Problem Statement:

I obtained the data for my dataset from cms.gov using two different csv files for the year 2016. The first file was titled ‘[medicare-physician-and-other-supplier-national-provider-identifier-npi-aggregate-report-calendar-year-2016.csv](http://localhost:8888/edit/Desktop/springboard/Springboard-20190719T193509Z-002/Springboard/Kagle_Datasets/Medicare_NPI/medicare-physician-and-other-supplier-national-provider-identifier-npi-aggregate-report-calendar-year-2016.csv).’ It included 55 different features for all Medicare contracted doctors. These features provided information ranging from professional background of physicans and on demographic data of each doctor’s patient population. This demographic information included percentage of patients for each doctor who fit into demogaphic groups (age, race, income), and each doctor’s percentage of patients with specific health designations and conditions. Because the files were CSV files I used the ‘pd.read\_csv()’ function.

The second file titled ‘ [Part D Prescriber PUF NPI Drug, CY2016](http://download.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Downloads/PartD_Prescriber_PUF_NPI_DRUG_16.zip)’ included drug prescription data for every medication that a doctor prescribed over over the year 2016. Consolidating the data to the most useful features I used pandas *iloc* function and pandas *drop* function to create a new dataframe including only ‘total\_drug\_cost’ (the total cost of medicare prescription) ’total\_30\_day\_fill\_count’ (total number of monthly prescriptions filled), ‘specialty description’, and each doctors’ ‘National Provider Identifier’ .

Using pandas’ groupby function I combined all data for different drugs into one row per doctor. By doing this I now had data on each doctors’ total number of 30 day prescriptions written and total cost to medicare of all medications written using medicare’s part D drug plan. After combing all of the rows from the Part D drug based dataframe described above into one row for each doctor, I merged this dataset with the first described dataset using pandas’ function ‘dataframe.merge()’.

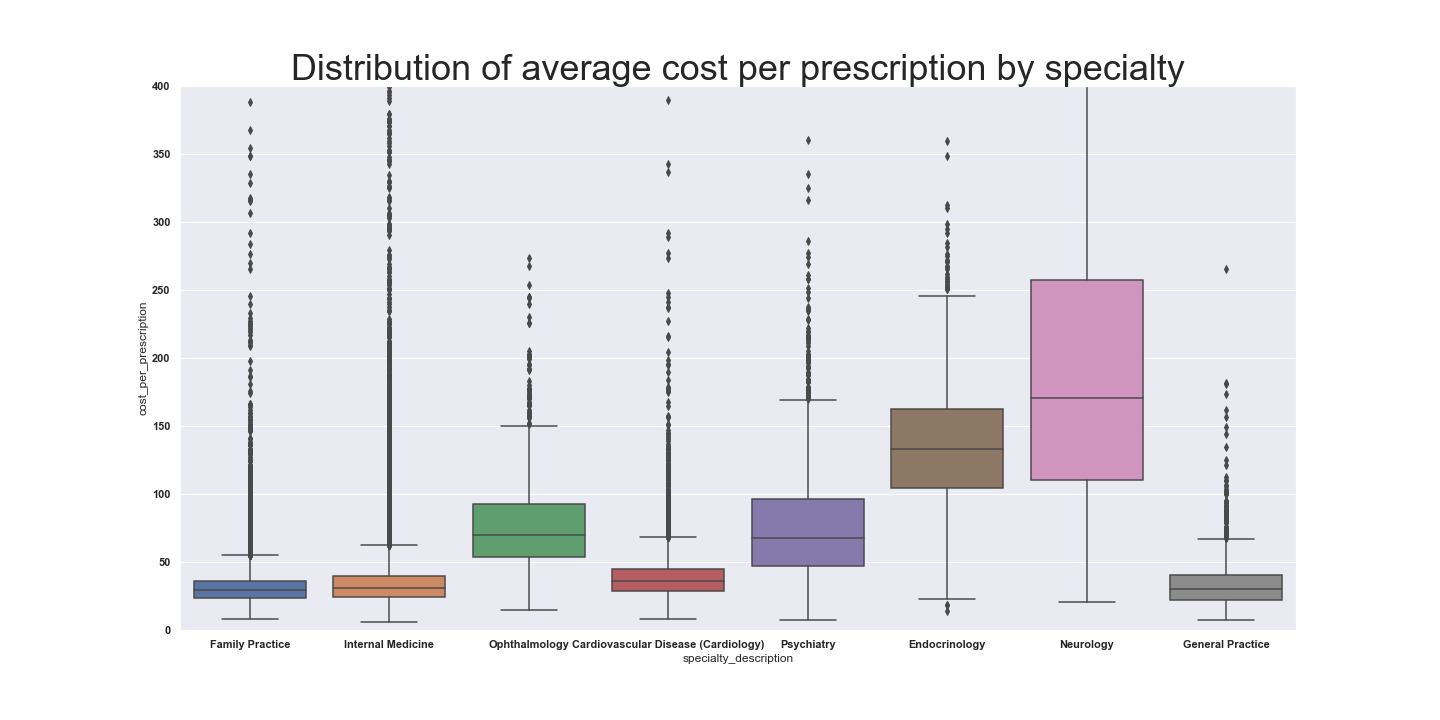
After merging the files I created a dictionary with each medical specialty as a key and each specialty’s respective average ‘total\_drug\_cost’ as a value. By using the map function I then created a feature that was equal to mean prescription cost of each physician’s reslective specialty. I created this feature as an initial method for recognizing and calculating the relationships between specialty and total drug costs.

Next I dealt with missing values. I did this because of learning algorithms’ inability to run data with nan values. In order to do so, I used the pandas function ‘dropna()’. I created a threshold of 44 values as the minimum amount of values a row could have without being dropped. This decreased the dataframe from 622705 to 397766 rows. For those rows that were not dropped but still had na values, I ran pandas’ ‘fillna’ function. I replaced each na value with the median value of it’s column. The code used was: DataFrame.fillna(DataFrame.median() as the fillna argument. (Is there a way to create the mean via the specialty?

Before using any machine learning on the dataset needed to replace any categorical values with numerical values. I did this for the feature ‘Gender of the Provider’. I created a dictionary (genderdict) where Male and Female were keys with values of 0 and 1 respectively. I then used remapped the feature with the code dataframe[‘Gender of the Provider’].map(genderdict).

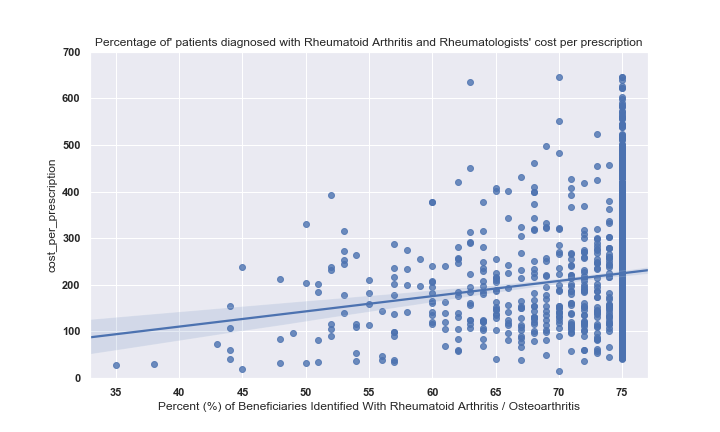
***Data Exploration***

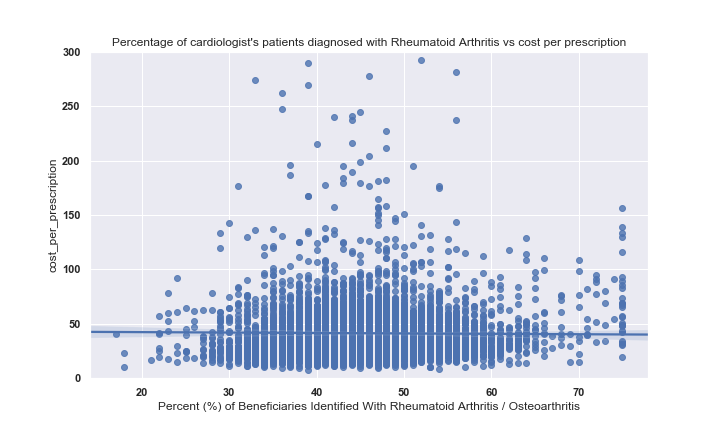
After obtaining and organizing the data I looked at Data exploration. The first visualization I created of the data, looked at how distribution of drug costs differed according to medical specialties.



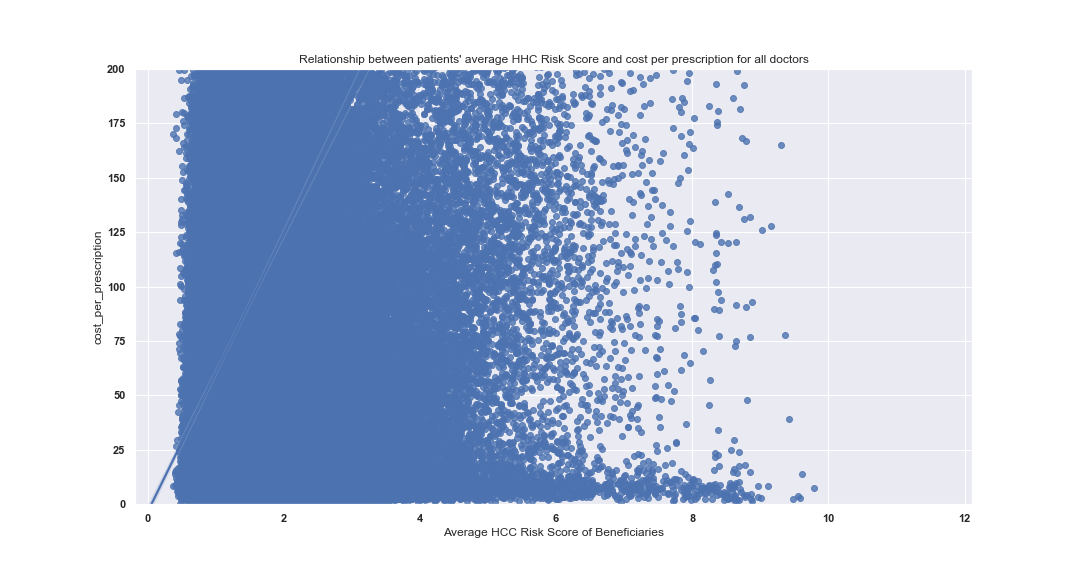
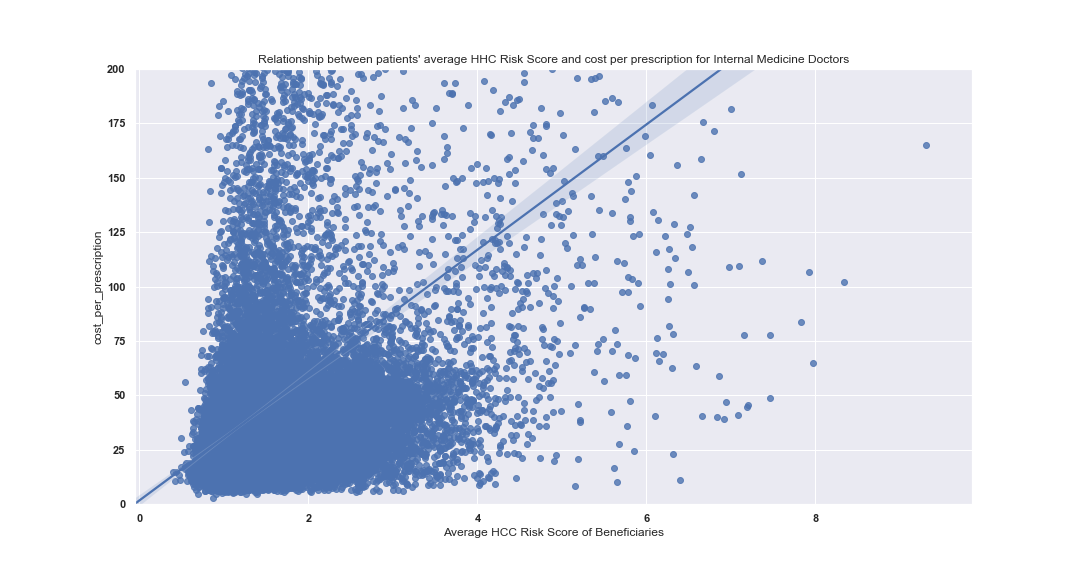
Looking at the graph one can see that a doctor’s medical specialty has a large bearing on the average cost per prescription he writes. Of these eight specialties, which are among the most common, it can be seen that Neurology and Endocrinology have the highest average prescription costs. Similarly, it can be seen that Psychiatry and Ophthalmology while not quite as high also had high average cost per prescription. On the other hand, primary care specialties and Cardiology had much lower average costs per presciption. This has to with the fact that besides cardiologists, specialist write prescription’s for a larger percentage of brand name medications, medications that are often targeted to smaller chronic illneses. Contrastingly Cardiological illness is so common, medications are often so widely prescribed and common that research has a long history and patents’ of effective medications’ have expired.

Next I tried to look at the relationship between the percentage of physician’s patients with specific health conditions and their average cost per prescription. I found however that it was necessary to isolate the distinct specialties to find a correlation. The reason for this is likely because individuals with chronic illness go to specialists that relate to their illness. Drugs that physician’s prescribe are related to the illnesses they treat and not the other illness the patient has. Therefore, illness that specialist treat are much more correlated with the drugs they prescribe and their costs. f you compare the effect of specific chronic illnesses to while not differentiating specialties you have difficulty finding correlation. This is due to the fact that chronic illnesses affect different related specialties drastically different depending on if the specialtist treats said illness. Additionally, considering the overall relation of specific specialties with average drug costs, the fact that specialty is so highly predictive with drug costs overrides the correlation with specific specialties. A supporting plot of this theory can be seen in the plots below showing how the percentage of patients with Rheumatoid Arthritis effect specific specialties. While percentage of patients with Rheumatoid Arthritis has a large effect on the average prescription costs of Rheumatologists. It has very little relation to the average prescription costs of Cardiologist, who do not treat Arthritis.



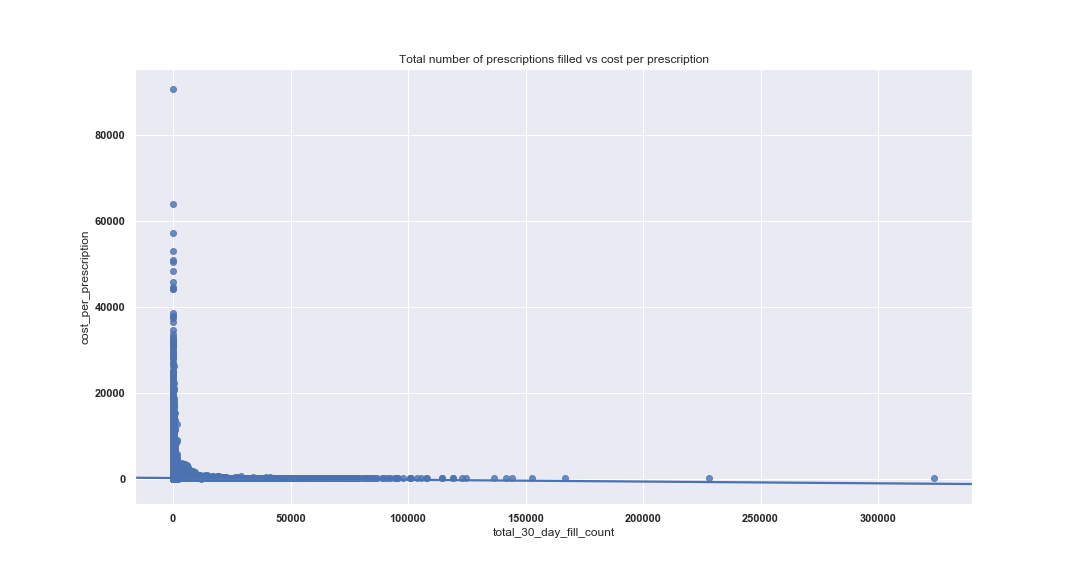


Finally I looked at how ‘HCC score’ affects drug cost. HCC score is “a risk adjustment model that is used to calculate risk per individual on a basis of how cumulative chronic illnesses  predict future healthcare costs. Essentially specific chronic illnesses have specific HCC values depending on their average costs and adding up these different chronic illness HCC values results in a specific HCC score. Physicians with Higher average HCC scores for their patient’s population were correlated with higher average costs. Of interest, family practice physician’s average prescription drug cost correlation with HCC score seemed much weaker. One could presume this is because patients who are very sick are going to be getting much of their care and prescription from specialists. Below are two charts demonstrating the relationships described above.



Supporting the theory above, you can see that the slope for cost per prescription is higher when comparing HCC risk score without filtering doctor types, while internal medicine (primacy care physicians) had a much weaker slope.

I also found that to demonstrate the correlations of chonic illness to average cost per prescription it was helpful to filter out doctor’s who barely prescribed any prescription. This could be a result of these doctors prescribing so little medications that just one or two expensive medication could dramatically increase the average costs of the medications the prescribe as a whole. Additionally, if they are prescribing so few medications they’re really not important to the underlying problem of high drug costs because they prescribe so little and make up a small amount of overall drug costs. Regardless of the reason it is clear that physicians who prescribed very little medications are more likely to have significantly higher average drug costs. Below is a plot demonstrating such.



This is a meaningful fact in that it demonstrates noise presented by doctors who do not prescribe many prescriptions in Medicare’s Part D drug plan, the machine learning algorithms are very powerful and can recognizing these trends and take them into consideration when making predictions.

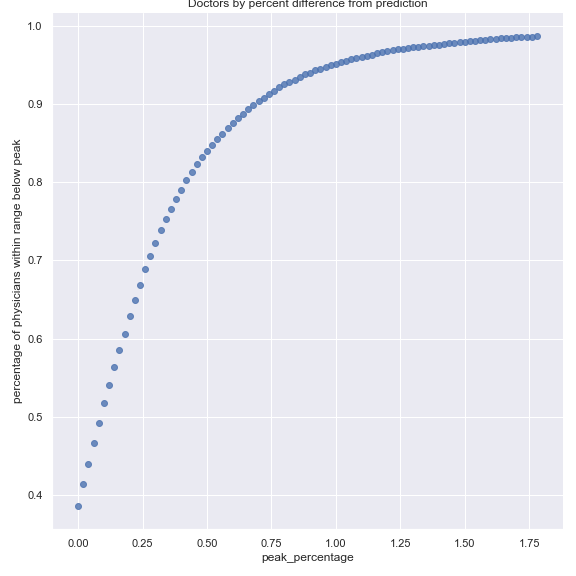
After recognizing how the percentage of patients with specific chronic illness had stronger or weaker relationships with average drug cost with specific specialties, I felt that I needed to improve my models performance by making the model recognize the difference in specialties as categorical variables. To be able to do so, I had to deal with there being over 70 different specific and the machine learning models wouldn’t be able to recognize the different specialties. My solution to this was to decrease the data so that only the 10 specialties were included. I also chose specialties with the most doctors prescribing over 3,000 medications. I did this because I wanted to choose specialties with the largest number of doctor with meaningful data. As opposed to simply the largest amount of doctors who had made any prescription regardless of how many of the physicians were creating random noise due to the lack of prescription volume.

Once I had filtered the data to ten specialties I then used pandas’ get\_dummies function to create dummy variables for each of the 10 specialties so that my algorithms would be able to recognize relationships between specialties and other things.

Once I had completed this, I began putting together the code for scikit learn’s RandomForestRegressor. First I ran GridSearchCV to find the best hyperparameters for my RandomForestRegressor object with my data.

After computing the best hyperparameters, I used random forest on the data set to create predictions for what the y\_values. I then used RandomForest.score() to find out what my accuracy score was. The score value for Random forest tells the Coefficient of Determination for the model. It tells how much of the variance for the dependent variable the model is able to explain by the independent variables. After running Random forest on the data we got a Coefficient of Determinitation of .65, this is a reasonable score.

Finally as a way of indicating how close my prediction’s were to the actual y\_test values, I calculated the percentage of predictions that were within 20%, 30% and 40% of the actual y\_test values. The slope was not normal. A theory about this is that a large number of physicians are particularly affected by recommendations and kick backs from pharmaceutical reps.

****

As a means of figuring out how effective the model was, I calculated the percentage of physicians that whose total cost of prescriptions was less than a specific percentage of the prediction. Below are the values for these distinct percentages:

Within 120 percent of predicted prescription drug costs = .627 percent of physicians

Within 140 percent of predicted prescription drug costs = .781

Within 150 percent of predicted prescription drug costs = .832

Within 175 percent of predicted prescription drug costs t = .911

Within 200 percent of predicted prescription drug costs = .948

Within 250 percent of predicted prescription drug costs = .979

Within 300 percent of predicted prescription drug costs = .989