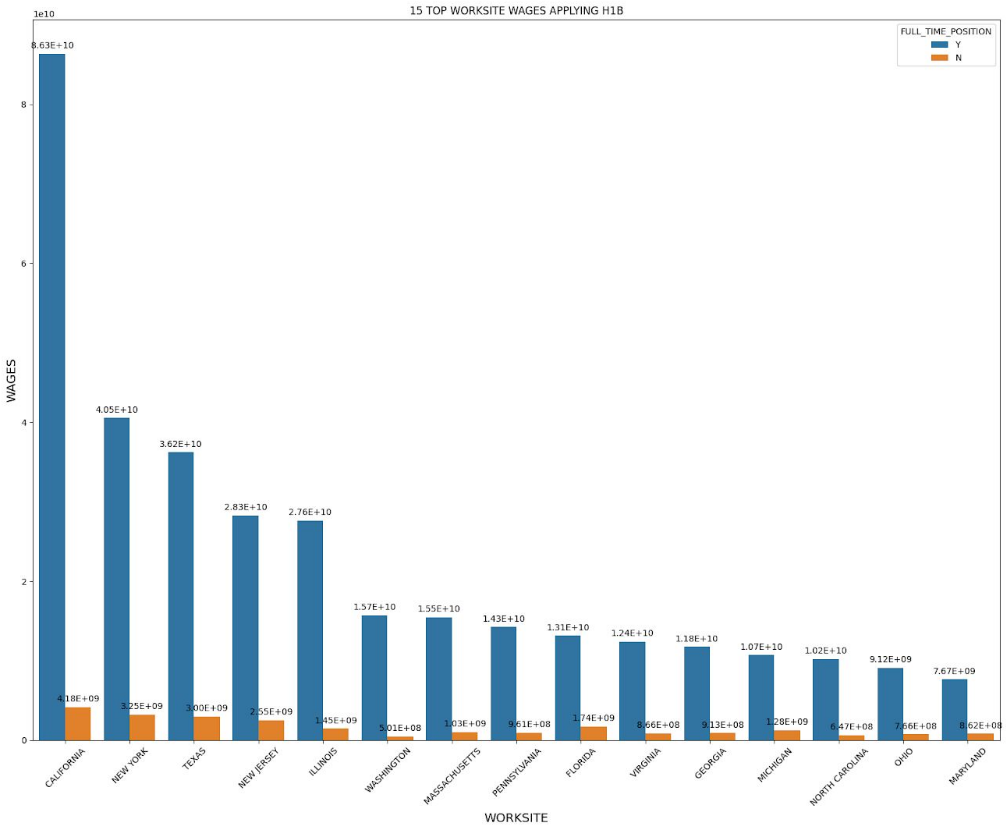
- Developer: Sophia Xiao @ 2020/1/16
- Dataset downloaded from Tianchi
  https://tianchi.aliyun.com/dataset/dataDetail?dataId=83994
- Goal:
  - To find the total wages of different worksites (US states)
  - Separate the wages by working position (full time/ part time)
- Website for study (cited):
  - https://stackoverflow.com/questions/49188960/how-to-show-all-of-columns-name-on-pandas-dataframe/49189503
  - https://stackoverflow.com/questions/40705480/python-pandas-remove-everything-after-a-delimiter-in-a-string
  - https://stackoverflow.com/questions/16958499/sort-pandas-dataframe-and-print-highest-n-values
  - https://stackoverflow.com/questions/48170867/how-to-get-the-common-index-of-two-pandas-dataframes
  - https://stackoverflow.com/questions/42532319/grouped-bar-chart-from-two-pandas-data-frames
  - https://datavizpyr.com/how-to-annotate-bars-in-barplot-with-matplotlib-in-python/
- Steps:
  - Read data from csv file
  - Because the worksite includes company address and city, I only need the name of state, and also because the format of data is clean, I use RE to keep only the state name

    - Then create a new dataframe consist of useful information, in this case, worksite, wages, and work time position
    - **I separate the dataframe from last step created by different status of work time position (full time as one, part time as the other) by setting condition**
    - I find the states with top wages for full time dataframe. Because the state of both work time position (full time & part time) have to be the same on the bar plot to avoid confusion, I use full time as the pivot because it spends more dollars
    - Then filter the same states from the part time dataframe because the state name of both part and full time are exactly the same except for the working time position and amount of wages
    - Then concat the two dataframe together for graphing
    - I research online for how to plot multiple bars and put x and y data accordingly, add values, labels, etc info. on the bars
- Problems I had during the project:
    - I had trouble separating the work time position by full and part time - which is an important part of this practice. I used
        - a natural join of copy of the dataframe
            - Ends up the resulting dataframe will have double rows of each status
        - sorting by index and wages
            - Ends up the resulting dataframe will only have full time states because their wages is way more higher than part time
        - and concating two original dataframe together

            - Does Not separate out full/part time status
        - So I finally use the condition to make another two dataframe, filter top 15 first from full time dataframe, then find the same states from part time by that
    - I have trouble labeling the bar plot, so I research online about it
    - I also have trouble add the hue of the bars (the color representation of the bars), I first use the way online, but 1 and 0 doesn't make sense, so I change it to its original value which is full time (Y/N) which makes more sense
- Saved plot



- Analysis:

- We can see CA is twice as much as NY spend on wages (for applying H1B, 2016), and the rest of the ranks pretty much match the more popular states for IT and business industries. There might be something to do with culture, population, law of states as well, will see if I get a chance to analyse deeper for these factors.