## Data Sets Overview

1. **Detailed\_Data.xlsx** : This is the file containing the “labels” for a subset of hospitals from all over the US. The labels could be the standardized price , total allowed amount or relative price(inpatient / outpatient or both). The study that details how exactly the variables in this file are calculated is found here:

[Hospital Price Transparency Study Round 2 | RAND](https://www.rand.org/health-care/projects/price-transparency/hospital-pricing/round2.html)

1. **RAND\_hospital\_data\_documentation\_2020\_05\_01** : This folder contains files pertaining to the above dataset. Please go through the **Readme.pdf** file in it to understand what each file contains.
2. **Rand\_hcris\_cy\_hosp\_a\_2020\_05\_01.csv.zip :** This is the zipped full hospital cost report information system (HCRIS) data file having all the financial metrics for all hospitals across several years .The detailed description of each variable in this file is provided in the files in the folder above.
3. **Medicare IPPS and charge data** : This dataset contains the utilization and charge data for providers participating in IPPS (from hereby referred to as ‘IPPS hospitals’) for 2011-2017, an example for 2017 can be found here:  
   [Inpatient Charge Data FY 2017](https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Inpatient2017). You can download files for other years similarly.
4. **Medicare IPPS Final Rule data files** : These datasets additional hospital level variables such as Wage Index for Provider , Resident to Bed Ratio - Indirect Medical Education , Disproportionate Share Ratio , Number of Cases , CBSA etc. and other relevant info on inpatient services performed by the IPPS hospitals such as Weights for **MS-DRGs**(codes used to identify the specific inpatient service), Average length of stay for each etc. You can download all relevant data files from this page:

[FY 2020 Final Rule and Correction Notice Data Files](https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/FY2020-IPPS-Final-Rule-Home-Page-Items/FY2020-IPPS-Final-Rule-Data-Files)

1. **Quality Data :** This data contains quality scores for all hospitals and providers in the US along with other variables like Readmissions rate , mortality rate etc. We are working on getting this delivered to you.
2. **Census Data** : The American Community Survey(ACS) website provides over 10,000+ metrics on the US population such as household income , age , healthcare spend etc. at different geographic levels such as county , census tract , zip code etc. They also have apis in different programming languages . We recommend using this particular data set called ACS 5 year data(2009-2018) which is the latest that they have. Check it out here:

[American Community Survey 5-Year Data (2009-2018)](https://www.census.gov/data/developers/data-sets/acs-5year.html). You’re free to include other datasets from them in your models though.

You can choose other datasets in your analysis if you conclude they are effective in improving the predictive accuracy/explainability of your models.

## 

## Problem Overview

The dataset called **Detailed\_Data.xlsx** has 726 “labelled” hospitals. The label/response field name is “Relative Price for Inpatient and Outpatient Care”. It is a continuous/real valued variable. However , there are around 4000 unlabelled hospitals for which you’d have to predict it. These hospitals , along with around 1100 financial metrics on them in each row, can be found in the dataset **3** for every year starting 1996 until 2019**\***. The “predictor” variables at a hospital level for these hospitals can be found in the datasets **3**,**4** and **5** in the above list. Stated simply in mathematical terms , you have to approximate the regression or conditional mean function,

Where y is the “Relative Price for Inpatient and Outpatient Care” variable , X is the matrix**\*\***(the so called “design matrix”) with rows as individual hospitals and columns as hospital level variables such as financials , medicare payments , case mix index etc. and is the regression parameters or the “weights” of your model. An example of a row(hospital) in X can be:

[avg\_net\_revenue , avg\_total\_medicare\_payments , avg\_num\_discharges]

Where the averaging is carried out over a chosen time frame(eg. 2015-2017). Note that this is merely an example and you can choose as many variables to pack into that vector depending upon the need as well as use the non aggregated versions of them.

You can also frame the above problem as a classification problem by bucketing the label into **k** classes eg. ( 0-100 : low , 100 - 400 : medium , 400 and above : high). The X remains the same, but the function now outputs a probability ( in the “One vs Rest” scheme) of y belonging to the **kth** class given X.

Aside from the hospital level variables in datasets **3 ,4 and 5,** you can include census variables aggregated based on for eg , the zipcode of each hospital. You can choose other aggregations like county , state , etc. if it improves your model performance metrics.

\*\*Encoding the data as a matrix is one way. You can also encode it as a tensor depending upon how your chosen modelling framework(eg. keras) accepts input.

\*Data For 2019 is not complete as CMS has only uploaded HCRIS data on 3102 hospitals for 2019.

## Deliverables

You can choose whatever language you’re comfortable with (even just excel) to parse all the data and build your models. However, at a minimum, you’re expected to submit the following three:

1. **Code**: source code of your data processing pipeline + models and validation.
2. **Documentation:** A detailed description of your methodology, feature engineering , qualitative and quantitative measurements of the model predictions and any other stuff you find relevant to this exercise.
3. **Data:** Datasets you used for your models including , but not limited to the ones mentioned on page 1.
4. **Report:** A presentation highlighting your work and conclusions in brief.