Who is buying dental insurance from the health exchanges?

By Justin Vena

Background:
The ACA (Patient Protection and Affordable Care Act) passed in 2010 and changed the way healthcare industry behaved. This project covers the consumers who were affected by the ACA introducing the Health Insurance Marketplace (or called the health exchanges) to the United States. Health and Dental Insurance products became available for sale in late 2013 on the exchanges and came into effect for the 2014 plan year.

Before the Health Exchange was established, most people received their insurance from their employer or buying directly from the insurer. Now that consumers can shop around on an easier centralized location, this lead to the question: what types of consumers will buy from the exchanges? The exchanges make it easy for a consumer to shop around between insurance products, knowing the offerings of price and plans available.

Project Question:
What characteristics can we successfully use to classify which types of ZIP codes are more likely to buy an ACA dental insurance product?

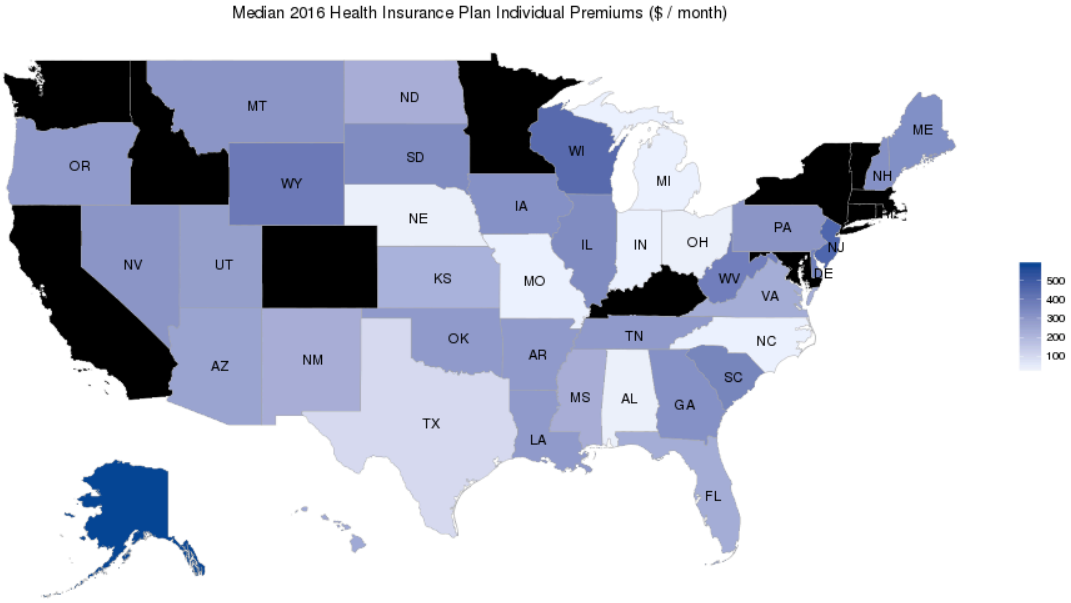
Data Collection:
In order to solve this problem, I needed to gather Insurance, Medicare, and Census information. The data came from two sources, Kaggle.com and Census.gov.

Kaggle.gov is a data science competition website. It has several public data sets that are available to look at. I used the Health Insurance Marketplace dataset. The dataset provides us with insurance and Medicare data from 2014 to 2016. At the end, only 2014 data was used, as the census data and the health data were most complete for that year. Kaggle only has data on 35 states and the health and dental insurance products sold in each state. Several different datasets are included, but I will primarily use the Service Area Data, which provides information about the exchange plan geographic information.

Census.gov has public datasets the US government publishes. I used the 2014 ZIP Code/Income dataset to pull together and match the geographic and income data.

Features used for Analysis:
Kaggle Data Features:

- Geographic: State (35 states in data, those black are missing)



- Geographic: County
- Geographic: ZIP
- Plan Standard: Low or High describes how “rich” the benefits are.
- Child/Adult Plan: If the plan is for a child and/or an adult

Census.gov Data Features:

- Geographic: ZIP
- Income

Data Clean-Up and Visualization:

Kaggle Data Clean-Up, Issues and Visualizations:
The biggest issue with the Kaggle data set is that with several of the plans listed, the ZIP code was listed as 0, or missing. I ran the model manipulating the data in a couple ways. First method was to exclude the missing ZIP codes from the model. The second option was to replace the missing ZIP with average income of all the ZIP codes per state. The second option was used in my model, as it allowed the model to run with a larger dataset.

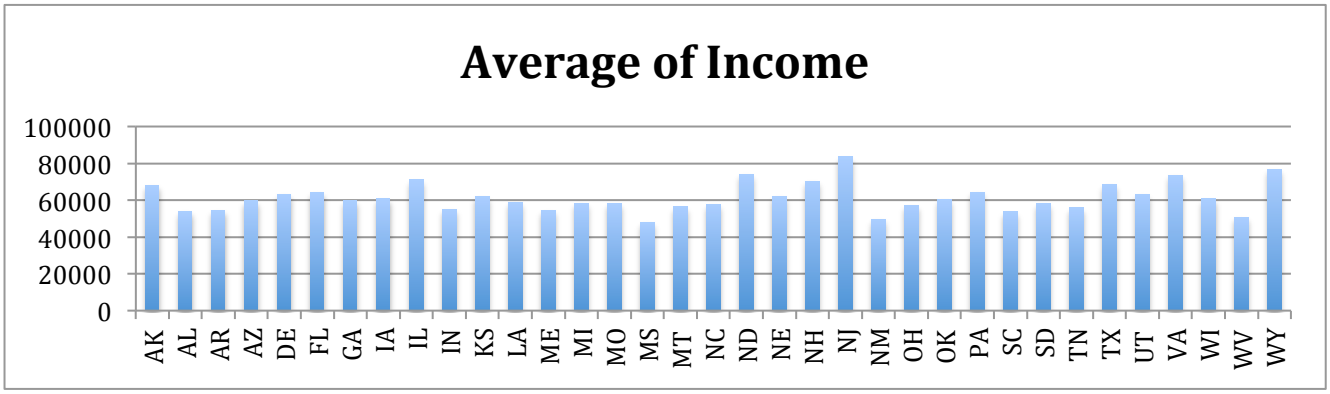
State by Plan Level

	State	High	Low	Grand Total
1	TX	4,254	5,778	10,032
2	GA	1,749	1,908	3,657
3	MI	1,680	1,752	3,432
4	VA	1,607	1,654	3,261
5	NE	1,419	1,419	2,838
…	…	…	…	…
31	NJ	84	126	210
32	AK	87	116	203
33	ME	64	64	128
34	NH	40	40	80
35	DE	12	24	36
	Grand Total	23,465	27,411	50,876

Since many of the data features used in the model were categories, the data needed to be prepared where it could be easily read into python and it would not be confused by the string value.

Census Data Clean-Up, Issues, and Visualizations:
The ZIP/Income information needed to be cleaned and organized. The most complete census dataset that was found was for total family income unadjusted and for individual income unadjusted. Total family income unadjusted was used since plans sold in the data set were children only plan.

State by Average Family Income



Average income for the data selected seemed reasonable compared to the national average ($62,242 for data set).

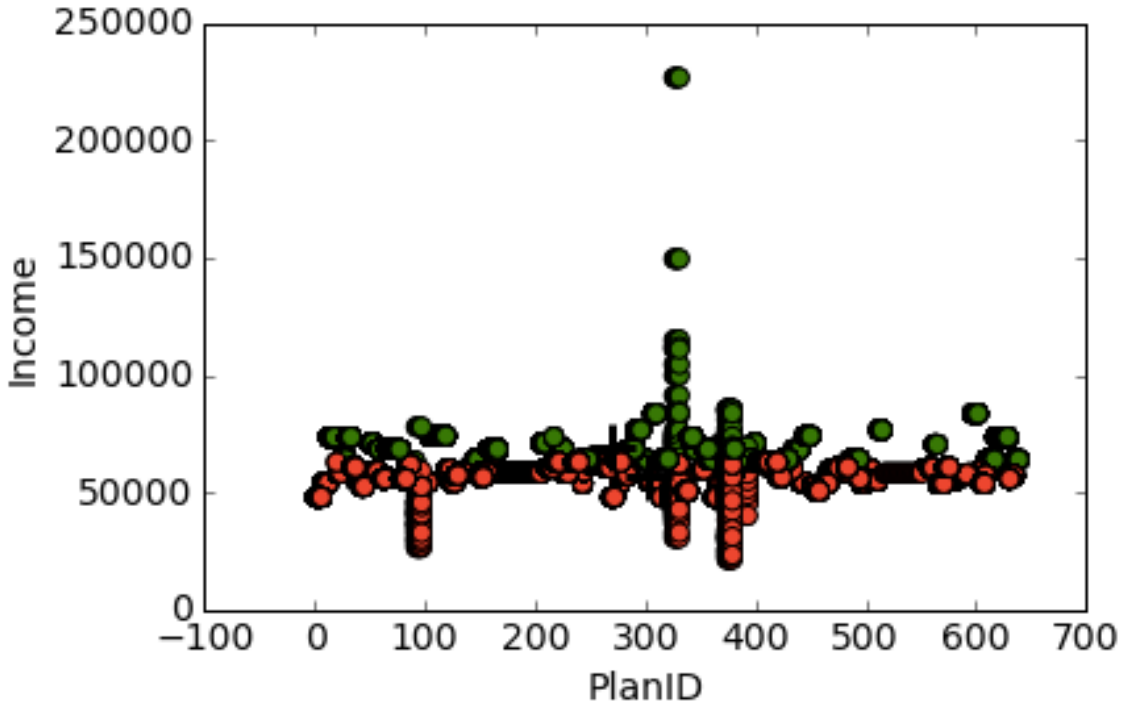
Selecting and Running the Model:

We want to solve our problem by segmenting the different types of plans into clusters that exhibit similar characteristics.

I used K-Means Algorithm to cluster the groups into clusters. I tested using a cluster from 2 to 5. The results were tested and validated using silhouette score. Silhouette scores rank from -1 to 1.

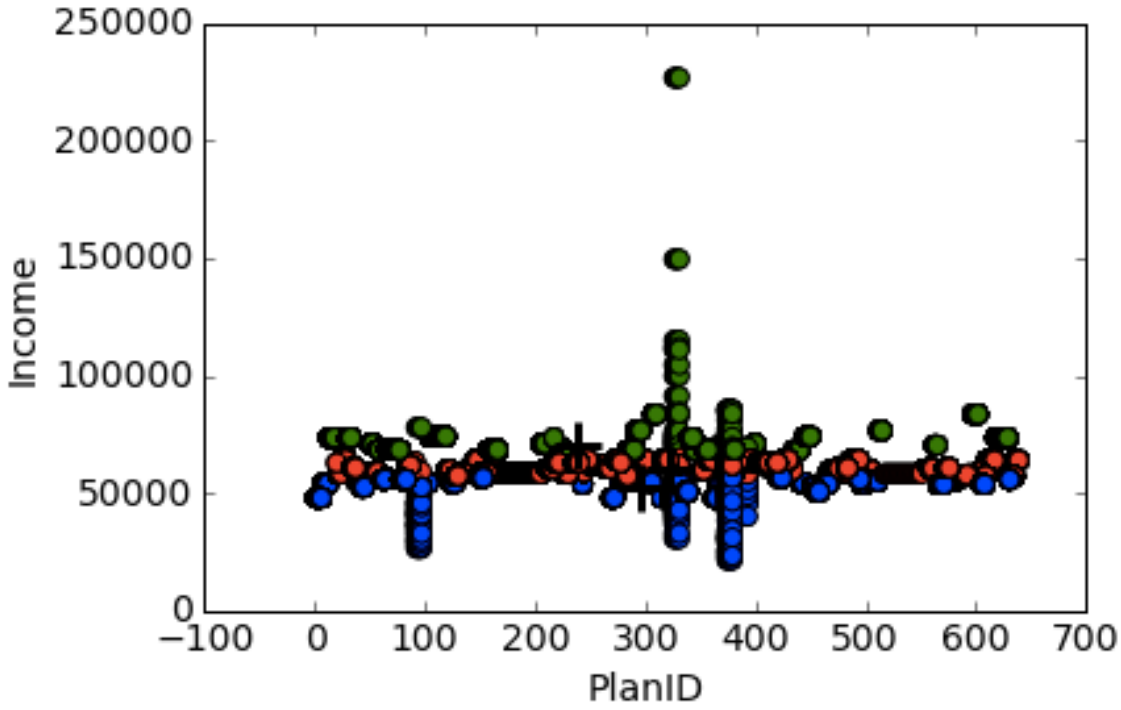
KMeans Clustering Results:

2 Clusters



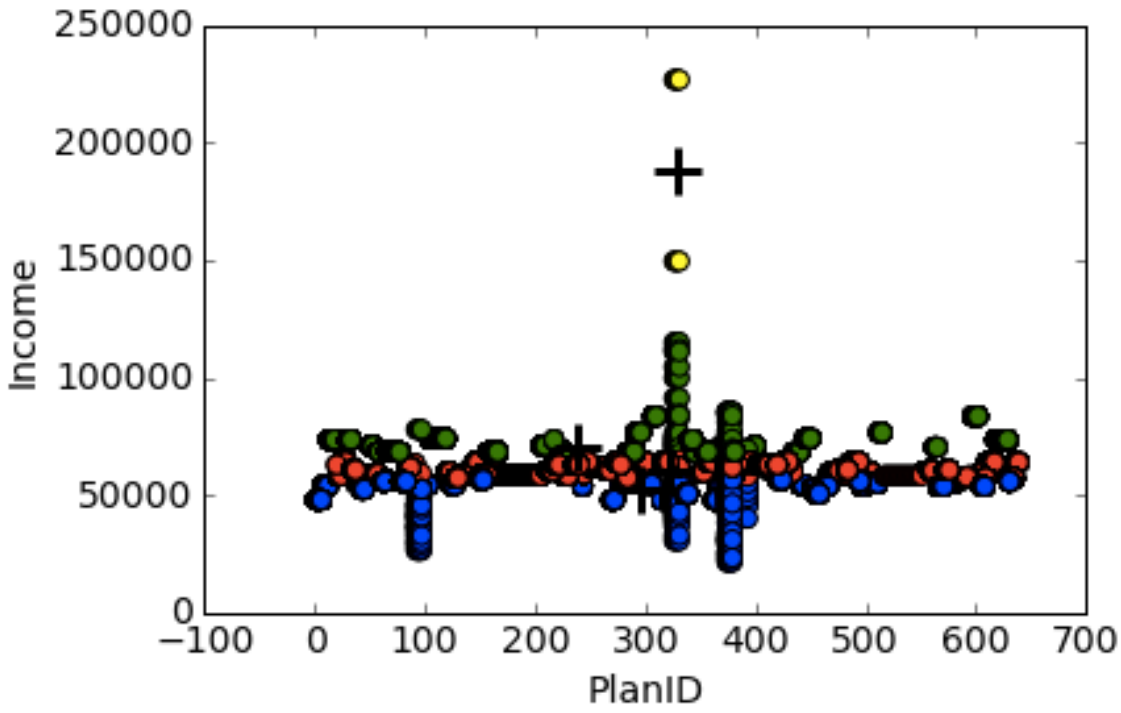
Silhouette Score: 0.60891

3 Clusters



Silhouette Score: 0.61235

4 Clusters



Silhouette Score: 0.61455

Clusters and Silhouette Scores:

Clusters	**Silhouette Scores:**
2	0.60891
3	0.61235
4	0.61455
5	0.62639

Average Income Per Cluster:

2 Clusters	3 Clusters	4 Clusters	5 Clusters
57,163	52,460	52,460	47,884
69,045	60,499	60,499	57,350
	70,536	70,452	62,487
		188,042	70,459
			188,042

Minimum Income Per Cluster:

2 Clusters	3 Clusters	4 Clusters	5 Clusters
21,521	21,521	21,521	21,521
63,258	56,710	56,710	52,644
	65,878	65,878	60,440
		114,958	66,698
			114,958

Maximum Income Per Cluster:

2 Clusters	3 Clusters	4 Clusters	5 Clusters
62,927	56,452	56,452	52,426
226,625	65,415	65,415	59,754
	226,625	114,912	65,988
		226,625	114,912
			226,625

Code:

Coding was written in python, data prepared in Microsoft Excel. Please see my Github repository (https://github.com/jv113090/DS5) for more information.

Business Applications and Conclusions:

This type of study is useful to see what types of plans should be sold in different ZIP codes. Corresponding how “rich” the benefits are for a particular plan to a certain income level, we could a better focus on how to sell individual and segment plans around the United States. To make this more successful, possible dental plan elements such as Maximums or Deductibles could be added to improve the model.

With the average income known about the plans sold to which ZIP codes, insurers can focus advertising on, or creating plans and networks that better suit these customers. Looking at the range the clusters place the incomes with the corresponding richness of the plans.

With extra time, I would try to explore the other data outside the geographic part of the Kaggle dataset. I would add more plan features that included more benefit elements. I would also try to use dimensionality reduction to see how I could improve the model. Many of these could be added to make the model more robust.