- Cleaning initial data set (HealthCost_data_cleaning.ipynb)
  - Downloaded from data.gov
  - Stripped whitespace from beginning and end of column names
  - Replaced spaces in column names with underscores and made all characters lower case. This is to make it easier to deal with columns in Ipython notebook
  - All dollar amounts were given with dollar signs so pandas read them as strings. I removed the dollar signs from every element in the cost columns and converted them to floats so that they could be manipulated mathematically.
  - Zip codes with leading zeros were in the csv files as 4 digit numbers, so I used zfill to pad them all to five digits strings for geocoding.
  - I removed characters from provider names and street addresses that gave geocoding problems (names ending in ", THE", commas, apostrophes)
  - All city names that were longer than 15 characters were cut off at 15. This needed to be corrected to improve the geocoding results.
    - I selected all unique providers by their id number and filtered on cities with 15 characters and wrote them to a csv file (85 total city names).
    - This csv file was edited by hand to add a second column that contains the corrected city names, which is used for the replacement code in the next step. After removing cites that had exactly 15 characters I ended up with 45 city names that were truncated.
    - I ran code on the entire dataset that, for each of the 45 truncated city names, replaced them with the full city name using the csv file described previously
  - Each procedure type (drg_definition) had a long name (e.g. 039 - EXTRACRANIAL PROCEDURES W/O CC/MCC), so in order to made the analysis easier I created a new string column (drg_id) that is just the three digit code describing the procedure (e.g. '039')

- Geocoding (HealthCost_providers_geocode.ipynb)
  - Read in the cleaned dataset described above to a dataframe
  - I filtered on the 'provider_id' column to select on only the unique providers (3,337 providers total)
  - Combined the address information into a single string ('provider_name', 'provider_street_address', 'provider_city', 'provider_state', 'provider_zip_code'). This was necessary as it is the input when using the GoogleMaps API to find the GPS coordinates of each provider.
  - I found that sometimes the GPS coordinates returned by GoogleMaps was more accurate when including 'provider_name', but most of the time not using the provider name gave the best results. Because of this I ran the geocoding for both cases. I plotted the difference in latitude and longitude between the two results, and for each case that had a very large difference I checked it on the google maps website to see which on was correct.
    - For providers where the difference was very large, sometimes with the name worked and sometimes not using the name worked. It was mostly even.
    - As the difference got smaller, not including the name consistently gave the better results.
  - The best results for each provider were combined and put into a single csv file
  - Created a heatmap of the provider locations overlaid on a GoogleMaps map.

- Exploratory Data Analysis (EDA) (HealthCost_EDA.ipynb)
  - Numbers
    - Number of rows: 163,065

        - Number of providers: 3,337
        - Number of different procedure ID's: 100 (not every provider had all 100)
        - Number of discharges: 6,975,318
    - Calculations
        - National median cost, median payments and median medicare payments for each procedure type (bar plot)
        - For each row calculated the fractional difference in cost, payments and medicare payments compared to the national medians for that procedure (histogram)
        - Number of a given procedure totaled over all providers (bar plot)
        - Total number of procedures in each state (bar plot)
        - Plot of average covered charges and average total payments for procedure 039 for all providers. Total payments has small scatter, total cost has very large scatter (scatter plot)
        - Looked at whether a provider had more procedures above the national median or more procedures below for total procedure cost (scatter plot)
            - Made separate plots for each case
            - It appears that providers with more procedures above national median are more tightly clustered near large cites.
            - Providers with more procedures below the national median are more evenly distributed.
        - Averaged the fractional differences for each provider over all procedures (scatter plot)
            - Saw some variation by state

- Random Forest Classification (HealthCost_RFR.ipynb)
    - I read in the full, cleaned dataset and merged it with the geocode data for each provider, so that there was a latitude and longitude associated with each row.
    - I attempted to predict whether the cost of a procedure would be above or below the national median using a random forest classifier.
        - I calculated the fractional difference of the procedure cost compared to the national median for each row.
        - I then added a column classifying the procedure at that provider as being above or below the national median.
        - Using the GPS coordinates as the input X and the fractional difference classifier as y with the Random Forest Classifier, I performed a test-train split with the training set as 2/3 of the total data set.
        - I obtained a R^2 score of 0.875 when using the training model on the test set
        - I plotted the confusion matrix for the results, and it was a symmetrical matrix with ~0.87-0.88 on the diagonal and ~0.12-0.13 on the off-diagonals.
        - I calculated an AUC score of about 0.95
        - **Adding in other variables such as the population and median income by zip code seemed to have no significant effect. I also calculated the number of other providers within 40 miles of each provider, but adding that variable had no effect as well.**
        - **I also found that I get the same results with just the provider_id tag as the input X using pandas' get_dummies to convert the variable to an indicator variable rather than numerical value. This makes me think that the GPS coordinates are essentially acting as an indicator for the specific provider rather than giving any useful information as far as relative position to other hospitals.**