*Name: Benjamin Gaines*

*GitHub: KentuckyBenjamin*

*Email: Benjamin.gaines@jackpotinsights.com*

Explanation of Work Performed

**What did I do?**

I identified a list of around ~34K companies who have shown reasonable evidence of growth & maturity. Around ~10K of those companies are certified Hub Zone businesses

**How did I do it?**

I used sample data from the contract award database from USAspending.gov to find a list of companies who have received federal awards every year for the last three years. Once I had the list, I filtered it for companies who were certified Hub Zone businesses.

The first part of this work involved going to the USAspending.gov award database download page. I downloaded the zip files for FY2019, 2018, and 2017 (note: I did not include these zip files in the repo because they are all around 2GB.) The zip files are quite large. Each year’s zip file has around 5-7 csv files in them with 1M rows worth of award information. Each award row has ~275 columns.

Because these zip files were huge, and most of the data in them wasn’t necessary for this, the next step I took was to take a sample of each zip file and reduce it in size to something I can easily use for testing purposes. Accordingly, I grabbed one csv file of 1M rows from each zip file (2019, 2018, and 2017), and deleted all but the 15 most obviously useful columns (things like award amount, name, etc.) All of this work can be found in the govt\_contract\_awards.py file.

Once I had a much smaller, workable csv file for each of the last three years, I created a dataframe from each file. From this dataframe, I made lists of all the companies who received awards from the government for each of the last three years. I then generated a ‘selected company list’ of companies who did business with the government every year for the last three years (meaning they were in each csv file, dataframe, and list). Finally, I used this ‘selected company list’ to get some basic info about each of the companies who were on the ‘selected company list’ (e.g. duns number, zip code, state, etc.). I aggregated all this info into a final dataframe called ‘all\_company\_df’.

Finally, I filtered the final dataframe to identify all the companies who were certified Hub Zone businesses. I then exported that info into a csv file, along with a csv file of all the companies I identified. For kicks, I also created a html file from the list of Hub Zone companies so I could open it in my browser and look around a bit more easily than in iPython or my IDE.

**If I had more time, what would I do next?**

With more time, I would do two things. First, I would repeat this work, but with all data for the last ten years. (As opposed to sample data for the last three years). Secondly, I would write a program to find the websites of each of these companies, scrape them, and filter them for key words like ‘family business,’ and ‘family owned.’ I could use that work to winnow down the list even further to include family owned businesses.

Regarding the first initiative of adding all data from the last ten years, I would not go about getting that data the same way that I did on this (downloading the CSV and reducing it through python scripts). Instead, I would use the usaspending.gov API, or possibly just recreate their entire database, which is an option they offer.

Regarding the second initiative of writing a scraper, I would pass all of the names from this list into the Bing API to get the URL for these companies. I would then write a scraper that iterates through the list of URLs, scrapes all raw text, and returns a value if the text contains any of the key words of interest. In this case, I would search for things that indicate they are family owned.

**If I had it to do over again, what would I have done differently?**

I don’t regret the scope of my work. I think trying to find reasonably mature companies in Hub Zones was a useful exercise. I do, however, regret that I didn’t just use the API or set up a copy of their database. By using sample data, I found about 10K companies of interest, but I now feel confident I could find another ~30-60K just by working with population data in lieu of the sample data.

**Closing Thoughts…**

This was a lot of fun! I spot checked companies here and there while doing this and found a few neat ones just though spot checks alone. Pioneer Machine & Tooling Co, for instance, is a family owned business which, given its age, must be at least in its second generation. Most likely, it is in its third generation.

Roberts Oxygen Company is another company that seemed interesting. They have around ~75 employees and have been steadily growing over the years. Like Pioneer Machine & Tooling Co, they are family owned. They are in their third generation currently. Their current CEO, Will Roberts, is in his mid-thirties, so would likely want to remain in place as the CEO if they took outside investment.