CUNY School of Professional Studies

**Title**

Predicting Hospital Readmission Using Medicare Hospital Spending by Claim Data and Choosing a Drug Utilization Model That is More Cost Saving

Student: Bin Lin

Course: DATA 698 – Master Research Project

Professor: Arthur O’Connor

**Introduction**

US healthcare system is one of the most complicated and unique among all the industrialized countries. MIT Medical listed top 5 things that we should know about this health system.

1. There is no universal healthcare.
2. Healthcare is very expensive.
3. Most people in the U.S. have health insurance.
4. Most Americans will get most of their care from a “primary care provider” (PCP).
5. Most Americans will usually need an appointment to get medical care.

For this study, it applies specialized knowledge of data science and machine learning to draw insight from healthcare data. It is well known that the healthcare in the United States is very expensive. According to an article published by CMS (Centers for Medicare & Medicaid Services), America spends around 3.5 trillion dollars on healthcare, which is approximately $10,739 per capita in 2017. The healthcare cost is approximately 17.9% of Gross Domestic Product. This figure is considerably higher than that of most other developed countries (around 10%) according to Department for Professional Employees. In addition, the cost keeps increasing as those baby boomers start hitting their retirement age. If America spends too many resources on healthcare, there will be less money and resources we can spend on elsewhere because the budget is always limited.

Using statistics from Department for Professional Employees, in 2014, 48 percent of U.S. health care spending came from private funds, with 28 percent coming from households and 20 percent coming from private businesses. The federal government accounted for 28 percent of spending while state and local governments accounted for 17 percent. The percentages tell us that public programs take up huge percentage of overall healthcare spending. Typical examples will be Medicare and Medicaid. This study will investigate the breakdown of spending by the hospital by claims to see if higher spending lead to better healthcare outcome. The hypothesis is that for those hospitals spend more on prior-to-index admission, index admission, and post-index admission do not necessarily have better quality of care. For this research in particular, it is using hospital readmission as the criteria for measuring the quality of care.

Per Mayo Clinic, hospital readmission is patient admission to a hospital within 30 days after being discharged from an earlier hospital stay. The reason that the Affordable Care Act (ACA) wants to reduce hospital readmissions is because it is one of the critical category of data used to evaluate the quality of hospital care.

Furthermore, I would want to know if Managed Care Organization (MCO) drug utilization that Affordable Care Act advocates has relatively lower cost than the traditional Medicare drug utilization model-Fee-for-Service (FFS), whether the result is statistically significant. Before speaking about drug utilization, we have to know the definition of FFS and MCO. Using definition from Healthcare.gov:

FFS: it is a method in which doctors and other health care providers are paid for each service performed. Examples of services include tests and office visits.

MCO: it is the managed care organization. They accept a set per member per month (capitation) payment to provide for delivery of Medicaid health benefits and additional services.

In order to rein the cost of prescription drugs, organization such as Missouri Department of Social Services conducted a study to compare the cost between two payment methods. The result showed that Managed Care Organization saves approximately 1.7% of the healthcare cost compared with FFS. However, the study is more of a descriptive analysis of the summary statistics. It does not offer p-values, therefore, we cannot generalize what has been studied. However, this research project will address the issue.

**Backgrounds**

Two very important legislations that were passed in history shaping the healthcare system in US: the Social Security Act Amendment (1965) and the Affordable Care Act (2010). Social Security Act Amendment created two public programs – Medicare and Medicaid.

### [Medicare](http://www.medicare.gov/):

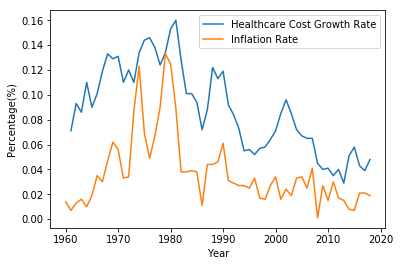
Medicare is a federal program offering insurances to people over 65 (without regarding their income), younger disabled people, and dialysis patients. Patients pay a small monthly premium to keep the coverage. They are still responsible for the deductibles and co-pays. It is run by a federal agency called Centers for Medicare and Medicaid.

### [Medicaid](http://www.medicaid.gov/):

Medicaid is a federal-state assistance program. It helps cover medical and prescription expenses for low-income people. Patients usually pay little to none for the healthcare services. It is run by the state and local government. Therefore, the program is different from state to state.

The data that was collected and used in this project are from these two programs, since data from private payers are proprietary and not very comprehensive. On the other hand, after establishments of these two programs, the healthcare spending skyrocketed with growth rate top inflation rate almost every single year based on the data given by Kaiser Family Foundation (See Below). With the tremendous amount of money spent by these two programs, it is reasonable that the taxpayers would like to see better outcomes such as lower infant and mater mortality rate, higher life expectancy, lower hospital readmission rate (the response variable) and so on.

## Health Care Costs by Year (Drew using Python Matplotlib):



Affordable Care Act:

From the same dataset above, the increase cost started gets curbed after around the year of 2010 when Affordable Care Act finally became effective. Affordable Care Act, which is also called Obama Care, was initially introduced by President Barack Obama and then approved back in the year of 2010. The law has three primary goals:

1. Make affordable health insurance available to more people.
2. [Expand the Medicaid program](https://www.healthcare.gov/medicaid-chip/medicaid-expansion-and-you/) to cover all adults with income below 138% of the federal poverty level.
3. Support innovative medical care delivery methods designed to lower the costs of health care generally.

The third goal of ACA is what this study will explore further. With determination to slow down the cost and flatten the spending curve, the Affordable Care Act established Hospital Readmissions Reduction Program (HRRP) in 2012 to reduce payments to hospitals for excess readmissions, which is the driving force for the healthcare cost to go up.

**Literature Review**

Literature I:

One journal article available online ‘Factors Associated With Increases in US Health Care Spending, 1996-2013’ was written by [Joseph L. Dieleman, PhD](https://jamanetwork.com/searchresults?author=Joseph+L.+Dieleman&q=Joseph+L.+Dieleman); [Ellen Squires, MPH](https://jamanetwork.com/searchresults?author=Ellen+Squires&q=Ellen+Squires); [Anthony L. Bui, MPH](https://jamanetwork.com/searchresults?author=Anthony+L.+Bui&q=Anthony+L.+Bui). The objective is to “quantify changes in spending associated with 5 fundamental factors related to health care spending in the United States: population size, population age structure, disease prevalence or incidence, medical service utilization and service price and intensity.” The study was utilizing the data on these 5 factors from 1996 to 2013. The data also span over 155 health condition, 36 age and sex groups, and 6 types of care. The study used a technique called demographic decomposition to measure the relative effects of these 5 factors as shown in the following equation. The 95% confidence interval (also called uncertainty interval) was calculated using bootstrap method by drawing out 1000 estimates from databases. The end results are reported by inferential statistics. They are composed of the mean and 2.5% and 97.5% of the 1000 estimates. The study drew out the conclusion that healthcare service price and intensity, population growth and population aging are positively associated with increase in health care spending, while disease prevalence or incidences had negative association.

Since the study is unable to determine what implication medical service utilization impose on the healthcare spending, this research project will be an excellent opportunity to fill out the gap.

Generally speaking, people are not able to control the population growth and aging population. While the service price and intensity is very difficult to measure using the datasets that are available. The scope of disease prevalence or incidence is too broad for this course. Therefore, investigating the correlation between medical service utilization and healthcare spending will become more practical and inform policy makers the way to slow down the spending.

Literature II:

There is another study named “Predictive Modeling of Hospital Readmitting Rates Using Electronic Medical Record-Wide Machine Learning: A Case-Study Using Mount Sinai Heart Failure Cohort” which is conducted in Mount Sinai back in 2016. It involves using machine learning techniques to build a predictive model to help guide “the clinical decision for evaluating readmission status”. This cohort study included 1068 patients (with 178 patients were readmitted within 30 days after discharge) and total 4205 variables from Mount Sinai’s EMR. Every patient has Heart Failure as his or her primary diagnosis. The variables that were extracted include diagnosis code (n = 1763), medication (n = 1028), laboratory measurements (n = 846), surgical procedures (n = 564), and vital signs (n = 4).

The methodologies that used in this study will be extremely helpful for my research study. The study used orthogonal validation with logistic regression as well as Principle Component Analysis (PCA) to compare relationship among variables. To reduce bias, the study randomly splits the data into 70% training set and 30% testing set for building the model. When trained and tested all the models, 5-fold cross validation is also performed. Correlation-Based Feature Selection method (CFS) is also very critical step to reduce the likelihood of overfitting on the training data. The final machine learning model that it used is Naive Bayes. Per the article, this model includes only 105 features with AUC = 0.78 and accuracy = 83.19%, which are both much higher than the existing predictive models.

This study is one of the many studies that concentrate on the features that were extracted from EMR. For this research study tackles the issue from administrative perspective instead of from clinicians’ perspective, with particular focus on money. Allocating the fund reasonably at the hospital level could have potential benefits to reduce the hospital readmission. In the meantime, the techniques offered by this study will still be helpful for my study.

**Hypothesis**

First Issue:

Null Hypothesis (H0): Higher Medicare Hospital Spending by Claim does not lead to lower hospital readmission

Alternative Hypothesis (HA): Higher Medicare Hospital Spending by Claim leads to lower hospital readmission

Second Issue:

Null Hypothesis (H0): FFS and MCO prescription drug utilizations are equally cost saving

Alternative Hypothesis (HA): FFS and MCO prescription drug utilizations are not equally cost saving

**Data Sources**

Data were collected from [www.data.gov](http://www.data.gov), [www.cms.gov](http://www.cms.gov), and [www.healthcare.gov](http://www.healthcare.gov). The first problem that was addressed will use Medicare Hospital Spending by Claim data as well as the Hospital Readmission Reduction Program data. The seven claim types (Home Health Agency, Hospice, Inpatient, Outpatient, Skilled Nursing Facility, Durable Medical Equipment, Carrier) over three admission periods (1 to 3 days Prior to Index Hospital Admission, During Index Hospital Admission, and 1 through 30 days After Discharge from Index Hospital Admission) are available for each individual hospital. After data tidying and transformation, the final dataframe will have 21 explanatory variables. And plus one Complete Episode variable, total 22 explanatory variables will be analyzed. The outcome variable will be the actual number of readmission or readmission rate. The second problem will use the State Drug Utilization Data 2017. Two categorical variables are FFS and MCO, while the Units Reimbursed and Medicaid Amount Reimbursed will both be numerical variables.

**Statistical Methods**

Across all the variables, correlation coefficients between any two variables will be calculated using Pearson’s Correlation method to summarize the strength of linear relationship between two variables. It can serve as a way for the feature selection as well. The variables that are highly correlated with each other may not add extra information to the final model, but act redundantly to slow down the training process. In addition, the data will be centered and scaled to help improve the accuracy of the data. The formula is showed below, where u is the mean and s is the standard deviation.

*Scaled Value = (x - u) / s*

In addition, to increase the validity of the model, dataset will be randomly divided into the training and testing dataset (80/20) with 10 fold of cross validation. A multivariable linear regression model will be built, and subsequently evaluated using Accuracy, Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Square Error (RMSE) et cetera. Other types of models such as Random Forest, Support Vector Machine, Neural Networks, K-Nearest Neighbors, Gaussian Processor, and Gradient Boosting models will be built as well. In the end, the best model will be picked up to make predictions on the testing dataset. If time permits, logistic regression model can also be built in order to investigate if certain hospital will have higher than average national hospital readmission or lower. The average national hospital readmission rate will be calculated as the decision boundary to determine the classification. Any value equal or higher than the average will receive classification of 1, while any value smaller that the average will receive classification of 0. Confusion matrix will be built to help calculate sensitivity, specificity, precision, FNR, FDR etcetera. ROC curve will be drawn to illustrate the accuracy of the model. The AIC, F1 score and area under the curve (AUC) are the parameters that will be used to evaluate the accuracy of the model.

For prescription drug utilization study, Two Sample T-Test will be performed to check if the Medicaid Amount Reimbursement per each unit of drug will be any different if MCO utilization and FFS utilization are adopted by the state Medicaid program. The differences will be measured in terms of T-score. Then it can be translated into p-value to tell us how significant is the differences, whether the MCO has higher Medicaid amount per each unit of drug or lower. Confidence interval will also be calculated. In this case, if the interval crosses zero, then the result is not significant and vice versa. Another approach is to resample the data points and calculate its bootstrap statistics.

**Data Acquisition and Wrangling**

Data Acquisition:

Three csv files were collected from internet. The first one is the “Hospital Readmission Reduction Program” which was downloaded from healthdata.gov. It is composed of 12 features and 19344 observations. Most of the features are self explanatory by the columns names, except the excess readmissions. According the CMS, “excess readmissions are measured by a ratio, calculated by dividing a hospital’s number of “predicted” 30-day readmissions for heart attack (AMI), heart failure (HF), pneumonia, chronic obstructive pulmonary disease (COPD), hip/knee replacement (THA/TKA), and coronary artery bypass graft surgery (CABG) by the number that would be “expected,” based on an average hospital with similar patients.”

The second dataset is downloaded from data.gov. The data of the dataset is called “Medicare Hospital Spending by Claim”. It is a dataset with 13 features and 67826 observations. Per CMS, the data displayed here describes average spending levels during hospitals’ Medicare Spending per Beneficiary (MSPB) episodes by Medicare claim type.

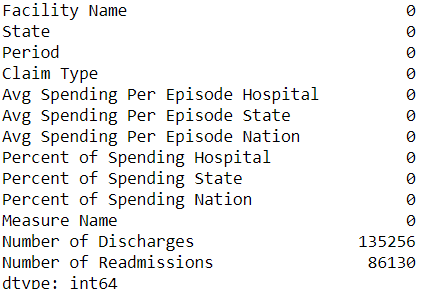
The third dataset is downloaded from data.medicaid.gov. The name is called State\_Drug\_Utilization\_Data\_2019. It is composed of 20 features and 3,578,549 observations.

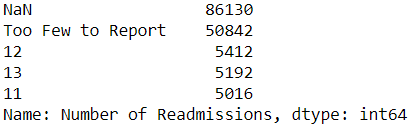
Per CMS, Drug utilization data are reported by states for covered outpatient drugs that are paid for by state Medicaid agencies since the start of the Medicaid Drug Rebate Program. The data includes state, drug name, National Drug Code, number of prescriptions and dollars reimbursed.

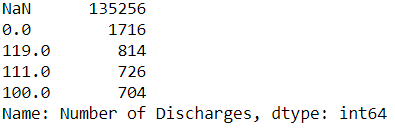
Data Wrangling:

After loaded the first two datasets into python Jupiter notebook and got the summary of the data frame, it is immediately noticeable that both data frames share several identical variables, such as Facility ID, Facility Name, and State. Therefore, these two data frames were merged together. The resulting data frame contains 22 features and 338,976 observations.

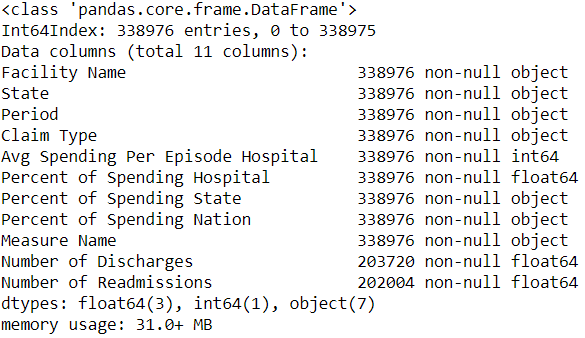
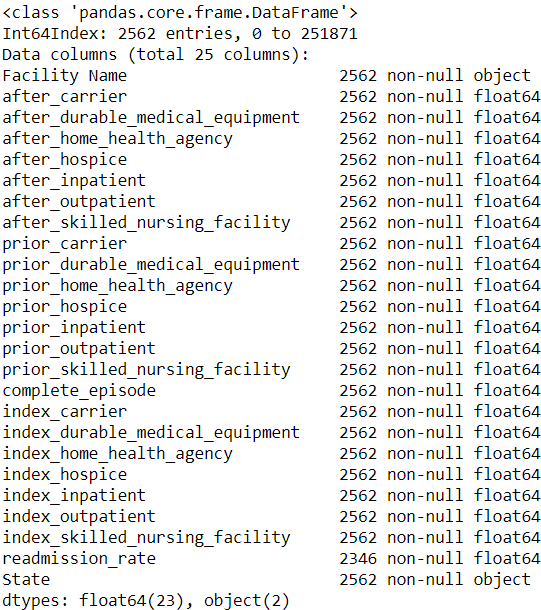
After dropping variables that are unnecessary or not really relevant to study of my hypothesis, the amounts of missing data were calculated as well as shown in the following figure. Number of Discharges and Number of Readmission have quite a lot of missing values. Since those variables are very critical for my analysis, imputation of the missing data will be performed as necessary. However, the exploratory data analysis will be performed first so that it will be useful to find out the best way to fix the issue of missing data, either dropping the observations or imputation.



****

****

Another critical step of data wrangling is to create two new variables: Period\_and\_Claim, readmission\_rate. Period\_and\_Claim was created by combining variables of Period and Claim together from original dataframe, in the meantime, readmission\_rate was calculated by dividing Number of Readmissions from Number of Discharges. Then the average hospital spending per episode was calculated grouped by each facility name as well as each type of Period\_and\_Claim. After transforming data from long to wide, the data table became significantly different from the original dataset (See below). The left one is the old data table while the right one is the new data table.

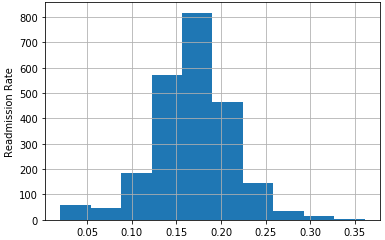
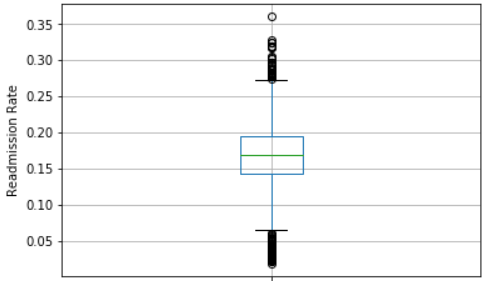
** **

As we can tell, the outcome variable radmission\_rate does have some missing value, but the amount was much less than what was initially thought. Therefore, those missing value rows will be dropped before the analysis.

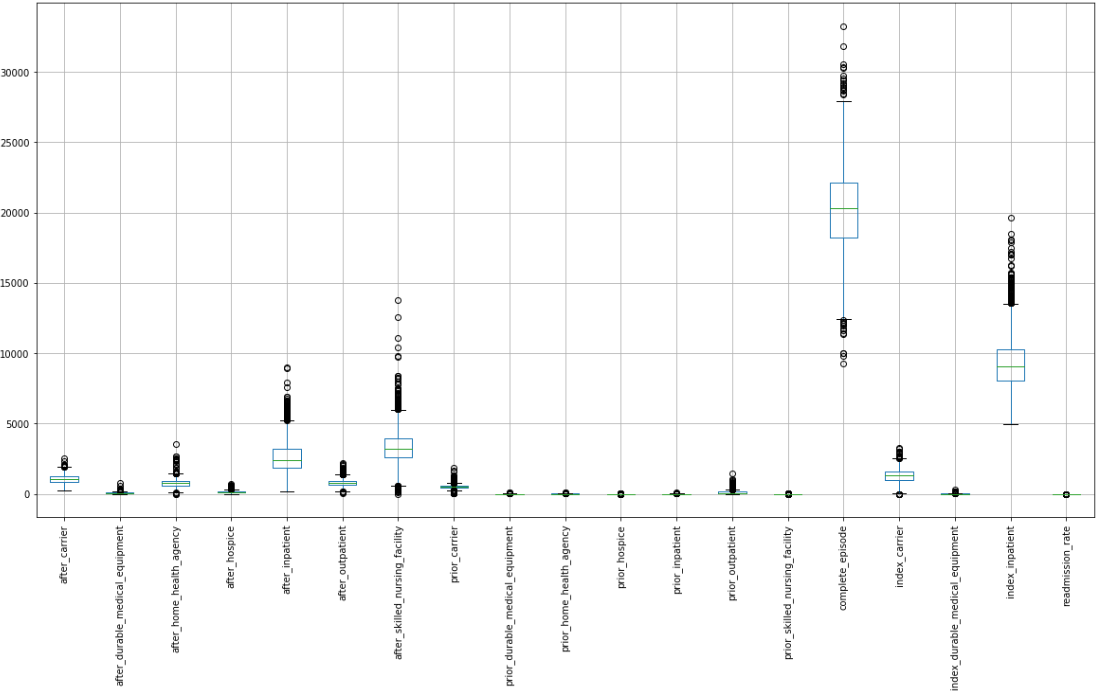
**Exploratory Data Analysis**

Exploratory data analysis can facilitate gaining general knowledge about the dataset. It usually involves using visualization tools to summarize the characteristics of each variable. Descriptive or inferential Statistics might be used as well. EDA will help formulate, hypnotize, and explore for more questions to ask and study.

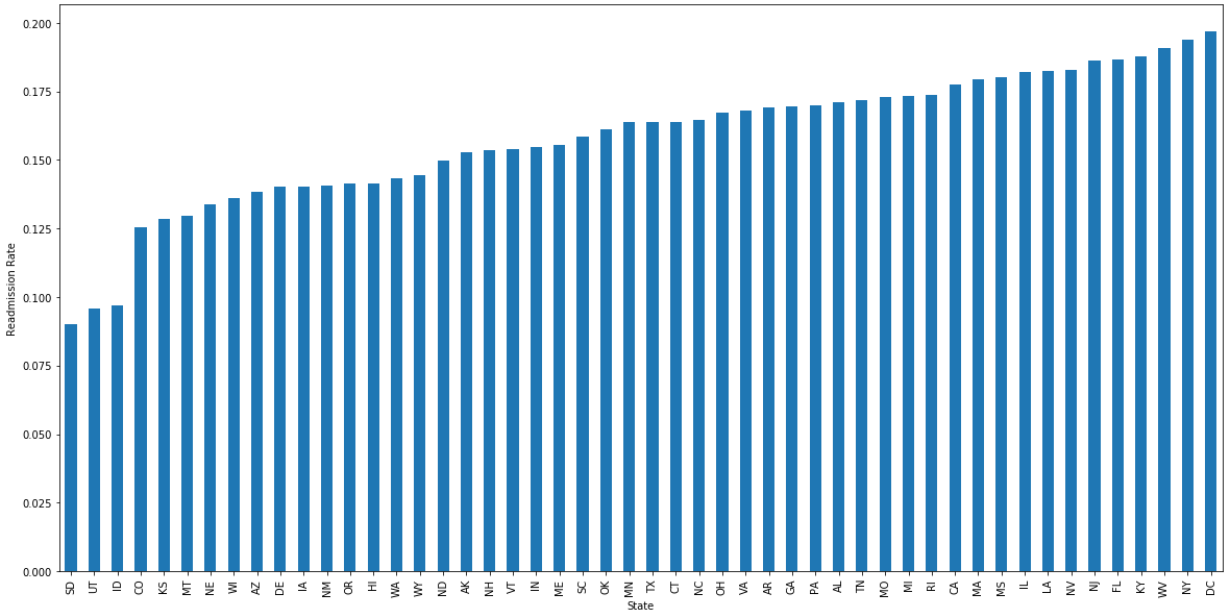
After investigating further into the variables, the boxplot was drawn to visualize the distribution of the outcome variable – readmission\_rate. The mean and interquatile range (IQR) are pretty much preside in the center. The distribution is basically symmetric. Same phenomenon was observed from the histogram. The distribution is unimodel, symmetric, and bell shaped. It is a typical distribution of normal distribution.



From the boxplot of other variables, many of them have outliers and are skewed to the right. It is reasonable since the odd is getting very unlikely as the claim amount for each facility for each claim type gets larger. The situation deems the preprocessing will be necessary for them before the data can be used to train and test the model. The method that was used is called StandardScaler from sklearn package.



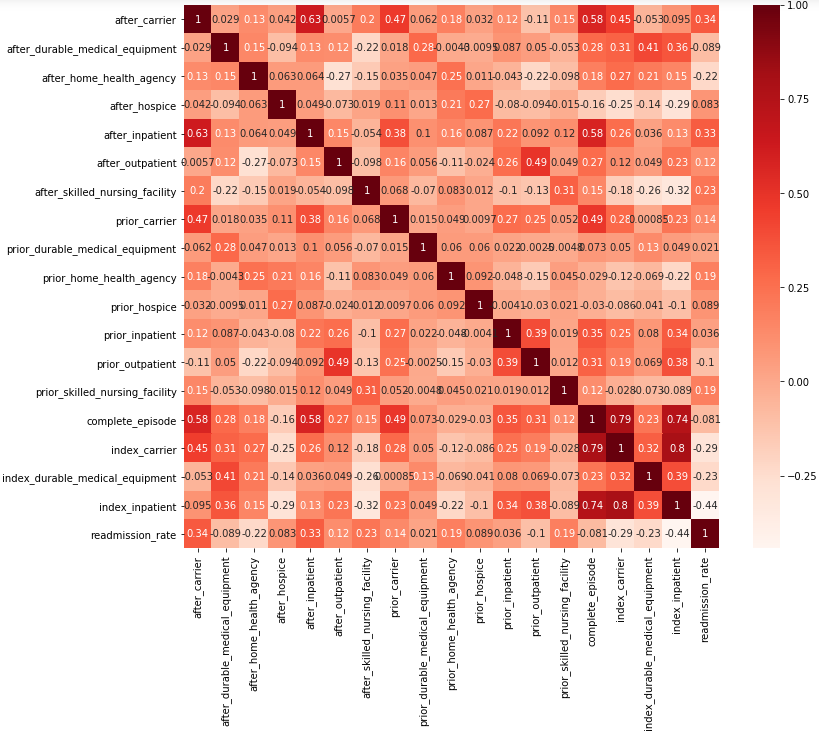
One interesting insight is by comparing the readmission rates among each state in America in bar graph arranged by ascending order. It is shown that West Virginia, New York, and DC have the higher readmission rate in the country. The rate is approximately 0.2. On the other hand, South Dakota, Utah, and Indiana have the lowest readmission rate in the country. Their rate is approximately 0.1. The national average is 0.167798 with standard deviation 0.045579 based on the summary statistics.



**Findings and Conclusion**

Feature Selection:

The following heatmap checks for collinearity between each variable. The redder the color is, the higher the correlation coefficient. In general practice, the feature variables that are highly correlated with outcome variables (readmission\_rate) should be dropped. In this case, there is no such variables exist. In addition, when we perform feature selections, whichever features are high correlated with each other can also be dropped as well, since additional feature won’t add extra information to the model but rather slows down the training process. From the following figure, we can tell complete\_episode, index\_carrier, index\_inpatient have very strong positive relationship because their correlation coefficients are approximately 0.8. However, supposedly, the spending of a hospital on each claim does not really affect each other; high correlation would most likely due to chance. Therefore, all the feature variables except index\_home\_health\_agency, index\_hospice , index\_outpatient, index\_skilled\_nursing\_facility, which are all zero across the board, will be used to build the regression model.



Model Building:

Several models were built and evaluated for the accuracy of predicting readmission rate for each hospital (see below). The data were splitted into training and testing data set in 80: 20 ratio. Moreover, 10 folds cross validation were performed as well in order to mitigate overfitting.

* Ordinary Least Square Regression
* Support Vector Machine
* K-Nearest Neighbors Regression
* Multi-layer Perceptron Regression
* Gaussian Process Regression
* Random Forest Regression (Ensemble)
* Gradient Boosting Regression (Ensemble)
* Voting Regression (Ensemble)

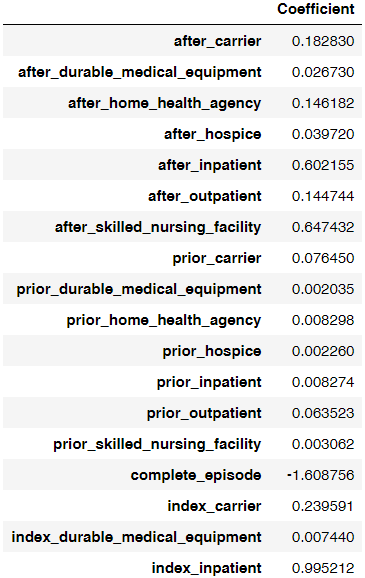
Ordinary Least Square:

The formula that was used for Ordinary Least Square method is from multi-variable linear regression (see below).

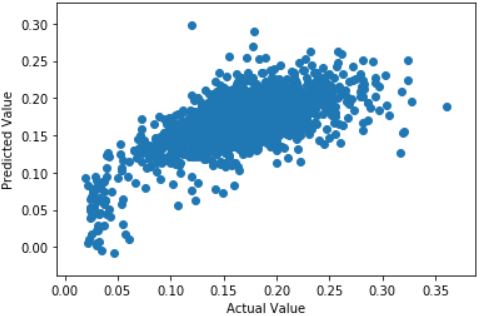


The intersection b0 = 0.1678751983882069. The coefficients for dependent variables are listed

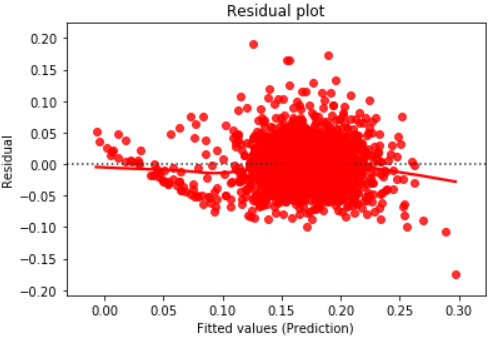
below.



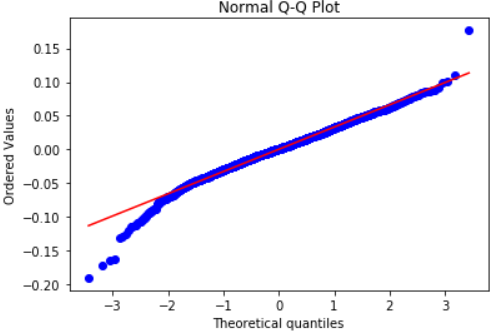
The following scatter plot between the actual readmission rate and the predicted readmission rate shows that both values are about to fall around the straight line y = x. So the prediction is approximately equal to the actual value, but there are still differences or what we called residuals.



The residual plot from below is the scatter plot between the actual values and residuals. If the figure shows any patterns then it means the relationship between target variable and feature variables is non-linear. On the other hand, if the figure shows no pattern then it means the relationship between target variable and feature variables is linear. From our example, the residuals fall almost evenly on both side of zero and display no pattern at all. It indicates the relationship is linear.



QQ plot generally depicts if the residuals are normally distributed. When the residuals falls closely on the normal line, it means the residuals are normally distributed, which is the case in the following figure.



There are total 4 measurements were calculated to evaluate regression performance.

* R² ([coefficient of determination](http://en.wikipedia.org/wiki/Coefficient_of_determination)): It provides a measure of how well future samples are likely to be predicted by the model.
* Mean Absolute Error (MAE): It is a risk function corresponding to the expected value of the absolute error loss or l1-norm loss.
* Mean Squared Error （MSE）：It is a risk function corresponding to the expected value of the squared error loss or quadratic loss.
* Root Mean Squared Error (RMSE): It is the square root of MSE

The result is shown in the following.

*Model Accuracy R^2: 0.4658689444977252*

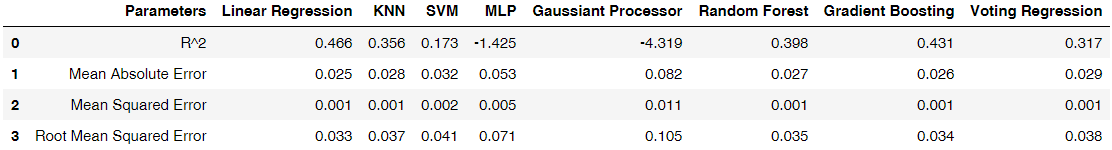
*Mean Absolute Error: 0.025316375788329945*

*Mean Squared Error: 0.0011091736487517767*

*Root Mean Squared Error: 0.033304258717944416*

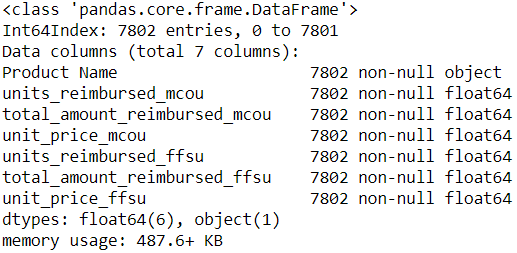
From the summary table for all the models, it is noted that Multi-variable linear regression model has the best performance as it has the highest correlation coefficient, and lowest MAE, MSE,

and RMSE.

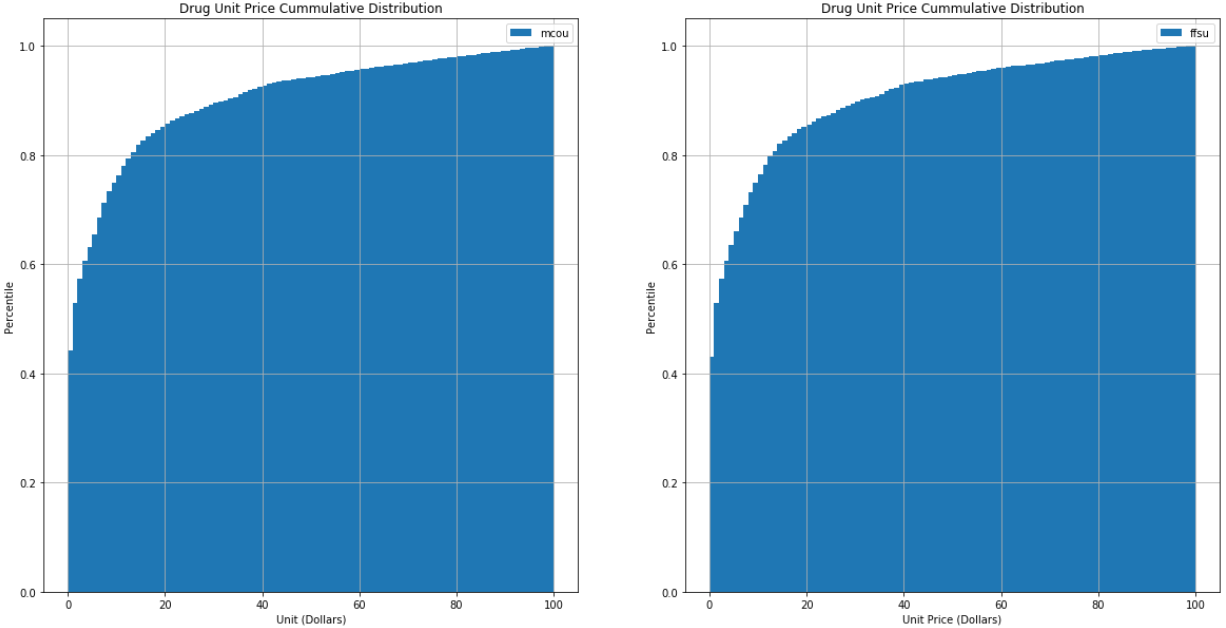


Drug Utilization Comparison:

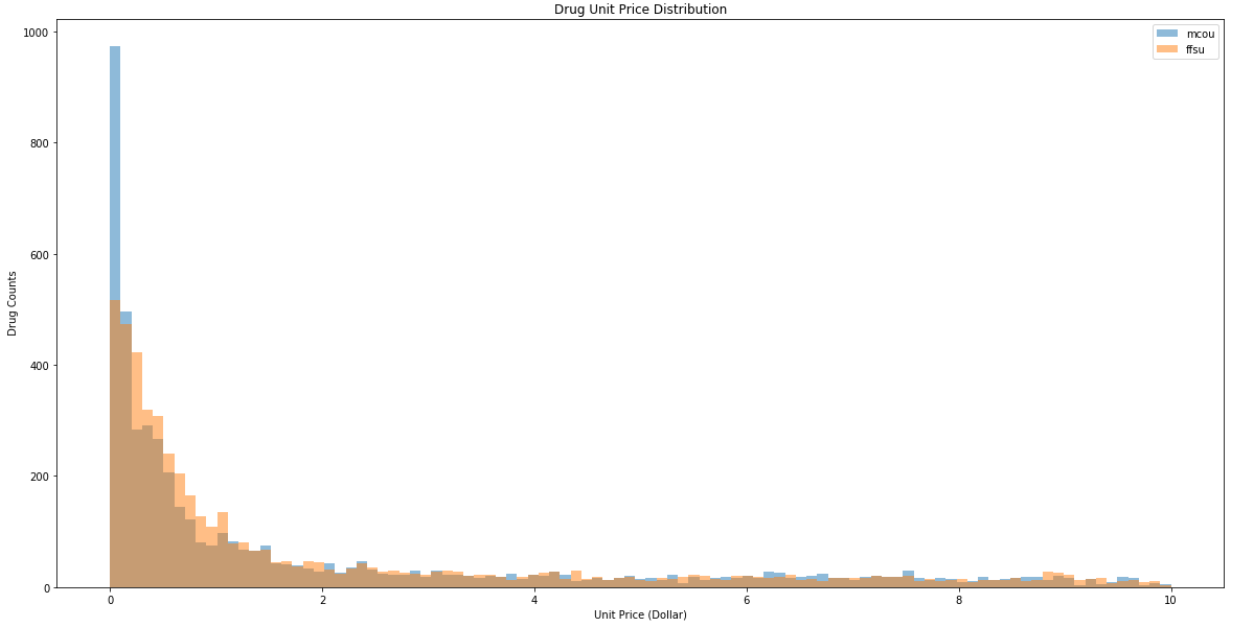
There is a separate dataset available that can be used for drug utilization studies. After data tidy and transformation were completed, the final dataframe is shown as follows.



The cumulative distribution of the unit drug price is very similar between FFS and MCO drug utilization models when the price is under $100. Since majority of the unit price is under $100, therefore, unit price which is greater than $100 are not show in the following figure. It is also very noticeable that more than 80% of drugs in the market are sold under $20 for each unit.



The histograms of drug unit price for these two utilization models display very interesting phenomenon. Drugs reimbursed under Managed Care Organization have large number of extremely low priced drugs as shown in the blue peak. While drugs reimbursed under Fee for Service utilization model have relatively larger amount of drugs whose prices are a little bit higher than those of their counterparts for the other model. The curve appears to be smoother for FFS model than MCO model. For drugs that are priced higher than $2 per unit, distribution become similar for both models.



Two sample T-test was conducted to compare the mean of drug unit price for both utilization model. The mean and standard deviation for both groups were shown in the following.

*Managed Care Organization mean unit price: 252.64771403609996*

*Fee for Service mean unit price: 284.3316189717122*

*Managed Care Organization unit price standard deviation: 1631.6444939455073*

*Fee for Service unit price standard deviation: 4152.241952381325*

T statistics was calculated to be -0.6272641341087188 and the corresponding p-value is

0.530495288515858. Since the p-value is way higher than the significance level alpha = 0.05, we failed to reject the null hypothesis, which means the two sample t-test accepts the null hypothesis.

Furthermore, the 95% confidence interval was calculated to be (-130.69596747600747,

67.32815760477901). When the confidence interval across 0, it means the null hypothesis is true.  
In short, the study could not find the significant differences in terms of mean between two

sample groups.

**Conclusion**

For the first issue this study is trying to address, it investigates the relationship between Medicare hospital spending by claim and the hospital readmission rate. Multiple machine learning algorithms were developed to make predictions on the readmission rate based on the data of what is available. After extensive evaluation, it is determined that the ordinary least square method

for multi-variable linear regression has the highest accuracy. According to the resulting coefficient for each feature variable, we can tell that most variables have positive relationship

with the outcome variable readmission rate. However, the claim on the complete episode of the

hospitalization has negative relationship with outcome variable. Since claim for the complete

episode is the conglomeration of all the other claim types combined, it might present a

phenomenon called Simpson’s Paradox. It is a phenomenon “in probability and statistics, in

which a trend appears in several different groups of data but disappears or reverses when these

groups are combined” according to the definition from Wikipedia. Overall speaking, the higher Medicare Hospital Spending by Claim does not necessarily lead to lower hospital readmission as demonstrated by the negative relationship exists in one of the feature variable and outcome

variable.

For the second issue this study is trying to address, the result failed to reject the null hypothesis

since the p-value is greater than 0.05 and confidence interval cross zero. Therefore,

the sample data show FFS and MCO prescription drug utilizations are equally cost saving at

significant level of 0.05.

**References**

#### Healthcare in the United States: The top five things you need to know | MIT Medical

[https://medical.mit.edu/my-mit/internationals/healthcare-united-states](javascript:openWebLink('https://medical.mit.edu/my-mit/internationals/healthcare-united-states'))

#### The U.S. Health Care System: An International Perspective — Department for Professional Employees, AFL-CIO

[https://www.dpeaflcio.org/factsheets/the-us-health-care-system-an-international-perspective](javascript:openWebLink('https://www.dpeaflcio.org/factsheets/the-us-health-care-system-an-international-perspective'))

#### The Facts on Medicare Spending and Financing

[https://www.kff.org/medicare/issue-brief/the-facts-on-medicare-spending-and-financing/](javascript:openWebLink('https://www.kff.org/medicare/issue-brief/the-facts-on-medicare-spending-and-financing/'))

Missouri Department of Social Services. COMPARING PERFORMANCE: *MANAGED CARE AND FEE-FOR-SERVICE,* January 2015

Dieleman, J., Squires, E., Bui, A., Campbell, M., Chapin, A., & Hamavid, H. et al. (2017). Factors Associated With Increases in US Health Care Spending, 1996-2013. *JAMA*, *318*(17), 1668. doi: 10.1001/jama.2017.15927

#### What is the difference between Medicare and Medicaid?

[https://www.hhs.gov/answers/medicare-and-medicaid/what-is-the-difference-between-medicare-medicaid/index.html](javascript:openWebLink('https://www.hhs.gov/answers/medicare-and-medicaid/what-is-the-difference-between-medicare-medicaid/index.html'))

[Shameer K](https://www.ncbi.nlm.nih.gov/pubmed/?term=Shameer%20K%5BAuthor%5D&cauthor=true&cauthor_uid=27896982), [Johnson KW](https://www.ncbi.nlm.nih.gov/pubmed/?term=Johnson%20KW%5BAuthor%5D&cauthor=true&cauthor_uid=27896982), [Yahi A](https://www.ncbi.nlm.nih.gov/pubmed/?term=Yahi%20A%5BAuthor%5D&cauthor=true&cauthor_uid=27896982), [Miotto R](https://www.ncbi.nlm.nih.gov/pubmed/?term=Miotto%20R%5BAuthor%5D&cauthor=true&cauthor_uid=27896982), [Li LI](https://www.ncbi.nlm.nih.gov/pubmed/?term=Li%20LI%5BAuthor%5D&cauthor=true&cauthor_uid=27896982), [Ricks D](https://www.ncbi.nlm.nih.gov/pubmed/?term=Ricks%20D%5BAuthor%5D&cauthor=true&cauthor_uid=27896982), et al. (2016). PREDICTIVE MODELING OF HOSPITAL READMISSION RATES USING ELECTRONIC MEDICAL RECORD-WIDE MACHINE LEARNING: A CASE-STUDY USING MOUNT SINAI HEART FAILURE COHORT. [Pac Symp Biocomput.](https://www.ncbi.nlm.nih.gov/pubmed/27896982) 2017;22:276-287. doi: 10.1142/9789813207813\_0027.