Is Your Hospital Bamboozling You?

Helping Patients Shop For Medical Care In Michigan

Team 54 (Cohort 2) – Nadia Ayub, Junting Huang, Amber Johnson, Brionna Jordan, Nancy Njuguna, Kasole Nyembo, Evelyn Wangai.

Contents

Business Problem

Effective January 1 2021, the US Centers for Medicare and Medicaid Services (CMS) promulgated a new rule for hospitals requiring them to disclose pricing on a wide range of hospital services. This transparency regulation is intended to help patients make informed decisions, increase competition, and drive down the cost of healthcare. Our project focused on exploring how this price transparency regulation can help patients understand healthcare costs, and what the relationships between hospital pricing and demographic data really are. We hope to uncover if health care services are indeed more shoppable across different communities and geographical regions.

Business Impact

Healthcare costs are one of the largest categories of consumer spending (8.1 percent of consumer expenditures; BLS 2019a). Consumers are becoming more invested in understanding the associated costs of healthcare services. Based on the research provided by the Robert Wood Johnson Foundation (2016), the figure below provides key insights on the price transparency in the U.S. and how American consumers expect it to evolve.

Figure 1: Key facts on healthcare transparency

Pricing transparency has been advertised as a movement to "help Americans know the cost of a hospital item or service before receiving it." (Centers for Medicare and Medicaid Services) Each hospital operating in the U.S is now required to provide a machine-readable file and a portal/display where patients can pick and choose services in a user-friendly format, allowing consumers to estimate costs. Based on this premise, we downloaded the pricing data from hospitals for two purposes:

1. To develop a tool which allows patients to research costs for various procedures; this tool will reduce their financial burdens and empower them to take a proactive approach with their health.
2. To provide an analysis on the variability of hospital pricing based on demographic and hospital statistics.

Based on our preliminary research, we chose the state of Michigan to explore its pricing data. Although we investigated additional states including Massachusetts, Florida, Illinois, and Ohio,

the data in Michigan was reasonably easy to find and arguably representative of U.S. demographics.

Data Sourcing and Cleaning

Data Sourcing

We pulled our data from three main sources: (1) pricing data from hospital websites, (2) hospital data from the American Hospital Directory, and (3) population/demographic data primarily from the Michigan State Department of Health and Human Services. After attempting to manually source the data, we discovered how ineffective and inefficient this was. We then developed a python script that extracts relevant information from individual data files. The organization of our dataframe is provided below:

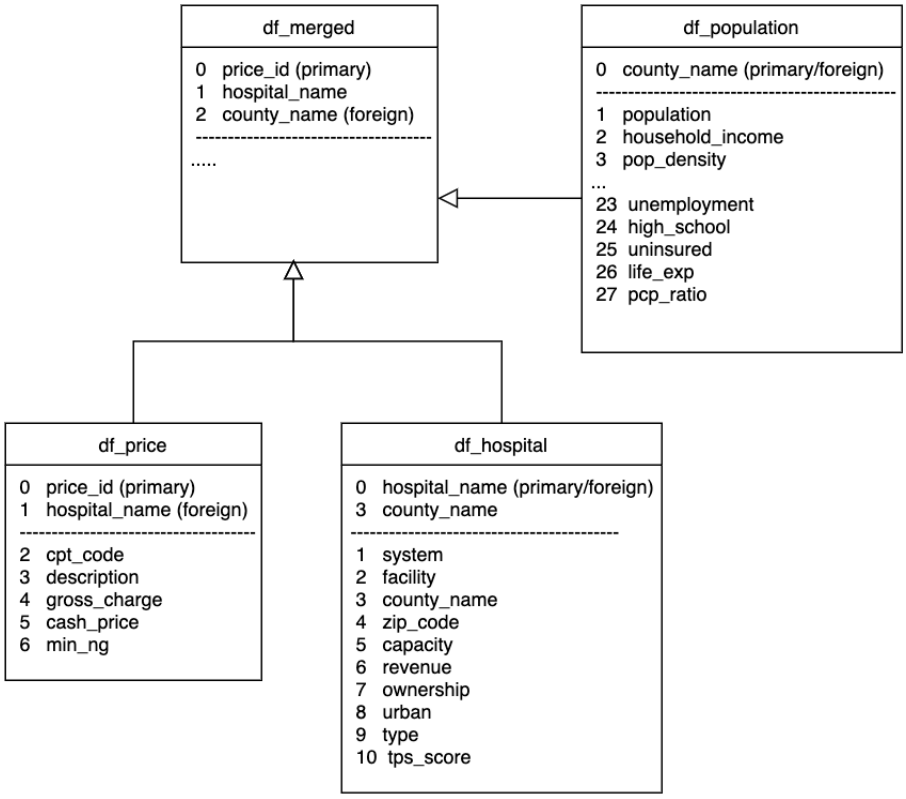

Figure 2: Data Schema

Overall, our dataset contained a total of 120 hospitals in 18 different hospital systems. We chose seven procedures with four pricing variables. These procedures were chosen because they cover a wide range of services on the pricing spectrum; they are also among the most coded procedures in our hospital data sets. We selected the codes from a CMS-specified list of 70 codes that hospitals are expected to publish. The procedures we chose are below (AAPC):

1. **Brain MRI (CPT code: 70551):** a medical procedural code under the range - Diagnostic Radiology (Diagnostic Imaging) Procedures of the Head and Neck.
2. **Vaginal Delivery (CPT code: 59409):** a medical procedural code under the range - Vaginal Delivery, Antepartum and Postpartum Care Procedures.
3. **Colonoscopy (CPT code: 45378):** a medical procedural code under the range - Endoscopy Procedures on the Rectum.

4. **First Patient visit for 30 minutes (CPT code: 99203):** a medical procedural code under the range - New Patient Office or Other Outpatient Services.
5. **Head CT (CPT code: 70450):** a medical procedural code under the range - Diagnostic Radiology (Diagnostic Imaging) Procedures of the Head and Neck.
6. **Blood Count Test – A (CPT code: 85025):** a medical procedural code under the range - Hematology and Coagulation Procedures.
7. **Blood Count Test – B (CPT code: 85027):** a medical procedural code under the range - Hematology and Coagulation Procedures.

Data Cleaning

For data cleaning, we used OpenRefine as our main platform. Below are the high-level steps we took to clean the data:

- **Identifying duplicates:** Duplicates exist due to multiple reasons. Most commonly, multiple prices are assigned to the same CPT code because of service category (outpatient/inpatient/emergency), insurance payers, or procedure complication. We decided to take the outpatient and low-end complication entries for all these duplicates.
- **Identifying errors.** Errors occur for example when hospital names from different sources do not match. We double-checked the sources to eliminate these errors.
- **Identifying inconsistencies:** Formatting inconsistencies included whitespace, letter case, or even typos, etc., and they were all corrected in OpenRefine.
- **Identifying missing data:** This was addressed on a case-by-case basis. Since we are dealing with pricing, we wanted to note any blanks as NaN. For certain data points such as the quality score, we wanted to note that a 0 meant it was not available and not a score of 0.
- **Identifying outliers:** Outliers may or may not be errors. We examined statistical anomalies to make sure they can be reasonably explained. One error was corrected, where a decimal point (.) was mistaken for comma (,).

Figure 3 shows the contents of our main data frame. Despite our best efforts, by the end of our data sourcing and cleaning phase, we had approximately 5-19.5% of missing data.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 590 entries, 0 to 589
Data columns (total 25 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   price_id          590 non-null    string
 1   cash_price        500 non-null    float64
 2   gross_charge      569 non-null    float64
 3   max_ng            435 non-null    float64
 4   min_ng            435 non-null    float64
 5   capacity          590 non-null    int64
 6   revenue           562 non-null    float64
 7   tps_score         590 non-null    float64
 8   pop_density       590 non-null    float64
 9   household_income  590 non-null    int64
 10  unemployment      590 non-null    float64
 11  sex_ratio         590 non-null    float64
 12  children_percent  590 non-null    float64
 13  senior_percent    590 non-null    float64
 14  white_percent     590 non-null    float64
 15  high_school       590 non-null    float64
 16  pcp_ratio         590 non-null    float64
 17  life_exp          590 non-null    float64
 18  uninsured         590 non-null    float64
 19  cpt_code          590 non-null    string
 20  urban             590 non-null    object
 21  ownership         590 non-null    object
 22  system            590 non-null    object
 23  hospital_name     590 non-null    object
 24  zip_code          590 non-null    string
dtypes: float64(16), int64(2), object(4), string(3)
```

```
price_id              0
cash_price           90
gross_charge         21
max_ng              155
min_ng              155
capacity              0
revenue              28
tps_score             0
pop_density           0
household_income      0
unemployment          0
sex_ratio             0
children_percent      0
senior_percent        0
white_percent         0
high_school           0
pcp_ratio             0
life_exp              0
uninsured             0
cpt_code              0
urban                 0
ownership             0
system                0
hospital_name         0
zip_code              0
dtype: int64
```

Figure 3: Dataframe contents

Exploratory Data Analysis

Hospital Overview

We started with a pricing comparison of different hospitals and systems. Figure 4 shows a breakdown of average cash charge. Based on the graph below, one hospital was an outlier and charged a higher than average price for most of the services.

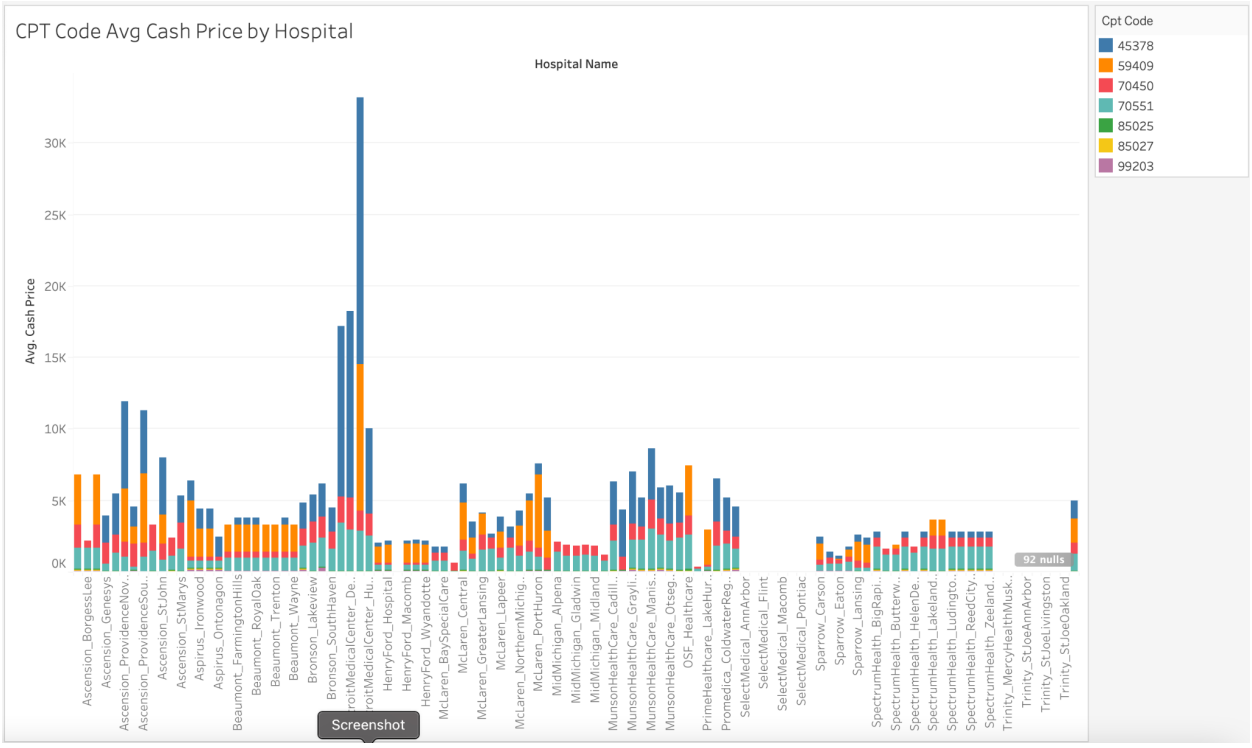

Figure 4: Average cash price divided by CPT code

Price Index

To better estimate pricing data, we divided the data frames into individual procedures and observed density plots. Figure 5 (pictured below) shows the results for each of our pricing variables.

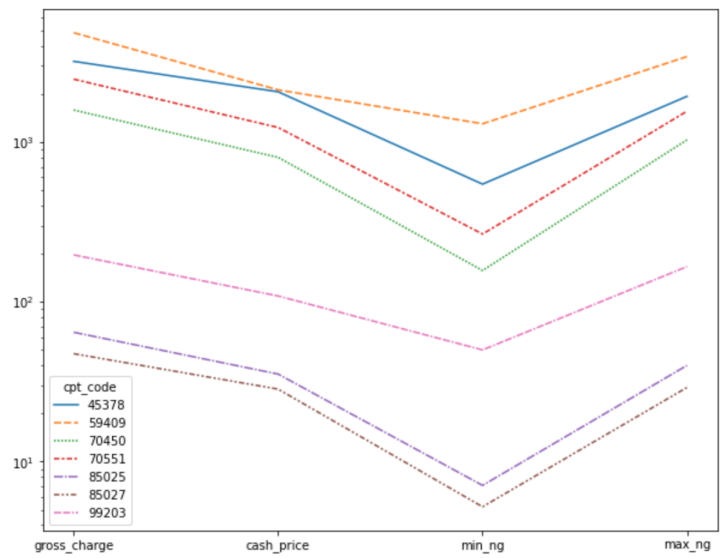

Figure 5: Multiple line plot comparing price variables.

We decided to use cash price as our main price variable for the following reasons: (1) gross charge is the price that a hospital bills, however, it is not the price that patients pay; (2) max negotiated and min negotiated are highly situational because they are dependent on specific payers and insurance plans; (3) cash price is an attainable price point that a patient can negotiate as their bill.

Due to these seven selected procedures having vastly different ranges of prices, we divided all pricing data into seven groups and standardized the cash price within each group. We used SKlearn's StandardScaler by removing the mean and scaling to unit variance. The formula used is as follows: $z = (x - u) / s$

To better understand the pricing level of specific hospitals, we took the mean of all standardized cash prices and assigned a price index score to the hospitals. For this score, 0 was the mean, and anything above 0 would be above the mean (more expensive), and below 0 would be below the mean (cheaper). Our dashboard uses the price index score as a primary metric to inform patients about the hospital's pricing standard.

Cash price below mean = Cheaper		Mean Cash Price	Cash price above mean = Expensive	
-2	-1	0	1	2

Correlation Analysis

We were then interested in the correlation between cash_price (z score) and other demographic and hospital variables. In our original analysis, we used multiple linear regression and ordinary least squares (OLS), but the results were not strong. Therefore, we plotted a correlation matrix (Figure 6) and a biplot of Principal Components Analysis (Figure 7) to better understand the interactions between variables.

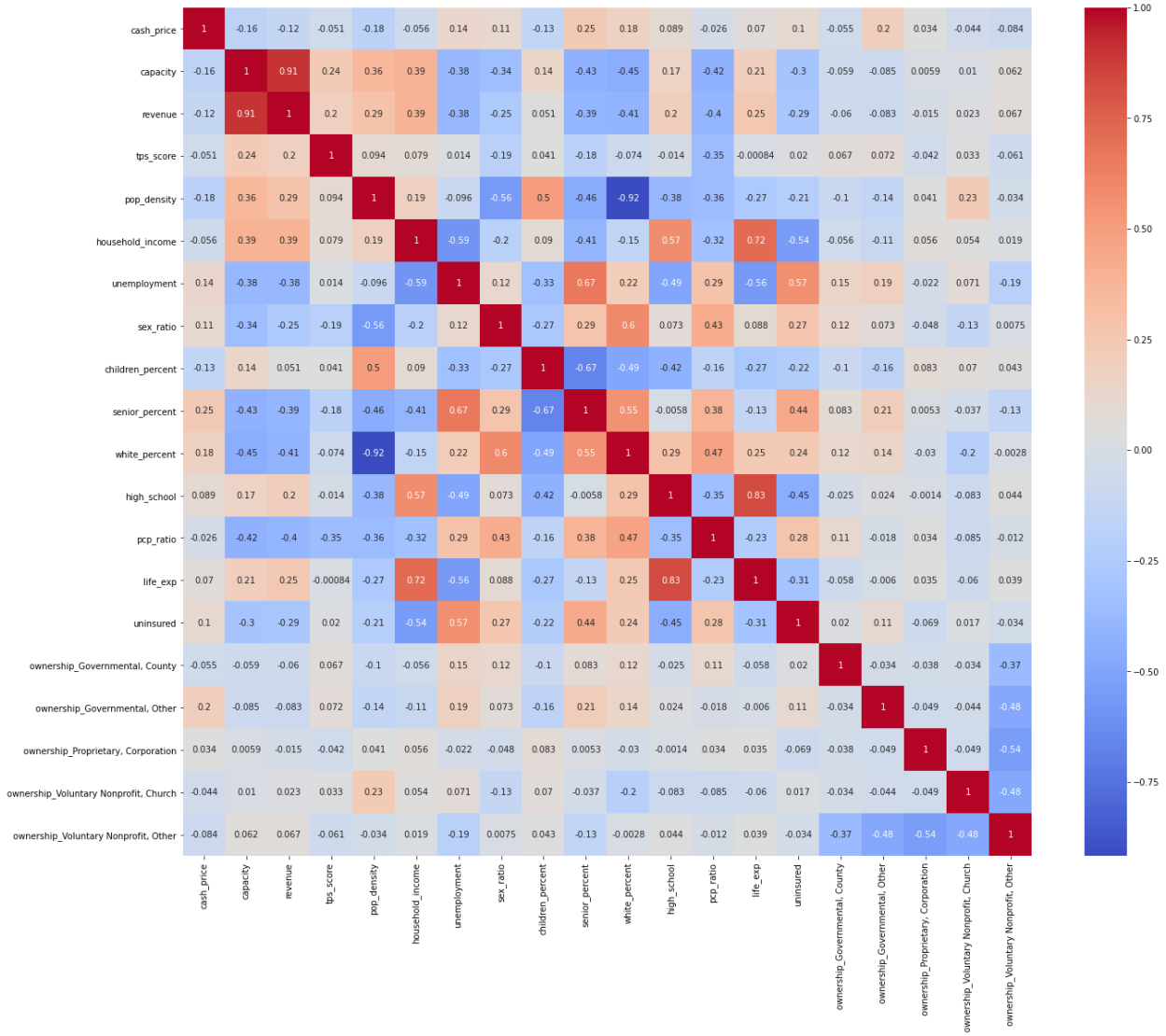

Figure 6: Correlation Matrix with all variables.

As seen in Figure 6, there is a small negative correlation with capacity (-0.16), revenue (-0.12), and TPS score (-0.051). In addition, specifically with demographic data, senior (0.25), white_percent (0.18), unemployment (0.14), and pop_density (-0.18) were also correlated to the cash price. Based on the correlation matrix, we compared trends with cash price and its casual predictors with details provided in Figure 6 above and Figure 7 below:

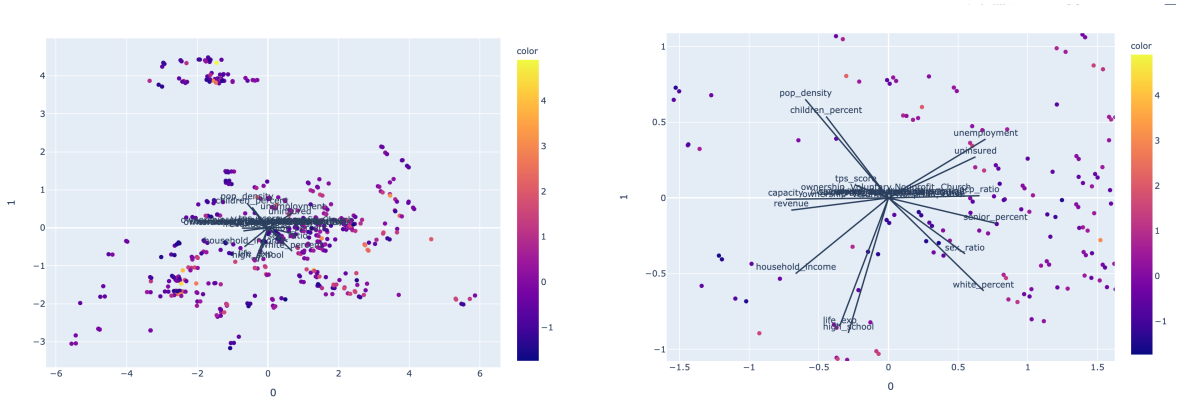

Figure 7: Biplot of Principal Components Analysis (PCA)

The Biplot helps us group correlated variables in order to analyze the possible causes for price discrepancies.

Conclusion

Below are our key EDA conclusions:

1. **Hospital pricing is weakly correlated to a set of key demographic and hospital variables:**

   In Figures 6 and 7, while there were some moderate correlations that could possibly explain some observable trends there are no strong correlations between cash_price and other variables.

   a. **Age: cash_price is higher in counties with higher senior_percent (0.25) and lower children_percent (-0.13)**
   The combination of senior_percent and children_percent indicates the county's average age group. This correlation may suggest that hospitals charge elder population more because they can afford higher price points.
   b. **Demographic: cash_price is higher in counties with higher white_percent (0.18) and lower pop_density (-0.18)**
   Counties with a higher percentage of white residents are typically more rural; these counties tend to have smaller hospitals. Alternatively, counties that are less white tend to have a larger proportion of large hospitals, and they also tend to be more urban. As such, hospitals in rural counties are generally more expensive because they are small, remote, and usually lack competitions.
   c. **Economic: cash_price is higher in counties with higher unemployment (0.14) and uninsured (0.10), but lower household income (-0.06)**
   This may suggest that hospitals charge higher cash prices when fewer people are covered by insurance. Residents in these regions have less power to negotiate the price. These areas also tend to have higher unemployment rates and lower household income.
   d. **Size: cash_price is higher with hospitals with lower capacity (-0.16) and revenue (-0.12)**
   This might mean that larger hospitals are probably charging on volume. Larger hospitals tend to be cheaper than private hospitals. They may benefit from increasing returns to scale, meaning that their costs are lower as they are shared among more physicians.

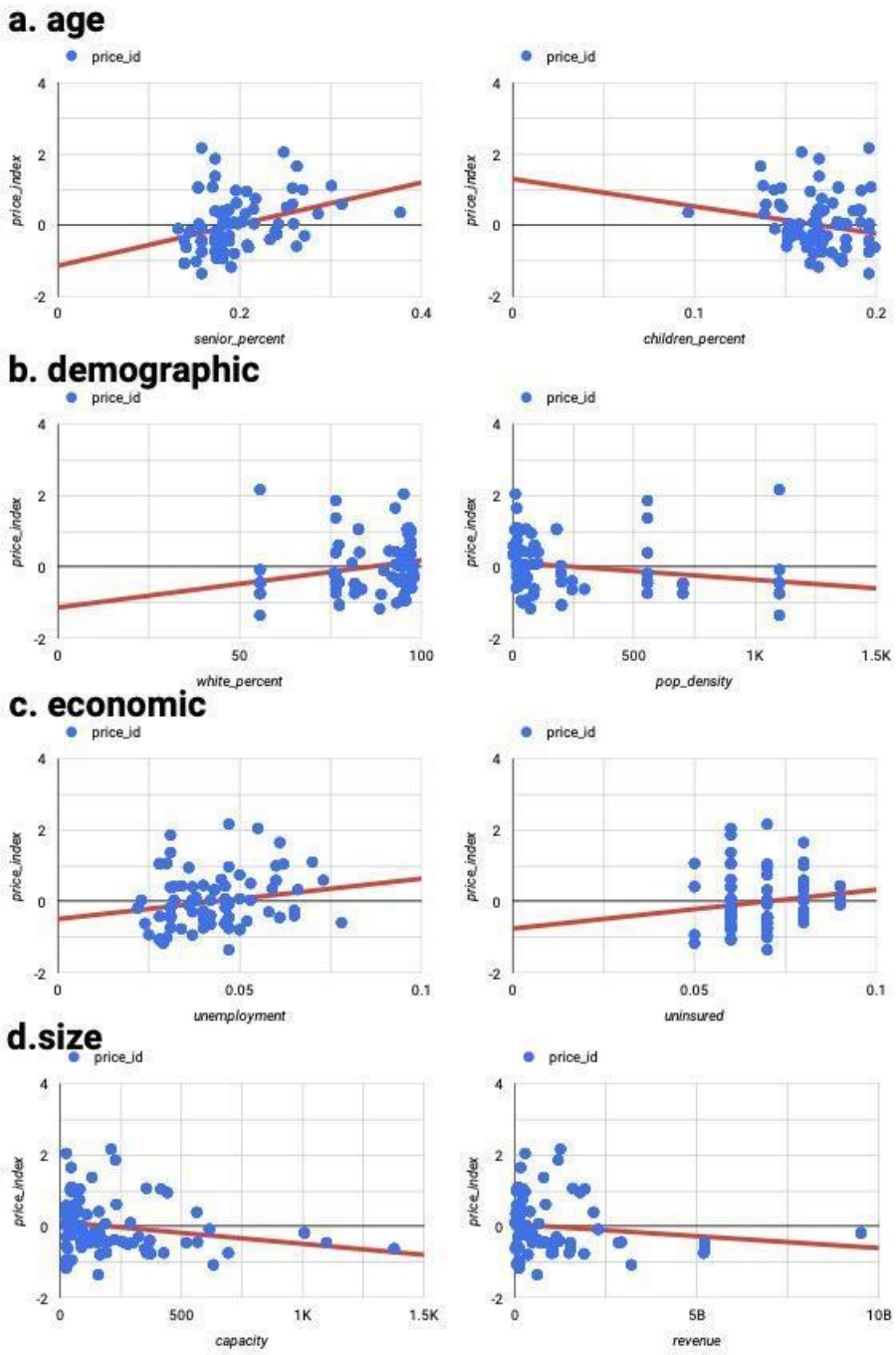

Figure 8: Scatter maps for key correlators

2. Demographic data provides insights into the discrepancies of the distribution of healthcare resources in Michigan

In Figure 7 and 8, we noticed a clear pattern of clustering, which forms distinctive groups of data groups. Although these differences do not seem to heavily determine the prices, the clusters offer insight regarding the distribution of healthcare resources in the state and gather

observable conclusions regarding resources and its effect on hospital's pricing strategies. We have provided geographical details of that clustering below in Figure 9:

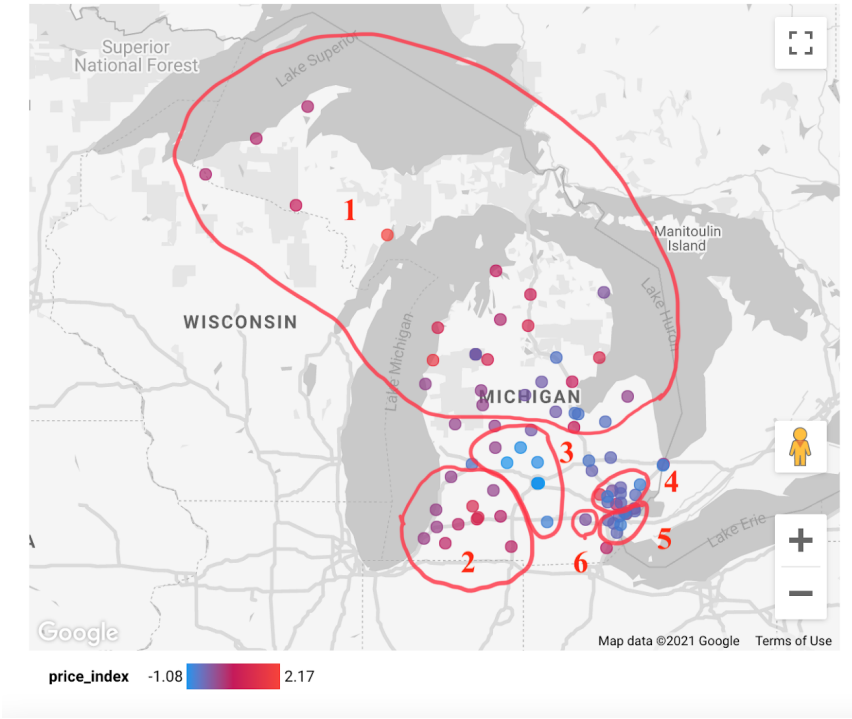

Figure 9: Geographical clustering

1. **Northern Michigan**: **Expensive**
   aging population, rural, very high uninsured/unemployed, small hospitals
   The pricing in this area tends to be expensive with fewer options. It is mostly rural with some urban areas. Northern Michigan has the largest senior percentage across the state and the highest uninsured/unemployed rates. It typically has a lower cost of living; fewer people spread out over a greater area, which can possibly discourage competition among healthcare providers.
2. **West Michigan (Kalamazoo Region)**: **Very Expensive**,
   middle-aged population, suburban, moderate uninsured/unemployed, small hospitals
   This region has the highest prices across the state. Although it is not entirely clear what the determining factors are for its high price, the region may suffer from a lack of large-sized hospitals.
3. **Central Michigan (Lansing Region): Very Cheap**
   young population, suburban, very low uninsured/unemployed, medium-sized hospitals
   This area has some of the most economical options in the state. It is home to the state capital Lansing, Michigan. It has some of the lowest uninsured/unemployed rates in the state probably due to many people working for the public sector.
4. **Oakland County (Detroit)**: **Cheap with Expensive Options**
   middle-aged population, urban, low uninsured/unemployed, large-sized hospitals
   This is one of the richest counties in the state. Although it enjoys the benefit of its urban setting, which means larger hospitals and more options/competitions for cheaper price, it also has incredibly expensive options for ones who can afford it.
5. **Wayne County (Detroit): Cheap**
   young population, very urban, moderate uninsured/unemployed, large-sized hospitals

This is the most densely populated county in the state. It is also the youngest county. As such, the options here are generally low-cost with a few exceptions.

Although the correlations between cash_price and demographic/hospital variables we observed are relatively weak across the state, it could be a result of Simpson's paradox. In each region of the state, certain trends and patterns play out differently, which may end up canceling out each other. For example, although white_percent is correlated with lower pop_density and thus higher cash_price, in Detroit (Wayne and Oakland), white_percent is however also correlated with lower unemployed/uninsured and thus lower cash_price, which was not the case. In other words, although Oakland has higher insurance coverage and low unemployment rate than Wayne, it is also more expensive; this is the opposite to the trend we observed across the state. In conclusion, we need to be careful about the specific contexts in which we apply these trends for interpretations.

Predictive Modeling

Preliminary Selection

After the EDA, we developed a model to predict hospitals' cash_price. We began with simple linear models but received poor results. We then decided to test a few more commonly used regression models. Our common approach for all models was to designate a scorer object with the scoring parameter. The score we specifically focused on included the below metrics:

```
from sklearn import datasets, linear_model
from sklearn import linear_model
from sklearn import neighbors
from sklearn import svm
from sklearn import tree
from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score
from sklearn.model_selection import train_test_split
from sklearn.metrics import explained_variance_score
from sklearn.neural_network import MLPRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.neighbors import KNeighborsRegressor
from sklearn.linear_model import SGDRegressor
```

Our model specifically focused on the below nine regression methods.

```
# regression model list
regr_list ={'Linear Regression': linear_model.LinearRegression(),
            'Bayesian Regression': linear_model.BayesianRidge(),
            'Least Angle Regression': linear_model.Lars(),
            'Stochastic Gradient Descent': SGDRegressor(),
            'K-Nearest Neighbors': neighbors.KNeighborsRegressor(),
            'Support Vector Machine': svm.SVR(),
            'Decision Tree': tree.DecisionTreeRegressor(),
            'Random Forest': RandomForestRegressor(),
            'Neural Network': MLPRegressor()}
```

The results each are below:

Model	Explained Variance Score	Max Error	Median Absolute Error	Mean Absolute Error	R^2 Score	Accuracy (Baseline model)
Linear Regression	0.218	3.467	0.672	0.465	0.1797	54%
Bayesian Regression	0.229	3.524	0.666	0.484	0.1960	55%
Least Angle Regression	0.198	3.509	0.678	0.486	0.1707	Did not test
Stochastic Gradient Descent	0.226	3.499	0663	0.479	0.1903	Did not test
K- Nearest Neighbors	0.3222	3.447	0.598	0.423	0.2804	59.22%
Support Vector Machine	0.318	3.521	0.596	0.403	0.2728	Did not test
Decision Tree	-0.04	4.004	0.646	0.395	-0.058	Did not test
Random Forest	**0.444**	**3.616**	**0.525**	**0.402**	**0.4243**	**66.60%**
Neural Network	0.399	3.046	0.5605	0.463	0.3874	63.80%

We took R^2 Score as the primary metric for consideration and selected 5 models to move to the accuracy rate test.

Model Selection (Accuracy Rate)

The accurate rate is modified from the metric of the mean absolute percentage error (MAPE). However, because the z score of cash_price can be negative after the standardization, we replace the mean with the distance between the mean and the minimum. And the formula is as follows:

1 - the mean absolute error / (the mean cash_price - the min cash_price)

We ultimately chose Random Forest as the best predictive model due to a combination of a high R^2 score and accuracy rate.

To mitigate the challenge of limited data, we also used K Fold Cross Validation (rather than the standard train/test method) when testing the accuracy rate.

Model Optimization (Hyper parameter Tuning)

Concerns around the predictive model are that If the dataset contains a large number of predictors that are uncorrelated to the outcome, the random forests algorithm will be forced to choose amongst only noise variables at many of its splits. This will lead to poor performance, so it is essential to test the model parameters to better optimize it as well.

Our efforts were then focused on increasing the predictability of the Random Forest model. By applying both Random Search and Grid Search, we were able to find which parameters were most conducive to our model and were able to see a 39% increase to our baseline prediction model. See details below:

```
{'bootstrap': False,
 'max_depth': 100,
 'max_features': 2,
 'min_samples_leaf': 1,
 'min_samples_split': 3,
 'n_estimators': 800}
Model Performance
Average Error: 0.5747 degrees.
Accuracy = 66.60%.

Model Performance
Average Error: 0.1287 degrees.
Accuracy = 92.52%.

Improvement of 38.93%.
```

The final model we developed can reduce the average error (the difference between the predicted and the test cash price) to 0.1 unit (standard deviation). And the accurate rate is above 90%, whereas other models are about 60% (as noted in the table above). Based on the above strategies, our random forest model is able to predict the missing prices relatively well.

Note: The accuracy rate is measured with the z score of cash_priace. This means that when they are converted back to the original cash_price, even 0.1 unit error (standard deviation) in prediction can be very significant in real cash price (especially when our raw data has a relatively large variance to begin with). See Figure 10 below for the comparison between original and predicted cash_price for all procedures.

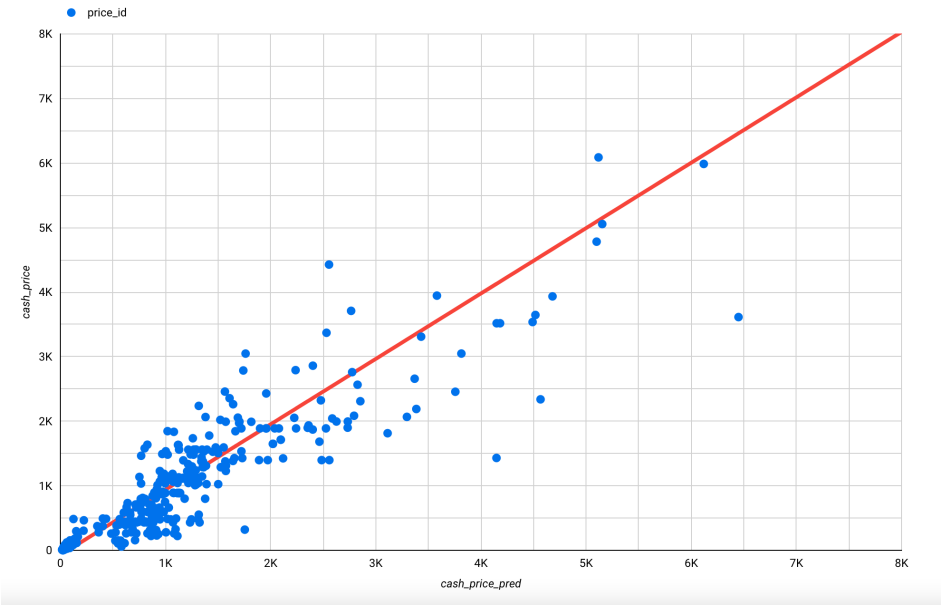

Figure 10: Predictability results

Patient Dashboard

Based on our EDA, we designed our dashboard based on the key variables that we believe patients would find most useful to make decisions. A link to our final dashboard is here.

DS4A GROUP54 | Tableau Public

Conclusions and Discussions

The price transparency movement debuted in January 2021 is advertised as a movement to" help Americans know the cost of a hospital item or service before receiving it." Based on our analysis of Michigan, the extent to which the hospitals adhere to the letter and spirit of the federal rule varied widely across the hospitals. Based on our analysis, we can confirm that hospital prices had a wide variation. No two hospitals in the same region charged the same for the same procedure. Neither demographic data nor hospital specific data showed a strong correlation to why these prices have this large variation.

We estimate that over half of our team efforts were spent on understanding and extracting data from the machine-readable files. Each machine-readable file varied in what they provided, and it is vital for us and the average patient to be able to distinguish between terms such as "cost" "charge" "payer specific price" etc. Oftentimes, we had multiple price points stated for the same procedure and realized that they were no standard as to how much detail in pricing each hospital provided. There is also no clear relationship as to what is quoted to the patient. "Charge" and "Price" are often terms that were used interchangeably. In addition, one of the main challenges in deciding costs is to know if the patient is insured or not insured. For patients who are insured their price would be dependent on the hospital negotiated rate (a rate negotiated by the hospital) which varies drastically as well. For patients who are not insured, they typically (most often) can call in and negotiate a cash price. Adding to this complexity is the fact that despite stating that these costs are typically going to be an estimate of what patients incurring the service will pay, it does not include costs for personnel performing the service (provider, nurse practitioners etc.) or any equipment used during the procedure/care. etc.

Future analysis can work to understand if health systems are facing any pressure to lower prices to compete for consumers shopping for health services, and if insurers face the same pressure to negotiate discounts (American Medical Association, 2015). Our analysis highlighted the wide variation, and our model can predict pricing where it is unavailable. However, as we did not find strong correlations with the variables that we hypothesized would predict costs, future work can focus on identifying additional variables to test for effect on pricing, as well as considering disaggregating the pricing data by county.

Future policy efforts can focus on:

a) Enforcing and standardizing pricing data availability
b) Improving the ability of patients to use this information by centering health literacy
c) Incentivizing affordable care by focusing on modifiable factors associated with higher prices

References

AAPC. HomeCodesCPTCPT Codes. *Codify by AAPC.* [Online] https://www.aapc.com/codes/cpt-codes.

American Medical Association. 2015. The Challenge of Understanding Health Care Costs and Charges. *Journal of Ethics.* [Online] November 2015. https://journalofethics.ama-assn.org/article/challenge-understanding-health-care-costs-and-charges/2015-11.

Centers for Medicare and Medicaid Services. Hospital Price Transparency. *CMS.gov.* [Online] https://www.cms.gov/hospital-price-transparency.

Michigan Department of Health and Human Services. 1990 - 2019. Michigan Population, 1990-2019. *Michigan Department of Health and Human Services.* [Online] 1990 - 2019. https://www.michigan.gov/mdhhs/0,5885,7-339-73970_2944_5325---,00.html.

Robert Wood Johnson Foundation. 2016. How price transparency controls healthcare costs? *Robert Wood Johnson Foundation.* [Online] March 1, 2016. https://www.rwjf.org/en/library/research/2016/03/how-price-transparency-controls-health-care-cost.html.

United States Census. 2020. SAIPE State and County Estimates for 2019. *United Census Bureau.* [Online] 2020. https://www.census.gov/data/datasets/2019/demo/saipe/2019-state-and-county.html.

Data Sources:

American Community Survey, 5-year estimates (2015-2019)

high_school https://data.census.gov/cedsci/table?q=s1501&tid=ACSST1Y2019.S1501

Small Area Health Insurance Estimates (SAHIE) using the American Community Survey (2018) uninsured

https://www.census.gov/data/datasets/time-series/demo/sahie/estimates-acs.html

Bureau of Labor Statistics (2020) unemployment

https://www.bls.gov/lau/tables.htm

National Center for Health Statistics - Mortality Files (2017-2019)

life_exp*

Area Health Resource File/American Medical Association (2018)

pcp_ratio*

Hospital Statistics

https://www.ahd.com

Hospital Pricing data

Individual hospital websites