2021

Identifying eNewsletter Audience Characteristics Driving Advertising Revenue

SPINGBOARD CAPSTONE II

STUDENT LIUDMILA KHALITOVA, SUPERVISOR VARUN BHATIA

FINAL REPORT

Bulletin Healthcare produces eNewsLetters for various medical professionals which are distributed to specialized medical associations. The NewsLetters are free for recipients, and their production costs are covered by money from advertisers who would like to target and reach physicians. Bulletin Healthcare is looking to expand its network of healthcare professionals and increase the reach of their briefings but is unsure what kind of physicians they should focus on first.

Advertisers can target BH recipients based on the following criteria or any combination of them:

- Professional characteristics
    - Area of specialization (e.g. OBGYNs, oncologists, podiatrists, etc)
    - Medical Associations in which a recipient is a member (some of them are based on the areas of specialization such as ACC - American College of Cardiology, others are based on location such as Texas Medical Association, etc)
- Professional lifecycle (eg. Student, Nurse Practitioner, Resident, physician, etc)
- Location: State, city, zip code
- Hospital characteristics:
    - Location
    - Size
- Prescription behavior
    - Specific drugs prescribed in the past
    - Prescription frequency/Prescription volume

One option for Bulletin Healthcare is to focus their efforts in trying to expand its network among doctors with rare specialties because advertisers tend to pay higher to reach recipients with rare specialties such as hematologic oncologists. On the other hand, expanding reach among more general specialties such as family medicine, internal medicine, etc. might also make sense because there may be fewer advertisers trying to reach highly specialized audiences than those advertising their products to general customers, and because there are more doctors in general practice than there are in specialized areas.

By using the data on recipients targeted by healthcare advertisers via BH briefings over the course of 2020, I created a tool that helps identify the most relevant recipient characteristics associated with advertising revenue. The model will help Bulletin Healthcare leadership to make better informed decisions regarding which medical organizations they should partner with for briefing distribution.

Data Wrangling

1. The data sets.

The dataset for analysis was comprised of two datasets.

a) The main dataset was pulled from BH database and contained 31 columns and 200,000 rows.

Each row contained data about an HCP opening eNewsletter at a certain point in time.

The columns included recipient's NPI (a unique number identifying healthcare professional in the U.S., their gender, city, state, zipcode, whether they were the sole proprietor or worked for someone else, their primary therapeutic area, the list of subspecialties within the main therapeutic area, how many procedures they've performed in the last year, the datetime when they opened the briefing (when an impression has occurred), what briefing/eNewsletter it was, and how much advertisers paid for that particular impression. In addition, there was information about a hospital/practice where the doctor worked. That includes city, state, zipcode, hospital size, and whether it was a stand-alone hospital or a part of a larger hospital system.

npi	first_name	last_name	gender	address	city	state	zipcode	sole_proprietor	medicare_ind	m	specialty	derived_specialty_list	event_ts	adiolineitemi
1.71E+09	Catherine	Cheney	N/A	7 Arthur Av	Marblehe	MA	1945	X	N	Y	Internal Medicine - Gastroentero	GE,IM	2/13/2021 17:54	6
1.77E+09	Suzelle	Luc	N/A	N/A	MEDFORD	MA	2155	N	N	N	Internal Medicine	RHU,IM,MPD	2/7/2021 12:17	6
1.29E+09	George	Waters	N/A	N/A	Attleboro	MA	2703	N	N	N	Internal Medicine - Cardiovascul	CD,IM	2/25/2021 5:47	5
1.29E+09	George	Waters	N/A	N/A	Attleboro	MA	2703	N	N	N	Internal Medicine - Cardiovascul	CD,IM	2/25/2021 5:47	6
1.29E+09	George	Waters	N/A	N/A	Attleboro	MA	2703	N	N	N	Internal Medicine - Cardiovascul	CD,IM	2/25/2021 5:47	6
1.81E+09	Olivera	Boskovska	N/A	N/A	BROCKTO	MA	2301	N	N	N	Family Medicine	FM,FP,MDM,OBG	3/7/2021 23:19	5
1.81E+09	Olivera	Boskovska	N/A	N/A	BROCKTO	MA	2301	N	N	N	Obstetrics and Gynecology	FM,FP,MDM,OBG	3/7/2021 23:19	5
1.81E+09	Olivera	Boskovska	N/A	N/A	BROCKTO	MA	2301	N	N	N	Family Medicine	FM,FP,MDM,OBG	3/12/2021 5:17	5
1.81E+09	Olivera	Boskovska	N/A	N/A	BROCKTO	MA	2301	N	N	N	Obstetrics and Gynecology	FM,FP,MDM,OBG	3/12/2021 5:17	5
1.81E+09	Olivera	Boskovska	N/A	N/A	BROCKTO	MA	2301	N	N	N	Family Medicine	FM,FP,MDM,OBG	3/6/2021 22:14	5

b) The second dataset contained information about prescriptions written by each individual doctor in 2018. The dataset was released by CMS.GOV and included information about Part D Medicare prescribers. The columns we were interested in included NPI, how many drugs a doctor prescribed, and what was the overall cost of these prescriptions.

2. Data preparation

2.1. Merging and cleaning. The first step was to merge the two data set using NPI number, and calculate the total number of impressions registered for each unique NPI, and the total sum of money paid by advertisers for all of these impressions. The columns that had more than half values missing were dropped.

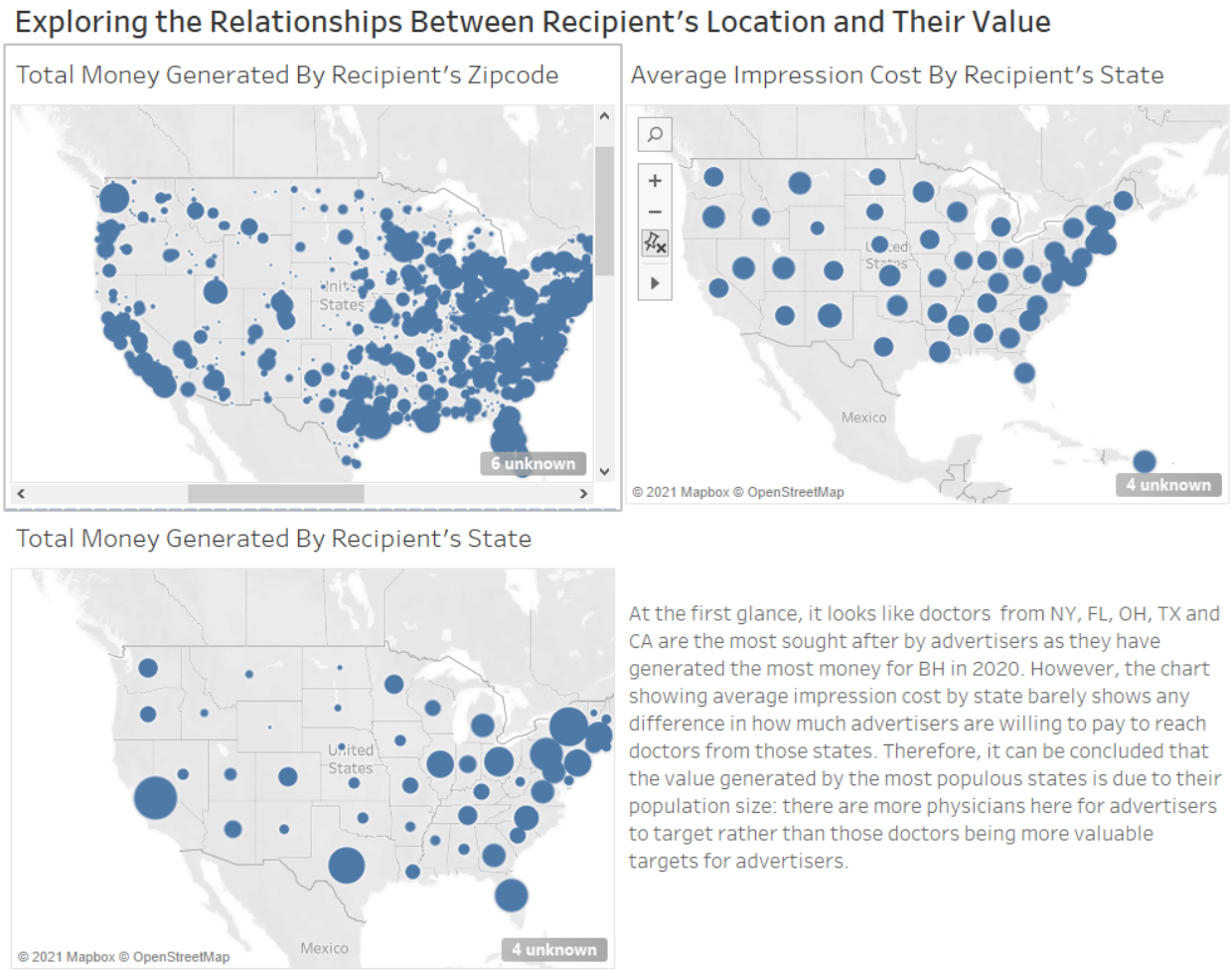
2.2. Variables with continuous data were explored for outliers, the outliers were removed, and the variables log-transformed to fix skewness.

2.3. The categorical variables were recoded wherever necessary.

2.4. The variable listing subspecialties was transformed to "subspecialty count" by counting each element in the list.

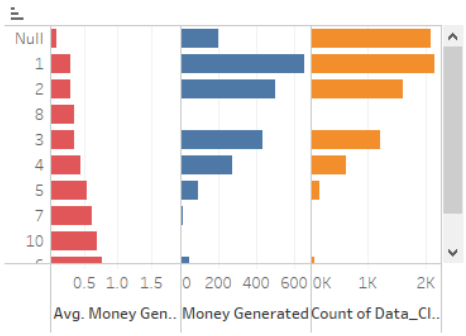
Exploratory Data Analysis

Exploratory data analysis was performed using Tableau. I explored the relationships between each of the features and the target variable, the amount of money generated by each individual doctor. The exploratory analysis suggested a relationship between the dependent variable and the following features: specialty, sub-specialty count, the association through which a briefing is delivered to a recipient, and whether the doctor works in a hospital that is a member of a larger system. These variables were selected for modeling.



Relationship Between Physician Specialty and Their Value

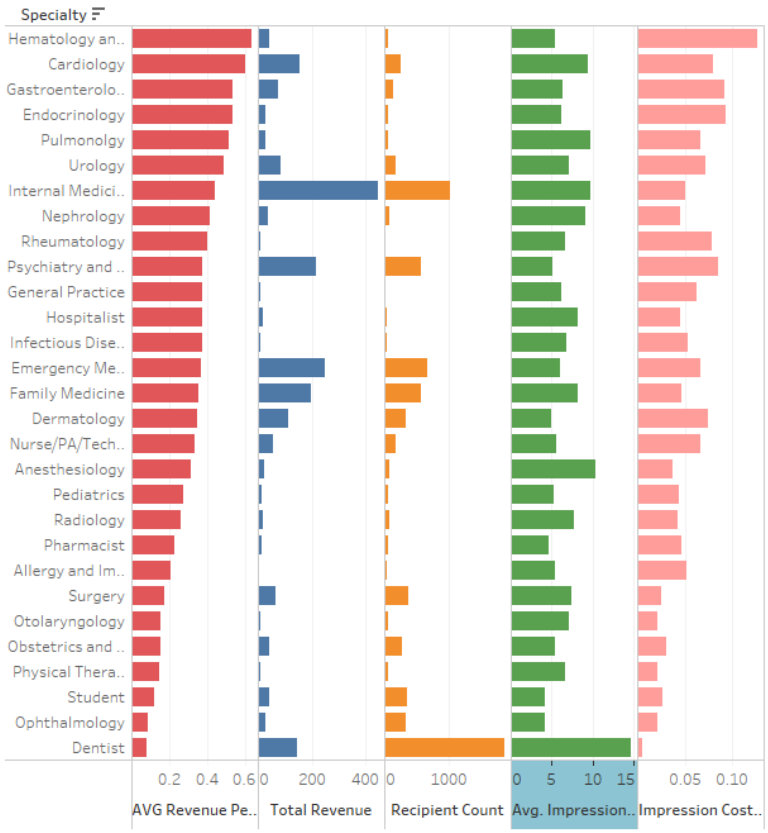
Avergage And Sum Money Generated By The Number of Specialties a Recipient Has



While most money were generated by recipients with 1 or 2 specialties, this is because most participants in the dataset have 1 or 2 specialties. The averages, however, show that with an increase in the number of specialties, advertisers target the recipient more often thereby inceasing the amount of money these recepients earn for Bulletin Healthcare.

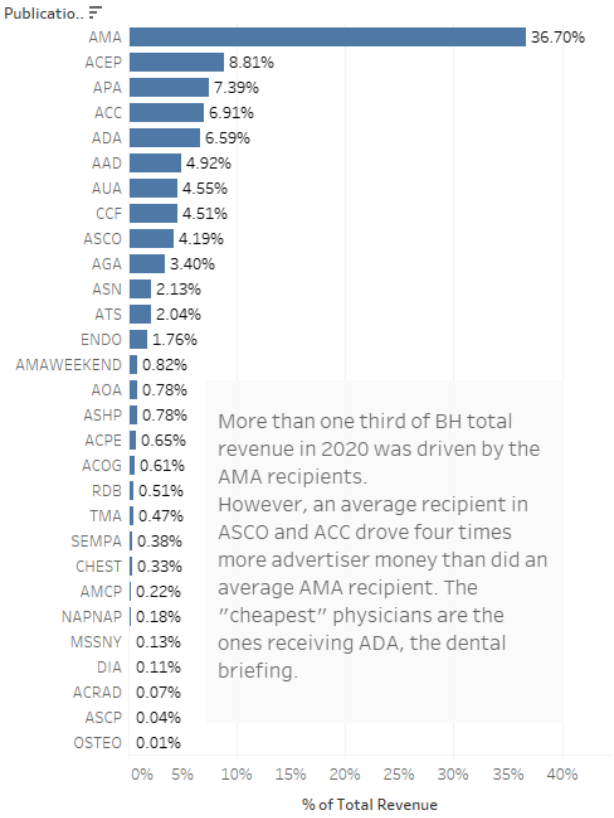
Furthermore, while dentists generated the most total revenue, this is due to the fact that this is one of the two largest populations in the sample (along with internists). They also generated the most impressions suggesting that advertisers target them more often, however their average impression cost is the lowest, meaning that advertisers pay little to reach them. On the other hand,on average, HemOns, cardiologists, gastros, endos and pulmonologists drive way more revenue. Interestingly, as the Impression Count chart indicates, this seems to be not due to the fact that these specialties are targeted more often but because advertisers tend to pay more to target them.

Advertising Revenue By Specialty

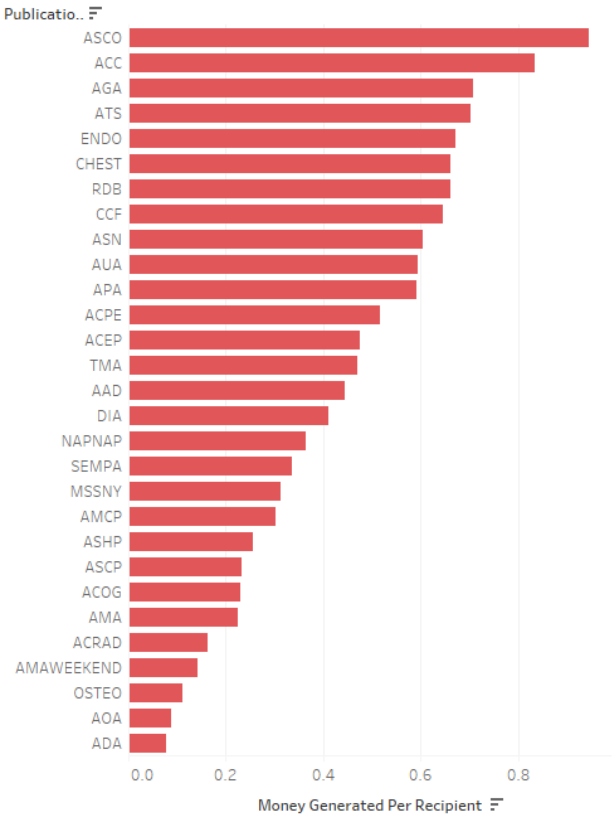


Relationship Between Briefings and Revenue

Revenue Generated By Each Publication As a Percentage of Total Revenue

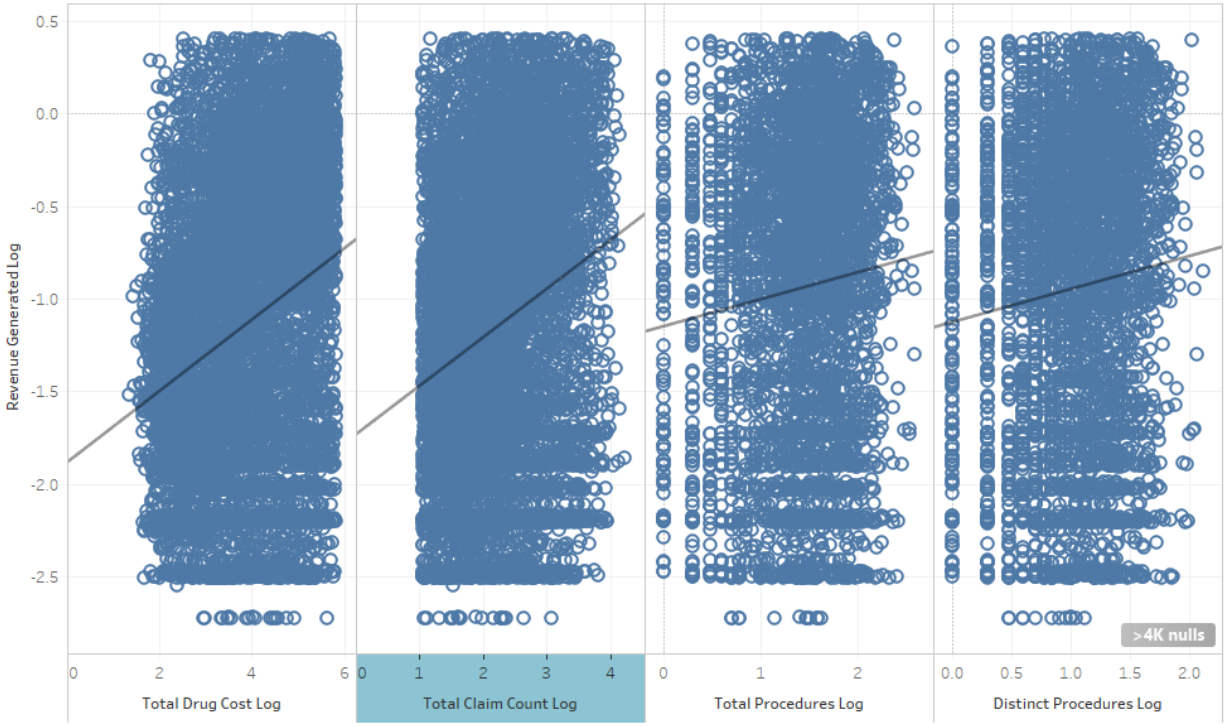


Most Valued Publications



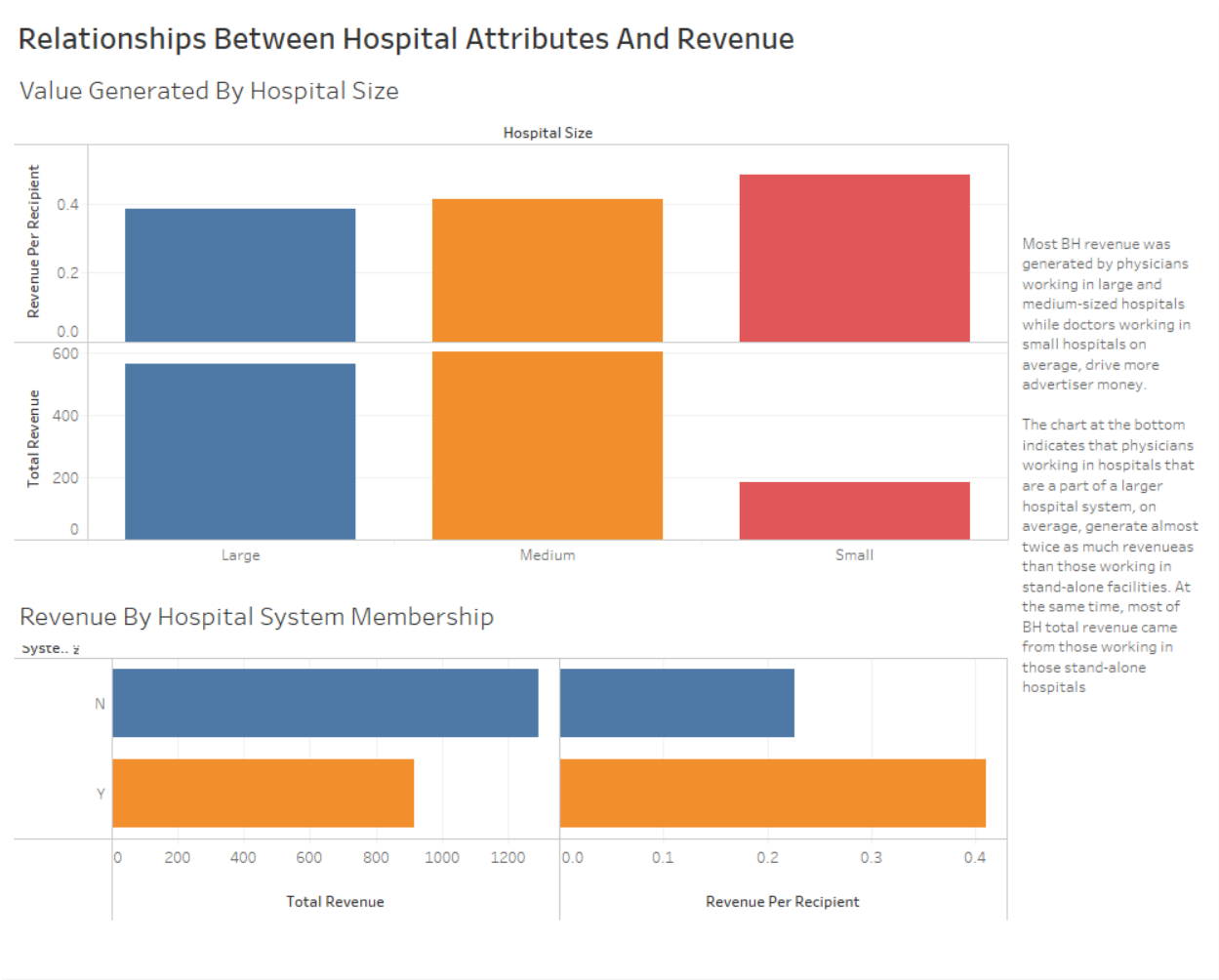
Relationship Between Revenue and Drugs Prescribed/Procedures Performed

Relationships Between the Amount of Prescriptions/Procedures Performed and Money a Recipient Generates for BH



While the charts indicate statistically significant relationships between a physician's ability to generate revenue for BH and the amount of drugs prescribed (as measured both in terms of count and in terms of the total cost of those drugs), the relationship is weak in magnitude explaining less than 1% of the total variance in the total revenue generated by a recipient.

The relationships between the number of procedures performed and revenue is even weaker while also statistically significant



Modeling

1. Data Preprocessing.

a. Dealing with missing values. Sub-specialty count was a integer variable, so missing values were replaced by the median. For categorical variables, missing values were replaced by the most frequently occurring values.

b. Infrequent values for categorical variables including publication/briefing and specialty were combined into group 'other'.

c. Dummy variables were created for all categories using pandas.get_dummies() method, dropping the first category to avoid multicollinearity

d. The data was split into a test and a training set with a test set equal to 25% of the sample.

2. Model selection.

Because the dependent variable (revenue) was a continuous one, the natural choice for this type of data was a regression model.

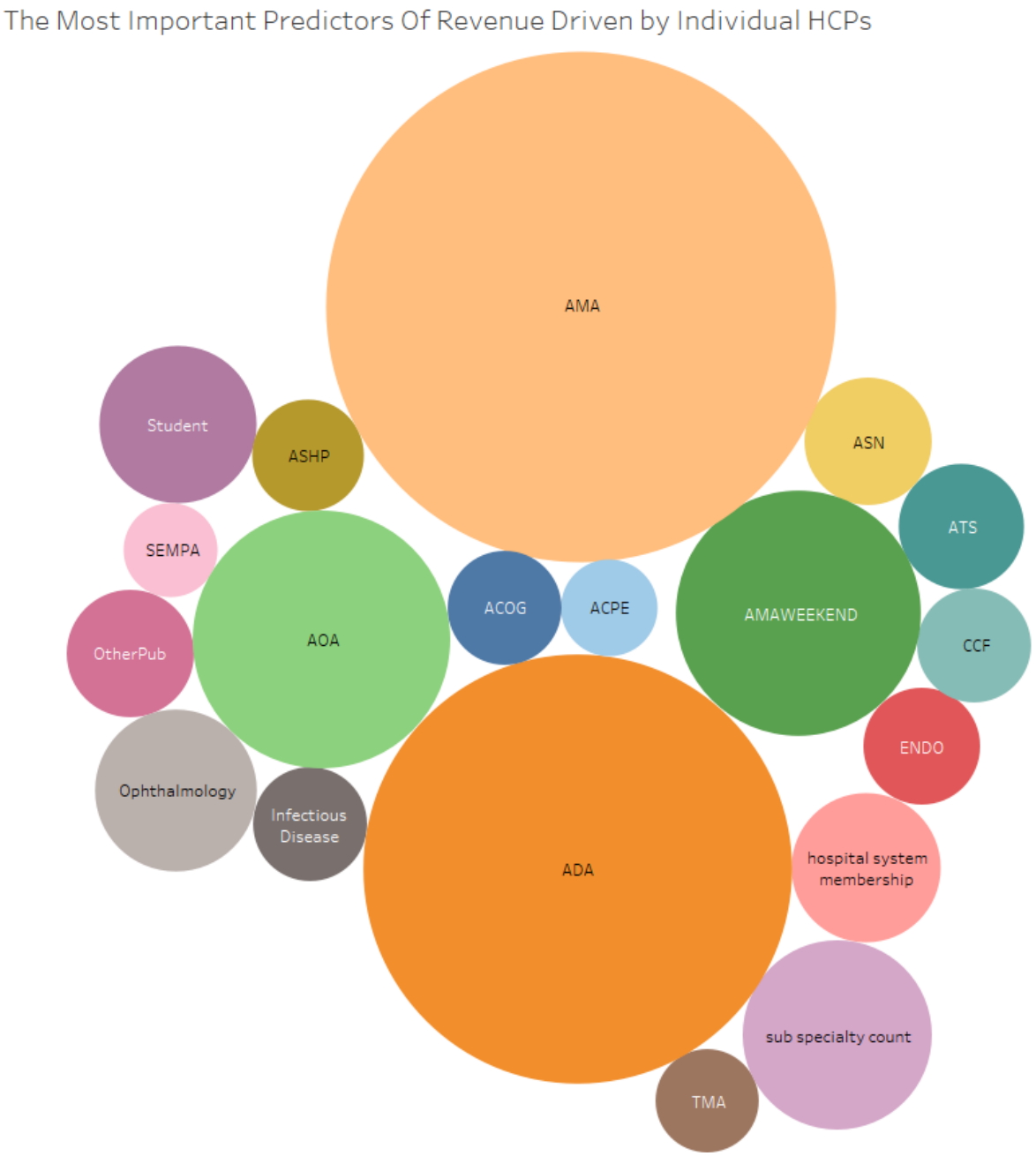
1. For the first model, I used scikit learn's LinearRegression() model in a pipeline with StandardScaler() to ensure comparability among the features. I performed a five-fold cross validation, the model's mean R-squared 0.365 meaning that it explained 36,5% of the variance in revenue just from recipients' characteristics alone. Given the fact that advertising prices influenced by other factors such as the type of advertiser, various discount programs, association size and so forth, which weren't taken into account, this was a good result.
2. For the second model I used OLS() from the Statsmodel package. The model explained 37.4% of the variance in the DV.
3. In an attempt to boost the model's predictive power I included an interaction term into the next model, but the interaction turned out to be not a significant predictor of revenue and did not improve the model's performance, so the second model was chosen. Its R-squared dropped to .34 on the test set, but still the model was quite useful in identifying the relationships between recipients' characteristics and advertising revenue.

Results

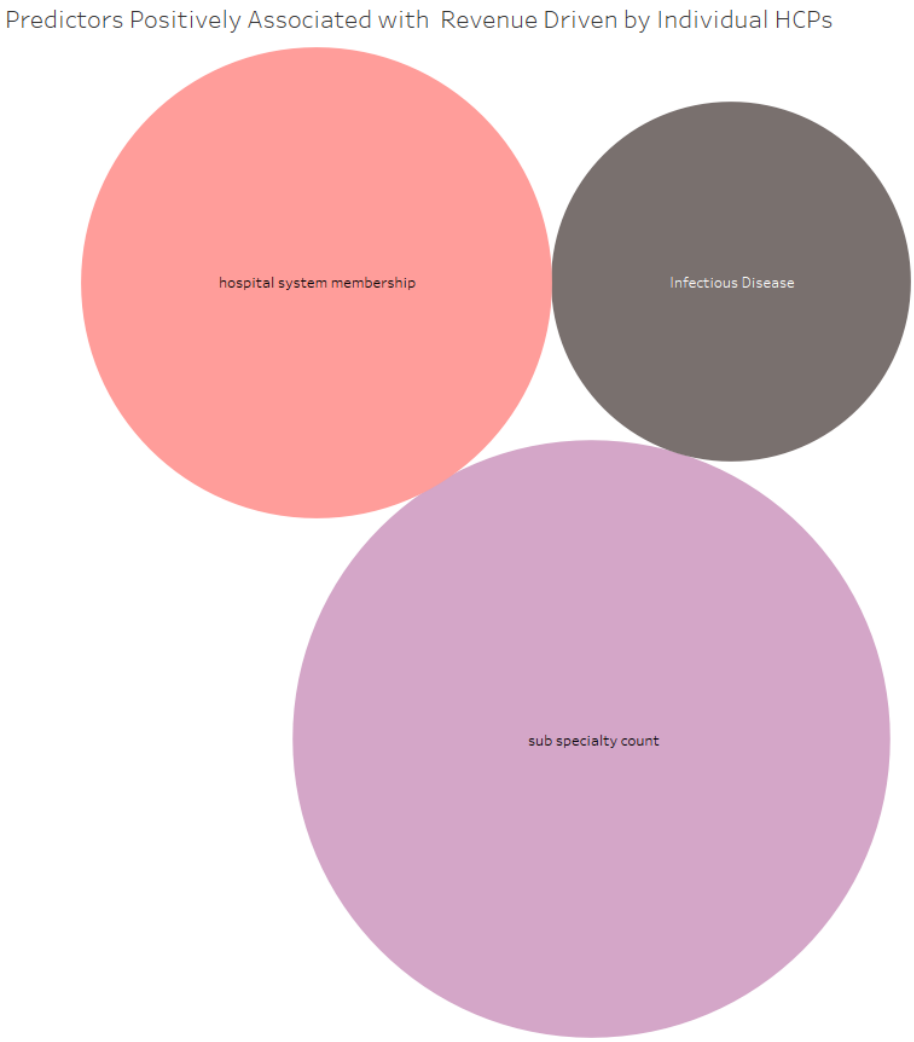
I explored the standardized b-coefficients of the predictor variables to determine which ones were the most important in predicting revenue.

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.7823	0.156	-5.016	0.000	-1.088	-0.477
hospital_system_membership	0.1081	0.021	5.187	0.000	0.067	0.149
total_claim_count_log	0.0310	0.030	1.027	0.305	-0.028	0.090
total_drug_cost_log	0.0464	0.024	1.910	0.056	-0.001	0.094
sub_specialty_count	0.0735	0.011	6.877	0.000	0.053	0.094
DM_ACC	-0.1560	0.102	-1.528	0.127	-0.356	0.044
DM_ACEP	-0.1315	0.092	-1.427	0.154	-0.312	0.049
DM_ACOG	-0.3251	0.129	-2.514	0.012	-0.579	-0.072
DM_ACPE	-0.3449	0.164	-2.106	0.035	-0.666	-0.024
DM_ADA	-0.9424	0.135	-6.991	0.000	-1.207	-0.678
DM_AGA	-0.1603	0.128	-1.252	0.211	-0.411	0.091

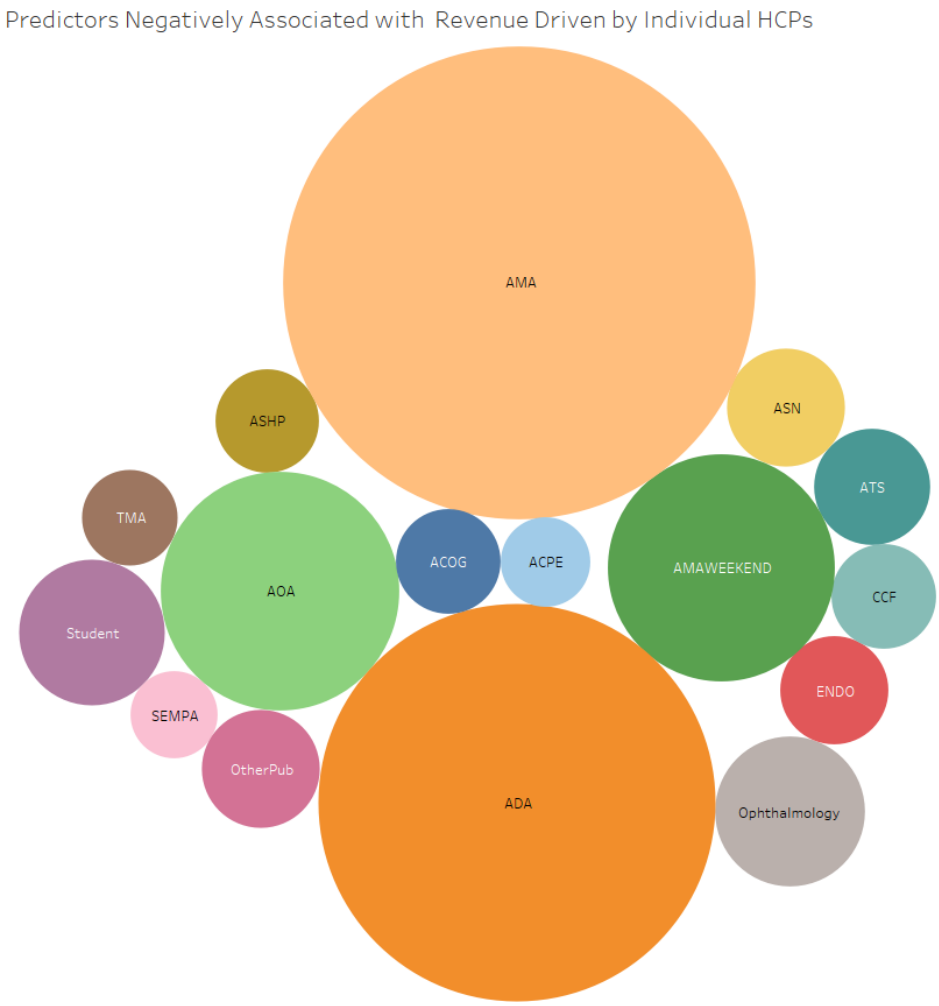
I have filtered out those predictors where p-values for their b-coefficients did not reach statistical significance at p>0.05.



As we can see, membership in associations including AMA, ADA, AOA, and AMAWeekend are the most important factors impacting the revenue driven by individual HCPs. This is not surprising given the ads are being priced differently, depending on the size of an association and its medical specialization.



Among the features that are positively associated with revenue, the number of subspecialties a recipient has plays the most important role. This gives an advantage to doctors with narrow specializations because their broader specialties are also included into the count, in addition to subspecialties. In addition, doctors from large hospital systems tend to drive more revenue, compared to local hospitals. Finally doctors specializing in Infectious Disease also drove more revenue than doctors with other specializations. However, this migh be the effect of 2020 and COVID-19.



Lastly, the most important negative predictors of revenue were membership in the American Medical Association (AMA) whose briefings are priced lower than more specialized publications, as well as memberships in AOA (American Optometric Association) and ADA (American Dental Association), perhaps due to the smaller amount of advertisers or products worth advertising to these audience (e.g. medications).

Takeaways

- Bulletin Healthcare should focus on expanding its partnerships with large hospital systems such as Cleveland Clinic, Mayo Clinic, John Hopkins, etc instead of focusing on medical associations.
- Recipients who specialize in therapeutic areas involving multiple specializations are the most promising targets for advertising.
- In the midst of the pandemic, it might makes sense expanding its reach to HCPs specializing in infectious disease.

Future research

The dataset used data for 2020, which was a very unusual year in terms of attention to healthcare news which might have affected the results. Further analysis of 2019 and 2021 could help clarifying the effect of the pandemic on the conclusions made in this study.

In addition, the model only used data on advertising targets which is not the only factor affecting pricing for advertising. This includes the type of product advertised, audience size, as well as the types of ads (e.g. banner vs advertiser-supplied content). The inclusion of these missing pieces into the model could help it better predict revenue.