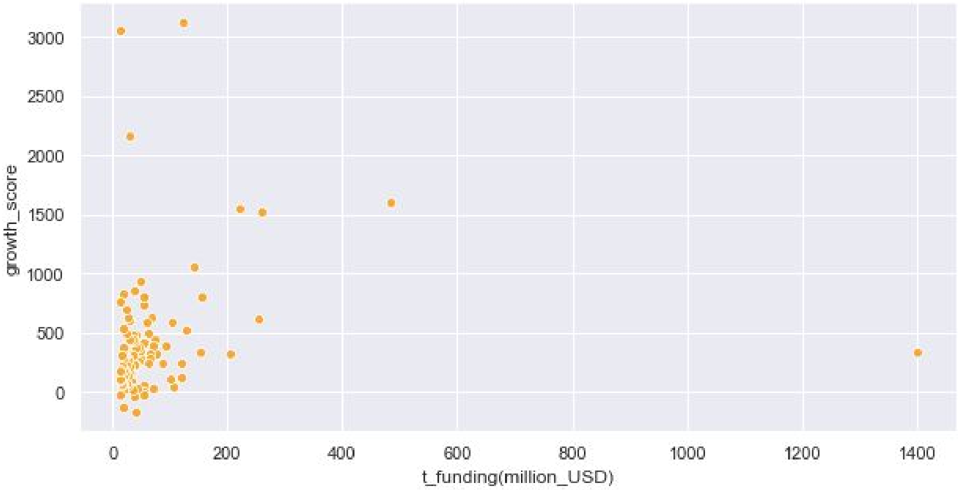


Figure 3 Relationship between Total Funding and Growth Score (100 startups)

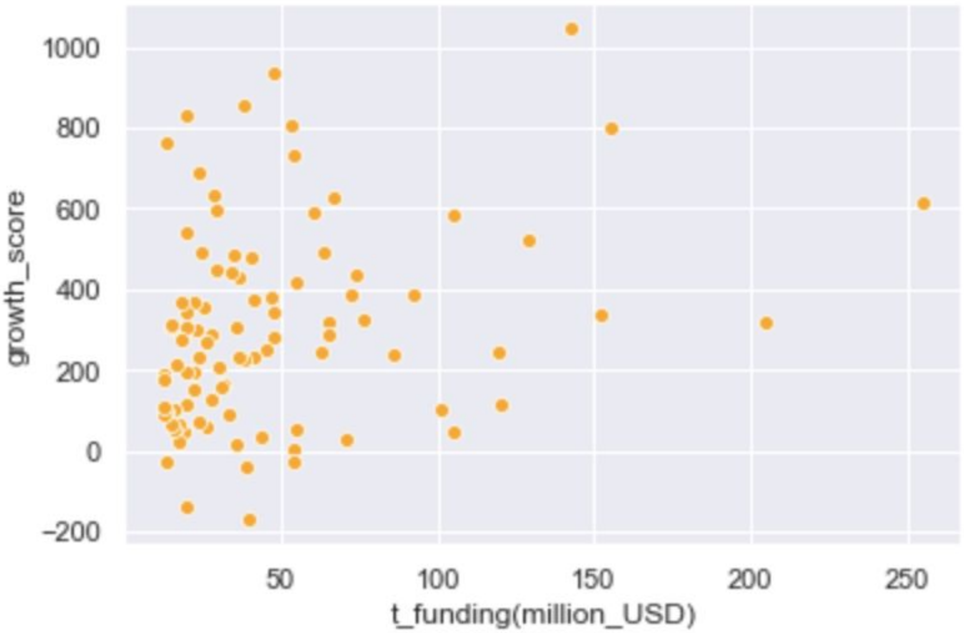


Figure 4 Relationship between Total Funding (<500) and Growth Score (<1500) (93 startups)

In addition, the count of startups in different funding stages and locations are calculated (Figure 5 and 6). Most of these startups have completed funding stage A. The number of startups which are going through stage B and C are 40 and 23 respectively, while 23% have entered the late funding stage. Among the 100 startups, 49 are located in the Bay Area, which accounts for nearly half of the sample. The locations with the next highest number of startups are New York and Boston, which have 11% and 10% startups respectively. As a result, in our following analysis, we grouped the startups by whether they are located inside or outside the Bay Area.

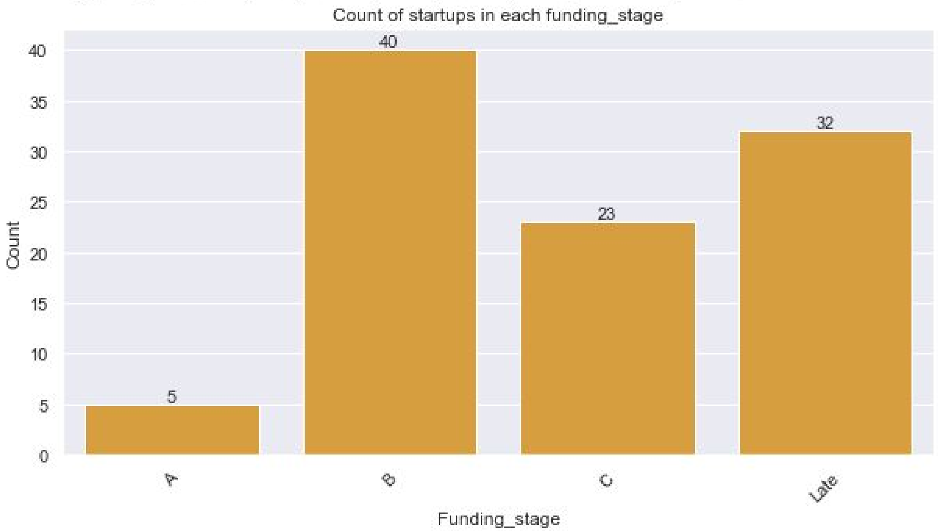


Figure 5 Count of Startups in Different Funding Stage

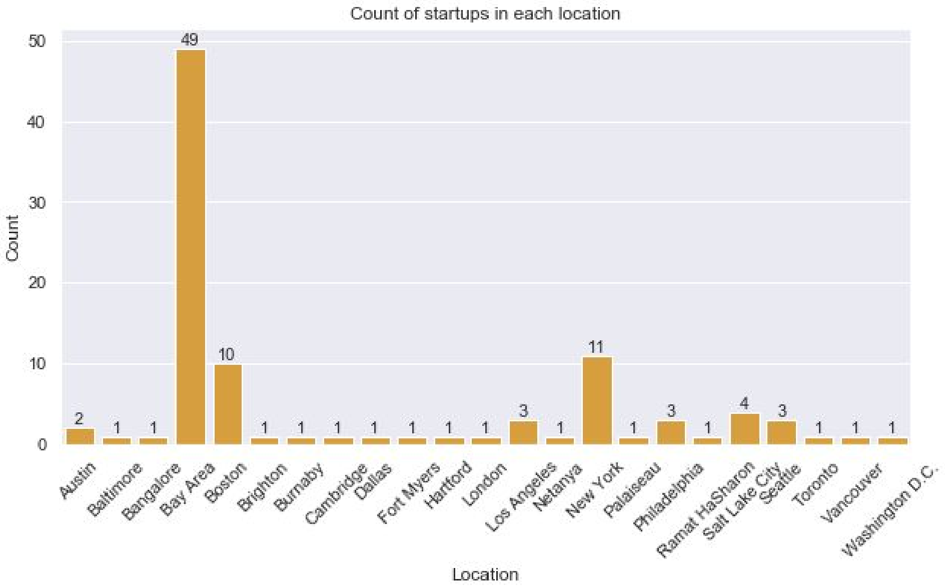


Figure 6 Count of Startups in Different Location

Similarly, when exploring the count of startups which received their last funding in different years, we found that the year that they received the last funding are concentrated in 2013, 2014 and 2015, almost accounting for 90%. Therefore, we focused our following analysis only on these three years.

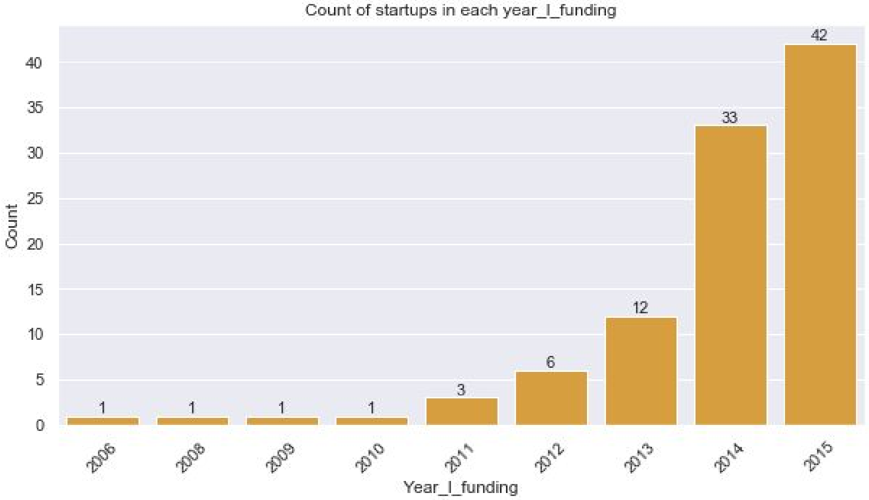
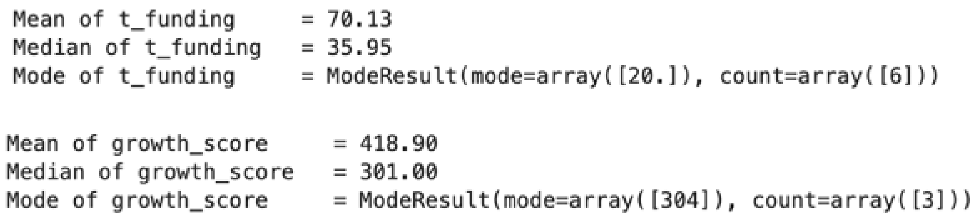


Figure 7 Count of Startups which Received Last Funding in Different years

Measures of central tendency and variability

As shown in Figure 8, the mean and median of total funding are 70.13 and 35.95 respectively. Its mode is 20 million USD, which is received by 6 startups. Growth score has a mean of 418.90 and a median of 301. Its mode is 304, which appears 3 times in the sample. Overall, the central tendency of total funding and growth score is high through the comprehensive analysis of Figure 2 and 8.

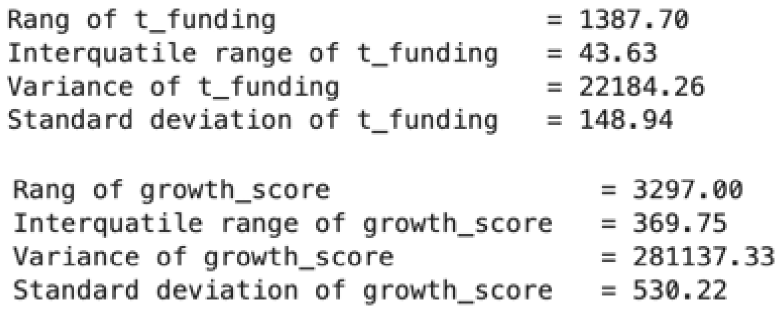


```
Mean of t_funding     = 70.13
Median of t_funding   = 35.95
Mode of t_funding     = ModeResult(mode=array([20.]), count=array([6]))

Mean of growth_score     = 418.90
Median of growth_score   = 301.00
Mode of growth_score     = ModeResult(mode=array([304]), count=array([3]))
```

Figure 8 Central Tendency of Total Funding and Growth Score

Figure 9 indicates that the range and the interquartile range (IQR) of total funding are 1387.70 and 43.63 respectively. The variance and standard deviation (std) of total funding are 22184.26 and 148.94. For growth score, its range and IQR are 3297 and 369.75 while its variance and std are 281137.33 and 530.22. As a whole, the variability of total funding and growth score is not low from Figure 2 and 9.



```
Rang of t_funding                  = 1387.70
Interquatile range of t_funding    = 43.63
Variance of t_funding              = 22184.26
Standard deviation of t_funding    = 148.94

Rang of growth_score                  = 3297.00
Interquatile range of growth_score    = 369.75
Variance of growth_score              = 281137.33
Standard deviation of growth_score    = 530.22
```

Figure 9 Variability of Total Funding and Growth Score

Analysis under different scenarios

Total funding and growth score in each funding stage

Figure 10 implies that the top ten most-funded startups are all in stage B or Late, of which 8 are in stage Late and 2 are in stage B. What’s more, most of the total funding is received by startups which are in the funding stage B or Late, up to 34.18 billion USD which is 80% together (Figure 11). Figure 12 suggests that half of the startups having top-ten growth score are in stage B and the other half are in stage Late.

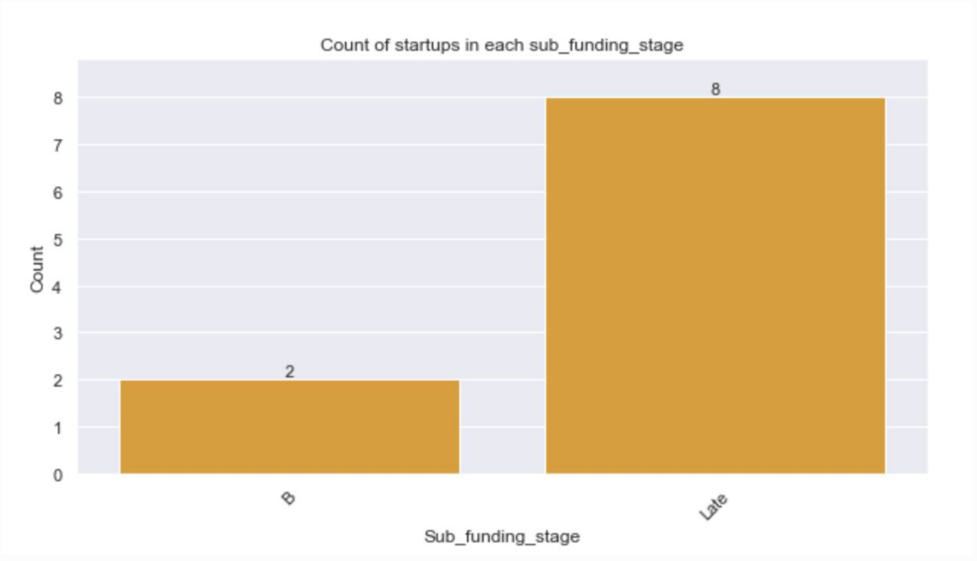
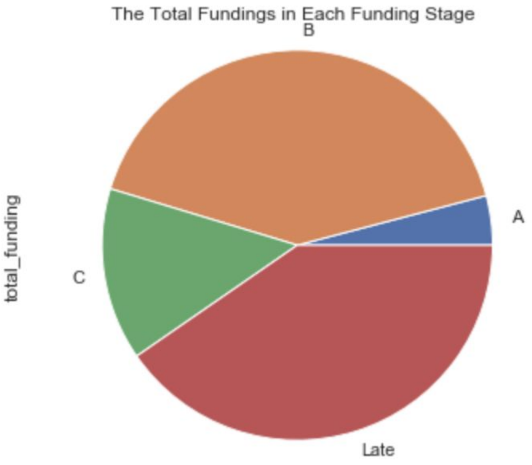


Figure 10 Count of Top-ten Funded Startups in Different Funding Stage



```
The total fundings in stage_late: 16910.00 millon USD which is 40%
The total fundings in stage_A:    1715.00 millon USD which is 4%
The total fundings in stage_B:    17273.00 millon USD which is 40%
The total fundings in stage_C:    5992.00 millon USD which is 14%
```

Figure 11 Total Funding in Each Funding Stage

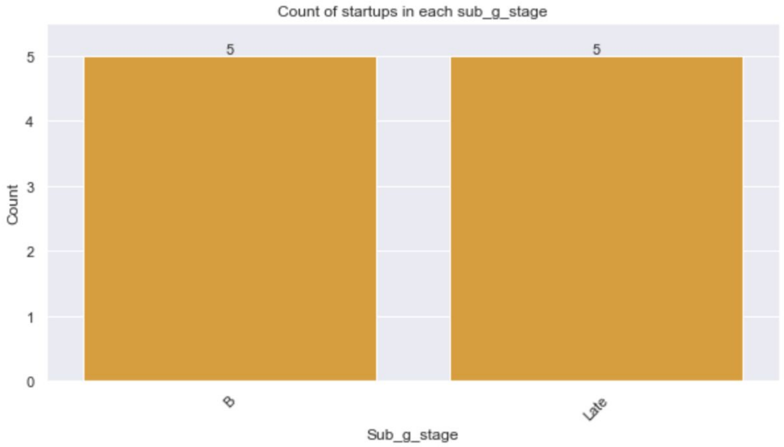


Figure 12 Count of Top-ten Growth Score Startups in Different Funding Stage

Figure 13 shows that the median funding allocated to a startup which is in stage Late was obviously higher and that the range in funding allocated was much higher. What's more, funding received by startups who are in stage B are more centralized compared with that received by those in stage Late, even though these two types of startups have similar counts and total funding.

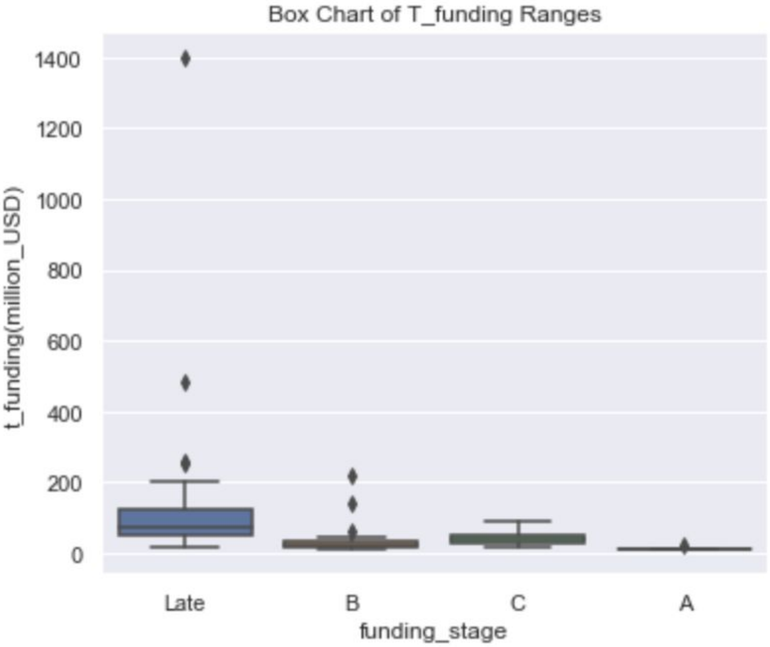


Figure 13 Range of Total Funding in Each Funding Stage

Total funding in Bay Area and Non-Bay Area

Figure 14 shows that the median funding allocated to a startup was slightly higher in the Bay Area and that the range in funding allocated was much higher. The maximum funding allocated was 1400 million USD in the Bay Area compared with approximately 500 million USD outside the Bay Area. Figures 15 to 16 show that most of the total funding received by the startups in this dataset was allocated to startups in the Bay Area who received over 60% of the total funding. There are 49 startups in the Bay Area and 51 outside of the Bay Area, so this means that 62% of the funding is going to 49% of the total startups which are in the Bay Area and 38% of the funding is going to the 51% of startups which are outside the Bay Area.

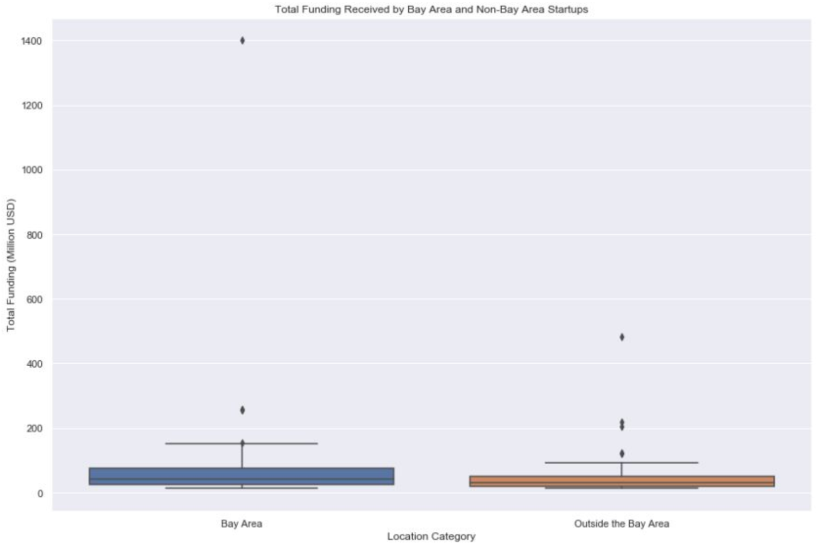
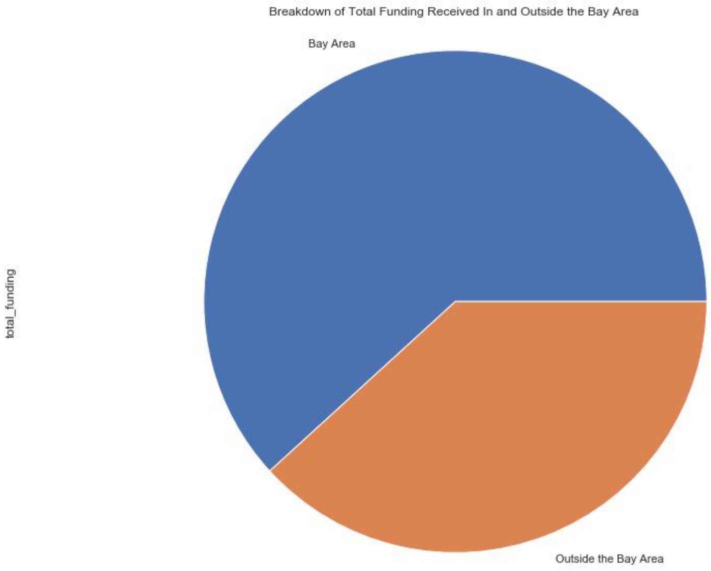


Figure 14 Range of Total Funding Received by Startups Inside and Outside the Bay Area



```
Total Funding Received in the Bay Area: 4333.60 million USD which is 61.80%
Total Funding Received Outside the Bay Area: 2679.20 million USD which is 38.20%
```

Figure 15 Total Funding Received by Startups Inside and Outside the Bay Area

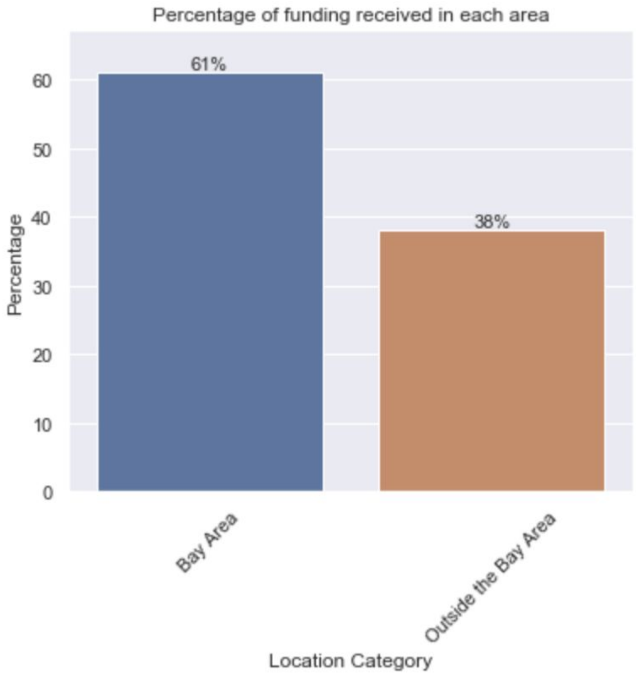


Figure 16 Percentage of Funding Received Inside and Outside the Bay Area

Total funding and growth score in the year of startups received last funding

Figure 17 implies that the majority of the top ten most-funded startups received their last funding in 2015, accounting for 80%. The total funding allocated to startups which received their last funding in 2014 or 2015 is up to 60 billion USD, which accounts for 93% (Figure 18). Figure 19 suggests that 8 out of the 10 startups having top-ten growth score received their last funding in 2015.

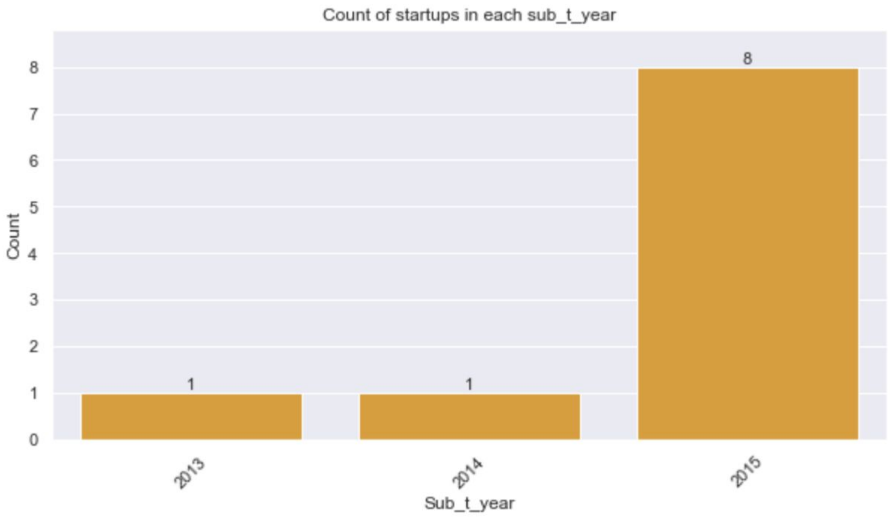
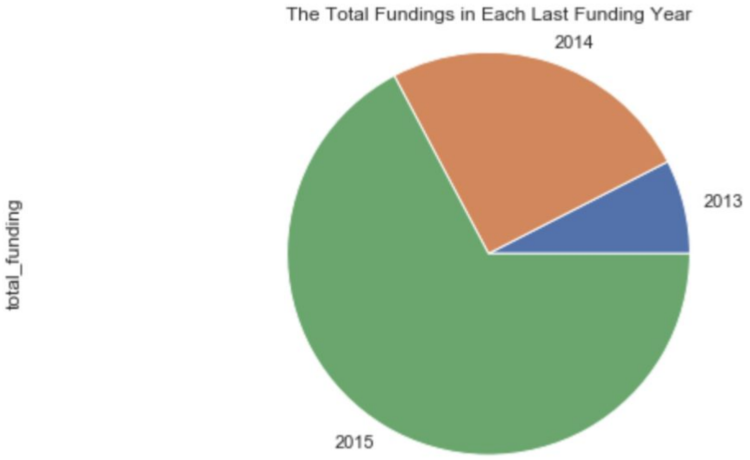


Figure 17 Count of Top-Ten Funded Startups by Year of Last Funding



```
The total fundings in 2015:    4363.70 millon USD which is 67%
The total fundings in 2014:    1635.70 millon USD which is 25%
The total fundings in 2013:    489.90 millon US which is 7%
```

Figure 18 Total Funding by Year of Last Funding

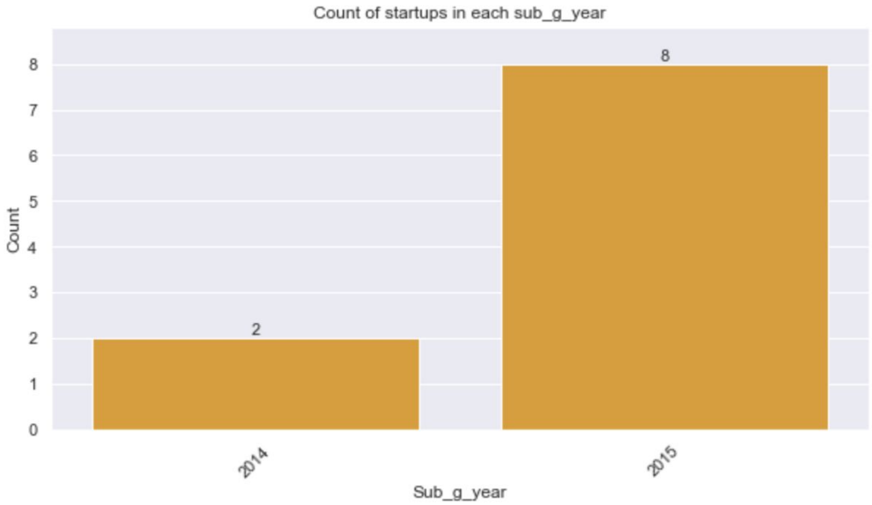


Figure 19 Count of Startups Ranking Top-Ten Growth Score by Year of Last Funding

Figure 20 shows that the range of the total funding allocated to a startup which obtained its last funding in 2015 was obviously higher but the medians are similar in the three years.

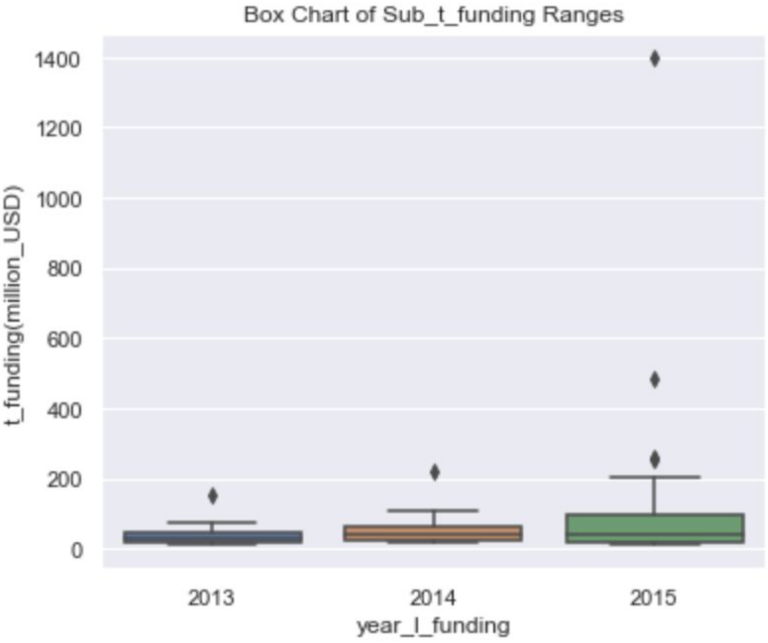


Figure 20 Funding by Last Year of Funding Received

Total startup count by location and funding Stage

As shown by figure 21, when the number of startups at each funding stage is compared for the Bay Area and outside the Bay Area, the pattern is not consistent across the stages. For stages A, B and C, a larger number of startups are found outside the Bay Area than inside the Bay Area. The graph also shows that most of the late stage startups are found in the Bay Area.

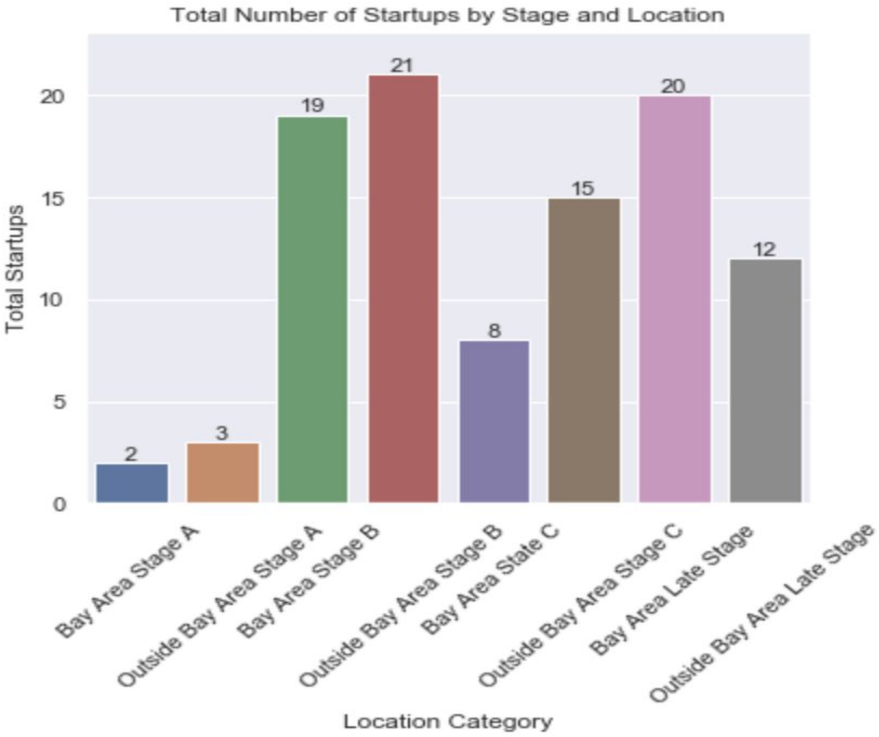


Figure 21 Total Count of Startups at Each Funding Stage Inside and Outside the Bay Area

Findings

Our hypotheses for this analysis were:

1. Total funding is positively related to funding stage
2. Total funding is positively related to growth score
3. Location is related to funding stage
4. Location and funding are related

For hypothesis 1, we found that funding is related to the funding stage. Figure 13 shows that the median funding increases with stage. Interestingly, the range of funding received also increases with the funding stage. The majority of the top-ten funded startups have entered into the funding stage Late (Figure 10).

Hypothesis 2 was tested during the descriptive statistics portion of analysis, where we plotted total funding against growth score (Figure 3). There does not appear to be a significant relationship between funding and growth score, however our analysis showed a positive relationship but not obvious when outliers are excluded. This suggests that the amount of funding may not be the most important factor determining the growth of a startup.

To test hypothesis 3, we compared a count of the total number of startups inside and outside the Bay Area across funding stages. Interestingly we found that there were more earlier stage startups (stage A, B and C) founded outside the Bay Area compared to inside the Bay Area, but there were more late stage startups founded inside the Bay Area compared with outside. We were expecting most of the top startups in all stages to be located in the Bay Area, so this analysis shows that there are lots of successful early to mid stage startups located outside of the Bay Area. In other words, in the early years the Bay Area may be the most ideal place to start a business, while now other places except for the Bay Area are increasingly attracting entrepreneurs attention, given that the founding time of startups in late funding stage is much earlier than those in stage A, B and C.

To examine the relationship between location and funding (hypothesis 4), we plotted the total funding received inside and outside of the Bay Area. This analysis showed that although startups in the Bay Area made up 49% of the dataset, they received 62% of the total funding. This suggests that being located in the Bay Area may correlate with securing higher funding.

Conclusion

Some of our findings were inconsistent with our hypothesis, but provided some insight that could support decision making. Overall, our conclusion is that funding is positively related to stage, but doesn't appear to have a strong correlation with growth score. This suggests that although funding is important, it is not the only factor that impacts the growth score. With regards to location, although most of the late stage startups were found in the Bay Area, and more funding was allocated to the Bay Area as a whole, for all other stages there were more located outside of the Bay Area. This suggests that although there is a high density of successful startups in the Bay Area across all stages, there are lots of successful early to mid stage startups located outside the Bay Area and being in the Bay Area isn't a necessity for a prerequisite for a startup to be successful. With the development of other areas, there is a possibility that the Bay Area will gradually lose its appeal as a startup paradise.

References

Columbus, Louis. 2015. “The Top 100 Analytics Startups of 2015.” Forbes. Forbes. August 8, 2015. https://www.forbes.com/sites/louiscolumbus/2015/08/08/the-top-100-analytics-startups-of-2015/.

Attachments

AnalysisofTop100Startups.ipynb