Problem statement and hypothesis

USA government expenditure on healthcare is set to reach 20 % of GDP by 2025[1]. There are many contributing factors, but is there some way to target efforts to reduce spending?

Examining the specific geographical areas in the Centers for Medicare and Medicaid (CMS) spending data should reveal differences in spending between areas. Those differences should be able to be correlated with demographic features of those areas from US Census data.

It is expected that the percentage of elderly people living in an area (75+ rears old) will be positively correlated with increased spending. Also the proportion of low income earners will also be positively correlated with increased government spending on healthcare in an area.

Description of your data set and how it was obtained

The dataset is made up of three different data sources: 1) Centers for Medicaid and Medicare outpatient spending data (2011-2014)[2]; 2) USA Census bureau 2010 decennial census data[3]; and 3) US Dept of Housing and and Urban Development HUD USPS ZIP Code crosswalk Q2 2016[4].

The CMS spending data was downloaded directly as a flat file CSV. As was the the ZIP to county crosswalk table.

The 2010 Decennial census data was only available via the census API and suitable features were accessed using Laura Kurup's Census API Python Script[5]. The script was configured with my census user API key, and the location_type variable was set to 'county'. The variables requested are listed out in the census_variables.csv file. There are 78 in total covering male and female age ranges, total population, racial and ethnic demography, plus household size.

Description of any pre-processing steps you took

All CSV files - including the CSV file created by API calls from the Census API Python Script were imported into their own data frames. In each data set it was necessary to preserve the keys that the datagrams would later be merged under. This meant that ZIP codes and FIPS state and county codes were set to be imported as strings - not integers - to preserve leading zeroes. In come cases it was also necessary to explicitly add back leading zeros using the zfill() method.

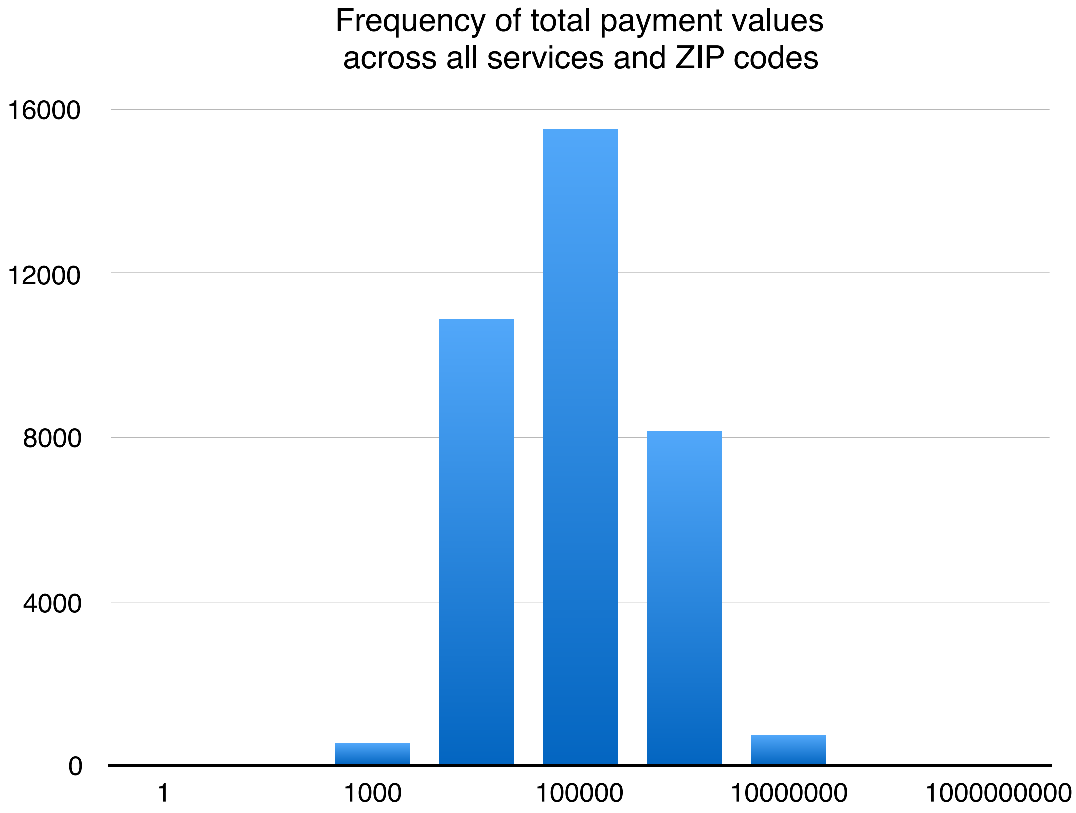
In the census dataframe state and and county FIPS codes were aggregated into a combined five digit FIPS code. The aggregated FIPS code was placed in a column called 'FIPS'.

The ZIP county crosswalk data had it's 'COUNTY' column renamed 'FIPS' so it could be used as a key for the pandas merge function when merging it with the census data dataframe. The CMS payment data had it's 'Provider_Zip_Code' Column renamed to 'ZIP' to be used as a key for a merge to the previously two merged dataframes.

All three data sets were merged together into one dataframe linking payment data at the ZIP level to census data at the county level.

What you learned from exploring the data, including visualizations
The ZIP county crosswalk has 52314 entries, the CMS payment data has 2806 unique zips, and my test census data pull had 10 rows (set to test API call pulled the right columns). The final census API data pull is expected to have one row for every county in the USA - around 3143 rows, possibly more due to inclusion of other county-like areas and territories such as Puerto Rico.

The 2014 CMS data has many different types of services listed out, the average payment for that service, the number of times that service was rendered. per zip code. Calculating the total payment per service per zip code (average payment multiplied by number of times service rendered) then plotting out a histogram of each total payment shows magnitude of expenses involved in CMS outpatient payments.



How you chose which features to use in your analysis

Initial Census API data pulls were based upon general descriptions of the population - to do an initial analysis including age, sex, and house size - easily available information in the decennial census information. These are also likely to be correlated with outpatient payment totals.

Details of your modeling process, including how you selected your models and validated them

The intention is to use regression to fit the continuous response value of total payments. The census values for each county will be the features. The result will be a model that predicts CMS outpatient payments based on demographic data.

Your challenges and successes

Data acquisition has been the biggest struggle. the Census API is very slow to pull all the data - over 12 hours to get 74 variables for all counties in the US. The Census API Python script definitely helped to get this going but initial configuration also required a bit of interpretation and set up - adding a utf-8 coding to the script and configuring the census_variables.csv file correctly. Neither of which were spelled out correctly in the documentation.

All of the pre-processing work required to get the data set up was surprising - managing data types, getting the merges to work correctly, preserving/restoring leading zeros.

Possible extensions or business applications of your project

Actually use the ACS survey information. Find areas of high medical service demand for medicaid and medicare services - predict based on demographic changes potential locations of new high medicaid/medicare spending.

Conclusions and key learnings

Get the data as early as possible and do not underestimate how much work it will take to arrange it into something analyzable. This is doubly the case if linking two data sets together - the effort required to interpret that connection is significant, especially if you need a third data source to be the cross walk.

References

1. Sean P. Keehan, John A. Poisal, Gigi A. Cuckler, Andrea M. Sisko, Sheila D. Smith, Andrew J. Madison, Devin A. Stone, Christian J. Wolfe, and Joseph M. Lizonitz National Health Expenditure Projections, 2015–25: Economy, Prices, And Aging Expected To Shape Spending And Enrollment Health Affairs Vol 35 No. 7
2. Medicare Provider Utilization and Payment Data: Outpatient https://www.cms.gov/research-statistics-data-and-systems/statistics-trends-and-reports/medicare-provider-charge-data/outpatient.html
3. Decennial Census (2010, 2000, 1990) http://www.census.gov/data/developers/data-sets/decennial-census-data.html
4.HUD USPS ZIP Code Crosswalk Files https://www.huduser.gov/portal/datasets/usps_crosswalk.html
5. Census API kit https://github.com/laurakurup/census-api
6. American Community Survey 1 Year Data (2014, 2013, 2012, 2011) https://www.census.gov/data/developers/data-sets/acs-survey-1-year-data.html