Approach, Assumptions, Modification and Extension

Approach: Read data in with Pandas for exploration. Noted there was a lot of missing wage data but felt that it was less valuable for the questions I was asking so it was dropped. My first two questions involved the number of postings over two years. One by location, and the other by position. So I created 4 DataFrames(DF) so each question was broken up by year. After cleaning I performed groubys on each DF so I had count’s for every location and position. For the location DFs I performed them on MSA\_ID and YEAR. Then LONG\_OCC\_CODE and YEAR for the postings by position DFs. After this I concatenated the corresponding DFs back together. I will get to the reason for this later when I talk about modifications I would make. After further cleaning I plotted all the results in Plotly. Looking at the spread of results, it was clear there were a lot of low numbers. These graphs contained unnecessary information, so I decided 50 postings annually was a good cut-off for both graphs.

For the third question I decided that running k-means on skill\_data.csv was the best option. I started by taking the corresponding DF and stripping it of everything but the skill scores and setting LONG\_OCC\_CODE as the index so it can be merged after clustering. I then wrote a preprocessor pipeline using PCA to reduce linear dimensionality so the data is easier to cluster. I ended up choosing 10 clusters simply because I was getting low silhouette scores, and I felt that was the best I could do without having too many clusters. I then plotted the clusters in matplotlib. After this I combined the DF back with the original skill\_df and posting\_df so that I could plot the cluster in Plotly to be interactive and link them with their corresponding jobs. I also exported all of the Plotly charts as HTML files for interactive use during presentation (3 of 5 will be used).

Assumptions: I assumed that more postings mean more need for focus, and less postings needed less focus. Since there was no indicator that lack of hiring was a problem, and there is no data to analyze anything involving factors involved with a lack of employment. Assumed wage is not a good indicator of advancement, partly because the wage data was rather poor. Finally I assumed that a lot of entry level positions mixed with middle skilled and non-middle skilled jobs was a good indicator of advancement. This is because someone can get an entry level position with the company, then move to another entry level, middle skilled job within the same cluster.

Modification and Extension: For starters (let us assume it was a little rust) I came back later and split the data by date for my first two questions by year later. This handicapped my visuals a lot, as I was simply out of time at that point. I would also go back and include the job group column as an indicator of potential department diversification as well as position diversification. I would also like to use a better ML algorithm than k-means. While a lot of the results “felt right”, I still had weak clusters. I also wanted to use the education data to help possibly eliminate some outliers such as the CEO position. Those jobs are not good focus for inclusivity and likely caused unnecessary weakness to the algorithm.