THE DATA OPEN REPORT

TEAM 9 REPORT

Abhimanyu Nag
Cecilia Luo
Saud Iqbal
Chase Robertson

TABLE OF CONTENTS:

Non-Technical Executive Summary

Key topic/Question:

As firm believers in the phrase: *"We are what we invest in"*, the first area of investigation that crossed our minds is how to understand people (such as physicians as per the data) through their financial earnings and decisions. This naturally led to a study into the primary care types of Physicians in relation to their overall finances.

Q: Is there a difference in payment across physicians and what are the factors that account for the discrepancy in primary care type of physicians? Can we differentiate between the different types of physicians based on the payments that they earn?
More specifically this can be answered through the following questions:

- Are any physicians outliers in their specific industries?
- Are they outliers in terms of value and volume?
- To what extent do geographical differences play a role in this?
- What physicians with high ownership and investment interest are outliers?

Q: The second question that arises on a more fundamental level as a result of the first, is **can we predict the primary care type of a physician based on their investment and ownership decisions, research payments and general payments as factors?** Our intuition leads us to explore this question on the basis that there *really is a way* to predict the above.

By addressing these questions, it can be concluded how these outliers affect the primary goal of the CMS and where CMS should be allocating/reallocating its distribution of wealth to provide fair and in turn better service. This would also help CMS to look into physician payments by providing a predictable outcome to a certain amount of investment.

Key Findings:

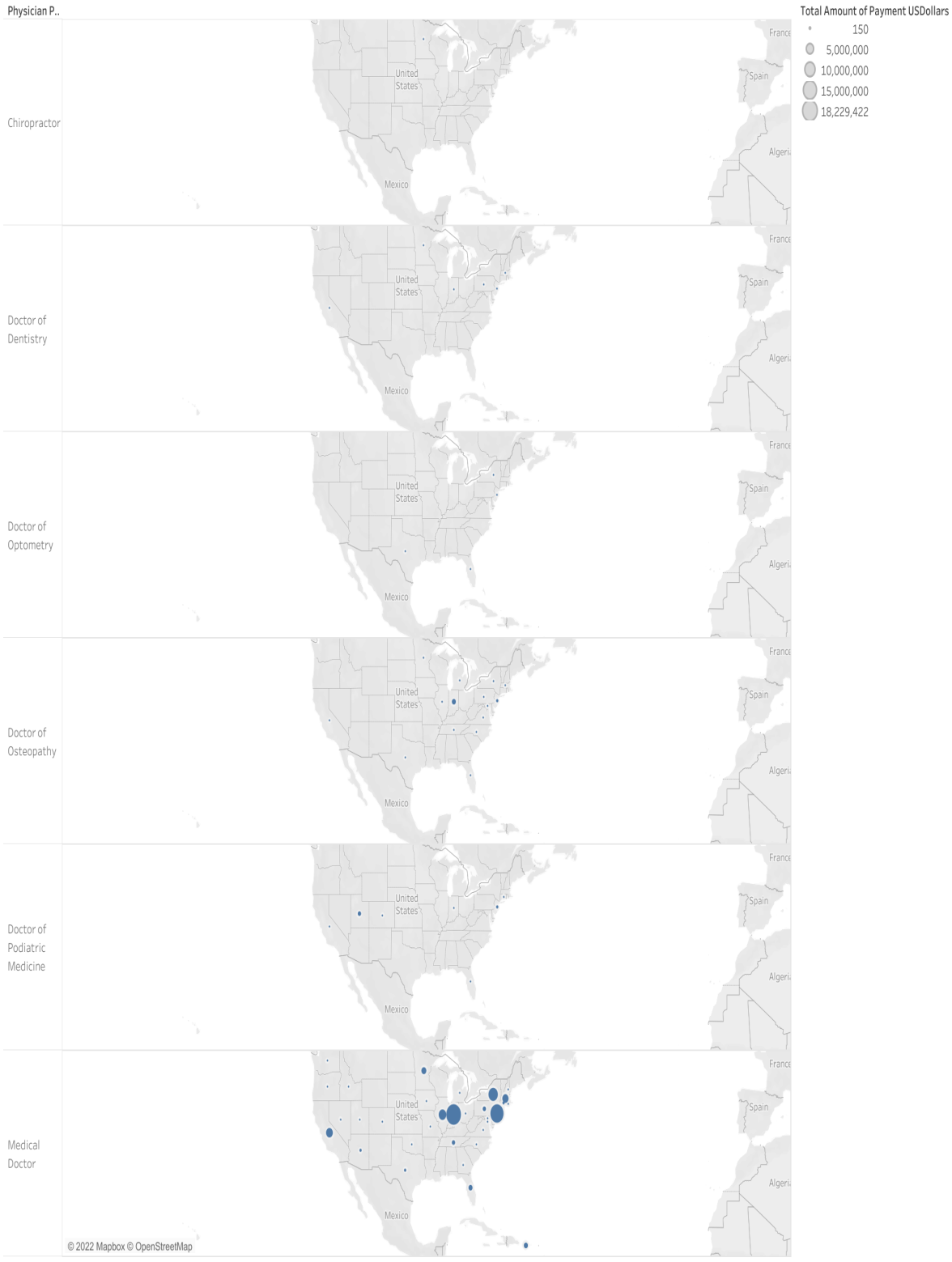
The data sources from the open payments database of CMS analyzed are of great significance in depicting the general payments, research payments and ownership/investments related to both the Medicare and Medicaid programs. Medicare and Medicaid provide medical care and payment assistance for 200 million Americans nationwide, many of which would not be able to access the care they require without it.

The findings of potential outliers for payments provide valuable insights that can be used by the government and health care workers to improve the quality of health services delivered to improve the overall health of the nation.

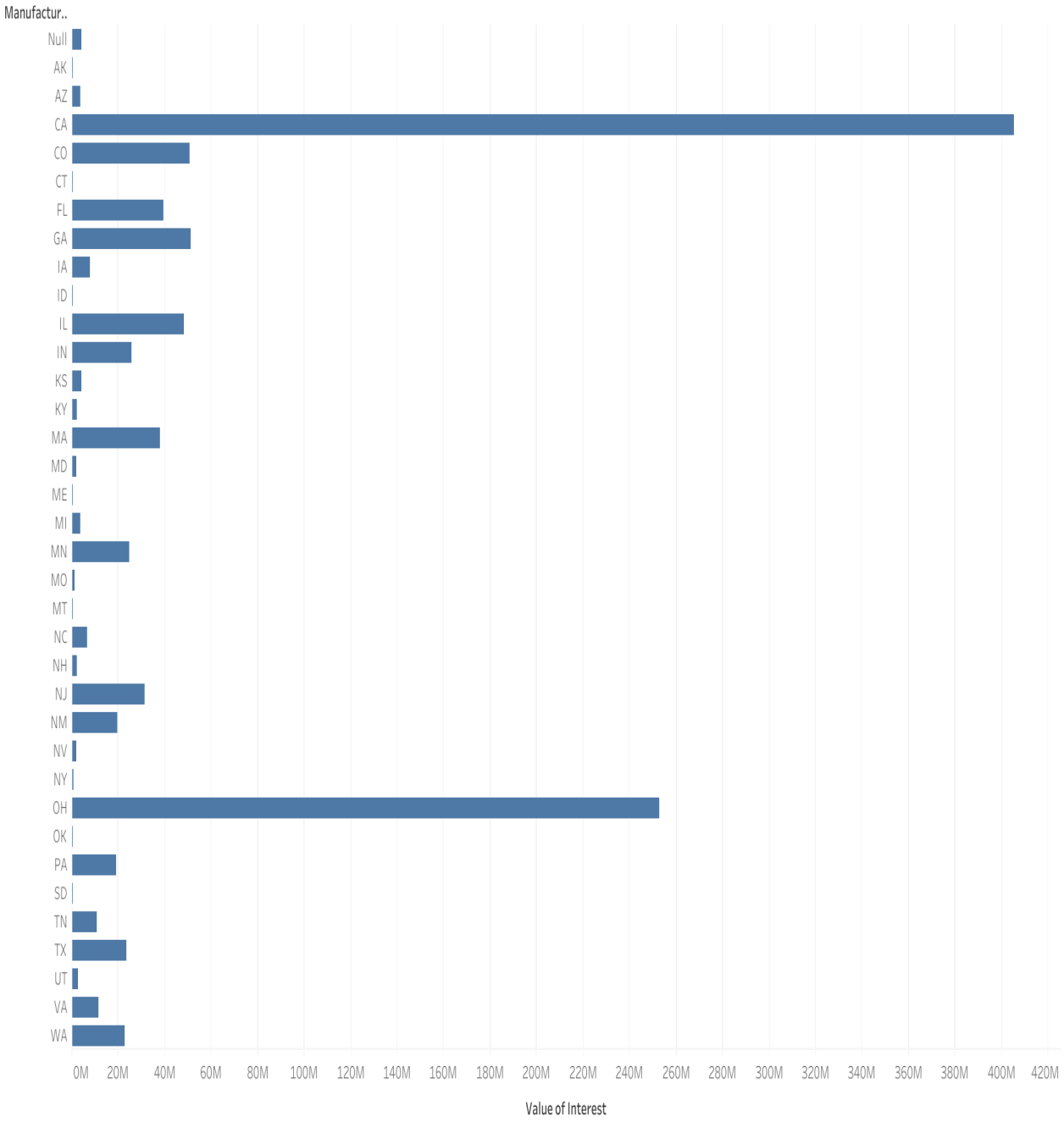
A thorough analysis of the data provided the following insights :

There are physicians who are outliers in their specific industry. Namely, by comparing physician counts, it was observed that the most number of physicians have primary care type as Medical Doctors, most of which come from California. As most physicians are from California, their ownership is predominantly tied up in the Lagua Beach company. The company Eli invested most in research in physicians such as medical doctors and doctors of osteopathy.

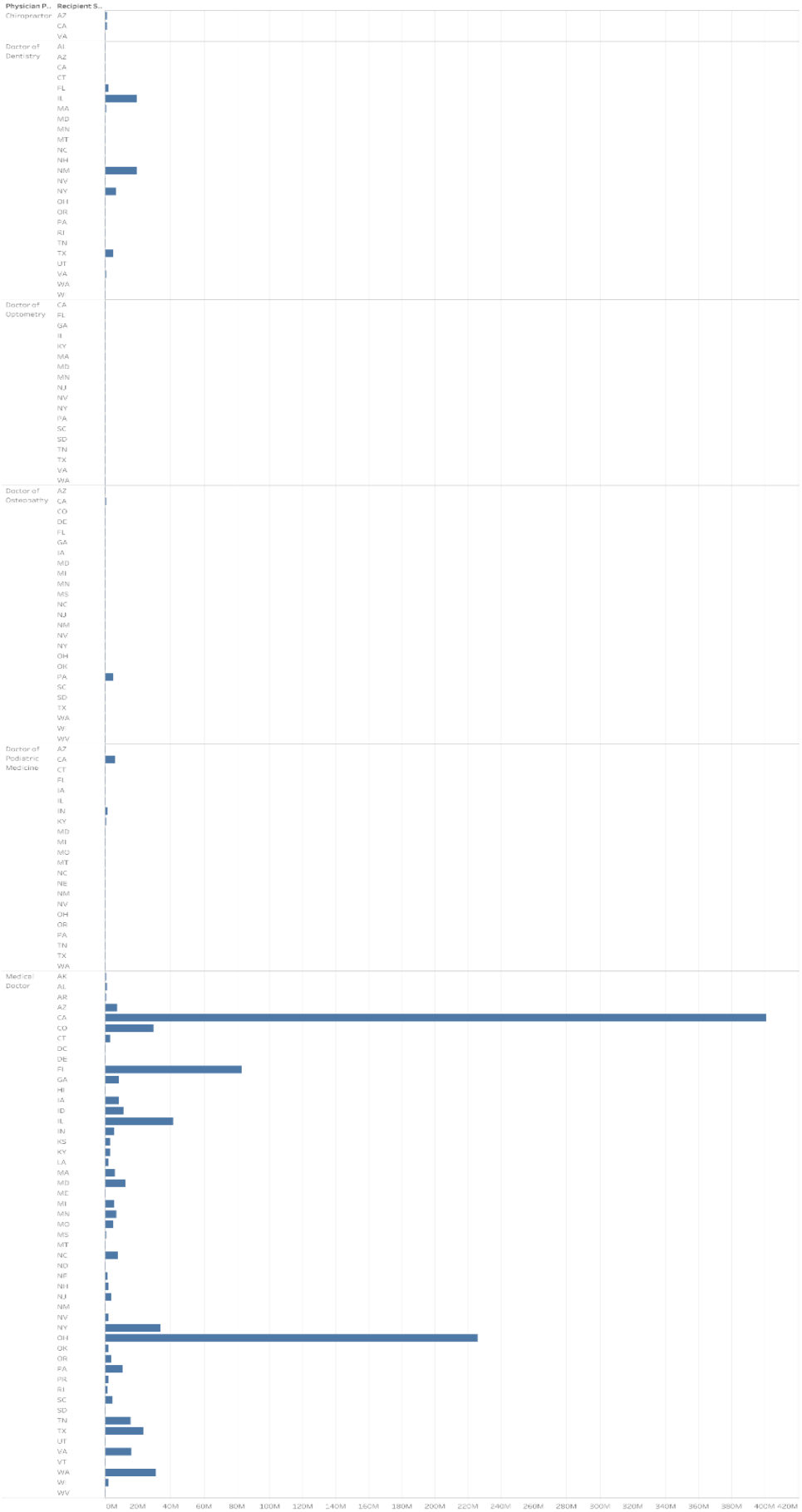
Sheet 3


Physician P..
Chiropractor
Doctor of Dentistry
Doctor of Optometry
Doctor of Osteopathy
Doctor of Podiatric Medicine
Medical Doctor
United States
Mexico
France
Spain
Algeria
© 2022 Mapbox © OpenStreetMap
Total Amount of Payment USDollars
150
5,000,000
10,000,000
15,000,000
18,229,422

Sheet 4


Manufactur..
Null
AK
AZ
CA
CO
CT
FL
GA
IA
ID
IL
IN
KS
KY
MA
MD
ME
MI
MN
MO
MT
NC
NH
NJ
NM
NV
NY
OH
OK
PA
SD
TN
TX
UT
VA
WA
0M
20M
40M
60M
80M
100M
120M
140M
160M
180M
200M
220M
240M
260M
280M
300M
320M
340M
360M
380M
400M
420M
Value of Interest

Sheet 3


Physician P..
Recipient S..
Chiropractor
Doctor of Dentistry
Doctor of Optometry
Doctor of Osteopathy
Doctor of Podiatric Medicine
Medical Doctor
0M 20M 40M 60M 80M 100M 120M 140M 160M 180M 200M 220M 240M 260M 280M 300M 320M 340M 360M 380M 400M 420M
Value of Interest

Furthermore, it was found that high earning physicians consist of physicians offering services that are high in value or have a high total number of payments. Low earning physicians are outliers both in value and volume. This trend is observable in a variety of cities, with some physicians being outliers because they have increased ownership and investment interest.

Additionally, it was observed that outliers for value are predominantly from royalty or license, and current or prospective ownership or investment interest. Whereas, the data for volume are distributed more evenly across the different natures of payments. This trend can also be noted when comparing by the recipient city where Los Angelos and New York were clear outliers in terms of value, but volume was more evenly dispersed.

Technical Executive Summary

Cleaning Process:

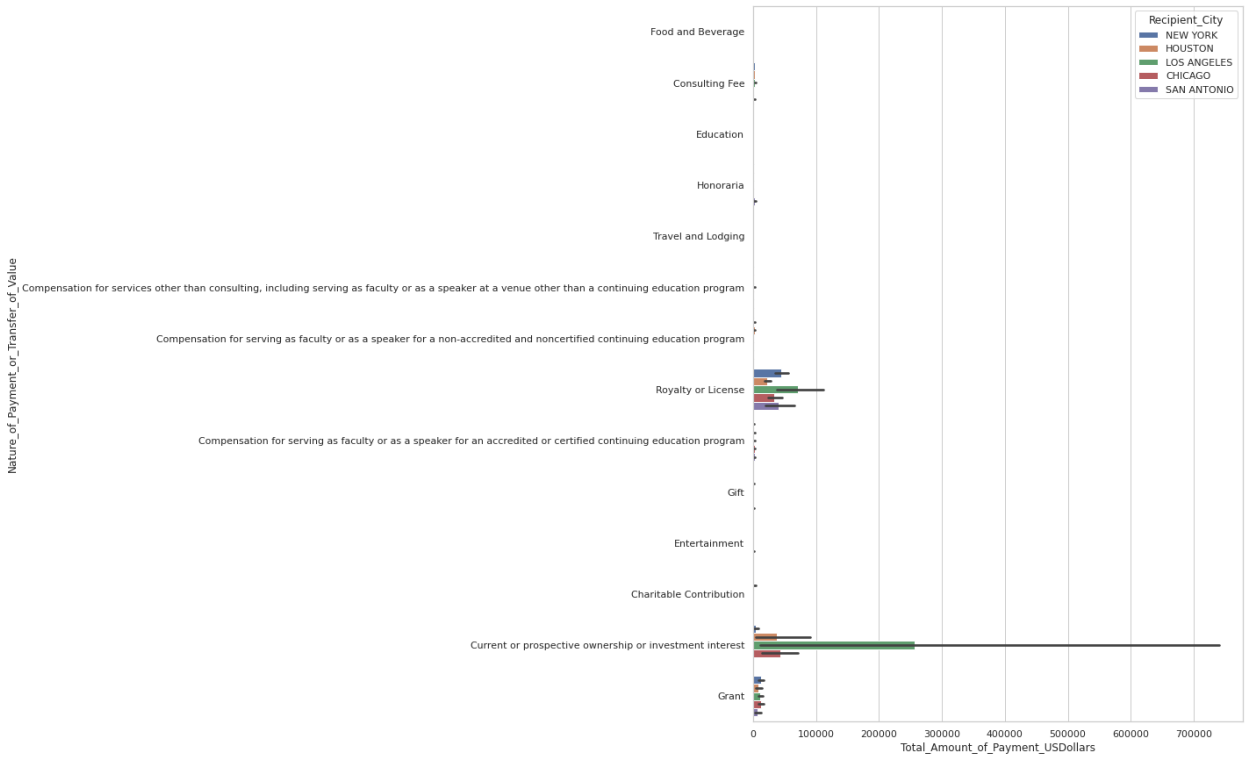
The dataset was cleaned using python and the pandas module. We sorted the information into relevant dictionaries containing only the columns that were necessary to address the question. Having a cleaned dataset with only the relevant data made analysis easier and helped to prevent mistakes. The dataset was transformed by integrating information from the dataset based on the common variables, for example, physician id, this made a comparison between the data more efficient. It also allowed data to be compared across datasets so insights can be drawn

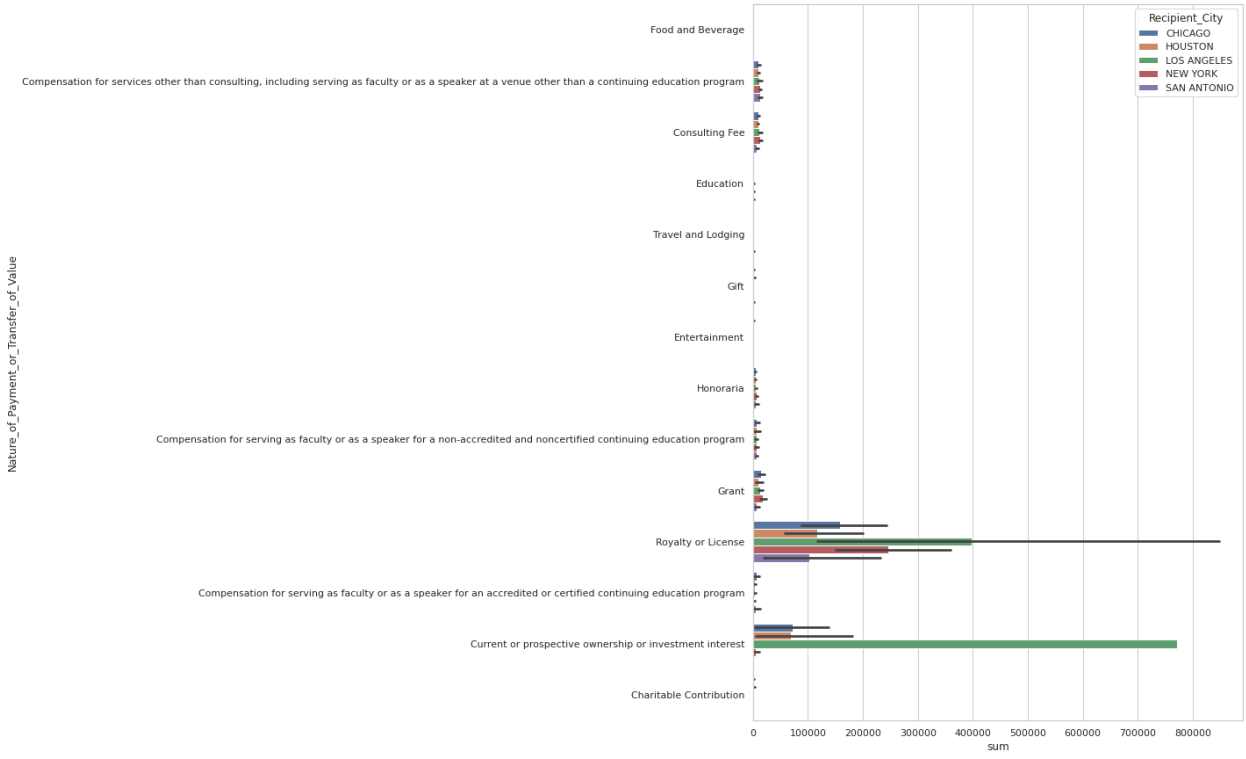
By grouping the recipients based on their state and city and creating a count for each combination, it allowed an analysis of how geographical location affected payments and the type of physician. Quality control was ensured by checking for missing values, and ensuring all data provided was correct, for example, the total payment must be a positive quantitative value etc.

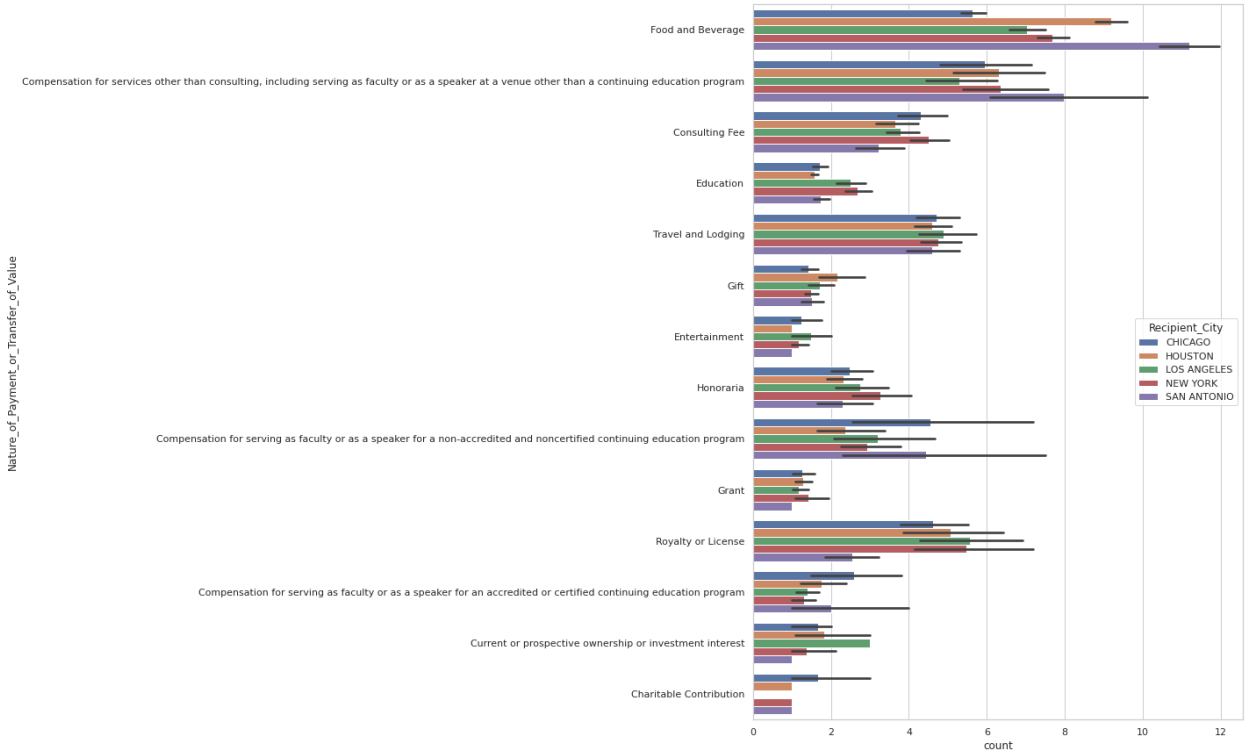
Exploratory Data Analysis and Modeling:

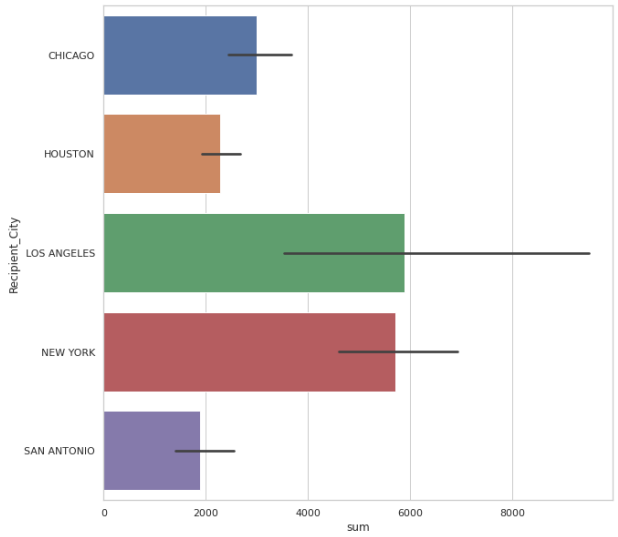
Aggregating the values of the general payments for each physician in every state using their unique physician id allowed the total amount earned to be calculated. From this data, the outlier physicians, specifically picking the top and bottom ten physicians with respect to the total amount earned, were analyzed by comparing the industry and the calculated median, maximum, mean and count. Barplots were generated from the data, to draw conclusions regarding the total amount of payment and which nature of payment had the highest sum and count, for each recipient city. Additionally, bar plots were also produced to highlight the different values and columns for each recipient city to determine the extent to which geographical location affected payments. Bar plots were utilized to show the distribution of data, to allow efficient comparison across different subgroups. Some of the produced bar plots are included below for reference.

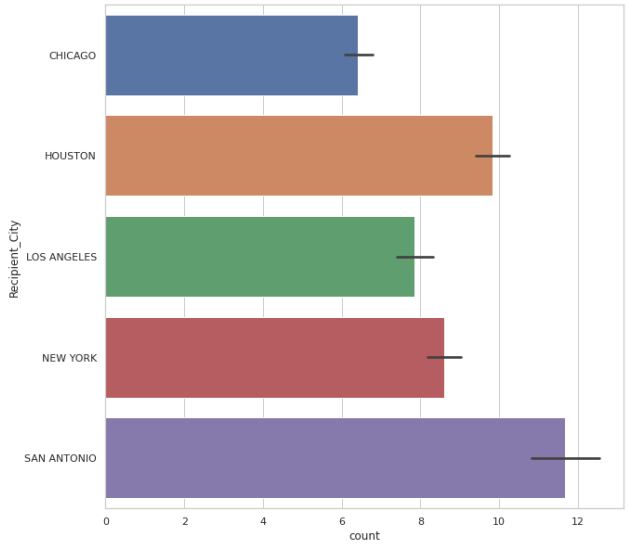
Most of the work has been done in code : Exploratory Data Analysis


Recipient_City
NEW YORK
HOUSTON
LOS ANGELES
CHICAGO
SAN ANTONIO
Food and Beverage
Consulting Fee
Education
Honoraria
Travel and Lodging
Compensation for services other than consulting, including serving as faculty or as a speaker at a venue other than a continuing education program
Compensation for serving as faculty or as a speaker for a non-accredited and noncertified continuing education program
Royalty or License
Compensation for serving as faculty or as a speaker for an accredited or certified continuing education program
Gift
Entertainment
Charitable Contribution
Current or prospective ownership or investment interest
Grant
Nature_of_Payment_or_Transfer_of_Value
0
100000
200000
300000
400000
500000
600000
700000
Total_Amount_of_Payment_USDollars


Recipient_City
CHICAGO
HOUSTON
LOS ANGELES
NEW YORK
SAN ANTONIO
Food and Beverage
Compensation for services other than consulting, including serving as faculty or as a speaker at a venue other than a continuing education program
Consulting Fee
Education
Travel and Lodging
Gift
Entertainment
Honoraria
Compensation for serving as faculty or as a speaker for a non-accredited and noncertified continuing education program
Grant
Royalty or License
Compensation for serving as faculty or as a speaker for an accredited or certified continuing education program
Current or prospective ownership or investment interest
Charitable Contribution
Nature_of_Payment_or_Transfer_of_Value
0
100000
200000
300000
400000
500000
600000
700000
800000
sum


Nature_of_Payment_or_Transfer_of_Value
Food and Beverage
Compensation for services other than consulting, including serving as faculty or as a speaker at a venue other than a continuing education program
Consulting Fee
Education
Travel and Lodging
Gift
Entertainment
Honoraria
Compensation for serving as faculty or as a speaker for a non-accredited and noncertified continuing education program
Grant
Royalty or License
Compensation for serving as faculty or as a speaker for an accredited or certified continuing education program
Current or prospective ownership or investment interest
Charitable Contribution
Recipient_City
CHICAGO
HOUSTON
LOS ANGELES
NEW YORK
SAN ANTONIO
0
2
4
6
8
10
12
count


Recipient_City
CHICAGO
HOUSTON
LOS ANGELES
NEW YORK
SAN ANTONIO
0
2000
4000
6000
8000
sum


Recipient_City
CHICAGO
HOUSTON
LOS ANGELES
NEW YORK
SAN ANTONIO
0
2
4
6
8
10
12
count

Prediction of Primary Care Type of Physicians using Payments in terms of value of interest and research as factors

Background and Intuition

Statistically speaking, the research and ownership datasets had more to offer than just an exploratory analysis on the primary care type and payments to the recipients. The analysis of relationships done in the previous section led us to realize that there might be patterns to mine that the naked eye is not enough to see. The non-existence of quantitative values did not help either. So with the goal to mine patterns from the data about primary care type of physicians and lost in the frequentist terms, we turned to Bayesian Inference and hoped to be able to define a prior and post distribution on categorical variables. We had fitted a multinomial model with a dirichlet prior and hoped for the best. Interestingly enough, we ended up deriving a better model which is typically known as the **multinomial logistic regression.**

Model

The model, as any google search would tell us, is a generalized linear model which is used to predict the probabilities of various outcomes of a dependent categorical random variable using independent variables which may have any values, quantitative or qualitative.
We provide a generalized outline of the mathematical working of the model

Outline

We used a predictor function:

$$f(m, n) = \beta_m . x_n$$ where β_m is the vector set of regression coefficients associated with outcome *m* and x_n is the vector set of explanatory variables associated with observation *n*.

We turn to **binary regressions** to be able to explain each term in the polynomial vector.
Since there are 6 outcomes (6 categories of primary care type for physicians), we take m = 6 and thus:

$$ln\frac{P(X_n=1)}{P(X_n=6)} = \beta_1 x_n$$

$$ln\frac{P(X_n=2)}{P(X_n=6)} = \beta_2 x_n$$

.

.

.

$$ln\frac{P(X_n=5)}{P(X_n=6)} = \beta_5 x_n$$

If we exponentiate both sides and solve for each probability we get

$$P(X_n = 1) = P(X_n = 6)e^{(\beta_1 x_n)}$$

$$P(X_n = 2) = P(X_n = 6)e^{(\beta_2 x_n)}$$

.

.

.

$$P(X_n = 5) = P(X_n = 6)e^{(\beta_5 x_n)}$$

Since we know that $\sum_{i=1}^{6} P(X_n = i) = 1$

We can say that

$$P(X_n = 6) = 1 - \sum_{i=1}^{5} P(X_n = i)$$

$$= 1 - \sum_{i=1}^{5} P(X_n = i)e^{(\beta_i x_n)}$$

$$P(X_n = 6) = \frac{1}{1 + \sum_{i=1}^{5} e^{(\beta_i x_n)}}$$

Thus using the given we get

$$P(X_n = 1) = \frac{e^{(\beta_1 x_n)}}{1 + \sum_{i=1}^{5} e^{(\beta_i x_n)}}$$

$$P(X_n = 2) = \frac{e^{(\beta_2 x_n)}}{1 + \sum_{i=1}^{5} e^{(\beta_i x_n)}}$$

.

.

.

$$P(X_n = 5) = \frac{e^{(\beta_5 x_n)}}{1 + \sum_{i=1}^{5} e^{(\beta_i x_n)}}$$

And this is the model that we will be using

Cleaning Process

We clean the datasets Ownership_payments and Research_Payments (Since these are the datasets which have the exact values we want) to include only those values pertaining to what we want i.e Primary Care Type, Manufacturer_Or_GPO_name, Recipient_State, Value of Interest, total amount of money paid, GPO from research dataset with Primary care type being our target(dependent) variable.

Assumptions that need to be checked to be able to use Multinomial Logistic Regression

1. The dependent variable should be a nominal variable - check (Primary care type of Physician)
2. One or more independent variables that are continuous, ordinal or nominal - check (Recipient_State, Value of interest etc.)
3. Independence of observations and the dependent variable should have mutually exclusive and exhaustive categories - check (6 mutually exclusive categories and independent observations (both datasets have separate tests))
4. No multicollinearity - check
5. There needs to be a linear relationship between any continuous independent variables and the logit transformation of the dependent variable - check (checked via SPSS)
6. No outliers, high leverage values or highly influential points - check (Outliers were removed prior to testing)

Model Working

Using the inbuilt SPSS tool of multinomial logistic regression

- The Goodness of Fit Test worked for both datasets (with specified parameters) with a P-value of 0.321 and 0.577 for both which is definitely > 0.05. Hence at the 5% significance level the data does not provide sufficient evidence to reject the null hypothesis and hence we conclude that the model is a good fit for predicting a physician's primary care type.
- The P-value of the final model is 0.004 and 0.000 for both datasets which are < 0.05. Hence at the 5% significance level, the data provides sufficient evidence to reject the

null hypothesis and hence we conclude that the final model is a better fit for predicting a physician's primary care type than the baseline model of just the intercept only.

Confusion Matrix for Classification

For Predicting Primary Care Type of Physicians using their investments/ownerships,

Classification

		Predicted		
Observed	Chiropractor	Doctor of Dentistry	Doctor of Optometry	Doctor of Osteopathy
Chiropractor	3	0	0	0
Doctor of Dentistry	3	59	4	0
Doctor of Optometry	0	0	49	0
Doctor of Osteopathy	6	1	0	18
Doctor of Podiatric Medicine	1	0	0	0
Medical Doctor	31	7	23	1
Overall Percentage	1.4%	2.1%	2.3%	0.6%

Classification

	Predicted		
Observed	Doctor of Podiatric Medicine	Medical Doctor	Percent Correct

Chiropractor	0	0	100.0%
Doctor of Dentistry	0	19	69.4%
Doctor of Optometry	0	8	86.0%
Doctor of Osteopathy	2	66	19.4%
Doctor of Podiatric Medicine	47	2	94.0%
Medical Doctor	26	2862	97.0%
Overall Percentage	2.3%	91.3%	93.8%

The given confusion matrix denotes a success of 93.8% between observed and predicted value which is pretty good and hence a fine indicator for predicting primary care type of physicians using their investments/ownerships.

For Predicting Primary Care Type of Physicians using their research payments,

Classification

	Predicted			
Observed	Chiropractor	Doctor of Dentistry	Doctor of Optometry	Doctor of Osteopathy
Chiropractor	1	0	0	0
Doctor of Dentistry	41	6	3	0
Doctor of Optometry	0	0	0	0
Doctor of Osteopathy	3	0	3	1
Doctor of Podiatric Medicine	0	0	0	0
Medical Doctor	376	0	17	0

Overall Percentage	1.7%	0.0%	0.1%	0.0%

Classification

	Predicted		
Observed	Doctor of Podiatric Medicine	Medical Doctor	Percent Correct
Chiropractor	0	0	100.0%
Doctor of Dentistry	0	1207	0.5%
Doctor of Optometry	0	381	0.0%
Doctor of Osteopathy	0	605	0.2%
Doctor of Podiatric Medicine	0	216	0.0%
Medical Doctor	2	21942	98.2%
Overall Percentage	0.0%	98.2%	88.5%

The given confusion matrix denotes a success of 88.5% between observed and predicted value which is pretty good (but not as good as the previous one) and hence a fine indicator for predicting primary care type of physicians using their research payments.

Concluding Remarks

As per the above results, we have been able to accurately predict the Primary care type of Physicians through their Ownerships/Investments and Research Payments and we also claim that this model can be used to chart patterns even in other datasets which could be used to gather more insights into CMS and their workings in general. All in all, we leave our readers with fresh insights into the quote: “We are what we invest in”.