**Technology Characteristics of Business : ETL Report**

Colin Beveridge, Abbey Guilliat, Scott Partacz, & Lucas Stefanic

***Introduction:***

For our Census API Data Project, we decided to perform a comprehensive analysis on a single dataset: Technology Characteristics of Business from the US Census 2019 Annual Business Survey (ABS). We are interested in exploring how limiting factors on the development of technologies vary across industries, particularly, we hope to determine which factors seem to be causing the greatest constraints for businesses. We are also interested in performing a cross-comparison of two distinct industries to see how limitations may differ across a couple different technologies. To explore these questions, we need to clean and re-structure the data in order to more easily access and manipulate different variables. We also need to check for duplicate and null values as well as investigate potential errors that may be present in the dataset to ensure that we produce accurate, viable results. The following document details our process of obtaining and transforming the data.

***Data Source:***

United States Census Bureau Annual Business Survey (ABS) APIs. (2019). Technology Characteristics of Businesses. Retrieved from api.census.gov/data/2018/abstcb.html.

Data was originally accessed from [https://www.census.gov/data/developers/data-sets/abs.2019.html](https://www.census.gov/data/developers/data-sets/abs.2019.html%20) on 01/12/2021.

***Extraction:***

To access the API we had to first request access keys from: <https://www.census.gov/data/developers/guidance/api-user-guide.Help_&_Contact_Us.html>. After this, we opened up a new .py file and:

1. Imported requests and pandas
2. Made a call to the API by referencing the provided address, including all the required and desired variables
3. Converted API call into a pandas dataframe
4. Set the values in the first row as column headers and then dropped the first row
5. Dropped all columns unnecessary for our analysis
   1. .drop(['GEO\_ID',"NAME","YEAR","NAICS2017","SEX","ETH\_GROUP","RACE\_GROUP","VET\_GROUP","NSFSZFI","FIRMPDEMP\_F","FIRMPDEMP\_PCT\_F","RCPPDEMP\_F","RCPPDEMP\_PCT\_F","EMP\_F","EMP\_PCT\_F","PAYANN\_F","PAYANN\_PCT\_F","FIRMPDEMP\_S","FIRMPDEMP\_S\_F","FIRMPDEMP\_PCT\_S"]
   2. + .drop(tech\_df.columns[16:],axis = 1)
6. Checked for duplicate values, null values, and proper datatypes – found that data was already well cleaned
7. Saved clean dataframe as ‘TechImpactCensusData.xlsx’ which each member transformed for their respective visualizations

***Transformation:***

General Analysis

1. Read cleaned ‘TechImpactCensusData.xlsx’ into a pandas data frame.
2. Filter out all fields that are not aggregates over all industries with below command.
   1. tech\_df = tech\_df[tech\_df[‘NAICS2017\_LABEL’] == ‘Total for all sectors’]
3. Filter out all other demographic breakdowns with several similar filter statements on categories [‘SEX\_LABEL’], [‘ETH\_GROUP\_LABEL’], [‘RACE\_GROUP\_LABEL’], [‘VET\_GROUP\_LABEL’], and [‘NSFSZFI\_LABEL’]. All must equal ‘Total’ except the last which should be set to ‘All firms.’
4. Make data frames to split out each technology. Example:
   1. AI\_df = tech\_df[tech\_df[‘FACTORS\_P’].str.contains(‘T1’)]
   2. Each technology data frame has the same structure, with different codes in the str.contains method from ‘T1’ to ‘T5’
5. The cloud based technology frame has the questions in different order, so reorder them with the following lines:
   1. topCloud = CloudB\_df.iloc[:4]
   2. CloudB\_df = CloudB\_df.iloc[4:]
   3. CloudB\_df = CloudB\_df.append(topCloud)
6. Recast all numeric columns as floats in a for loop:
   1. numeric\_colnames = AI\_df.columns[8:]
   2. for colname in numeric\_columns:
      1. AI\_df[colname] = AI\_df[colname].astype(float)
   3. That same line is repeated in the loop for all five technology data frames.
7. Next, we replaced the values in the ‘FACTORS\_P\_LABEL’ column with abbreviated versions of the labels in all five technology data frames.
   1. replace\_labels = ['Too expensive', 'Tech not mature', 'Improper Data', 'Unreliable Data', 'Insufficient Talent', 'Laws and Regulations', 'Security Concerns', 'Insufficient Capital']
8. After replacing labels, for all five technologies we repeat the following steps three times plotting percentage of firms, annual payroll of firms, and marketshare of firms responding each limiting factor. We set the data frames to only include the relevant limiting factor labels and their corresponding number of firms. This :
   1. AI\_df\_firmpct = AI\_df[[‘FACTORS\_P\_LABEL’, ‘FIRMPDEMP\_PCT’]]
   2. AI\_df\_payann = AI\_df[[‘FACTORS\_P\_LABEL’, ‘PAYANN’]]
   3. AI\_df\_marketshare = AI\_df[[‘FACTORS\_P\_LABEL’, ‘RCPPDEMP\_PCT’]]
9. We set the ‘FACTORS\_P\_LABEL’ column as the index for each new data frame.
10. Finally, for each separate field graphed, we recombine the tech data frames to be easily plotted. We add columns names that are nicely named for clean legends
    1. TotalInd\_df\_firmpct.columns = [‘Artificial Intelligence’]
    2. TotalInd\_df\_firmpct[‘Cloud Based Technology’] = CloudB\_df\_firmpct[‘FIRMPDEMP\_PCT]
    3. Repeat above step for the remaining three technologies.
11. From here it is quite easy to make the three general plots examining responses overall industries for all technologies. These general charts were used to guide further analysis.

Health + Finance Industry Comparison

1. Read cleaned ‘TechImpactCensusData.xlsx’ into a pandas data frame.
2. Create lists of limiting factor responses using values in the ‘FACTORS\_P’ column for each of the three technologies: specialized software, AI, and cloud-based
   1. Example list creation for AI: ai\_factor = ['T1E36R01', 'T1E36R02', 'T1E36R03', 'T1E36R04', 'T1E36R05', 'T1E36R06', 'T1E36R07', 'T1E36R08', 'T1E36R09', 'T1E36R10']
   2. Note: for our original analysis we did not include the last two factors (‘T1E36R09/10’) corresponding to ‘Technology not applicable to business’ and ‘No factors adversely affecting the adaptation of this technology’ for the purpose of solely focusing on instances in which tech development was impeded
3. The following steps were completed twice: once for each the health and financial industries
   1. Three data frames were created by filtering the original dataframe, first by sector (health or finance) and then by whether or not ‘FACTORS\_P’ .isin() the corresponding factor list
      1. This gives us three dataframes for both industries, each corresponding to a different technology
   2. Next, we replaced the values in the ‘FACTORS\_P\_LABEL’ column with abbreviated versions of the labels
      1. replace\_labels = ['Too expensive', 'Tech not mature', 'Improper Data', 'Unreliable Data', 'Insufficient Talent', 'Laws and Regulations', 'Security Concerns', 'Insufficient Capital']
   3. Then we set the data frames to only include the limiting factor labels and their corresponding number of firms: df = df[[‘FACTORS\_P\_LABEL’, “FIRMPDEMP’]]
   4. Finally, we set the ‘FACTORS\_P\_LABEL’ column as the index for each data frame
4. We then combined the data frames so that we had one data frame for each sector, this resulted in two data frames each with the limiting factor labels as their index + three columns which we named ‘Specialized Software’, ’AI Technology’, and ‘Cloud-Based Technology’ respectively
5. The main transformation has been completed.
   1. To plot the ‘Distribution of Technology Limiting Factor’ bar graphs we created copies of these dataframes and dropped the ‘Too expensive’ indices to look at the distribution of less popular factors
   2. For the ‘% of Technology Limiting Factors Related to Expenses’ pie charts we grouped the dataframes into two limiting factors groups: ‘Too expensive’ and ‘Other’ which was a sum of all other responses
   3. Finally, the ‘Prevalence of Specialized Software Limitations’ pie charts were made after grouping the dataframes into three groups: ‘No Limiting Factors’, ‘Tech not applicable’, and ‘Limiting Factors’

Security concerns surrounding Cloud Based Technology

1. Read cleaned ‘TechImpactCensusData.xlsx’ into a pandas dataframe
2. Drop the following columns: ["SEX\_LABEL","ETH\_GROUP\_LABEL","RACE\_GROUP\_LABEL","VET\_GROUP\_LABEL","NSFSZFI\_LABEL"]
3. Removed the rows with these values for the column FACTORS\_P\_LABEL: Technology not applicable to this business, No factors adversely affected the adoption of this technology, and lastly the rows for total reporting.
4. Lastly depending on which sector and which technology group you will have to filter the data to only show those criteria.