Capstone Project #1: Project Report

Earnest Long, Jr.

Project Title: Deep Demographic Insights from Amazon Customer Data

INTRODUCTION

Businesses are constantly seeking new customers to purchase their products. They attempt to attract these customers by way of a discipline called marketing. But the questions at the forefront of the minds of any marketer are: who should I market to, and how? There are many ways to approach these questions, but statistical data analytics can often generate insights that even the most seasoned marketer may miss. As an entrepreneur myself, I would find these insights extremely valuable for my own business; so, for my first capstone project, I have chosen to perform analytics on our own Amazon customer data to generate marketing insights. Specifically, I plan to explore the demographics of our current customer base in order to determine better who to target with our marketing and which messages may resonate more soundly with them.

Amazon provides customizable order reports to its sellers through its Seller Central portal. These reports include data such as purchase items & quantity, name, address, city, state, and zip code. The reports come downloadable in a well-formatted .txt file that is tab-delimited for easy importing. The demographic data that will be used in concert with the customer data comes from data.gov. One dataset, found at https://catalog.data.gov/dataset/zip-code-data, details US Income Data at State & ZIP Code Level. This will be used to estimate the income / wealth levels of our customer base, which can be used to generate more effective marketing messaging. For example, if the majority of our customers come from very wealthy zip codes, branding the product as a premium or luxury item (including even raising the price!) may be a great business move. There also exist state-level datasets with more detailed demographic data, such as race and age. This data also comes from data.gov (e.g. Demographic Statistics By Zip Code - State of New York Demographic Data). State-level analysis can be performed on the most popular states for our product, as desired.

I will visualize the data in ways that make the insights most clear. Histograms, for example, will allow me to easily group customers into bins by wealth / income, which will showcase the income distribution of our customer base. A simple X-Y plot of order frequency vs. time could also reveal the efficacy of marketing events at various points in

our company's history. There are many variables to analyze in this data, and I will look for insight wherever some may be found.

DATA WRANGLING / CLEANING

The data for this project comes from two sources: Amazon.com (order fulfillment reports through their Seller Central portal) and Data.gov (providing geographical demographic data). Because these are both data-friendly sources, the data I needed was available for download in an easily accessible format, and was quite clean from the start. However, some data wrangling & cleaning was still necessary to prepare the data for analysis.

For the demographic dataset (https://catalog.data.gov/dataset/zip-code-data), there were some rows where the zip code was 0, and some where the zip code was 99999. These rows corresponded to the aggregates for each AGI bin and "ZIP codes with less than 100 returns and those identified as a single building or nonresidential ZIP code" (per the documentation), respectively. These rows were deleted from the loaded dataframe.

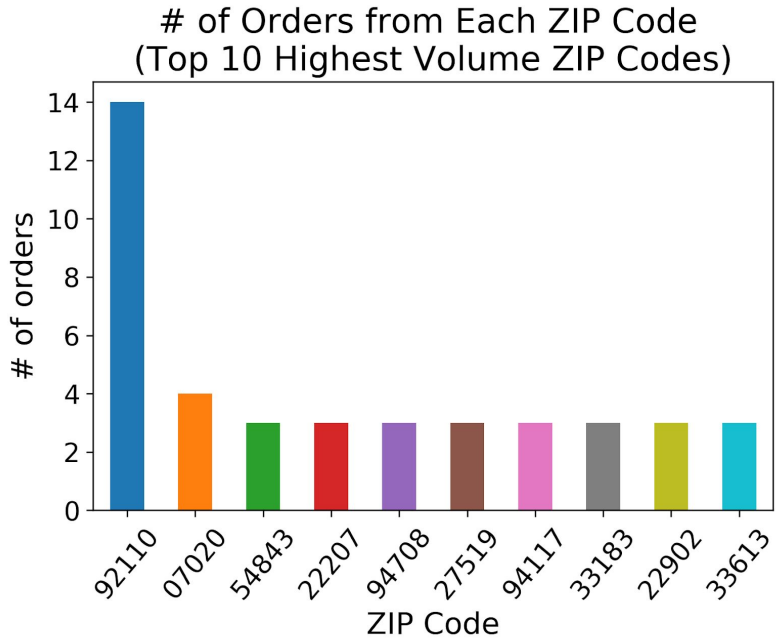
For the Amazon dataset, which consisted of many 30-day reports that were loaded individually and concatenated into a single dataframe, more cleaning was required. Some of the orders recorded in the reports were invalid - they had either been cancelled or were internal orders. These rows were deleted from the dataframe. The formatting of the zip codes was also inconsistent. Some zip codes were the standard 5 digit codes, while others were the full 9 digits. Using dataframe.apply allowed me to keep only the first 5 digits, while discarding the rest. Performing this operation uncovered another, unrelated, issue: some of the zip codes were only 4 digits long (e.g. 7047). After determining that relatively few rows (17) were malformed in this manner, I looked up some of the zip codes by the city and state of those rows. This revealed that those zip codes were merely missing a leading zero, which had been truncated by Excel due to number formatting. Another *dataframe.apply* call added the leading zeroes back to these zip codes with the *zfill* method (e.g. 07047). The same process needed to be applied to the data.gov dataset.

Exploratory Data Analysis

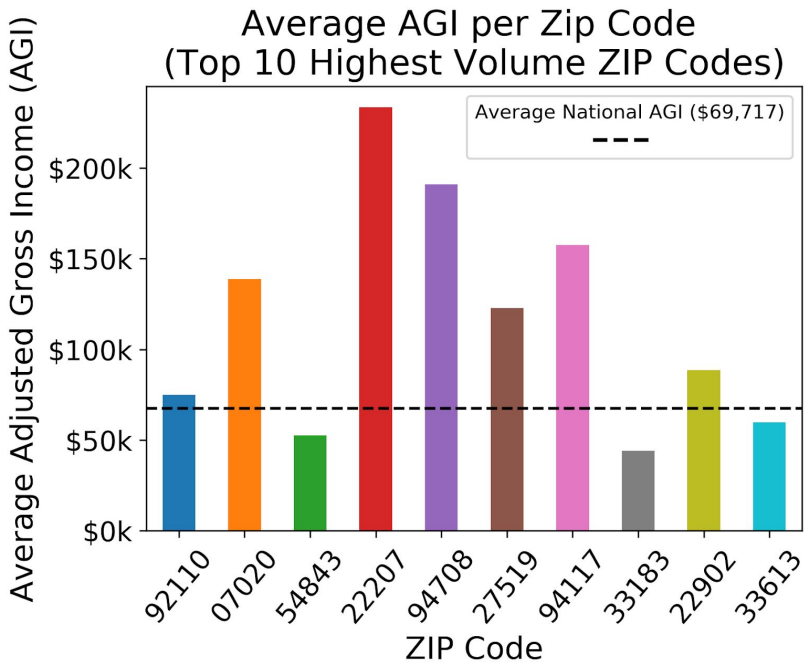
After compiling and cleaning both datasets, I set about the task of exploring and analyzing the data. My first goal was to analyze the ZIP code data to get an idea of my customers' demographics. Since both datasets contained ZIP code columns, I merged on that column to obtain a unified dataframe, and aggregated the values for each ZIP code to obtain total values for each ZIP code. I then divided by the total number of tax

filings to get average data for income, number of dependents, elderly filings, and types of households (via single, married, or head of household (HOH) filings). Counting how many orders came from each ZIP code allowed me to identify the top 10 ZIP codes and focus my analysis on those. I constructed a visualization for each of the following questions:

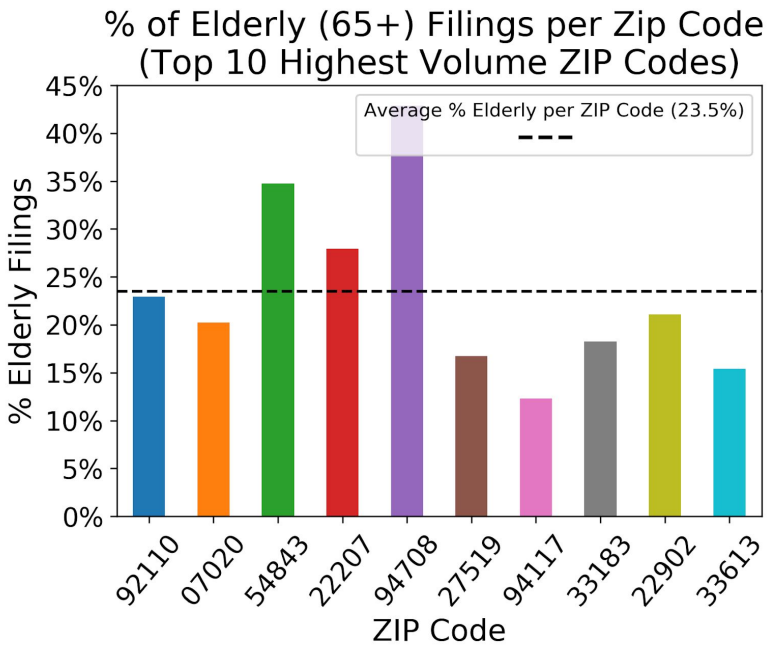
- Q: In which ZIP codes do most of our customers reside?
    - A: By far, most orders come from 92110, which encompasses San Diego, California. The next ZIP code is 07020 (Edgewater, NJ), followed by many other ZIP codes tied for third.



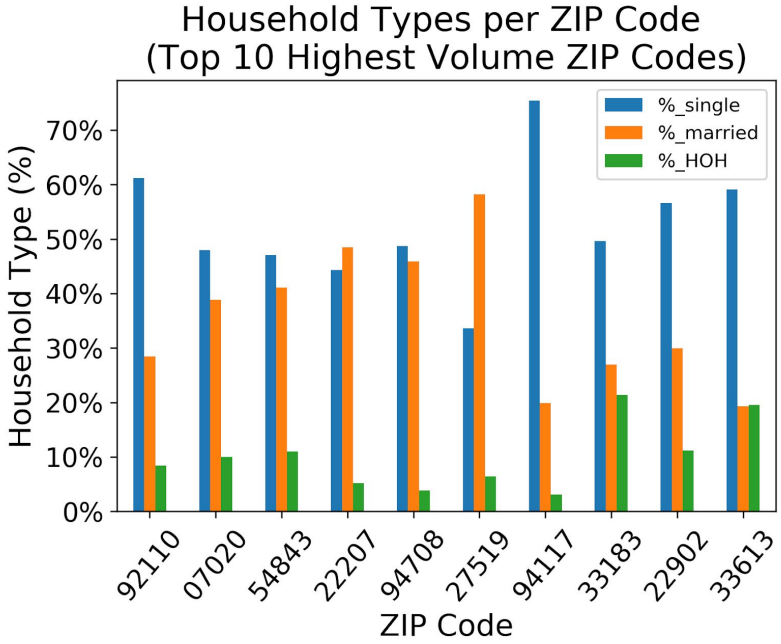
- Q: How wealthy are those ZIP Codes?
    - A: I calculated the national average Adjusted Gross Income (AGI) from the data and overlaid it on a bar graph with the average AGI of each of the top 10 ZIP codes. A few were below average, but 7 of them were above average, with 5 ZIP codes being well above average. This suggests that many of our customers may be relatively wealthy, which makes sense since our product is a bit more expensive than other products in our space.



- Q: Can we get some idea of the age distribution of these ZIP codes?
    - A: Again, I calculated and overlaid the national average percentage of Elderly filings per ZIP code onto the bar plot of each ZIP code's % elderly filings for comparison. Only 3 ZIP codes had a higher than average % of elderly filings. It's difficult to come to a definitive conclusion because most ZIP codes only average about 25% elderly filings, but since 2 of those ZIP codes had a much higher than average % of elderly filings, this may indicate that our customers are disproportionately in the 65+ age category. From other business analysis, we know that a good number of our customers do come from that age category, so again, this makes sense.

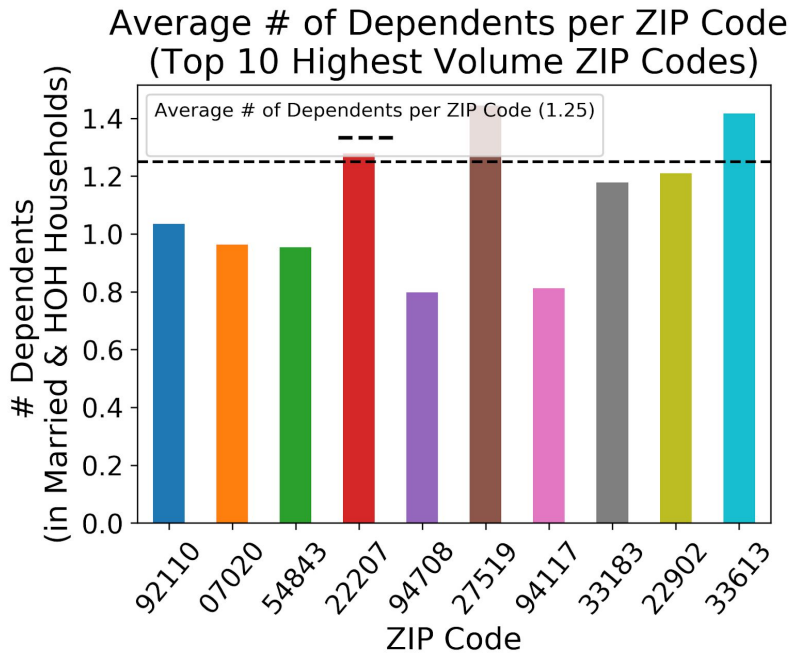


- Q: What is the distribution of single households vs. families (married or HOH filings)?
    - A: The distributions varied among ZIP codes, but our highest volume ZIP code (92110) was very disproportionately single, which suggests that single customers may be our target demographic.



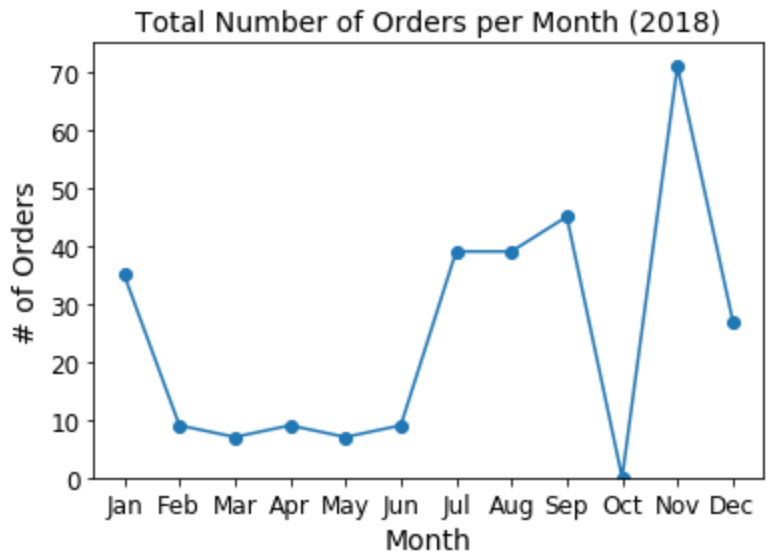
- Q: What is the average # of children in married & HOH households?

- A: For the married and HOH filings, it turns out that most of the ZIP codes had less-than-average average number of children per household.

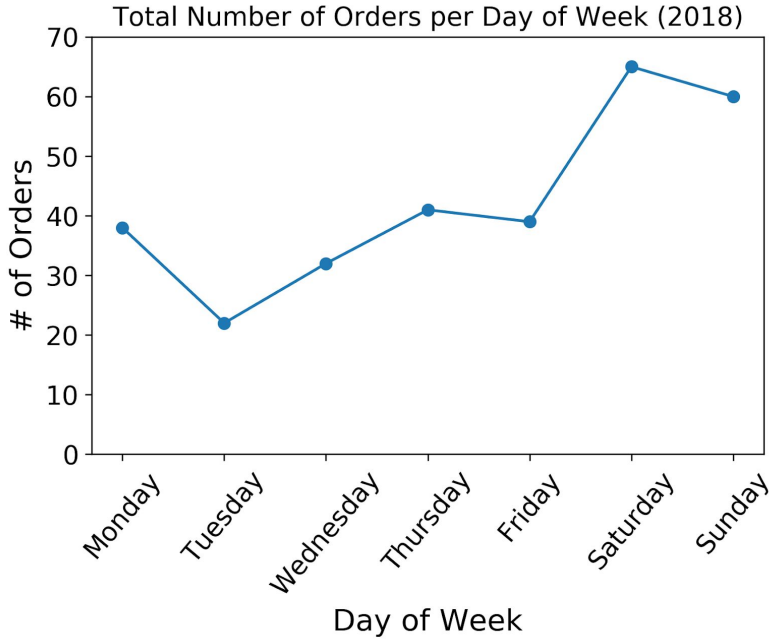


I then proceeded to analyze our order data by purchase date and time, allowing me to generate several behavioral insights. With the data restricted to only year 2018 (because we only have partial data from 2017), I asked the following questions:

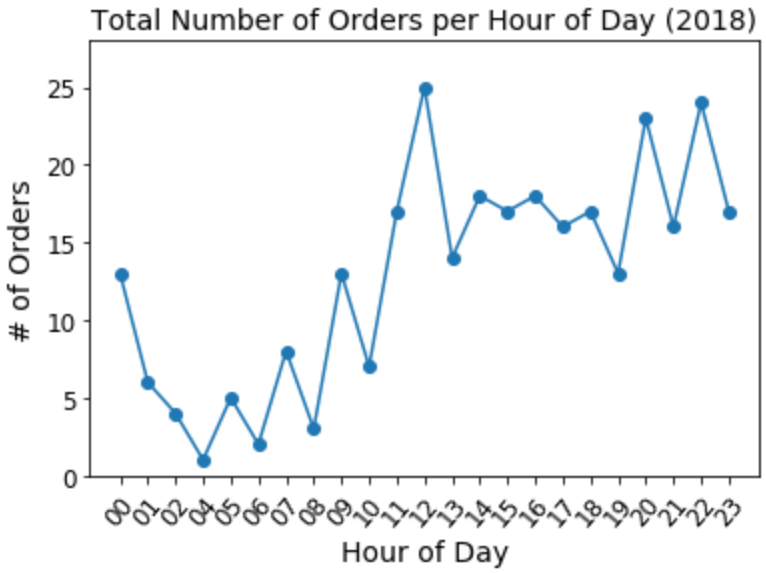
- Q: In what months do the majority of our customers make purchases?
  - A: November, overwhelmingly. This makes sense because it is the pre-Christmas shopping season, and December shows strong sales too. January is also popular, as are the summer months. We altered and adjusted our advertising at several points throughout the year though (stopping advertising in October, for example), so this analysis will need to be re-done with a consistent, sustained marketing effort.



- Q: On what days do the majority of our customers make purchases?
    - A: Saturday and Sunday, which makes sense being the weekend.



- Q: At what time of day do the majority of our customers make purchases?
    - A: At night mostly, but there is a mid-day spike as well. Do people place Amazon orders on their lunch breaks? It's difficult to establish firm conclusions due to varying time zones, though - this analysis needs to be re-done with each order's time in its own time zone.



One weakness of this analysis was the lack of order data - we haven't been on Amazon long enough to accrue substantial data. Other variables like variable advertising efforts and different time zones also need to be accounted for. It would also be useful to have order data for other Amazon products in order to determine what makes our customers unique in the aspects that were studied. Still, the plots reveal some interesting insights into our customer base that would be difficult to see by staring at the raw data alone.

Inferential Statistics

At this point in my capstone project, I endeavored to use my newly honed inferential statistics skills to answer some interesting questions about my data. The question I felt that was most pertinent to my data, in light of the analysis that I have performed so far, was the following: "I have chosen to analyze only the top 10 ZIP codes by volume... How representative are these ZIP codes of the order population from which they were taken?"

To answer this question, I chose 3 of the metrics I analyzed previously: average Annual Gross Income (AGI), % elderly filings, and number of dependents. To assess how representative they were of the total order population, I applied the Chi-Square Goodness of Fit test to each metric, where my null hypothesis was that the sample containing the Top 10 ZIP Codes has the same distribution as the total population. The results are as follows:

- Average AGI: $X^2 = 835.22$; $p \sim 0$; Null hypothesis (strongly!) rejected

- % elderly: $X^2 = 34.14$; $p = 0.00008$; Null hypothesis (strongly!) rejected
- # of dependents: $X^2 = 0.55$; $p = 0.99995$; Null hypothesis (strongly!) accepted

The results are what one would estimate from the visual plots of the metrics: average AGI and % elderly filings in the top 10 ZIP codes were significantly different than the overall order population, while the number of dependents was nearly identical. My recommendations in light of this analysis are mixed: On one hand, the fact that our highest volume ZIP codes differ so greatly in average AGI and % elderly suggests that those two metrics are prime candidates for targeted marketing efforts. On the other hand, the fact that our total order sample is still relatively small, and the top 10 ZIP codes, as yet, comprise only a small overall percentage of our total sales, it may be unwise to focus too narrowly on those metrics at the expense of others. Both more data and more analysis are required before firmer conclusions can be drawn.

Machine Learning

The next order of business for my capstone project was to select and apply machine learning algorithms to my data in an effort to uncover patterns and build predictive models. Between supervised and unsupervised machine learning, the choice was clear - for this project, I would use unsupervised learning. Supervised learning was ruled out immediately not only because there was no data to train a model on (I could not find any datasets containing Amazon order data), but also because it is unclear what manner of useful predictions could be made. Unsupervised learning algorithms can find distinct clusters in data, which, in this application, is perfect for further understanding our customer base. Going into this analysis, I hypothesized that I would find at least two clusters, at least one of which is clustered primarily by average adjusted gross income (AGI), as a proxy for customer wealth.

I tried several clustering algorithms, with mixed success. The features I used for the clustering were: price of items purchased (‘item-price’), number of filings (‘N1’) [proxy for ZIP code size], average AGI (‘avg_AGI’), average number of dependents (‘NUMDEP’), and average % elderly population (‘%elderly’). In my first attempt at identifying clusters, I used KMeans, one of the most popular clustering algorithms. KMeans requires that the number of clusters be provided as a parameter, and, since I did not know how many clusters our customer data had beforehand, I constructed an elbow plot of the sum-of-squares error vs. the provided number of clusters (K) that generated that error for 2 <= K <= 10. While the unscaled data generated a plot with a decent “elbow”, I quickly realized that, due to vastly differing scales in my data, scaling my data was essential. The elbow plot of the scaled data was far less useful, providing

no clear optimal value for K. I then tried another method of finding K: silhouette scoring. I calculated the average silhouette score for each value of K and hoped one would stand out from the others. Unfortunately, all of the silhouette scores were ~0.25, which indicates weak clustering.

Next, I applied Principal Component Analysis (PCA) to my data. This served one main purpose: to reduce the dimensions of the data to 2 in order to plot the data points on XY coordinates. I then moved on to other clustering algorithms, now with the ability to plot their results with their cluster labels. Plotting the data points plainly (without any clustering) showed a mass of points with some outliers around the mass, a.k.a. no clearly discernible clusters. This provided further evidence for the direction that my mind was headed - that there were no distinct clusters in our customer data.

I chose to try the clustering algorithm Affinity Propagation (AP) next because it determines the number of clusters rather than that parameter needing to be defined. I hoped that this would give me a useful number to compare my KMeans data to. Surprisingly, AP found a total of 50 clusters in my data! (Or, perhaps, unsurprisingly, since my previous experience with AP also generated a nonsensical number of clusters.) Just to explore further though, I calculated average silhouette scores for 11 <= K <= 20. The scores were all ~0.25 or less, giving no indication of an optimal K at higher values of K.

The final two clustering algorithms I tried were DBSCAN (Density-based spatial clustering of applications with noise) and HDBSCAN, an upgraded version of DBSCAN. Incredibly, after tuning its ‘eps’ parameter (a neighborhood parameter that determines how far away from a point the algorithm will search for a point to cluster it with), DBSCAN found 7 clusters that, when plotted, were reasonably distinct, along with many noisy points. Manual inspection of these clusters revealed similarities between the orders therein, suggesting that the clusterings are indeed legitimate. The characteristics of the clusters are in the table below:

Cluster	Count	AGI	Item-Price	NUMDEP	%_elderly	N1
0	405	High	Med. High	Average	Average	Average
1	28	Medium	Low	Average	Average	Average
2	5	Average	Low	Average	Average	Average
3	13	High	High	Low	V. High	Average
4	7	Med. Low	V. Low	Average	Average	V. High
5	5	V. High	Med. High	Average	V. Low	Average
6	5	V. Low	V. High	Average	V. Low	High

Most of the clusters are very small (5-28 customers), except for the largest cluster, whose features are fairly broad and generic, essentially representing the average of our customer base. More samples would hopefully enlarge these smaller clusters (thus increasing confidence in their legitimacy).

HDBSCAN produced a similar result to DBSCAN. While it should be all-around superior to DBSCAN (according to the creators, anyway), its superiority in this application is questionable. It disregarded one very distinct cluster found by DBSCAN, and essentially split another DBSCAN cluster into two very small, but reasonably distinct clusters.

Overall, it makes some sense that DBSCAN (and HDBSCAN) would perform better than the other clustering algorithms. For one thing, it is the only algorithm that attempts to deal with "noise", separating those points from the "meaningful" data in this 5-dimensional feature space. It is also very good at capturing non-linearly separable clusters (unlike many other algorithms which can only find globular clusters). The DBSCAN parameter eps requires some experimentation to tune, but it seems that if reasonable clusters can be found at all, varying eps will inevitably produce a gradient of results from nonsensical to reasonable, from which an optimal value (or values) can be chosen.

All in all, while it initially seemed that our customers were not clustered along the chosen features, one particular algorithm, DBSCAN, managed to pick out several small, interesting clusters. One obvious application of the results of this analysis would be to create marketing campaigns to target distinct clusters, as well as features shared by

clusters. One campaign could target young, lower AGI ZIP Code customers in densely populated (urban) areas w/ higher pricing, while another could target customers from very wealthy ZIP Codes (e.g. luxury branding) w/ higher pricing.

Conclusion & Recommendations

My various analyses over the course of this project have generated many insights that could (and should!) guide our business efforts in the future. The most basic visualization, the number of orders from each ZIP Code (sorted by frequency) reveals one ZIP Code that stands out from the others. Further investigation into why our product is so popular in this area is definitely warranted. Many of our customers' ZIP Codes are much wealthier than average, several have higher than average elderly populations, many are disproportionately single, and many have a fewer than average number of children per household. All of these facts contribute to an ideal customer profile that can guide our marketing campaigns. Care should be taken though: my statistical analysis indicates that these top 10 ZIP Codes are not very representative of the entire order population, so, recognizing that this is only a subset, we should continue to compile more data and be aware that there are other customer profiles out there as well. Finally, in keeping with my previous findings, the machine learning I performed discovered a several small clusters representing customer segments. More data is necessary to confirm their legitimacy, but aiming targeted marketing efforts at them may be a good idea.