ANALYSIS OF START-UP COMPANIES

**Group 2**

Ashraf Adel 196280

Farah Aymen 194233

Aissatou Aida 197346

Table of Contents

[Phase 1 2](#_Toc102858389)

[Introduction 2](#_Toc102858390)

[Datasets 2](#_Toc102858391)

[CSVs Used 2](#_Toc102858392)

[Features (Columns) 3](#_Toc102858393)

[Samples 3](#_Toc102858394)

[Acquisitions 4](#_Toc102858395)

[Additions 4](#_Toc102858396)

[Companies 4](#_Toc102858397)

[Investments 5](#_Toc102858398)

[Rounds 5](#_Toc102858399)

[Cleaning 5](#_Toc102858400)

[Phase 1 Questions 7](#_Toc102858401)

[Question 1 7](#_Toc102858402)

[Question 2 9](#_Toc102858403)

[Question 3 11](#_Toc102858404)

[Question 4 13](#_Toc102858405)

[Question 5 14](#_Toc102858406)

[Question 6 14](#_Toc102858407)

[Question 7 15](#_Toc102858408)

[Question 8 16](#_Toc102858409)

[Phase 2 17](#_Toc102858410)

[Scraping 17](#_Toc102858411)

[Questions 19](#_Toc102858412)

[Question 1 19](#_Toc102858413)

[Question 2 27](#_Toc102858414)

[Question 2.1 29](#_Toc102858415)

[Question 3 30](#_Toc102858416)

[Question 4 30](#_Toc102858417)

[Hypothesis testing using T test 31](#_Toc102858418)

[**Final Conclusion** 32](#_Toc102858419)

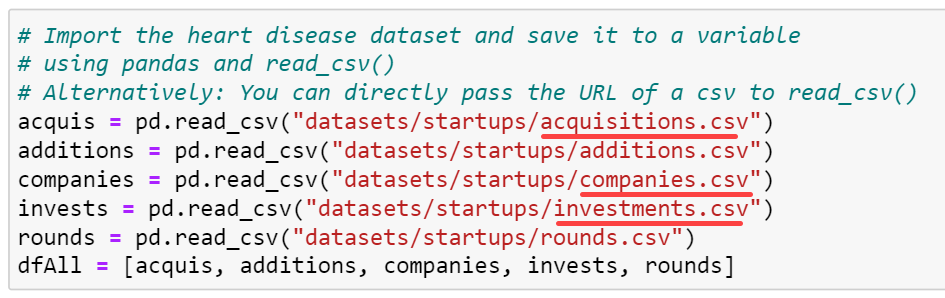
# Phase 1

## Introduction

This project is planned to use a data set on the acquisitions and funding of start-ups. It is aimed to focus on the analysis of the dataset to deduce the reasons behind the success and failure of a start-up. The goal from the analysis is to deduce how much funding on average does a start-up usually require, Reasons behind their failure in case of dissolving or reasons behind their success. The dataset is based on 66368 start-up companies, while some of these start-ups were acquired by other companies, specifically 18969. Moreover, 168647 investments were made in those start-ups. The dataset consists of 5 files: acquisitions, additions, companies, investments and rounds. Each file will contribute to reaching those goals. First, the acquisitions file shows the companies that got acquired, acquired by whom and when, and how much it was acquired for. Second, the companies file shows all information about the start-up companies including their name, location, whether operating or not, and its funding information. Lastly, the investments file shows which company was invested in by which company, how much was paid for it, and the type and amount of funding.

## Datasets

### CSVs Used

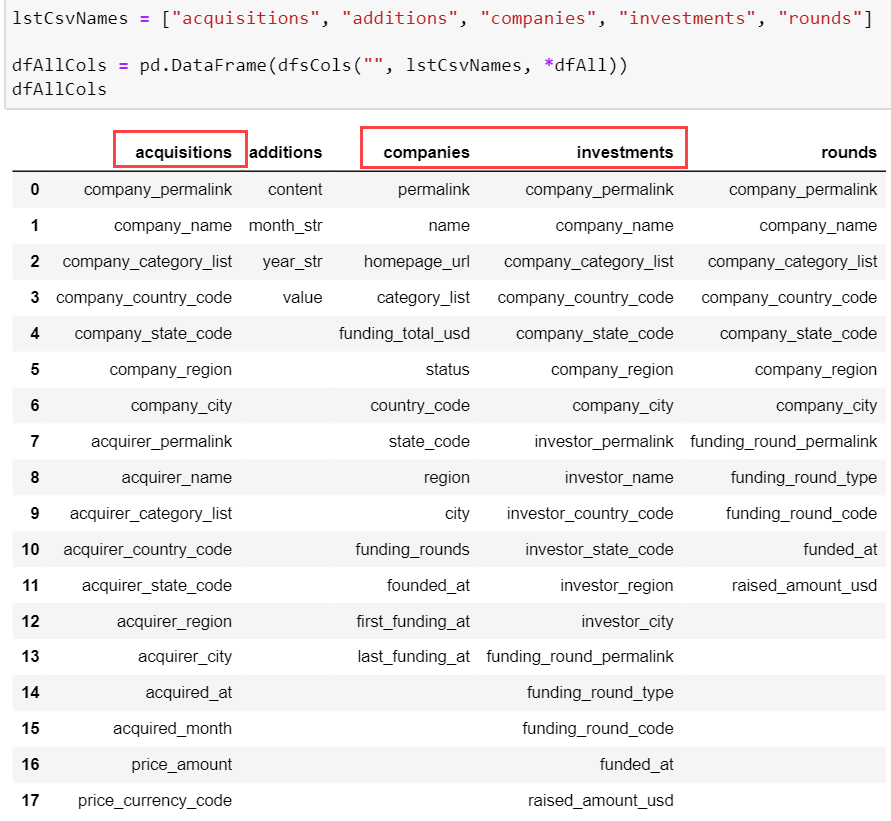




The analysis will mainly focus on the acquisitions, companies, and investments data sets while the others mostly contain redundant values that are already covered in the main datasets.

Python was used to represent samples of the data in a clear manner.

### Features (Columns)

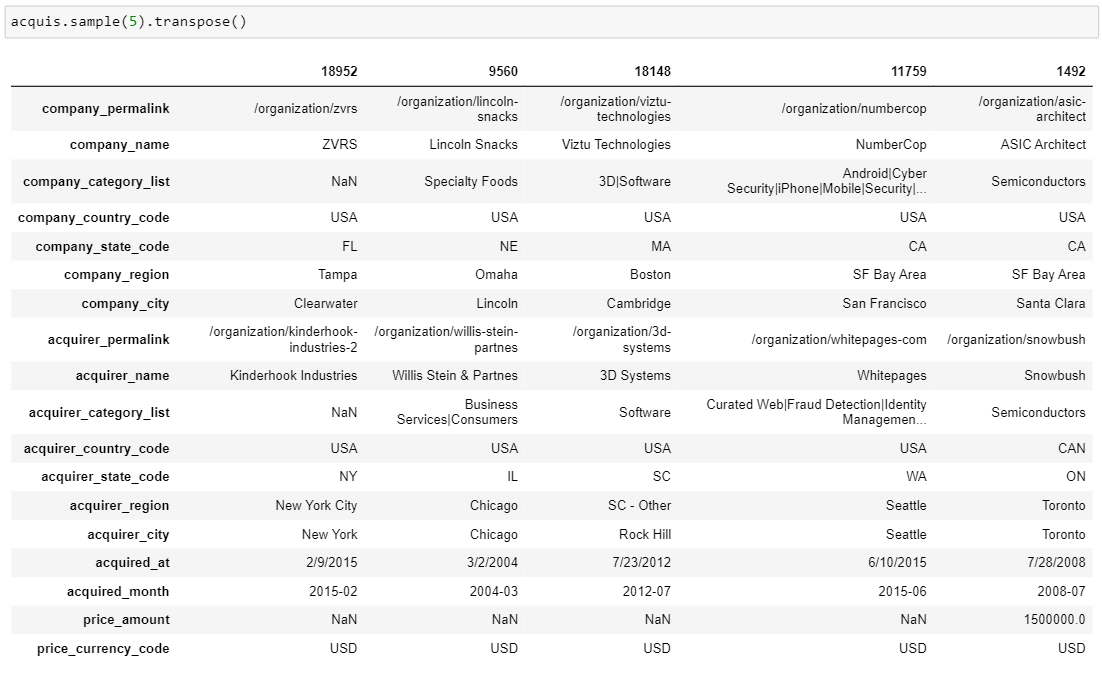


### Samples

5 random samples from each dataset:

The Tables were transposed to make the features clearer.

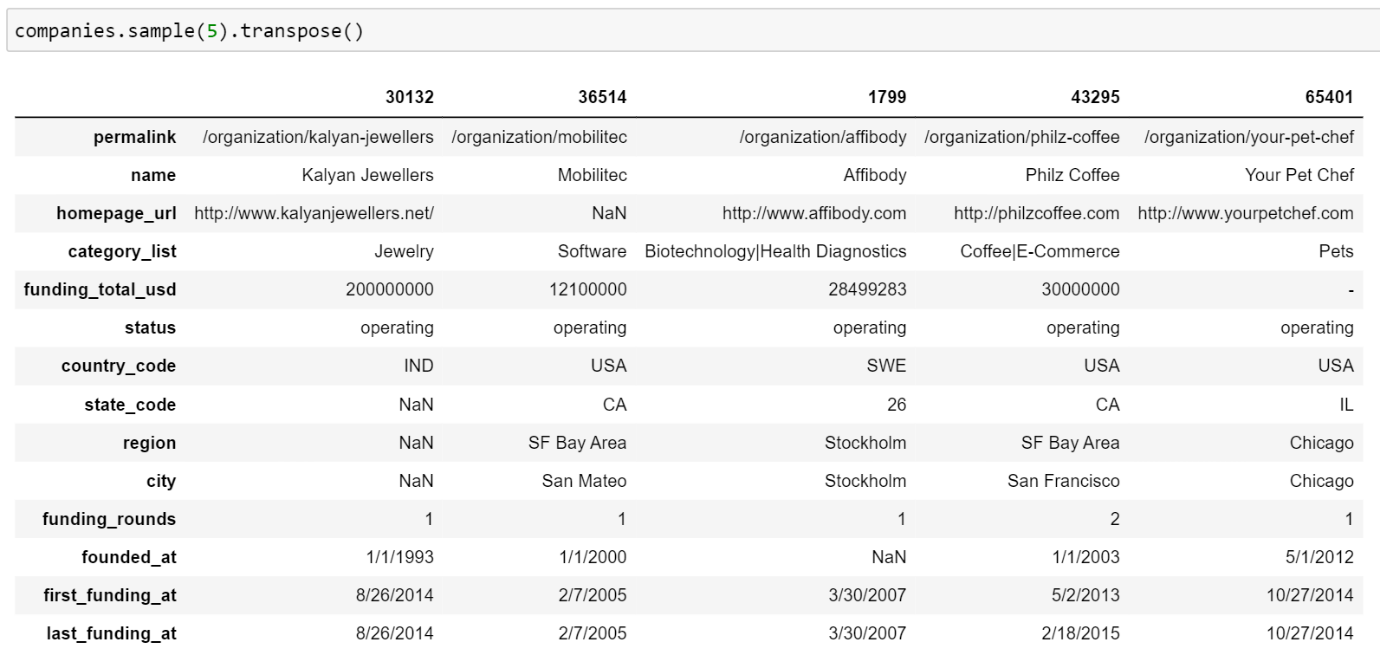
#### Acquisitions



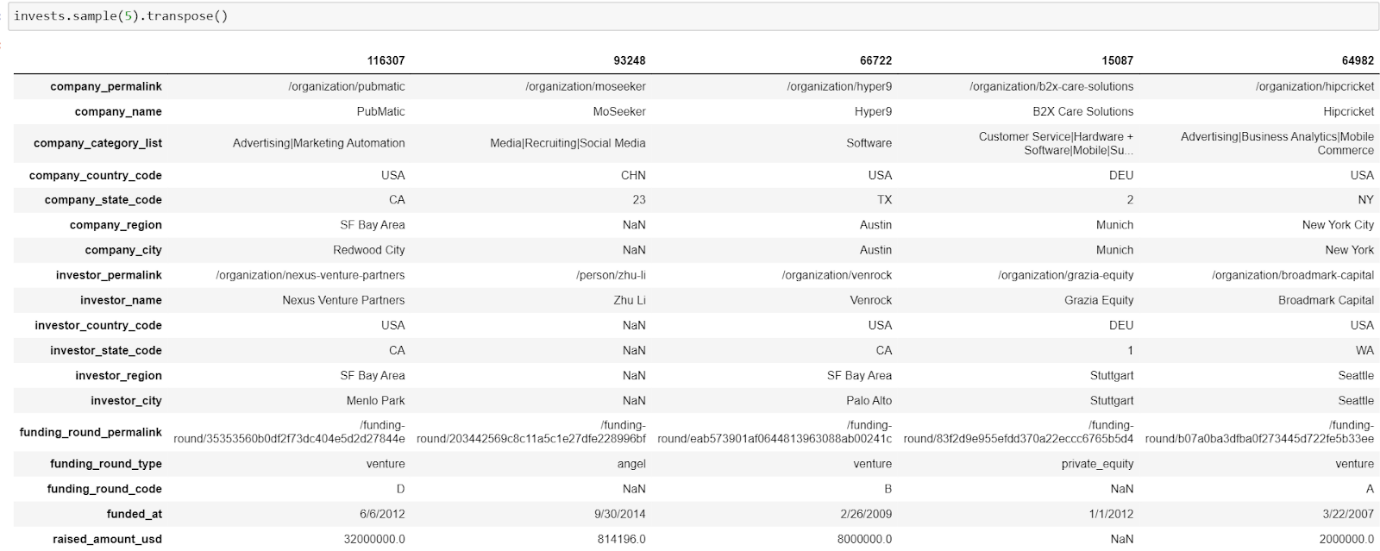
#### Additions



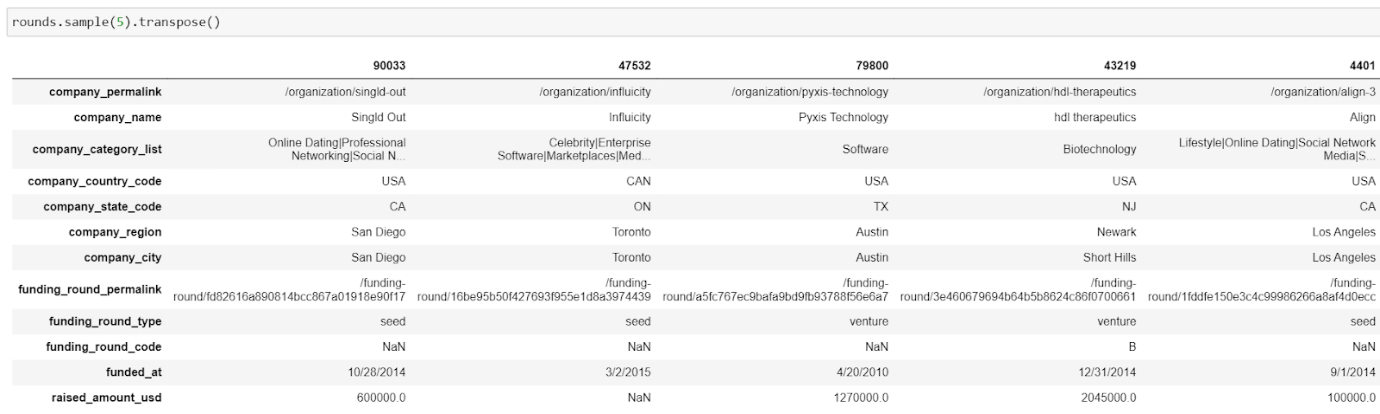
#### Companies



#### Investments

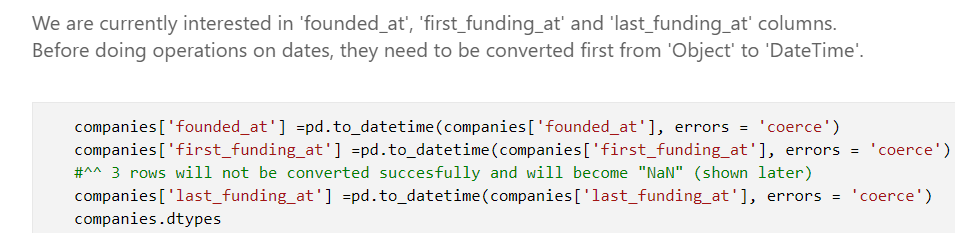
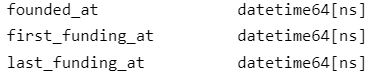
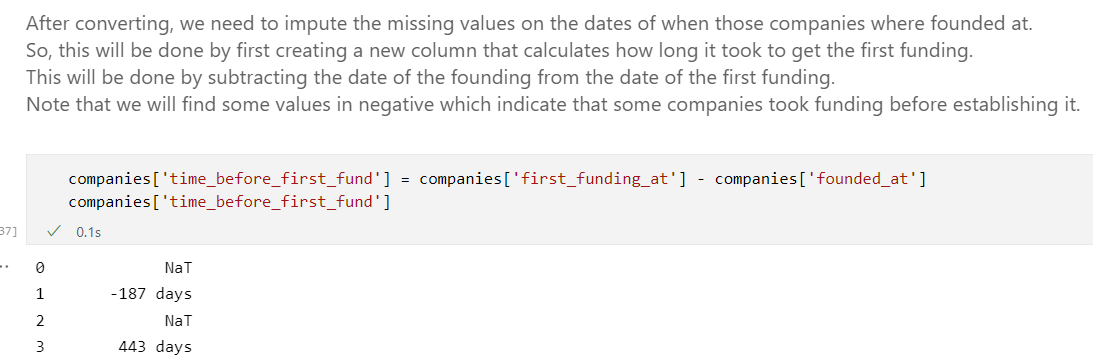
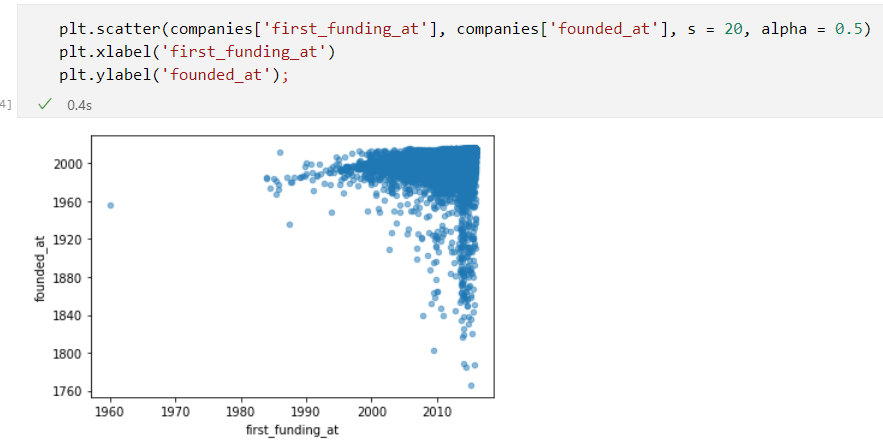
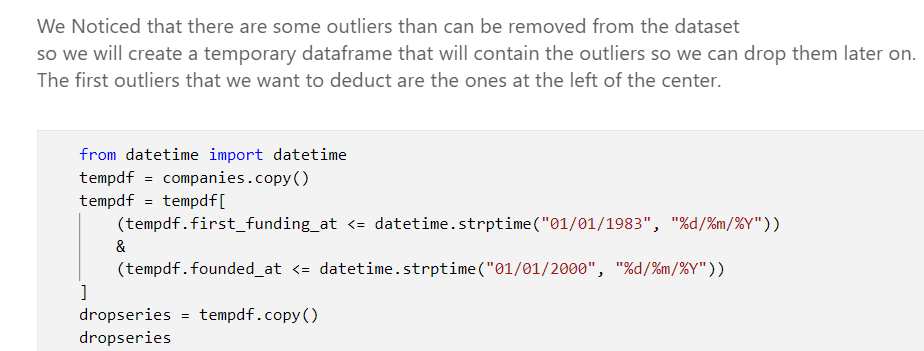
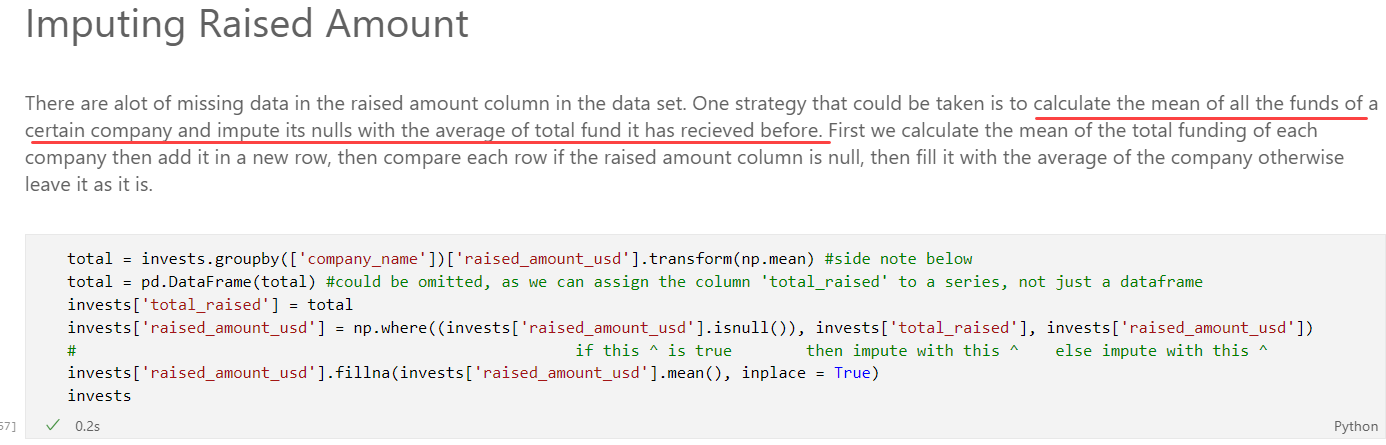
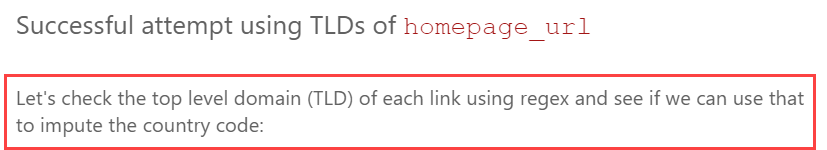
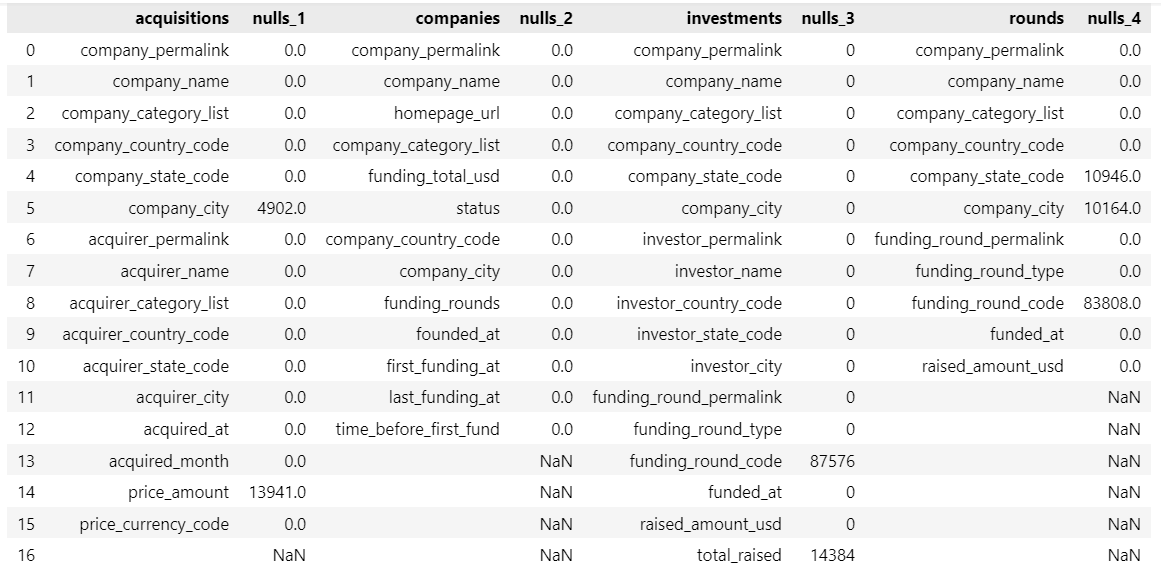


#### Rounds



## Cleaning

The following are the major cleaning steps done on the aforementioned datasets:

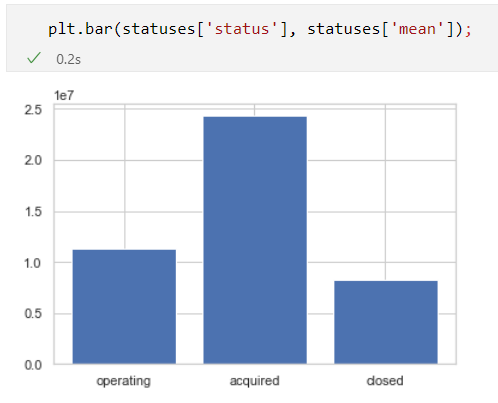
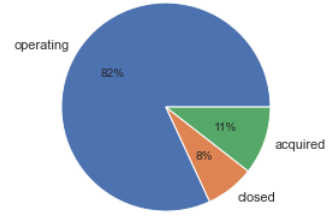
* Converted features to appropriate data types.  
  Example:  
    
  
* Created new feature 'time\_before\_first\_fund' for the reason specified in the markdown below:  
   
* Removing outliers.  
  Example:  
    
    
  
* Imputing values:  
  Example:  
    
  Example 2:  
  
* And other cleaning steps until we obtained the cleaned dataframes:  
  

## Phase 1 Questions

### Question 1

* **Can the factors that affect a start-up’s growth be determined?**
  + - Approach 1: Divide the companies into categories based on their status (either closed, operating or acquired) then look at the average of funding each category receives, then determine the correlation between the closed companies and their failure due to the low funds.
    - Approach 2: the second approach is finding the time it took for a company to receive its first funding regardless of how big or small this funding is. Funding after the company's establishment is assumed to be important on its success, so we'll try to verify this assumption.

Approach 1 main processing steps:

* Joined companies and rounds dataframes
* Grouped by raised\_amount\_usd column in rounds dataframe to sum the total funding for each company (an alternative method was to retrieve the total\_funding\_usd column in companies dataframe)
* Got the average of all funding in operating, closed, and acquired companies:  
    
  Analysis:  
  Initially, it seems that the closed companies have the lowest funding. However, it is important to realize that the data is not distributed equally:  
    
  And since the difference between the funding of operating companies to the closed ones is only 3 million USD, therefore one can deduce that the closed companies relatively had a lot of funding, yet they weren't able to continue operating and closed.  
  Conclusion:  
  Not getting enough **funding is not a critical factor on the company's chance of success**. This means that the companies shouldn't be demotivated if they received low funding.

Approach 2 main processing steps:

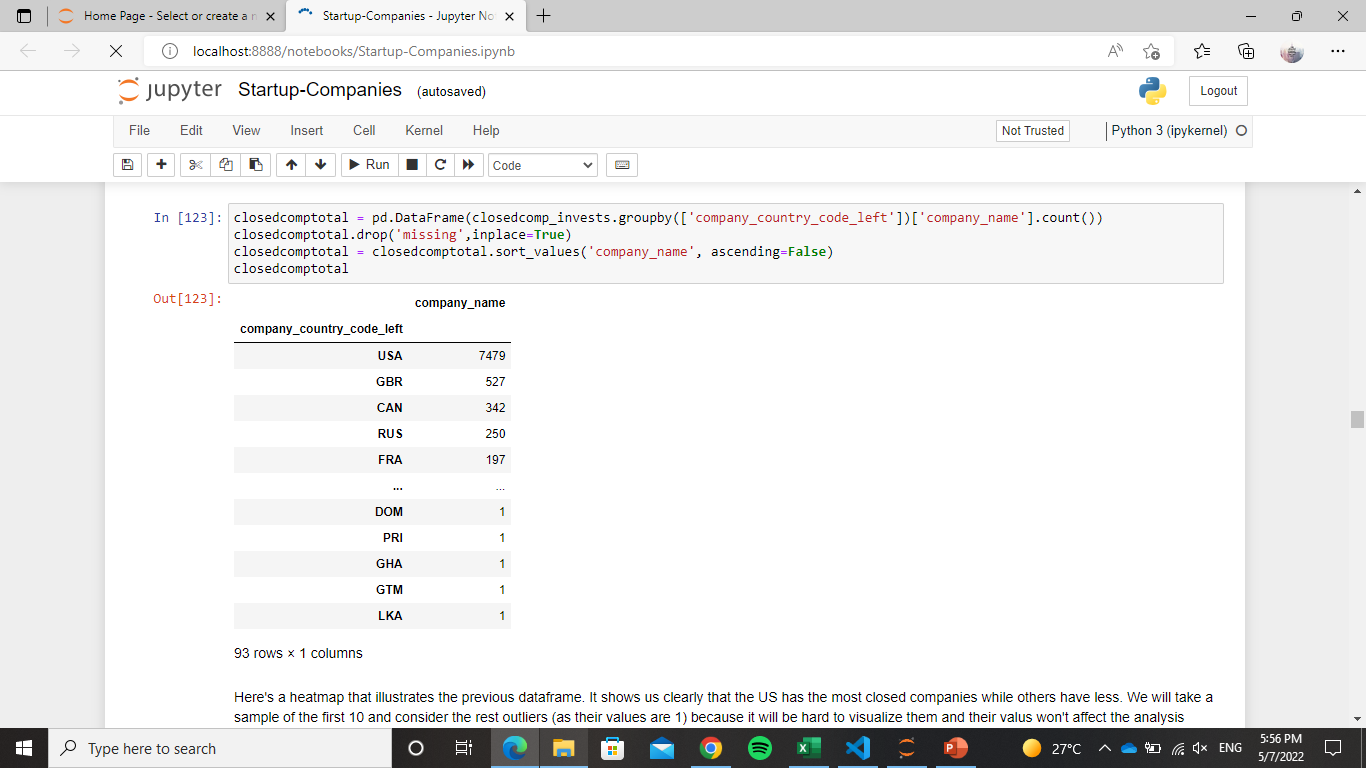
note: “time\_before\_first\_fund” column in “companies” dataframe is the difference (in days) between the first funding and the company’s establishment date.

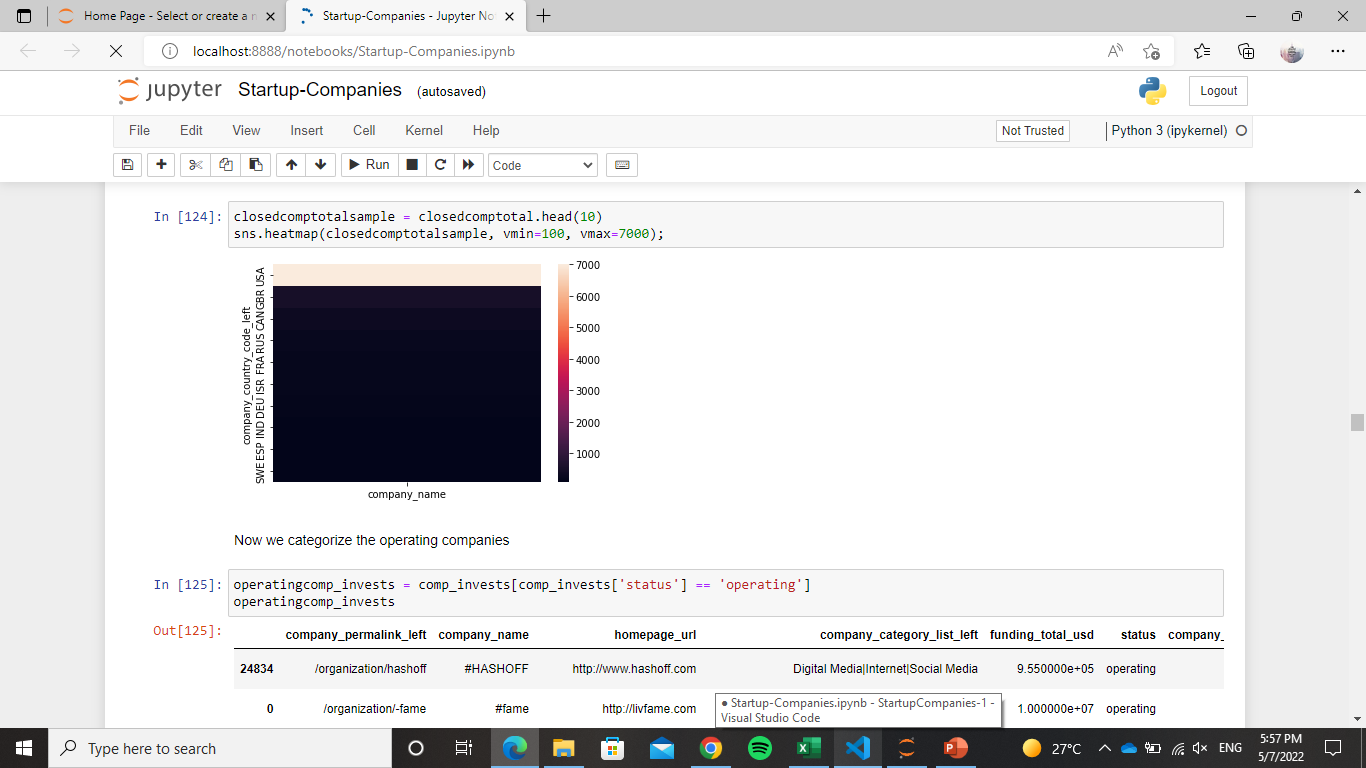
* Split the companies/rounds merged dataframe into 2: one for operating companies and one for closed companies
* sorted these 2 dataframes in descending order on “time\_before\_first\_fund” feature
* compared the top 5 companies from both dataframes:  
    
  (the X\_axis - 0.2 is to shift the blue and orange bars in order for them not to overlap)  
  Conclusion:  
  the operating companies had bigger gaps between funding and establishment. so one can safely assume that the time between the company's foundation and the first funding isn't an effective factor.  
  In short, the **time taken to get the first funding** can be considered a **non-critical factor** **in affecting the success of the company**. It means that the company doesn't heavily rely on that first funding to remain operating.

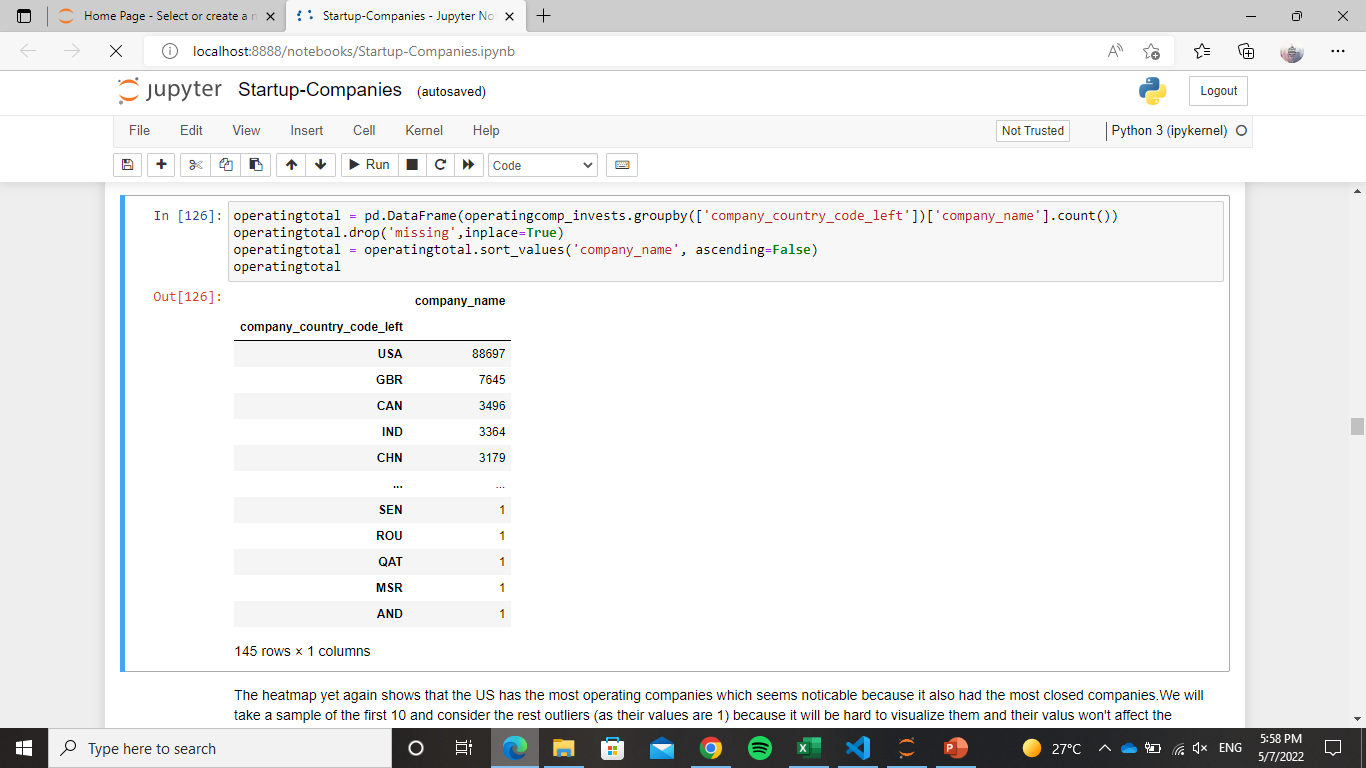
Therefore, the conclude that “raised\_amount\_usd” is an important feature that determines the company’s success, while “time\_before\_first\_fund” is not.

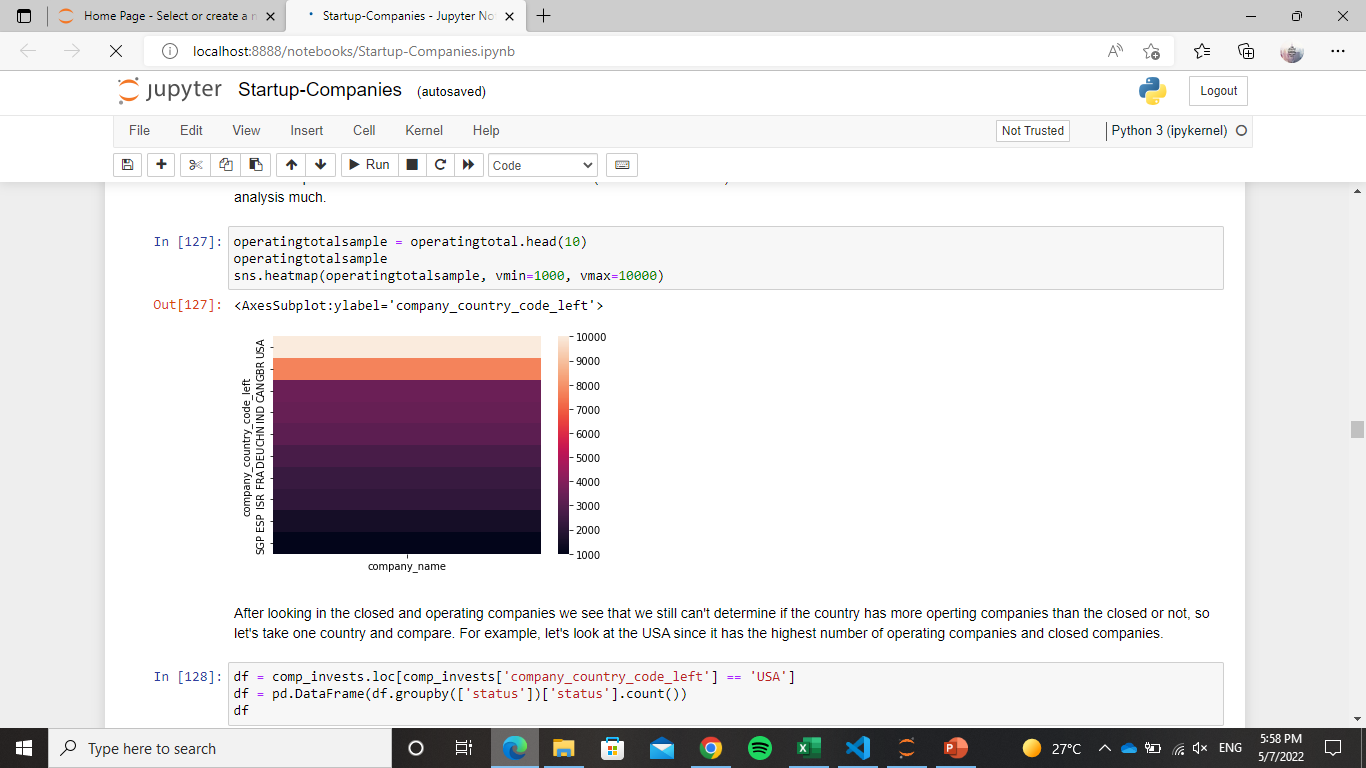
### Question 2

* **Which countries are most probable to have the most failed start-ups?**
  + Correlating the country and a start-up failure, the analysis can show which countries are the hardest to kickstart a company there that may lead to its acquisition or dissolving. Then, it can be linked with the reasons behind this difficulty that makes the start-up fail. Difficulties can be presented in lack of funding, low funding frequency, or other external factors. Additionally, a comparison can be deduced by forming a ratio of the number of closed start-ups to total companies in that specific region.
* **Analysis**
  + We will group by the country to see the number of companies (of each status) in each country to gain insight if there's a massive difference between the number of operating companies and closed ones in a single country.

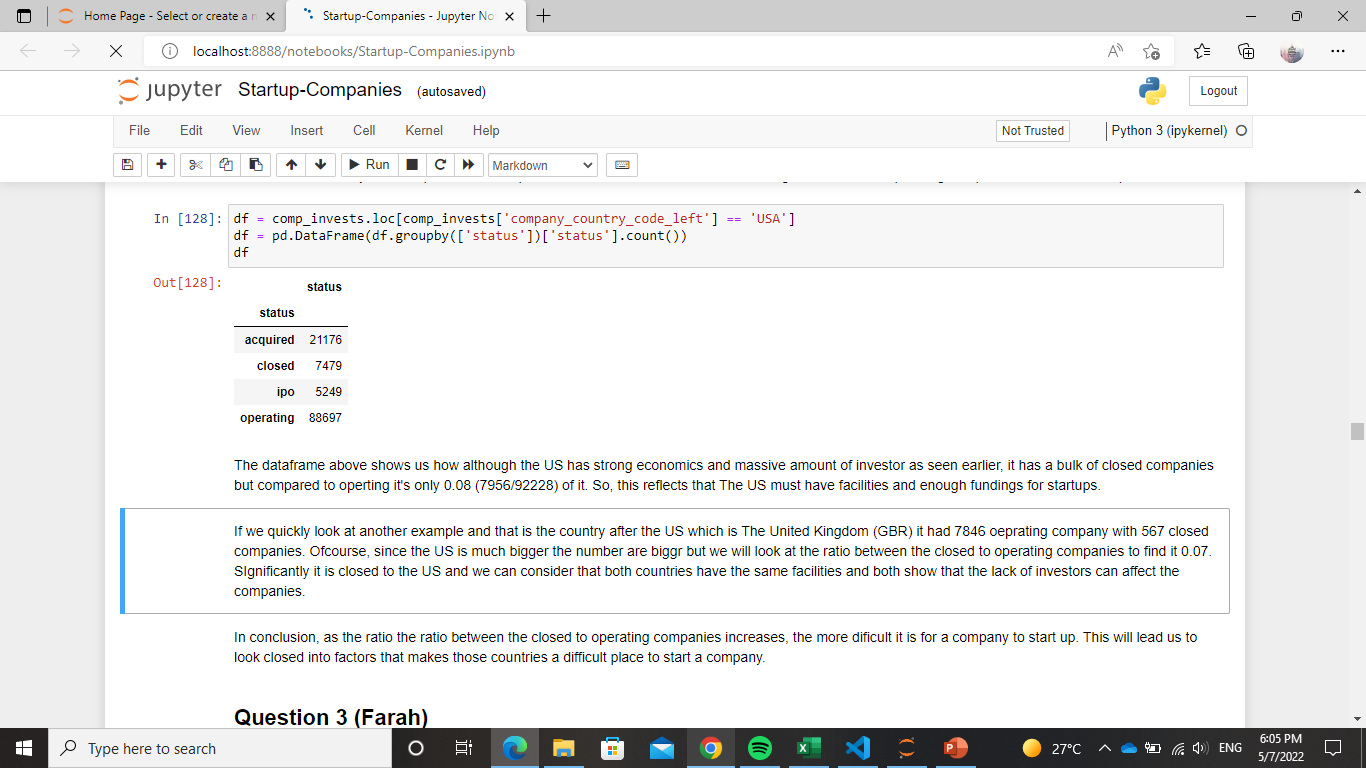








* + After looking in the closed and operating companies we see that we still can't determine if the country has more operating companies than the closed or not, so let's take one country and compare. For example, let's look at the USA since it has the highest number of operating companies and closed companies.



* + The dataframe above shows us how although the US has strong economics and massive amount of investor as seen earlier, it has a bulk of closed companies but compared to operating it's only 0.08 (7956/92228) of it. So, this reflects that The US must have facilities and enough funding for start-ups.

* + If we quickly look at another example and that is the country after the US which is The United Kingdom (GBR) it had 7846 operating company with 567 closed companies. Of course, since the US is much bigger the number are bigger but we will look at the ratio between the closed to operating companies to find it 0.07. Significantly it is closed to the US and we can consider that both countries have the same facilities and both show that the lack of investors can affect the companies.
* **Conclusion**

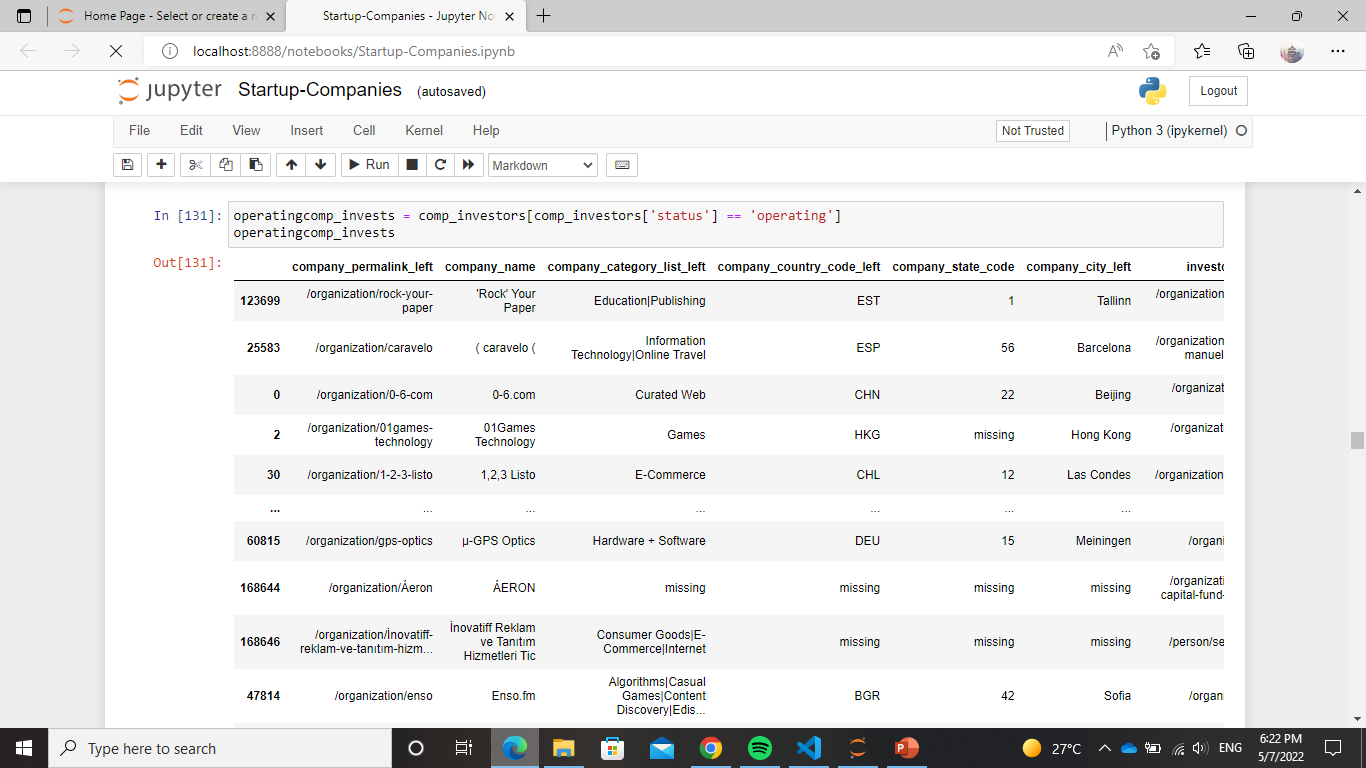
* + In conclusion, as the ratio the ratio between the closed to operating companies increases, the more difficult it is for a company to start up. This will lead us to look closed into factors that makes those countries a difficult place to start a company.

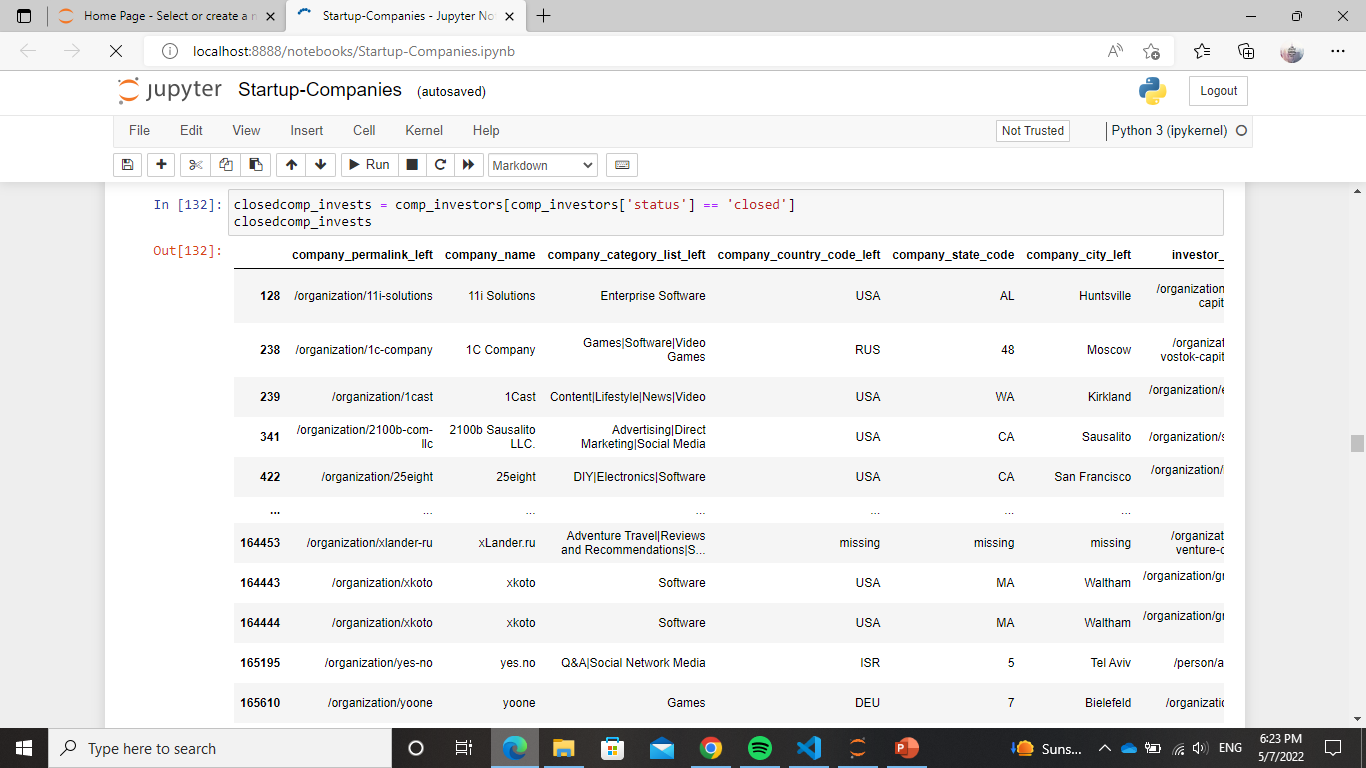
### Question 3

* **Does the number of investors in a country affect the number of successful companies in this country?**
* **Analysis**
  + First, We will narrow down to the investors that invested in the companies in the same country because there's a possibility that an investor invested in a company in a different country, and merge the new dataset with companies so we can get the status.

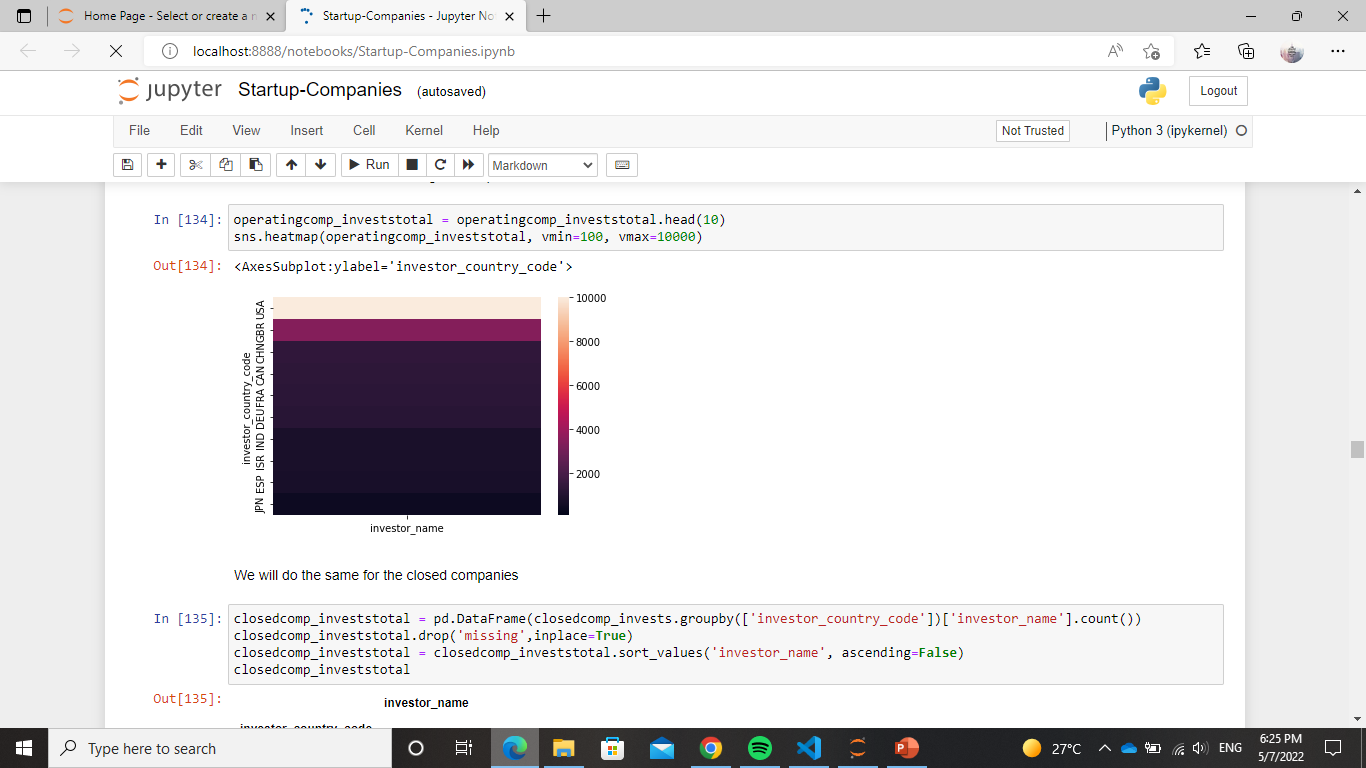


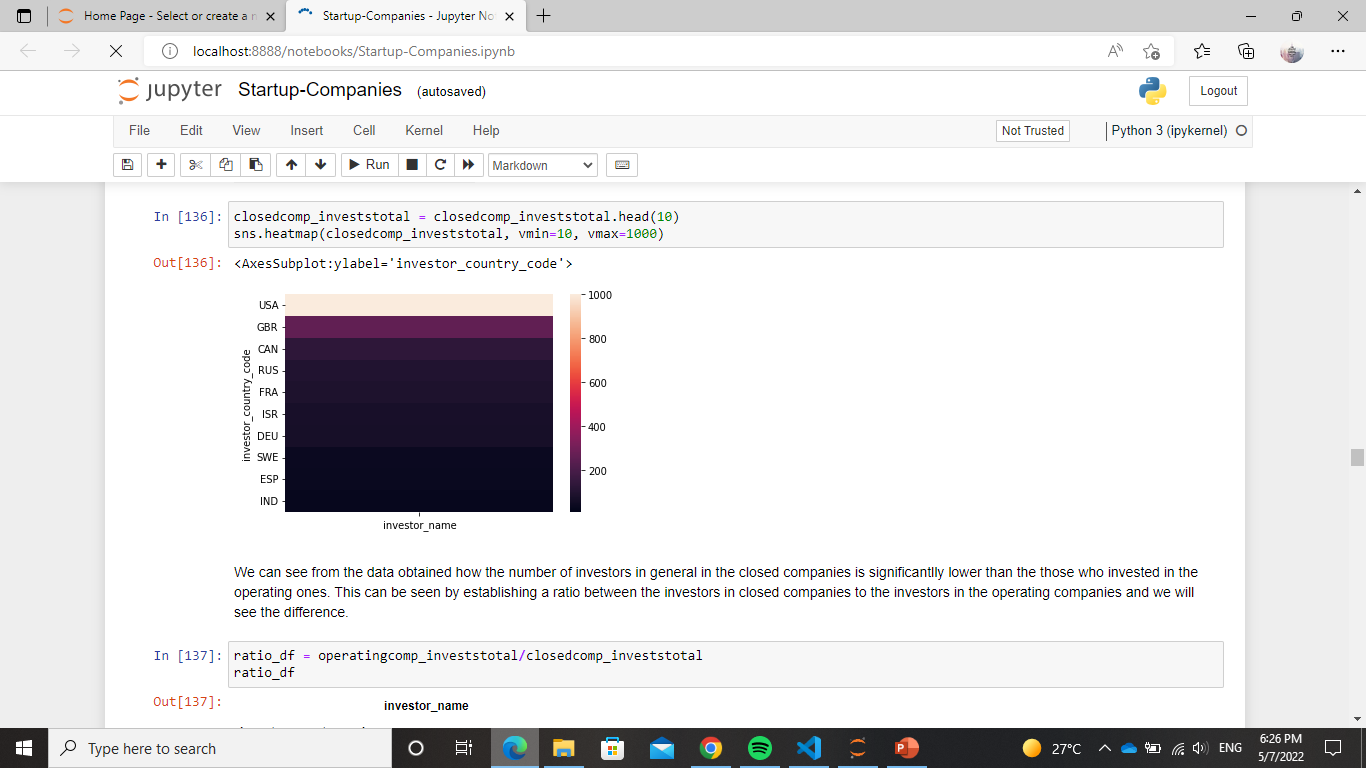
* + Categorizing the companies



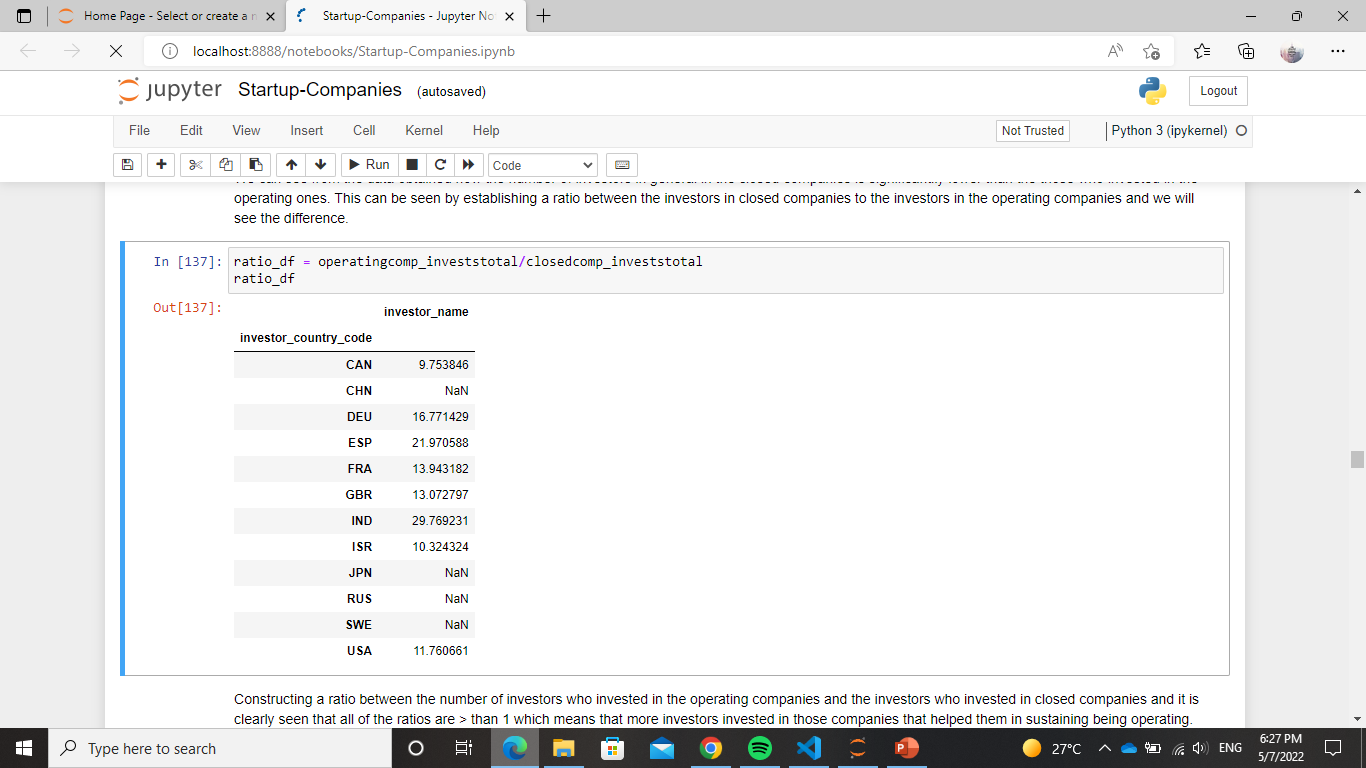


* + Next, we will group by the investor's country to see how many investors in each country invested in the operating companies.





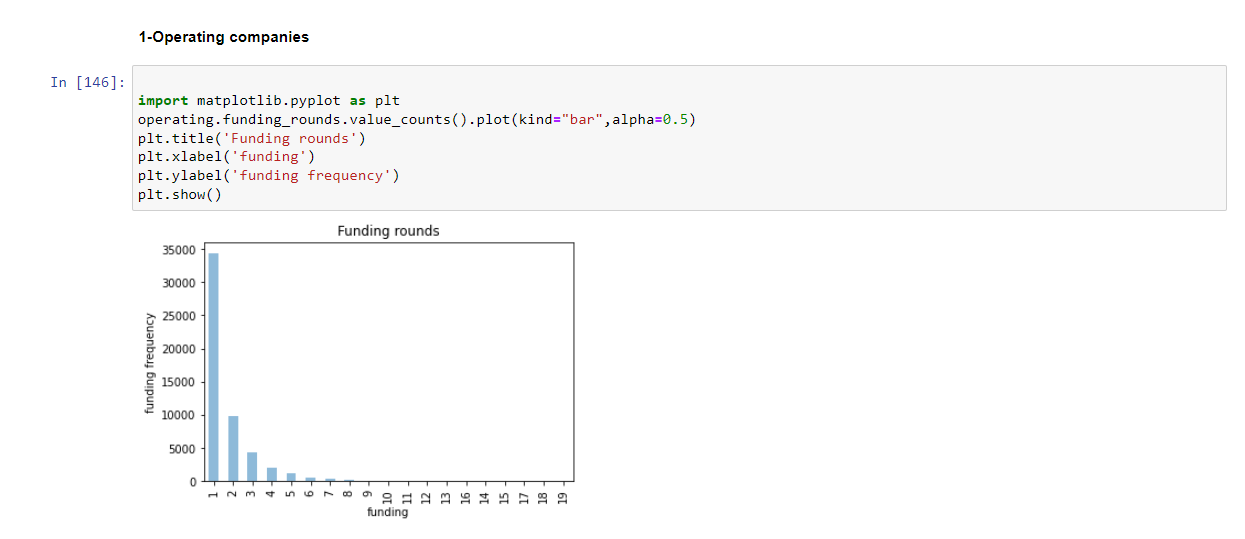
* + We can see from the data obtained how the number of investors in general in the closed companies is significantly lower than the those who invested in the operating ones. This can be seen by establishing a ratio between the investors in closed companies to the investors in the operating companies and we will see the difference.



* + Constructing a ratio between the number of investors who invested in the operating companies and the investors who invested in closed companies and it is clearly seen that all of the ratios are > than 1 which means that more investors invested in those companies that helped them in sustaining being operating.
* **Conclusion**
  + In conclusion, investing plays an important part in keeping the company operating. As the number of investors increases, the chance of success increases.

### Question 4

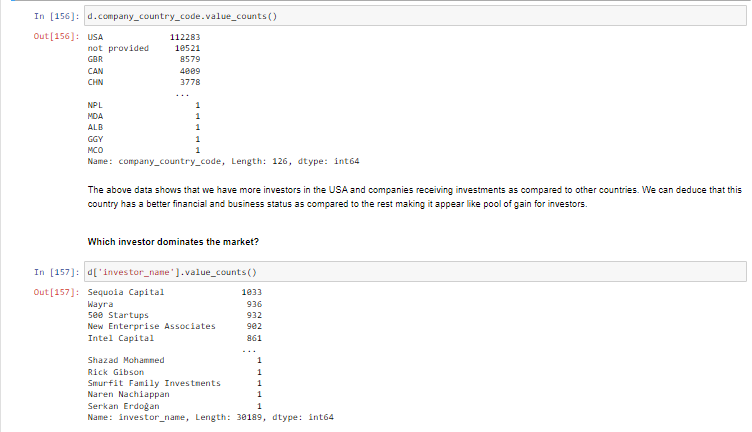
* **Does the amount of funding rounds affect a company's status?**
  + To find out we first checked the amount of companies found in each status category and then we visualised for each category its funding rounds.



After repeating the steps for the other statuses, we can conclude that operating companies exceed in founding rounds but all have higher one time funding.

### Question 5

* **How many investors do we have in different countries and which country dominates?**

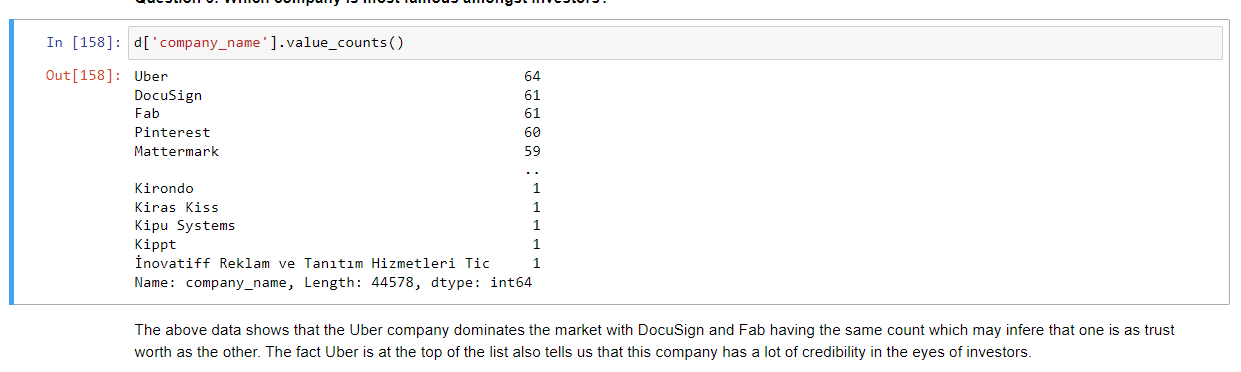
We were able to answer this question by checking the amount of times country codes repeat in the investor tables and the company country location names which could tell a little more about the country's financial and business status. From there we were also able to find out which of these investors dominates the market.

As we can see, although biotechnology is more of interest to investors, companies seem to have more interest in software which could be due to the expensive and delicate nature of biotechnology.

### Question 6

* **Which company is most famous amongst investors?**

Not always are companies which are famous on the market are of interest to investors. This was proved as shown below.



### Question 7

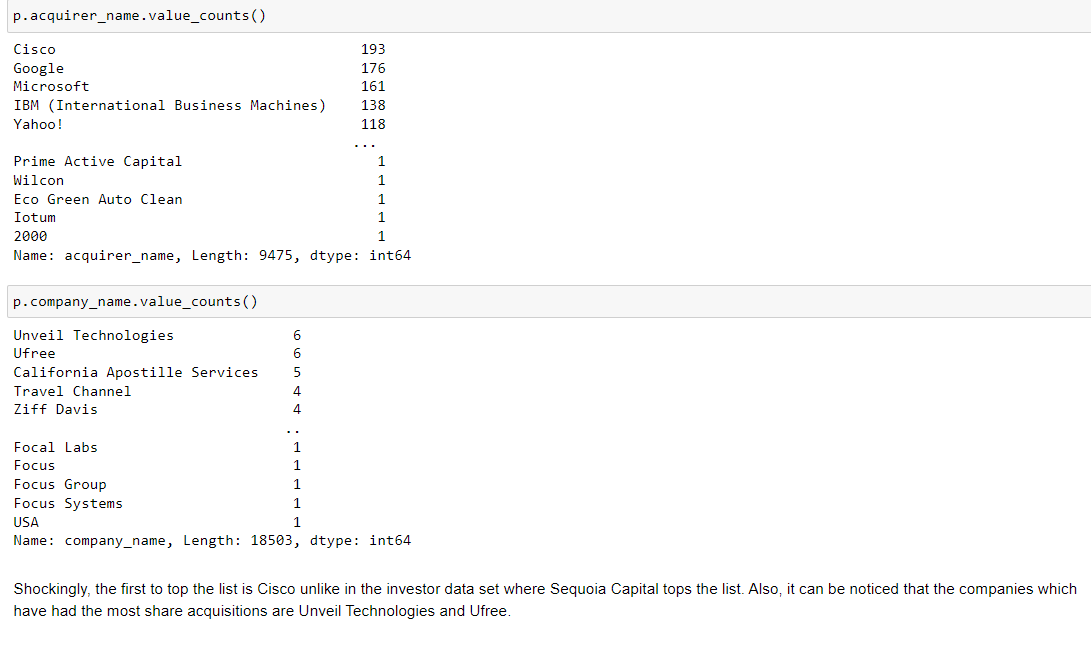
* **What about these company's specialisation as compared to investment? Which category (field) attracts more funding?** 
  + - Every start-up is built upon an idea or a category, and each category can attract interested investors. The data will be analysed to determine which companies under which category attract the most funding, and then it would show which category has the potential and is an interest to investors.
    - We will check the category list for both companies and investments dataframes to answer these questions.



### Question 8

* **Are most of the investors also acquirers?**

Business diversity being a skill most of the time is one of the reasons motivating investors to be diverse. To answer this question, we’ll count the frequency of appearance of acquirer names to determine which one of them is also an investor and compare it to the company names to figure out if the same companies are famous with investors too in this data set.

****

# 

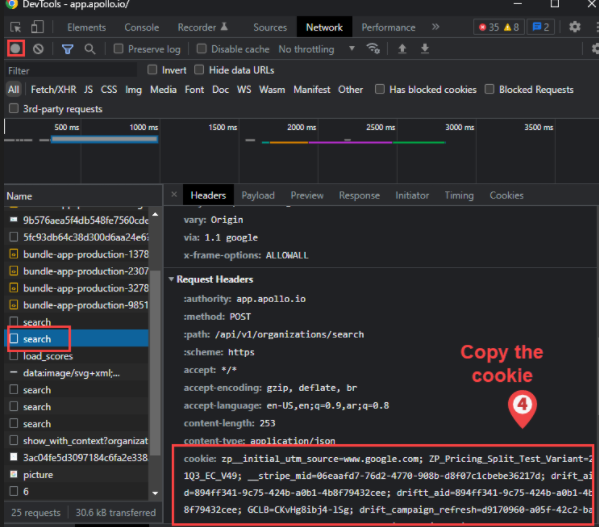
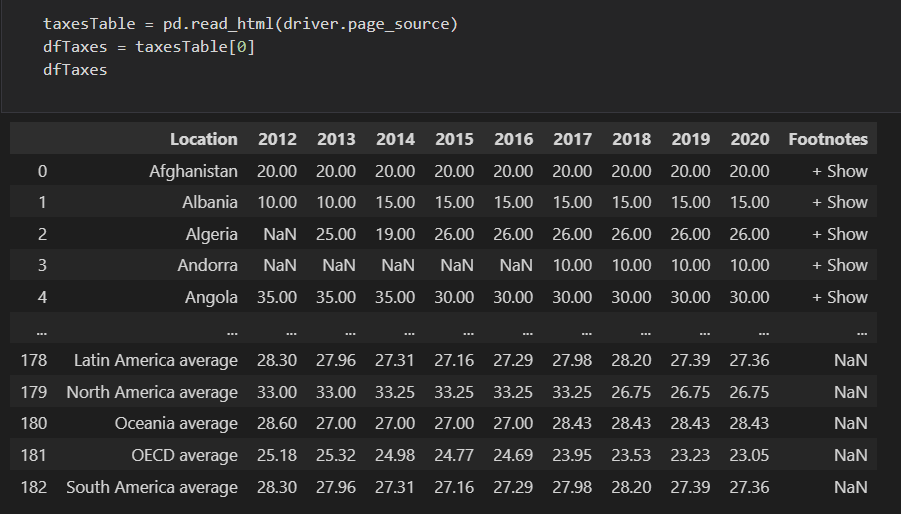
# 

# 

# Phase 2

## Scraping

The following are the scraping steps that will be used throughout the rest of the scraping questions:

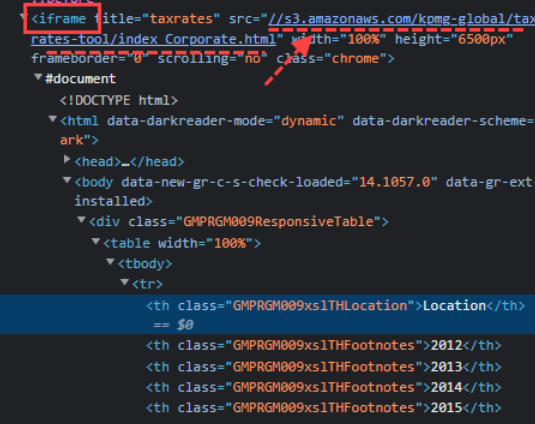
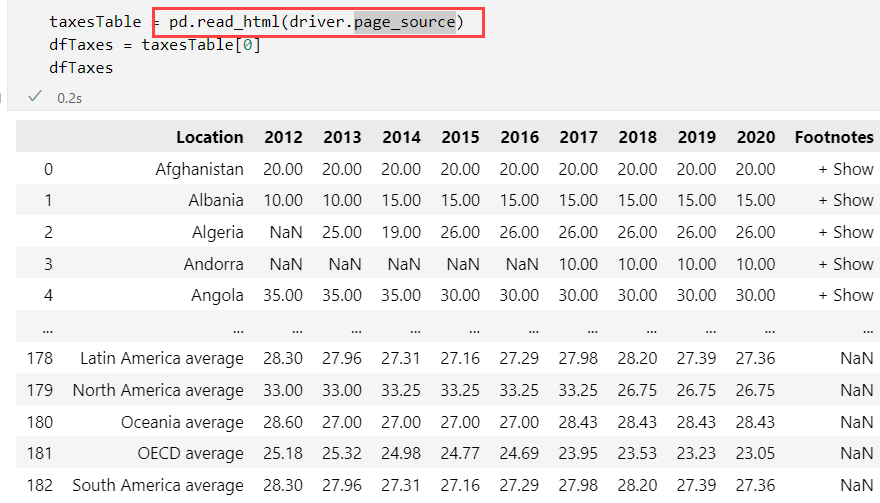
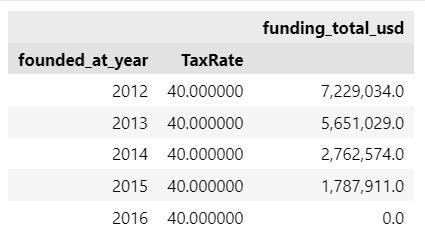
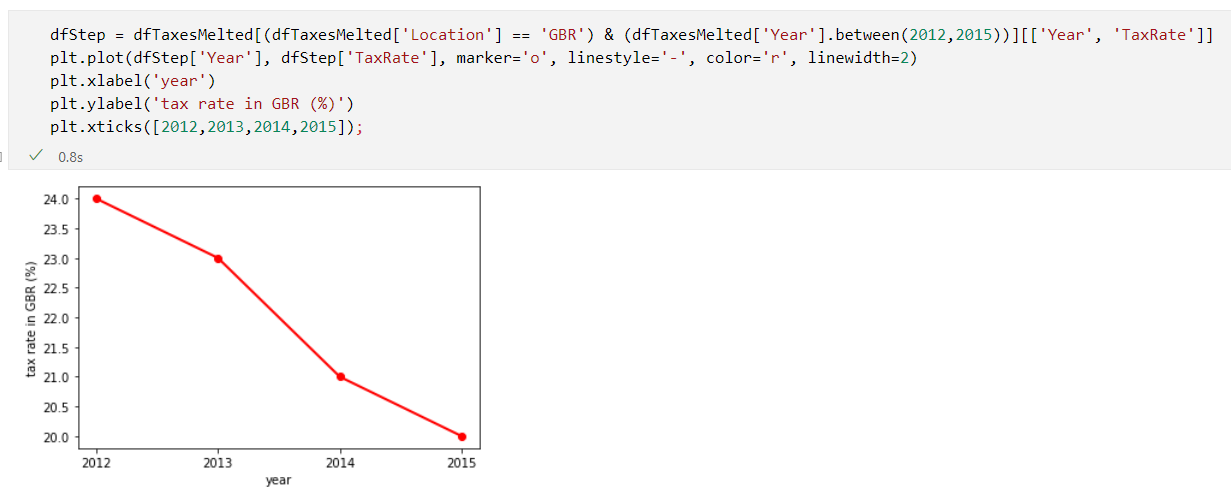
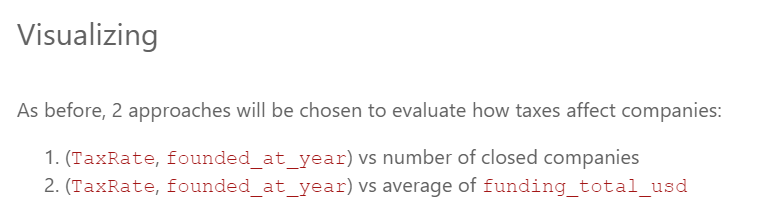
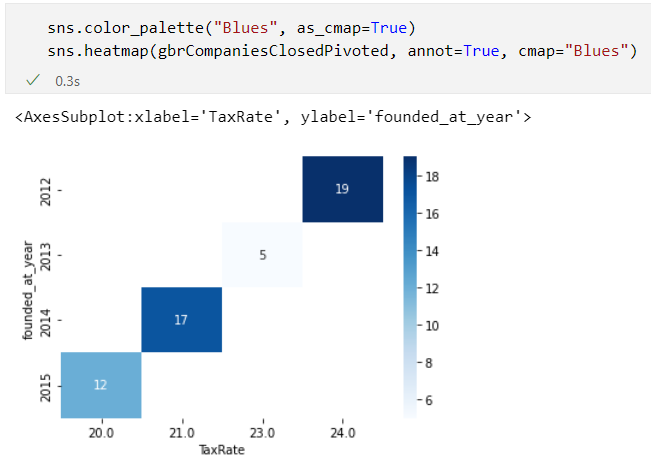
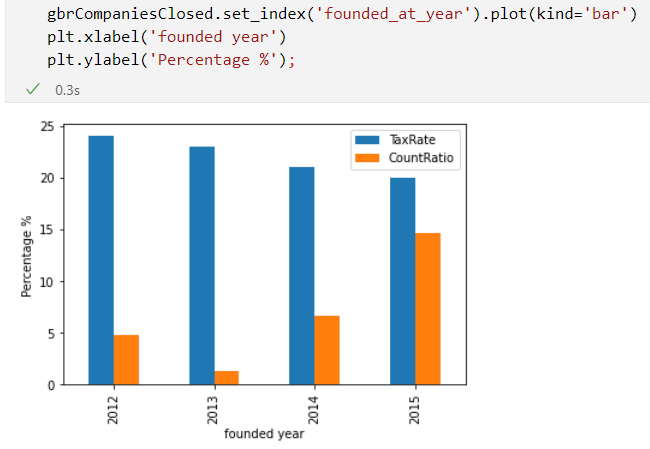
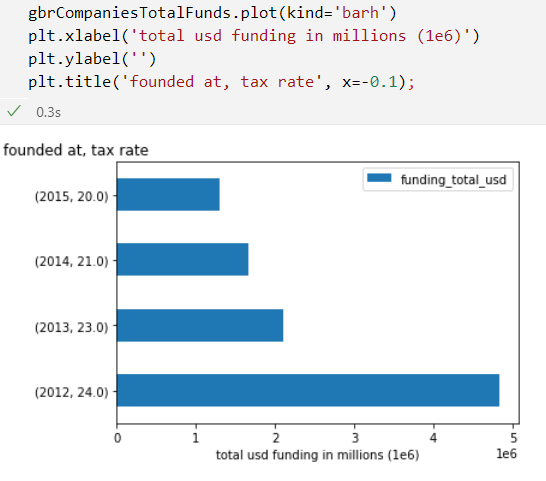
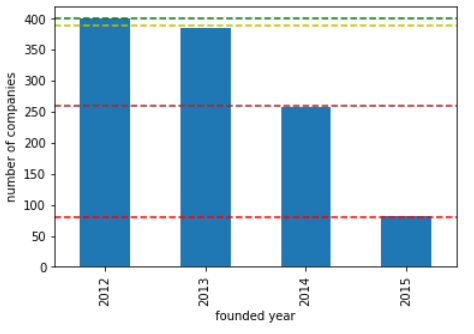
* signed up on Apollo to obtain the cookies used in post and get requests:  
    
    
  (note: full steps are displayed in the markdown of startup-companies.ipynb)
* Initial goal: split the 60k companies to 6 10k dataframes, where each one imputes the company’s number of technologies, the employee count, the company’s headquarters’ location, and its annual revenue by posting requests to Apollo and retrieving this info from the json of the company website retrieved. Also, to facilitate the process (in term of time) we requested each 10 companies simultaneously using threads. However, due to complications mentioned in the notebook file (e.g. getting blocked, one of the group members had poor internet connection, etc) we were only able to retrieve around 20k companies, and so questions 2 and 3 uses these 20k companies, while question 1 used the original cleaned dataset of 60k companies, and used the new 20k dataset just to impute some missing country codes.
* Also, regarding the first question, the KPMG website was used to retrieve information about the countries’ tax rates from 2012 to 2020:  
  part of the tax rate table found in KPMG:  
    
  after converting to dataframe:  
    
  (Note: again, the full steps are illustrated in the notebook’s markdown)

## Questions

### Question 1

* **Do taxes play an important factor in deciding a start-up’s success?**
  + This can be determined by gathering information about each country’s tax revenues (found in KPMG website) and correlating the results with the status of a start-up.

Main processing steps:

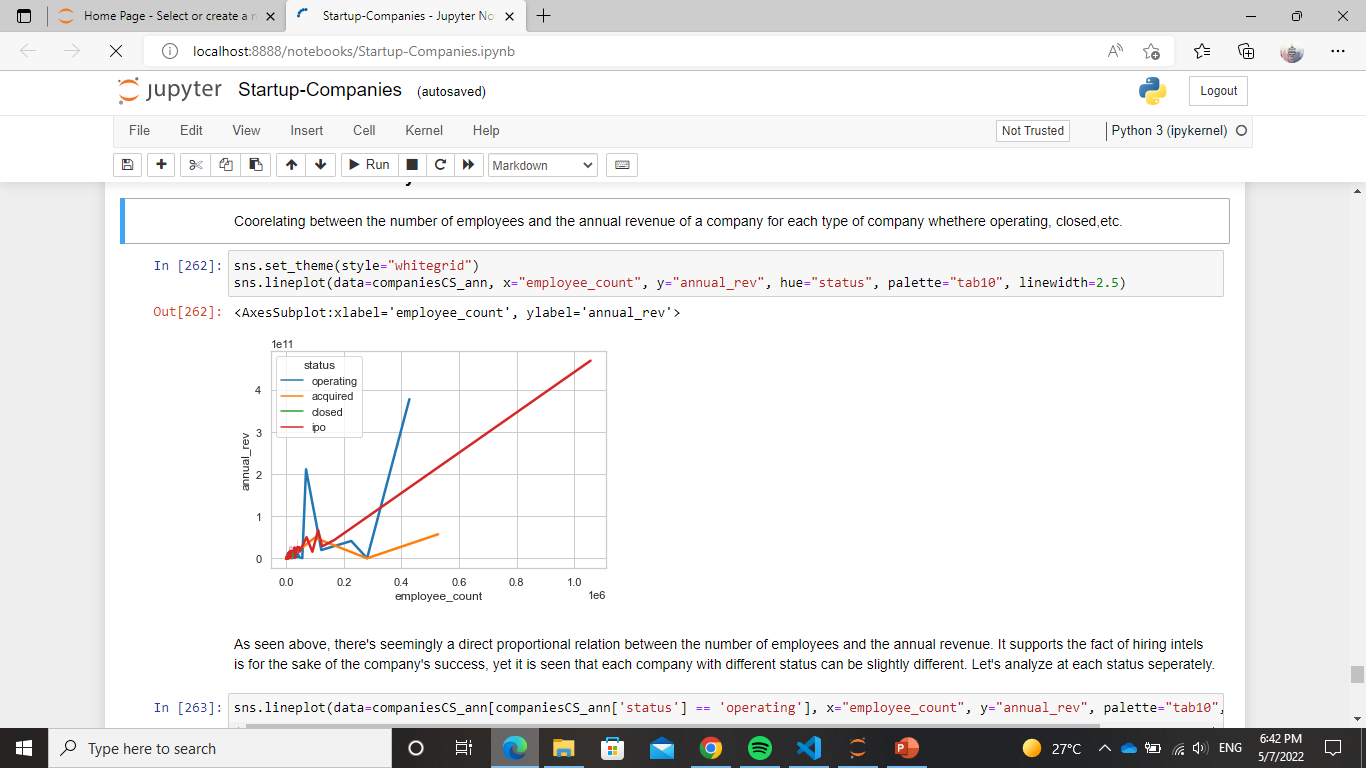
* Scrapped the tax rates from KPMG (as mentioned in the scraping section of this report):  
  However, note that first we retrieved the URL of the table from the iframe tag from main [kpmg website](https://home.kpmg/eg/en/home/services/tax/tax-tools-and-resources/tax-rates-online/corporate-tax-rates-table.html):  
    
  Also, the read\_html() function facilitated the conversion from html table to dataframe:  
  
* Cleaned the scraped data (e.g. removed “Footnotes” feature) and used melt() to make the years form a column instead of being in the dataframe’s headers
* Replaced the missing country codes with the scraped countries from Apollo:  
  
* Merged the companies dataframe with the taxes dataframe:  
    
  (note that left\_on and right\_on are the columns (keys) to be joined on)
* Tried to analyse USA initially, but found that there are no USA companies after 2016 in our dataset (and it was hard to scrape more USA start-up companies, due to getting blocked by Apollo). Also, because the taxes before 2016 were constant:  
  
* Analysed taxes and companies in The United Kingdom (GBR) instead (which are 1k companies founded in the range of 2012 to 2015). Taxes in GBR:  
  
* Visualized based on two features compared with “founded\_at” feature:  
  
* 1st approach:  
    
  Analysis:  
  Notice that even though the tax rate decreases each year, "a lot" of companies closed in 2014 and 2015 when compared to 2012 which had the highest tax rate.   
  However, we have to bear in mind that the 1k companies aren't equally distributed throughout the years:  
    
  Therefore we compared by the ratio of the closed companies to all companies in each year:  
    
  Analysis:  
  When we compare by ratios, still the same insight applies: "a lot" of companies closed in 2014 and 2015 when compared to 2012 which had the highest tax rate.  
  Conclusion:  
  **the welfare of companies isn't determined by the country's current tax rate**.
* 2nd approach:  
    
  Analysis:  
  Again, this is not accurate because the 1k companies aren't equally distributed throghout the years However if we look back at the distribution:  
    
  One can notice that even though 250+ companies were founded in 2014 and only 70+ were founded in 2015 The difference in the total funding is only around 100k USD, which is not much Therefore, these few companies in 2015 got relatively a lot of funding, and at the same time, the taxes went down by 1% Thus, there could be a slight correlation between decreasing the rate of taxes, and increasing the possibility that stakeholders give more funding to the start-up company

To conclude the whole question:

**taxes don't affect the welfare of companies (in terms of status), but have a slight inverse relation with the amount of funding given to companies**.

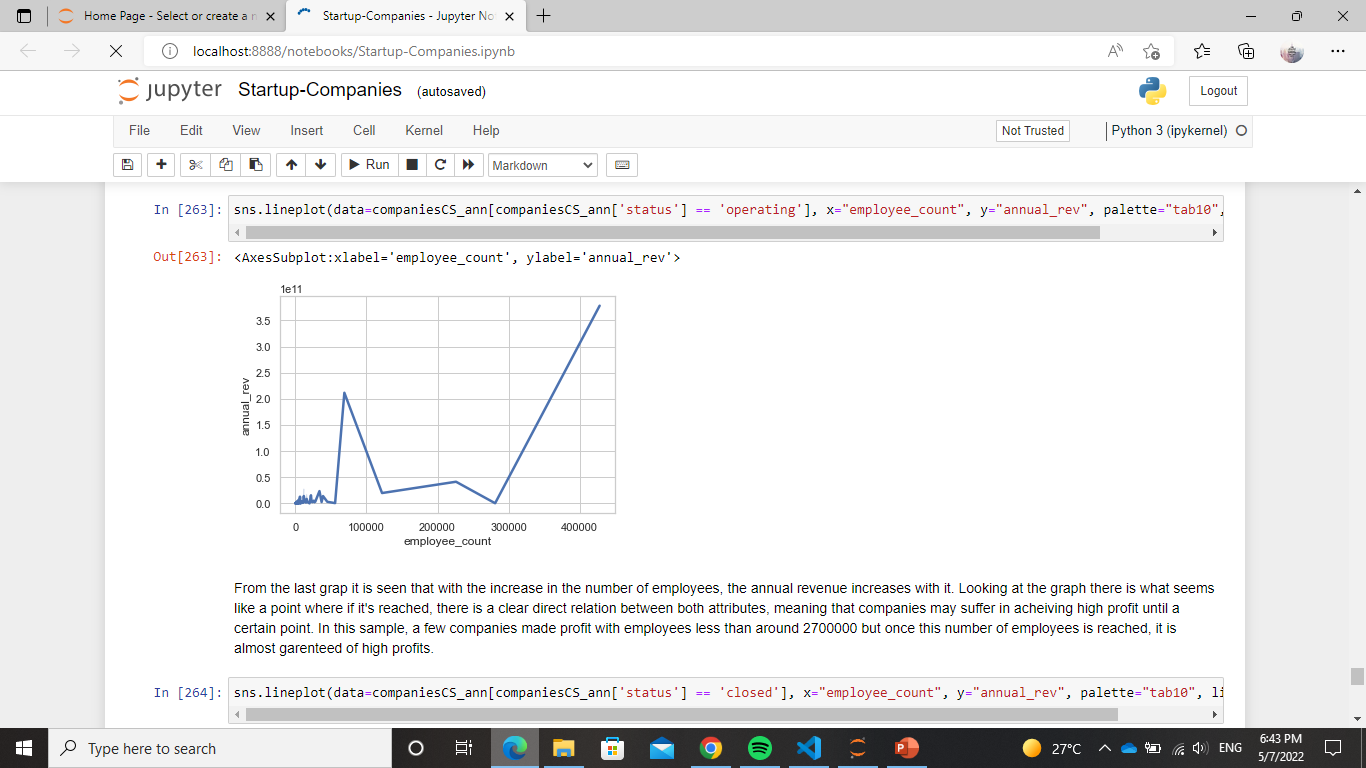
### Question 2

* **Is it possible to determine the suitable number of employees for each companies’ category lists**
  + This can be determined by gathering information about the number of employees from each company’s permalinks and integrating that data in a new column in the dataset.
* **Analysis**
  + Let's see if the relationship between the number of employees is proportional with the success of the company in other words operating or not. Unfortunately, we don't have details on employees' salaries but we have their number in each company. if it's directly proportional then the number of employees present an important asset to building a new company, while if it's directly proportional then it proves that employing many employees can be only a burden to kick off the company
  + Corelating between the number of employees and the annual revenue of a company for each type of company whether operating, closed, etc.



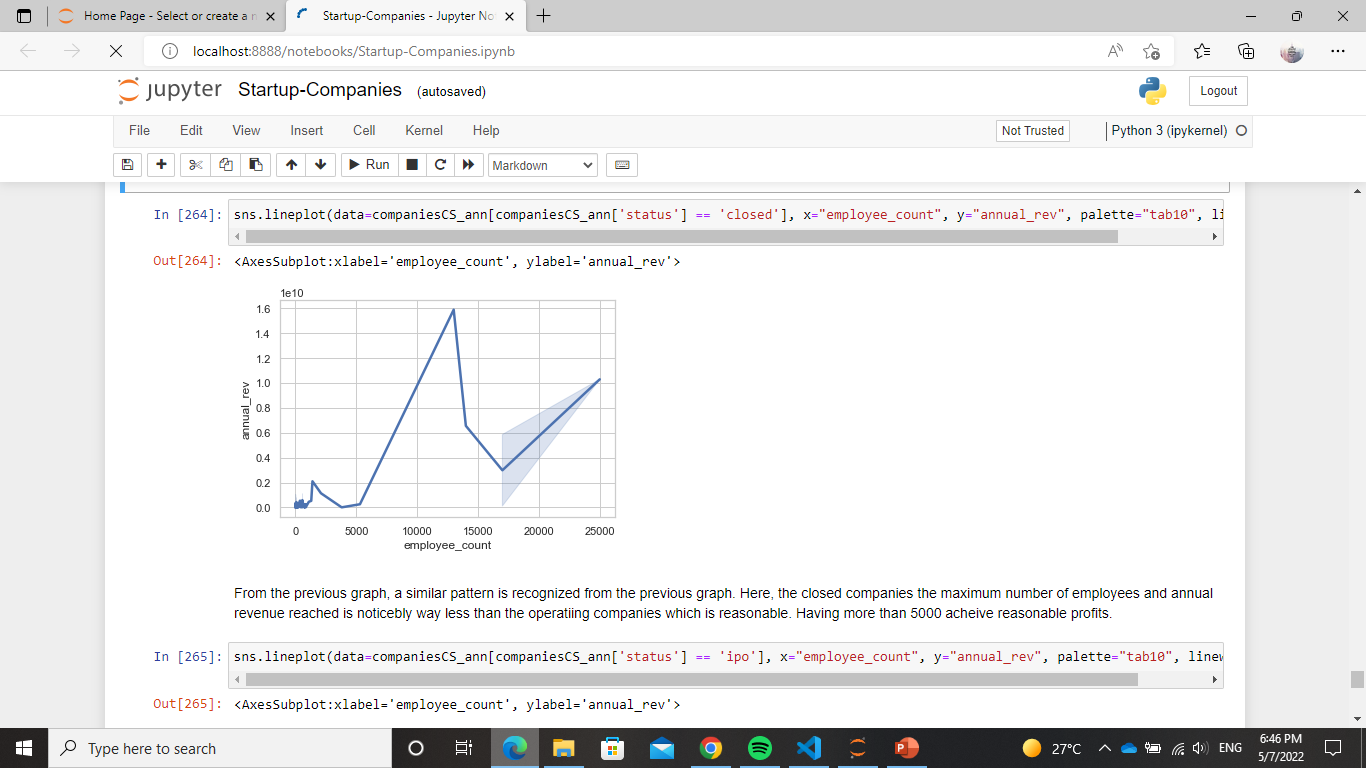
As seen above, there's seemingly a direct proportional relation between the number of employees and the annual revenue. It supports the fact of hiring intel is for the sake of the company's success, yet it is seen that each company with different status can be slightly different. Let's analyse at each status separately.

* + operating companies:



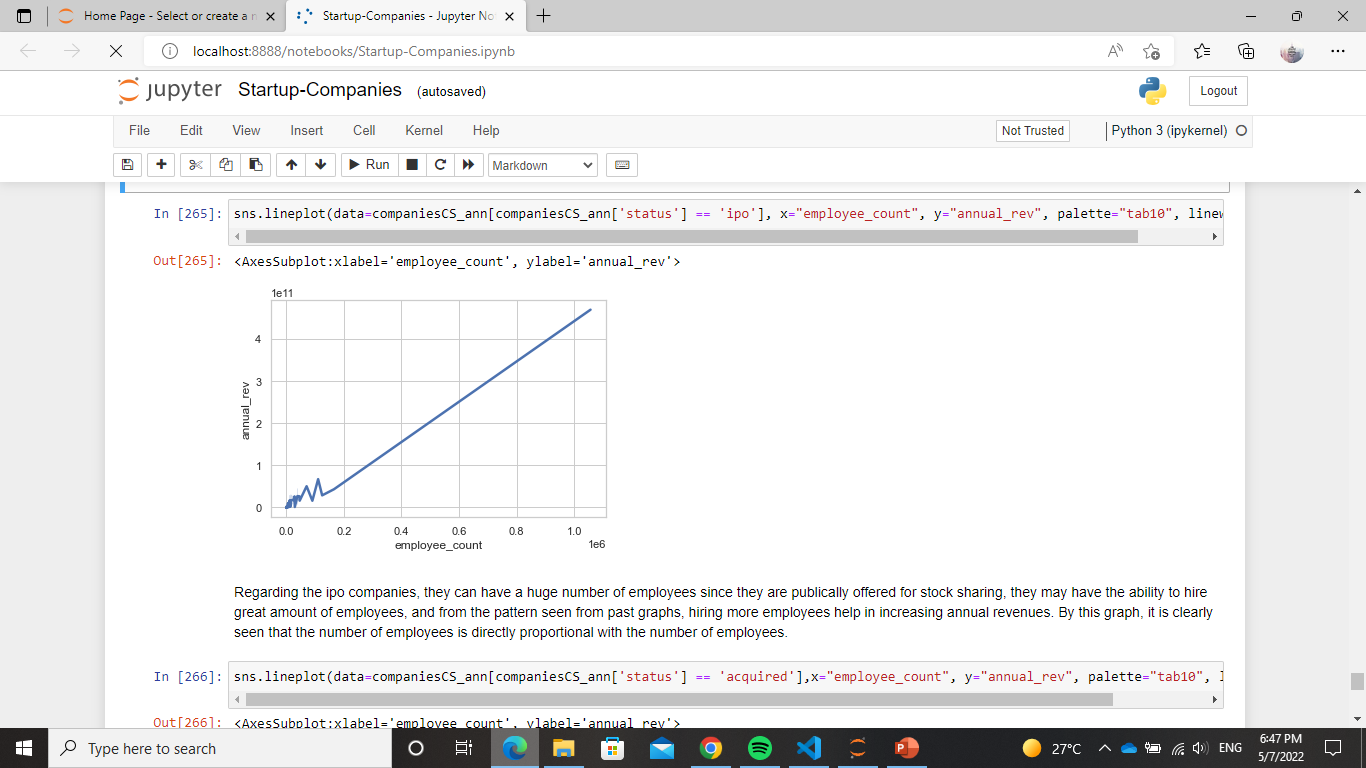
From the last graph it is seen that with the increase in the number of employees, the annual revenue increases with it. Looking at the graph there is what seems like a point where if it's reached, there is a clear direct relation between both attributes, meaning that companies may suffer in achieving high profit until a certain point. In this sample, a few companies made profit with employees less than around 2700000 but once this number of employees is reached, it is almost guaranteed of high profits.

* + Closedcompanies:



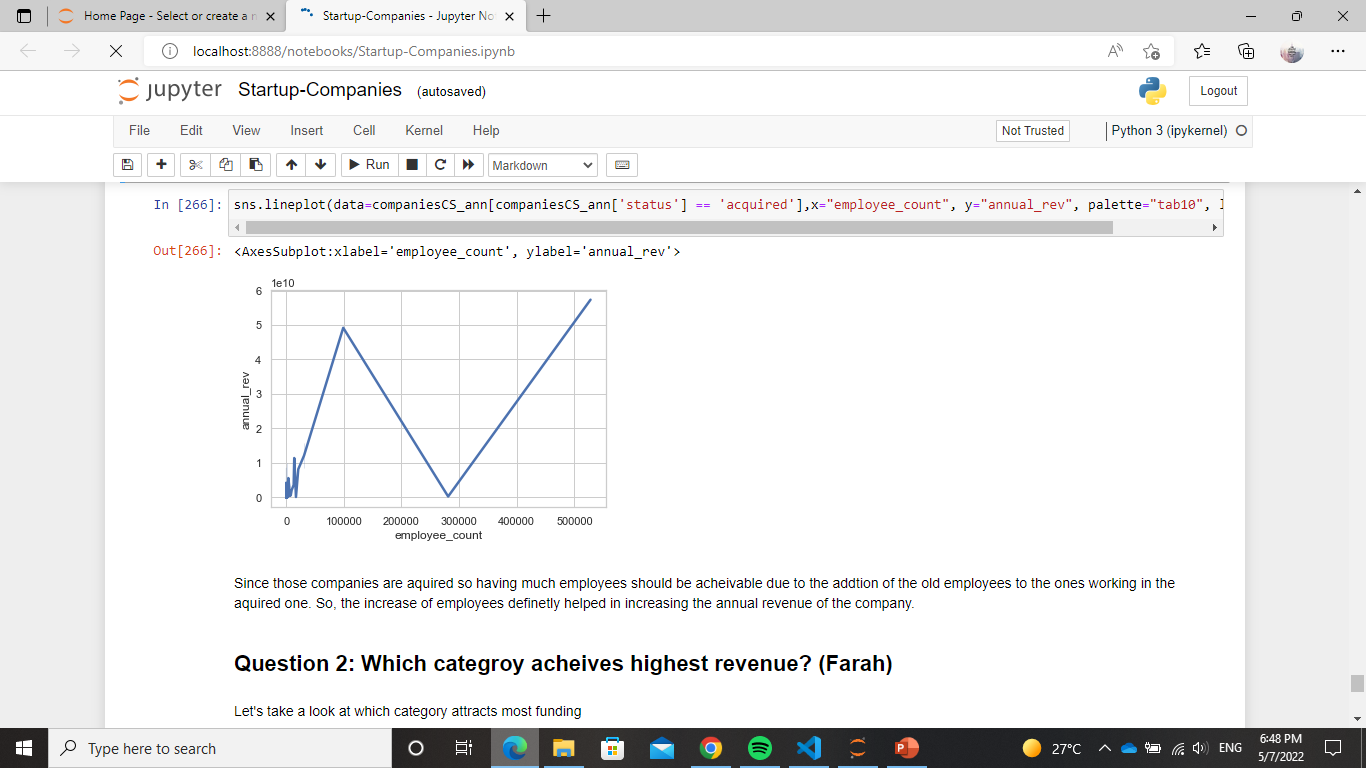
From the previous graph, a similar pattern is recognized from the previous graph. Here, the closed companies the maximum number of employees and annual revenue reached is noticeably way less than the operating companies which is reasonable. Having more than 5000 achieve reasonable profits.

* + IPO Companies:



Regarding the ipo companies, they can have a huge number of employees since they are publicly offered for stock sharing, they may have the ability to hire great amount of employees, and from the pattern seen from past graphs, hiring more employees help in increasing annual revenues. By this graph, it is clearly seen that the number of employees is directly proportional with the number of employees.

* + Acquired Companies:

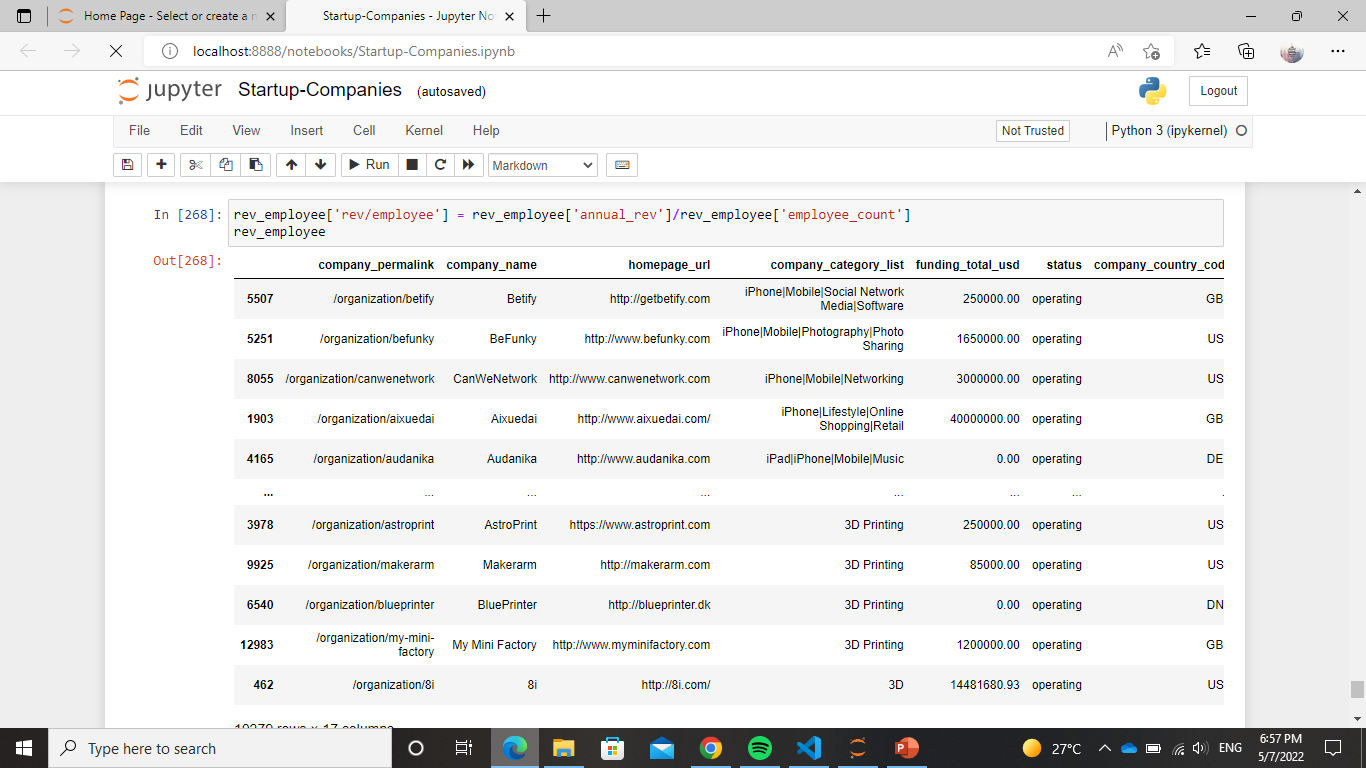


Since those companies are acquired so having much employees should be achievable due to the addition of the old employees to the ones working in the acquired one. So, the increase of employees definitely helped in increasing the annual revenue of the company.

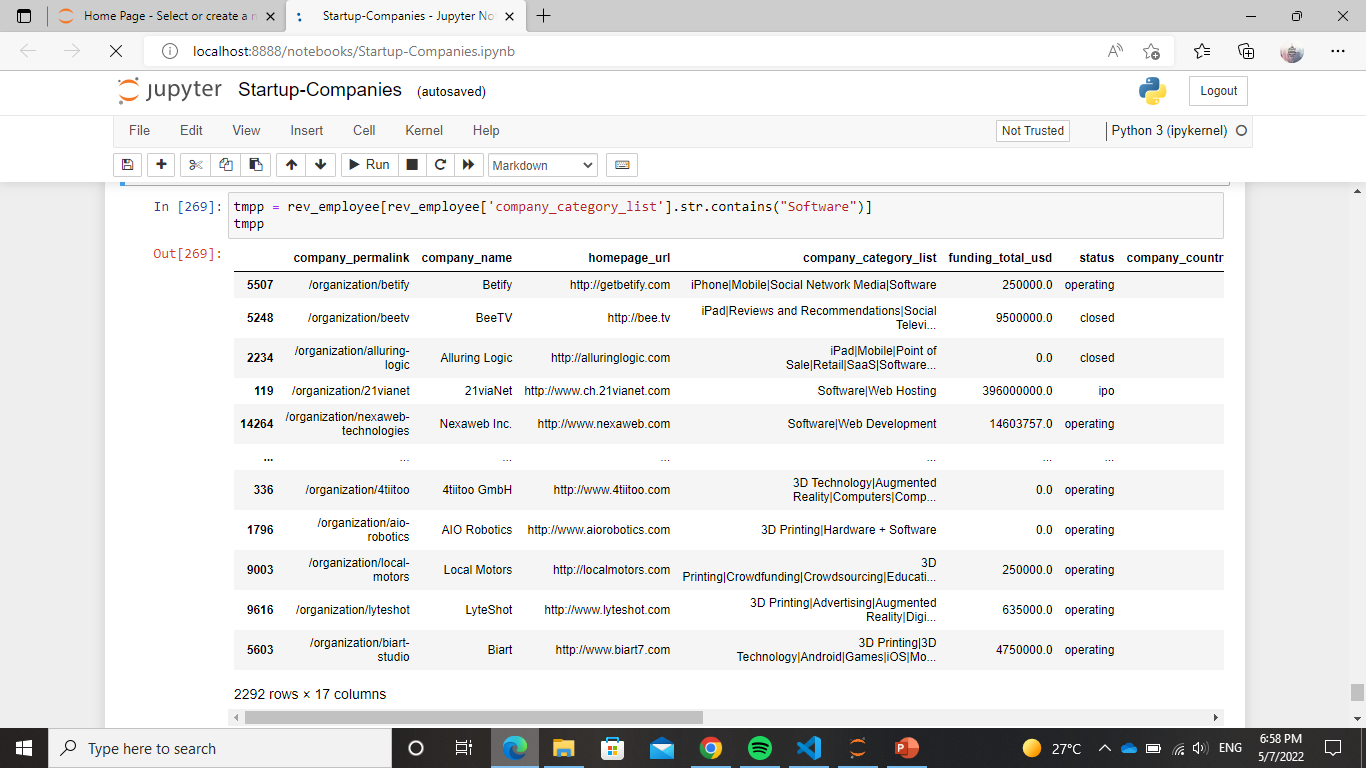
* **Conclusion**
  + In conclusion, the number of employees can be considered directly proportional with the annual revenue although of course there are other factors that affect the revenue, but hiring intels and labour can increase productivity and its cost doesn’t impact the company in a negative manner

### Question 2.1

* Which category achieves highest revenue?
  + Calculating the revenue per employee is a meaningful analytical tool because it measures how efficiently a particular firm utilizes its employees. Ideally, a company wants the highest ratio of revenue per employee possible because a higher ratio indicates greater productivity.

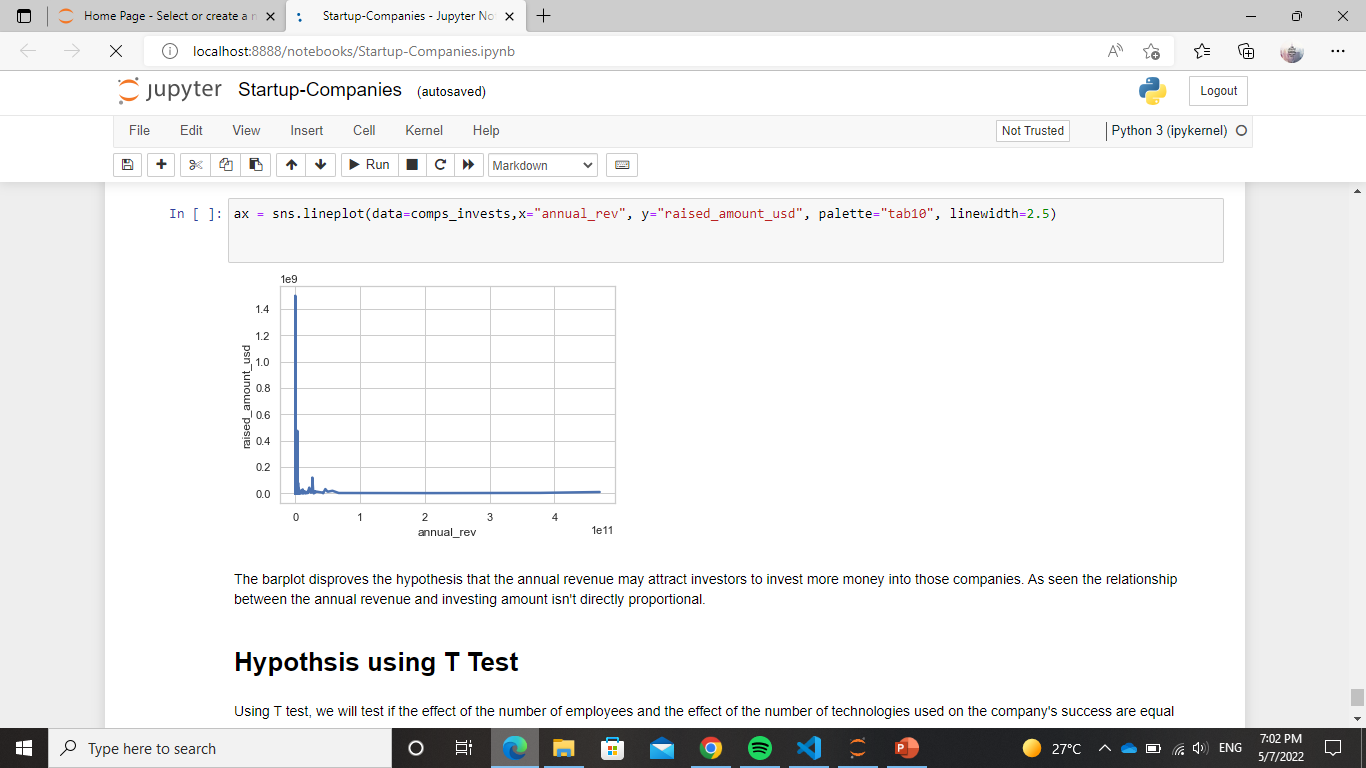


* + It is seen that companies working in software and tech in general has the most annual revenues. This shows that the market of this kind of field is profitable and has higher



### Question 3

* **Does a high annual revenue attract more funding?**
  + We might notice that there is a possibility of a direct proportional relationship between the amount of money invested in a company and its annual revenue. We will test it by visualizing the relationship.



The barplot disproves the hypothesis that the annual revenue may attract investors to invest more money into those companies. As seen the relationship between the annual revenue and investing amount isn't directly proportional.

### Question 4

* **Can the number of technologies that are used by the companies affect the rate of its success or failure?**
  + - This can be seen by checking the funding each company received based on their status and technology count.



The same procedure for the acquired and ipo companies which lead to the conclusion that,

* Companies with high funding and little or no technology count are more focused on products and other business services.
* Companies with high funding and high technological count imply that these technologies require high maintenance which the shareholders are willing to provide.
* Companies with low funding and high technological count imply that these technologies don't require high maintenance.

We can also conclude that some companies get acquired due to low funding and high technological demands whose maintenance the shareholders are not willing to pay extra money for.

### Hypothesis testing using T test

* T test is mainly used to compare between means of two groups to see if their effectiveness is equal or otherwise.
* Using T test, we will test if the effect of the number of employees and the effect of the number of technologies used on the company's success are equal
  + H0: The effect of the number of employees and the effect of the number of technologies used on the company's success are equal i.e., μ1 = μ2
  + H1: The effect of the number of employees and the effect of the number of technologies used on the company's success are not equal i.e., μ1 ≠ μ2

first, we will take a sample of 51 rows from the dataset



Calculating the t score



A negative t-value indicates a reversal in the directionality of the effect, which has no bearing on the significance of the difference between groups, so we take its absolute value

Calculating the degree of freedom





As seen above the absolute value of the t-score is > the t-value from the table so the null hypothesis is rejected

therefore, the effect of the number of employees and the effect of the number of technologies used on the company's success are not equal

### **Final Conclusion**

* Many Factors can affect the success of kickstarting a startup company.
* Gathering investments is one of the most critical aspects that can make a difference in a startup’s lifetime. However, the amount of funding or the time it takes for a company to get its first fund isn’t that important to the company’s success.
* Each country can have its own circumstances that provide a certain environment for a startup company either from laws, tax rates, etc.
* Companies need to spend more to gain more. For example, they need to hire enough labor and use sufficient technologies to achieve reasonable annual revenue.