About the Topic

The topic chosen covers the aquisition and analysis of data relevant to the "2017 SUSB Annual Datasets by Establishment Industry" that cover the US Metropolitan Statistical Areas (MSAs).

The topic consisted a basis for the investigation of "peripheral" datasets that would assist to further analysis and also come from web page scraping as well as web services. The NAICS code list is available through a public web service, while the MSA codes are available through a simple yet updated webpage for scraping.

Insights to gather from the data.

The topic may provide some interesting measures and statistical data regarding:

- Totals of companies per sector per MSA
- Geographical distribution of companies per size
- Distribution of companies per size per MSA
- Several average and median values of employement and payroll

About the Data

Dataset(s) and websites have chosen to utilize analysis

The datasets used for the analysis are:

1. The static dataset file comes from the US Census Bureau and is titled "2017 SUSB Annual Datasets by Establishment Industry" (https://www.census.gov/data/datasets/2017/econ/susb/2017-susb.html). SUSB annual or static data include number of firms, number of establishments, employment, and annual payroll for most U.S. business establishments. The data are tabulated by geographic area, industry, and employment size of the enterprise. The industry classification is based on 2017 North American Industry Classification System (NAICS) codes.

2. The North American Industry Classification System (NAICS) is used by the U.S. Federal Government to organize different types of businesses economic analysis purposes, and while there are different types of numbering systems out there that do this, a lot of local municipalities have also started adopting it for their own business classification purposes. The NAICS code list is available through a comprehensive RESTful API, which returns a JSON file with all NAICS code and their textual description as of year 2012. The API call is as follows: http://api.naics.us/v0/q?year=2012.

3. The third dataset comes from the Metropolitan Statistical Area (MSA)/Core-Based Statistical Area (CBSA) coding scheme used for NLSY97 geocode variables. It is available in a simple web page, where data extracted from the CBSAs, Round 8 section. The relevant URL

is: https://www.nlsinfo.org/usersvc/NLSY97/NLSY97Rnd9geocodeCodebookSupplement/gatt101.htm

Q5) Dimensions of the files in terms of row x columns and the file size (mb, gb, etc.) below. For example, 50,000 rows by 20 columns and 5.4 mb. If your file is a .json file, state the file size (mb, gb, etc.)
The static file has 424,247 rows x 15 columns and its size is 45mb

The static dataset has a very well documented structure and also needs some corrections to the data as well a need to exclude specific rows according to the noise of the data they include. The datasets from the NAICS API and the html scraping will create more comprehensive results since they will be used to add some more explaining columns to the initial dataset.

Insight 1 Explanation

Insight 1 groups companies per MSA area and projects the mean number of employees for companies that do business in those areas.

Insight 2 Explanation

Insight 2 proves if the size of enterprise is strongly correlated with the payroll of the company. The outcome - although positive - which means that the two columns are positively correlated, there in not a strong correlation, which leads to the outcome that larger companies pay lower salaries than smaller ones.

Insight 3 Explanation

Calculate correlations among all columns of the dataset. Negative correlations indicate that when the first column increases the second column decreases and vice versa

Insight 4 Explanation

Insight 4 orders sectors by the median of the employees working for them. A median represents more accurately the mean number of employees per sector, because it ommits outliers.

Insight 5 Explanation

Insight 5 calculates the sum of employees per county grouped by company size code.

Visualization 1 Explanation

Shows a heatmap of correlations among all columns of the dataset. Negative correlations indicate that when the first column increases the second column decreases and vice versa. Colors represent how strong a correlation is.

Visualization 2 Explanation

Creates a scatter clustering diagram which clusters together companies according to their number of establishments and total number of employees

Visualization 3 Explanation

Visualization 3 will create a bar chart showing the totals of companies per NAICS code in ascending order.