I propose to do an EDA and statistical analysis of a dataset from the Center for Medicare and Medicaid Services (CMS) on the medical and dental marketplace for 2014-2016.

The primary focus of my investigation will be what I subjectively consider sociological factors that occurred between 2014-2016 to see if there is objective evidence. First, although ACA was passed in 2010, Jan 1, 2014 was the date most of them actually went into full effect. Looking at the data, I've superficially noticed some factors that I suspect will change like pre-existing conditions and how long a child is carried due to the enforcement of the rules. I also want to look at what effect family size/type has on the plans. Is there a difference between 2 parent/2 kid household vs 2 parent/3 kid or 1 parent/2 kid?

Society's definition of a family broadened in the 2014-2016 time period. 2015 was a tipping point and there should be differences before and after. In 2015, the Supreme Court ruled states have to honor same-sex marriages from other states and society's acceptance transformed as well (ie Caitlyn Jenner announced her transformation and was the #1 story for weeks). Also in 2015, portions of the ESSA act went into effect regarding the medical care of foster children. In the 2014 data, I noticed there are some plans for non-related people. I would like to see if they are re-designed and/or named and whether the scope of them expanded.

Description of Data

There is a total of 6 *.csv files for each of the 3 years. The total dataset is 4.1GB

File	Function	Columns	Rows	Columns w/null	Numeric only
Benefits_Cost_Sharing_PUF.csv	Lists all US plans costs and what they don't pay for(ie Copay $20, Deductible $5000, cancer not covered)	32	1164869	18	6
Business_Rules_PUF.csv	the business rules for the plans (states allowed, dependents allowed, non-married cohabitation covered, etc); most are categorical (0 = nosmoke, 1=smoke)	23	2103	19	2(year and plan #)
Network_PUF.csv	Network ID (state, provider, etc)	14	937	2	5
Plan_Attributes_PUF_2014_2015-03-09.csv	detailed explanation of what is covered (wellness programs, can only children get braces, whats covered when having a baby, etc)	126	18718	not evaluated – python says too many for .info()	13
Rate_PUF.csv	premium increase amounts for many values ($ per kid, smoker, child under 2, over x age)	24	3796388	not evaluated – python says too many for .info()	14
Service_Area_PUF.csv	where is the plan offered(state, region of state, zip codes, etc)	18	8874	4	6

MVP: Number goal is to learn how to get meaning out of large healthcare related DBs. I have one file with 126 columns and two others with 1 million + rows. A) condense/consolidate/catalog the data from 18 files and 4.1GB to something more user friendly B) Try to locate data to graph and test hypotheses related to sociological factors listed above or others the data reveals. I'd like to see if I can find both spatial and temporal differences. It looks like there could be some limiting factors on spatial. Plan A is by state(NY), B is by zipcode(14603), and C is by region (NW NY).

MVP+: Get this into a queryable SQL database with documentation to make it easier for others to work. My SQL could be stronger and this could be a chance to work on it.

MVP++ A)There are lots of columns with open text values that I would like to breakdown to make usable. For instance, there can be up to 16 different wellness programs listed in one cell with the name varying between plans. A plan can have 6 different co-pays in the same cell for things like childbirth. I'd like to get experience parsing and categorizing unstructured.