**Biotech Project**

As discussed in our meeting, the idea for this project is to derive some descripting insights and mostly graphs for the companies that comprised the biotech industry.

The main dataset we will be looking at consists of the two excel files: Biotech Companies. Combines those files have 14,000 rows (hence the two files).

These files include all the biotech companies founded after 1976, in US, Canada, and Europe. I believe that most of the columns should be self-explanatory, but please let me know if anything needs further clarification.

Task 1: Produce a graph that shows how many new companies were founded each year.

The dataset already includes a column “year founded” but before producing the graph some cleaning-up is required. Specifically,

a) We need to remove companies that were not actually developing technologies or products that would be used for drug development (e.g., companies that providing consulting/legal services, or focusing on cannabis – there are quite a few of them in the dataset!). I have done this by basically filtering out any companies that had certain keywords in their name. In R, I have done that by using the following code (see keywords that I used).

# Remove companies that are LLC, Institute, Services, Cannabis, Consulting

finaldata2<-dplyr::filter(finaldata, !grepl("\*LLC$",finaldata$CompanyName))

finaldata3<-dplyr::filter(finaldata2, !grepl("\*LLC.$",finaldata2$CompanyName))

finaldata4<-dplyr::filter(finaldata3, !grepl("\*Institute\*",finaldata3$CompanyName))

finaldata5<-dplyr::filter(finaldata4, !grepl("\*Services\*",finaldata4$CompanyName))

finaldata6<-filter(finaldata5, PrimaryIndustry != "Life Sciences Tools and Services")

finaldata7<-dplyr::filter(finaldata6, !grepl("\*Cannabis\*",finaldata6$CompanyName))

finaldata8<-dplyr::filter(finaldata7, !grepl("\*Cannabis\*",finaldata7$ProductName))

finaldata9<-dplyr::filter(finaldata8, !grepl("\*Consulting\*",finaldata8$ProductName))

b) The second critical cleaning up bit is to remove duplicates that appear in the form of subsidiary companies.

My method of doing that was to remove companies that have the same first word in their name e.g., Abbvie Ltd and Abbvie AS. As you will see there are quite a few of Abbvies we only need tha parent company which is Abbvie Inc.

Note: Just picking focusing on the parent company (from the corresponding column) would not work because many companies that were created and operated many years as independent, were later acquired by another company.

For example, Medarex was created in the 90s, and it was acquired by Bristol Myers Squibb (BMS) in 2009. In the dataset BMS shows us the parent company of Medarex which is correct as of today, but for decaced Medarex was operating as an independent company. As such, Medarex and BMS should be counted as two separate companies for the purposes of counting the number of new companies founded.

In short, my approach was to find subsidiaries by finding companies that had the same first word in their name.

Once, I had done that I realized that I had also removed some companies that should have stayed in the dataset. For example, Cambridge X , Cambridge Y start with the same word but they are different companies (they just like to use Cambridge in their name for branding reasons). The next step would be to bring those incorrectly removed from the dataset back. I hadn’t done that yet. As Adam suggested, one way would be to look whether those companies have the same location, in which case they would not be subsidiaries. I might have to go back to Capital IQ to check if we can get more refined data on the location (e.g., France vs UK rather than just Europe).

Task 2: Public vs Private companies

Once we have completed task 1. We can classify the companies in public and private. As you will notice there is already a column in the dataset with this classification. At the moment this information is a bit misleading because of the subsidiary problem mentioned above. For example, Genmab A/S is the parent company and is correctly listed as a public company. Genmab Inc., however (the subsidiary in the US) is listed as a private company. For our purposes, if the parent company is public, the company should be counted as public. That’s why, once we remove the subsidiaries, I think the current classification to public and private companies should be accurate, so Task 2 should be trivial once Task 1 is done.

Task 3: Classify companies based on the technology they were working on

Drugs are in general classified to chemicals and biologics. Note here that biotech companies develop both types of products (and not necessarily just biologics).

For companies developing biologics, we have 4 fixed categories which are the following:

monoclonal antibodies, recombinant DNA (R-DNA), antisense and gene therapy.   The obvious way to classify those companies would be to look at their business description and search for these words. E.g,, if the text has “monoclonal antibodies” (mAbs) that would be a mAbs company, and so on.

Then within our list of biotech companies we will have many that developed chemicals. We do not need to break these down in smaller categories, we just need to group them as “chemicals”. I am not sure yet what the right keyword search for those would be. I ‘ll have a look and also took to my co-author and get back to you.

Task 4: Clinical development stage: Identify companies that have managed to send products in to clinical trials (Phase I, II, III) or even to the market, and classify them according to how far down the value chain they managed to get. E.g., out of the 8000 companies, 1000 had products in Phase I, 500 in Phase II, 200 to Phase III, and 100 made it to the market. We can talk more about this task later on.

Task 5: Background of biotech companies executives

There is another dataset that I showed you in our meeting which has information about the educational background (degrees, area of study, and university) of key biotech companies (CEO and board members). All we need to do here is to see how many of the CEOs or board members had advanced degrees (e.g., Master/PhD, MD) and whether there was diversity in the backgrounds (e.g, biotechemistry and biology were not the only kind of degrees). This dataset would not require much cleaning up.