First Homework Assignment

The US Government released Medicare data. You can obtain the file on the site: https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Physician-and-Other-Supplier.html

The name of the link on the web page is: *Medicare Physician and Other Supplier Data CY 2015* and then *Medicare Physician and Other Supplier PUF, CY2015, Tab Delimited format.*

Note that the size of the compressed file is large (so make sure you have a reasonably good connection).

There is an accompanying pdf file that describes the data. I suggest you read the entire page carefully.

You have to identify a clustering related problem and then 'solve' it. Suggested process to undertake:

1. Understand the data (features of the data)
2. Perform data exploration (histograms of various features, correlations, outliers)
3. Pose a business question that you want to answer
4. Solve the problem (this step must be clustering)
    a. You have to select a subset of features for your clustering
    b. Select the most appropriate algorithm
    c. Select the number of clusters
    d. Assess the quality of your clustering

Note that the data set is relatively large (not "big data"). If you will have problems querying the entire data set, be creative (in a clever way select a subset of data – based on the knowledge of the problem).