SCB Business Analytics Competition – Final Report

Fardin Ahsan (Team Leader) - [fxa8225@g.rit.edu](mailto:fxa8225@g.rit.edu)

Dren Gara - [dg9901@rit.edu](mailto:dg9901@rit.edu)

Harshit Kumar - [hxk8955@rit.edu](mailto:hxk8955@rit.edu)

**Executive summary**

The success of the business in terms of profitability was strongly correlated with the Average Gross Adjusted Income, School Test Performance, and Median Household income in the zipcode it was in, it also had a strong anti-correlation with the Unemployment Rate, in the area.

Zipcode 14618 was the best zipcode to buy or build a pharmacy in, as it ranked highly in all the metrics that correlate with a shops success and had the highest average profit of any zipcode.

As for buying vs building a pharmacy. We recommend building a pharmacy in the 14618 zipcode. As the struggling pharmacy in that area does significantly worse than what the data would suggest and in terms of profitability it is middling. Which suggests it might be doing many things wrong, which might be hard to undo. However, building a pharmacy and not doing those things, does set up the business for significantly more success, according to our analysis.

**Data preparation**

Our data preparation process consisted of two main steps. Web Scraping, and Data cleaning and processing.

**Web Scraping**

We scraped demographics data from [www.zipdatamaps.com](http://www.zipdatamaps.com) for the zipcodes in the dataset, using pythons web scraping modules. The purpose of gathering demographic data is to find correlations between demographic factors and a zipcode's profitability.

**Data Cleaning and Processing**

We modified the initial datasets provided {‘Pharmacies in ROC’ (PROC) and ‘Pharmacy Sales’(PS)} in a few ways, mainly to make analyzing the data easier for us, as most of the data was already formatted in a way that would work with our choice of software without any modifications. [(Appendix, Data Cleaning)](#Datacleaning)

The modifications were;

* Dropped **['telephone','address1','address2','state','website','zipcode4']** from PROC as these columns did not provide any useful data.
* Dropped all unidentified pharmacies (no store number), as store number was the primary key that connected the two datasets. We thought of assigning random store numbers in PROC for entries that didn’t have a store number, but PS did not have any entries that didn’t correspond to a store number that didn’t exist.
* Added new fields: **[‘total\_cost’,’profit’,’inventory\_cost’,’overhead\_cost’]**.
* Merged the two datasets (outwards merge) on **‘Store Number’** as the primary key.
* Dropped all stores that did not have any sales in 2021, assuming that those shops closed down.
* Dropped all sales data before 2018. (refer to: Why did we normalize the data on time?)
* Divided inventory cost by 0.68 to get a rough estimate on total costs. As there was no way to find the total costs with the information provided in PS (refer to Buying a struggling pharmacy vs building a new one.).

**Data analysis**

**Why did we normalize the data on time?**

At our initial attempt at exploratory data analysis, we discovered a discrepancy in sales data. Not every shop's sales data was in the same time frame. This rendered aggregate functions futile, without data normalization. Normalizing the data based on the number of sales doesn’t remove shops that might have closed down (despite having proportionately high profits when it was open). Therefore, we decided to normalize the data on time. However, that means that we will lose parts of the dataset. Therefore, taking sales only post 2018 made the most sense to us as it was a compromise between having a relatively normalized dataset and losing too many entries. We also realized that sales from 2020 and 2021 would capture patterns that are relevant during a respiratory pandemic but not any other time, but dropping that came at the cost of losing more recent data, thus we included entries from 2018 onwards, to act as a counter to that. [(Appendix, Units sold by the year)](#UnisSoldYear)

**Methodology and philosophy**

Our goal behind the analysis was to find the most profitable location, where BAC could begin expanding its operations and subsequently assist with suggesting either building a pharmacy or buying a struggling pharmacy. However, the data in the dataset was not sufficient to answer either of the questions. The solutions to those problems were making some educated guesses about the cost of running a business in Monroe County (To compare the profitability of shops), and scraping demographic data from [www.zipdatamaps.com](http://www.zipdatamaps.com) to find correlations between demographic markers and profitability. Most of the analysis we did was using a ‘group by’ method. This means that when we wanted to find something concerning something else (for example average profit per zipcode), we grouped the merged dataset based on that field (zipcode), then applied an aggregation function (such as count, sum, mean, or, cumulative sum), then accessed the field (average profit) we need in the grouped and aggregated table.

To answer the first question as to which factors drive the success of a business and predict future success. We decided that the best level of analysis was analyzing on a zipcode basis. A city-wide analysis would produce trends that are not precise enough and street-wise analysis would result in statistical “overfitting” where our models might capture patterns that are not necessarily there. Thus, going by zipcodes meant the ideal tradeoff between too much and too little granularity, given the size of our dataset. To find what makes a specific zipcode profitable, we looked at the correlations of various demographic data with the average profitability of that zipcode, which is nothing more than the total profits in that zipcode divided by the number of shops in that zipcode.

Coming up with a product portfolio simply consisted of finding the top 500 overall profitable products and sorting them in ascending order (setA). Doing the same thing for the recommended shop (set B). Then replacing products that appeared in set A but not set B, setA minus setB are the recommended new products. We used the same method for selecting vendors.

As for recommending between building a new store or buying a currently existing struggling store, which option, in either case, is informed by the data analysis. We weighed the costs and benefits of both options and made our recommendations. (Discussed in detail in the Results section).

**Results**

**What factors drive the success of existing pharmacy stores in the area, and predict future success?**

The demographic factors that correlated strongly with the profitability of a zipcode were [(Appendix, Correlation matrix)](#Corrmatrix);

* **Average adjusted gross income** had the strongest correlation with average profit (61% correlation)
* **School test performance** had the second highest correlation with average profit (35% correlation)
* **Median household income** had the third-highest correlation with average profit (27% correlation)
* **Unemployment rates** had the highest inverse correlation with average profit. (-26% correlation)

Based on the above correlations, A location that has high average adjusted gross income, better school test performance, relatively high household income, and low unemployment rates, would be the best location to buy or build a business in. **14618** was the most profitable zipcode by a large amount (12 times the average). And also, was consistently within the top 5 of all of the positively correlated metrics. [(Appendix, Zipcode Choropleth map)](#Choroplethmap).

However, it is worth noting that there was a slight negative correlation (~ -5.3%) between the number of shops in a zipcode per square mile and the zipcode’s profitability, we believe it is a risk worth taking given the otherwise strong correlations of average profit with the zipcodes demographic factors

**Buying a struggling pharmacy vs building a new one.**

None of the data provided allowed us to calculate the cost of maintaining a pharmacy, we could only calculate inventory costs. Thus, we divided the inventory cost by 0.68 (because roughly 68% of a pharmacy's total cost is inventory cost) to get a better idea of the total costs. We had to do this as comparing pharmacies is impossible without knowing its total cost. For instance, a pharmacy might have very high profits (without taking overhead cost into account) but once that is taken into account it might not prove to be as profitable). [1]

1. **Buying an existing pharmacy.**

Store number ‘4641’ (CVS - Albany - 4641) is a shop in **14618** that is not necessarily “struggling”. It’s in the 61st percentile of profits among all shops, with a total profit of roughly 73348 USD over the 40 months we analyzed. However, it is “struggling” compared to store ‘2633’ which finds itself in the 100th percentile (i.e. it's the most profitable store in the dataset).

**What changes should be made to the existing pharmacy?**

The shop has a rather limited product selection. It has only 159 unique products (Compared to shop number 2633 for example which has over 2000 unique products). Therefore the most obvious change would be expanding the product selection. Unfortunately, every invoice in the dataset was unique and that didn’t allow us to see which products were sold together. If we had that information we could have recommended adding discounts to either one of those products or selling them in bundles with a discount[. (Appendix, Product recommendations for shop number 4641)](#ProductRec4641)

We also think the pharmacy could consider advertising more or having more discounts during specific months as sometimes sales are off by up to 70% depending on the month. [(Appendix, Sales of 4641 by month)](#MonthSales)

We also recommend changing vendors to a list of the most profitable ones. [(Appendix, Vendor recommendations 4641)](#VendorRec4641)

**Cost of buying the pharmacy.**

The cost of a pharmacy is dependent on multiple factors. Its location, its size, its profitability, current market conditions, etc. Therefore, without having all that information we had to make an educated guess about how much the pharmacy would cost.

We estimate that it would cost around 750,000 USD to buy a pharmacy in **14618**. (We estimated this averaging ‘Pharmacy for sale’ postings in Upstate New York, on [www.bizquest.com](http://www.bizquest.com)) [2]

1. **Building a pharmacy.**

Predicting as to whether BAC should build a pharmacy as opposed to buy one is an even more difficult task, as we have no idea as to how to big the pharmacy should be, how many staff it will have, or how much of the margins to the shareholders plan to keep, or how long as they willing to incur losses (if it means it pays off in the future), etc. Therefore, we would have to make some heroic assumptions to answer this question.

To keep the comparison congruent, we will be using a number that we extracted out from the data set as much as possible, unless necessary to use a value from a secondary source.

If BAC is willing to spend the same amount of money if they were to build or buy a pharmacy, then the deciding factor would be which option pans out to be more profitable in the long run.

**Our Recommendations**

According to our analysis, we think that it's best to build a pharmacy in the zipcode **14618** because the average profit per day shop number 4641 makes is significantly lower than the mean when compared to other shops. This means there is probably something about the shop that makes it less profitable even though the other shop in the zipcode is the most profitable in our dataset and that it is in a zipcode that has the strongest predictors for success. Therefore, we think that if BAC builds a new pharmacy in **14618** and can avoid the pitfalls that make shop number 4641 comparatively less profitable, it might have the potential to be a much more successful business. [(Appendix, Decision comparison plot)](#Decision)

For the product stack of the built pharmacy, we recommend selecting the top 500 products from the dataset. (Appendix, Top 500 profitable products).

As for the vendors we recommend choosing 64 vendors from the 161 available, as the profitability of the vendors follows a Pareto distribution where 20% of the vendors are responsible for 80% of the profits. 20% of 161 would be roughly 32, we recommend it chooses the top 64 vendors as that would cover almost 100% of the profit. [(Appendix, Vendor recommendations buying a pharmacy)](#VendorRecBuy)­­.

**References**

**[1] Walthall, R., 2021. *What Does It Really Cost to Start a Pharmacy?*. [online]** Blog.shelving design systems.com. Available at: <https://blog.shelvingdesignsystems.com/what-does-it-really-cost-to-start-a-pharmacy> [Accessed 26 May 2021].

**[2] Sale, B., Sale, N., Sale, N. and Sale, N., 2021. *New York Pharmacies for Sale | Buy New York Pharmacies at BizQuest*. [online]** Bizquest.com. Available at: <https://www.bizquest.com/pharmacies-for-sale-in-new-york/?q=aT0xMyZvPTQ=> [Accessed 26 May 2021].

**Appendix**

Since we used pandas to do the data analysis, the syntax referring to a table is df (short for data frame).

Pharmacies in ROC is df\_info and df\_info\_final (final suffix implies the version that we used for the analysis)

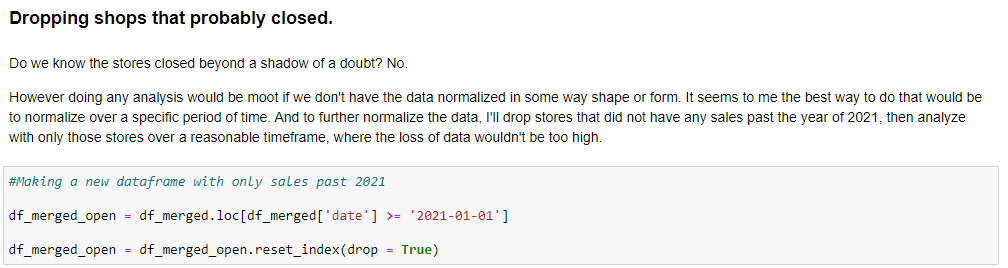
Pharmacy sales is df\_sales.

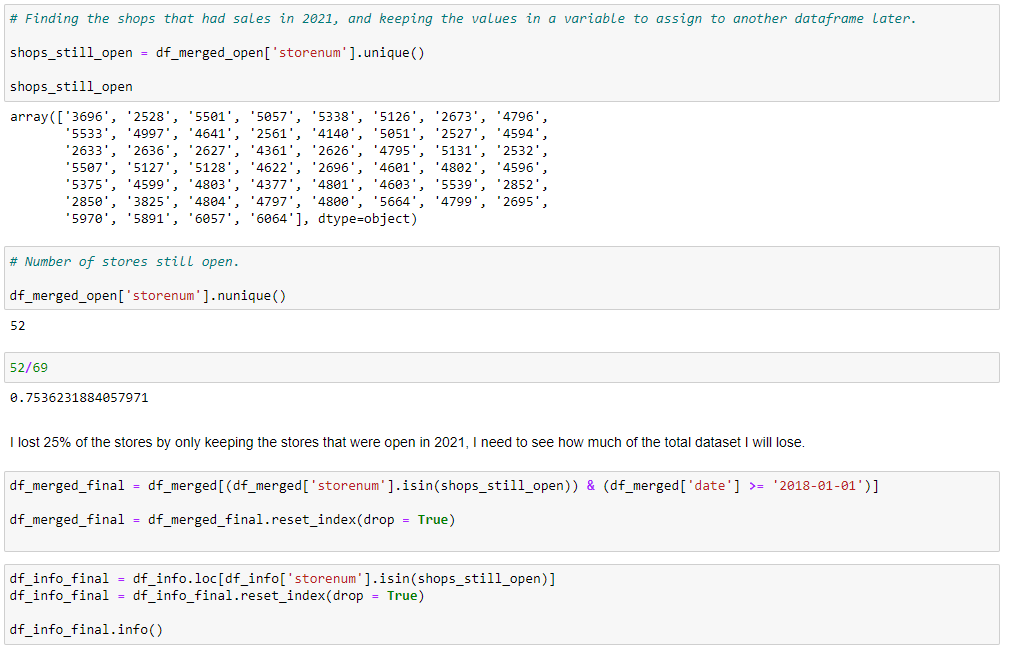
Merged sales and pharmacies are df\_merged and df\_merged\_final.

**Data cleaning**

****

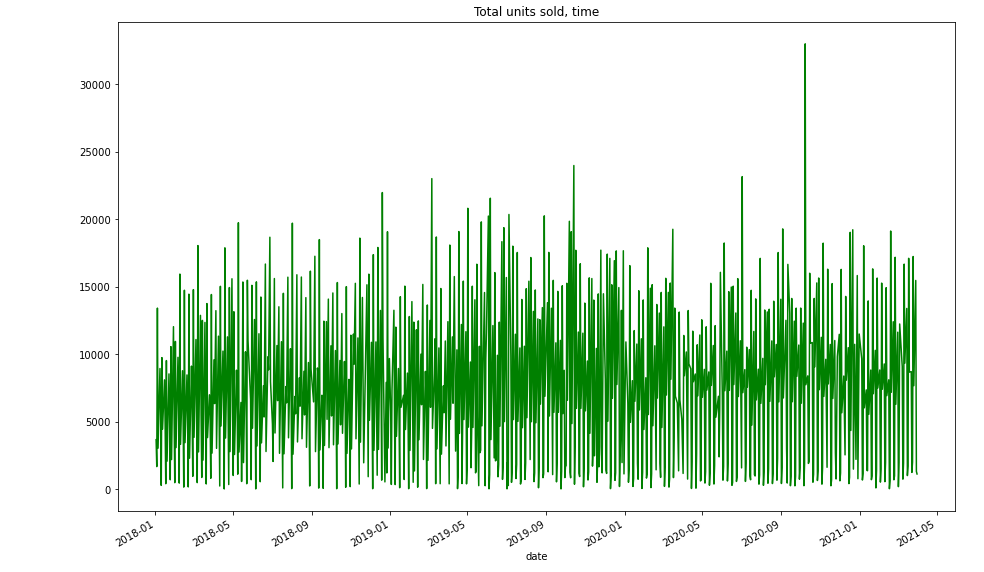
****

****



**Units sold per year**

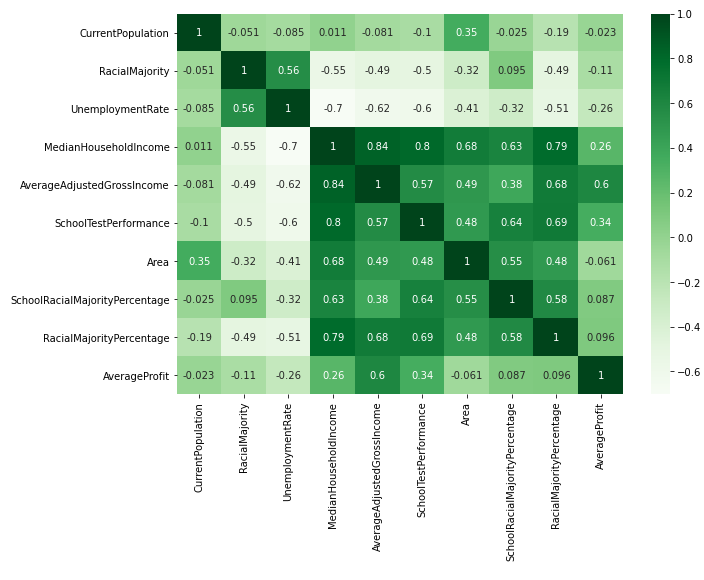
Even though there was a spike in the number of sales in late 2020, it could have been people buying things for the Covid pandemic (e.g. masks), However that doesn’t make all that much sense given the pandemic-related mandates started well before October of 2020. Nonetheless, it tells us that, post-2020 buying trends were erratic (total sales/year were the same), and should be accounted for when analyzing recommendations for the future.

****

****

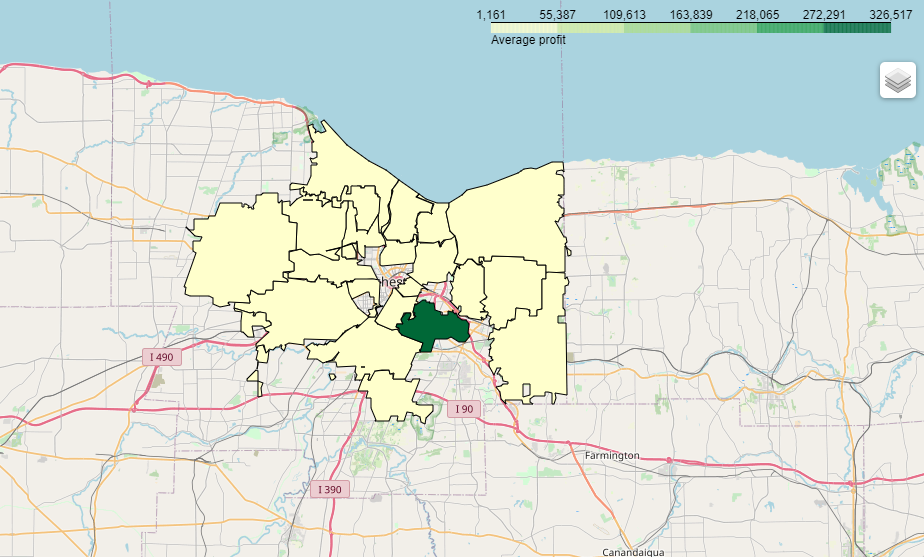
**Correlation matrix**

Correlation between average profit in a zipcode and demographic factors.

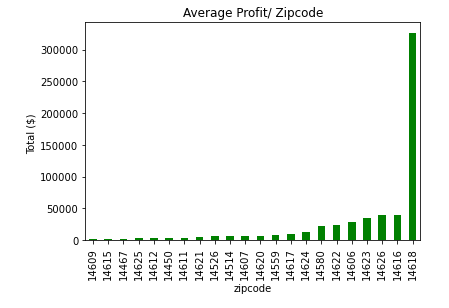
****

**Choropleth map**

The choropleth map shows profit density per zipcode.

****

The bar chart below shows the same information as the choropleth map, which is the Average Profit per zipcode.

****

**Top 500 profitable products**

Found by grouping df\_merged\_final by products, then sorting by profit and choosing the last 500 values.

setA = {34819, 34820, 69637, 34821, 36873, 36874, 36875, 36878, 77842, 77843, 43026, 43028, 36886, 36887, 69657, 77852, 43036, 43038, 49185, 49186, 36900, 49189, 36903, 36904, 43050, 43051, 36908, 88116, 34871, 100413, 34881, 100423, 26710, 43095, 86112, 36965, 36967, 36968, 36969, 36970, 36971, 36974, 36976, 43121, 43120, 36978, 43124, 41077, 43126, 43127, 12408, 43128, 88186, 36989, 43136, 43137, 77956, 34972, 34995, 45248, 28866, 28867, 26820, 26821, 26823, 26824, 26826, 84171, 84172, 26828, 26827, 45276, 45277, 45278, 82147, 88294, 88296, 88298, 86251, 43244, 22784, 89196, 22786, 4356, 22788, 89197, 266, 43285, 16676, 10548, 43316, 10550, 69946, 69947, 43328, 43331, 43333, 43334, 43336, 43337, 43338, 74086, 45419, 43387, 10627, 10628, 37258, 27025, 27026, 68022, 29119, 68036, 68037, 989646, 16850, 989653, 37336, 88536, 37338, 27102, 37346, 37347, 88548, 37348, 88556, 35314, 35316, 27125, 35318, 35317, 918010, 4626, 35354, 10784, 10791, 10792, 37418, 10802, 10803, 10804, 10805, 10807, 10808, 12856, 10809, 39492, 80456, 35416, 35418, 29287, 19061, 19063, 19064, 19066, 19067, 8828, 19068, 41604, 64136, 86670, 86672, 64147, 86691, 86692, 86693, 47786, 82607, 33467, 76478, 17087, 17086, 82627, 76487, 82636, 82637, 41692, 41693, 41694, 17127, 41705, 31475, 68340, 4866, 4867, 86796, 27410, 19226, 37663, 37665, 27433, 17206, 72504, 27453, 45886, 101187, 82787, 86884, 86886, 86887, 41846, 33658, 33663, 5006, 15248, 27544, 82846, 82847, 64418, 31658, 5036, 5037, 82867, 27605, 48099, 48102, 31719, 48105, 48106, 48107, 48108, 27629, 39918, 87026, 48122, 37886, 64513, 25604, 25606, 25607, 25608, 19476, 19477, 11290, 19486, 11294, 27680, 89121, 11298, 11299, 48164, 11296, 11297, 74801, 37938, 64573, 89154, 89164, 66636, 35918, 89175, 64601, 52314, 21596, 52316, 52318, 21598, 21597, 89191, 37991, 37993, 89193, 35947, 89194, 37994, 87150, 89199, 89200, 37998, 89198, 37997, 37996, 40053, 38006, 89207, 38008, 64636, 89215, 27780, 40071, 27783, 33937, 89242, 64676, 38056, 38058, 44217, 38088, 64715, 5326, 5327, 5329, 34001, 34003, 34004, 34006, 34007, 34008, 15582, 34014, 64736, 5346, 5347, 81124, 34029, 34030, 87280, 64752, 34036, 40186, 89339, 40193, 64776, 15626, 15627, 15628, 23828, 66836, 15644, 38174, 38176, 38177, 38178, 38179, 38180, 89387, 64816, 15667, 5430, 81208, 11586, 11588, 34116, 5446, 34117, 87380, 64858, 73053, 64863, 73055, 64865, 64864, 64867, 64866, 89445, 17766, 89447, 64868, 64870, 5486, 87408, 52593, 916850, 34162, 52594, 87409, 52596, 66936, 44419, 40327, 64904, 28043, 64914, 75165, 46504, 75181, 75183, 28086, 28087, 87485, 67006, 75210, 36301, 36304, 36305, 36306, 36307, 36308, 87510, 87511, 58838, 64996, 44516, 32233, 32234, 89577, 32236, 32238, 58868, 65013, 58872, 11771, 58875, 17916, 11773, 11774, 11776, 11777, 5635, 89610, 11786, 11788, 85526, 89626, 17956, 87589, 17958, 89646, 34359, 34366, 5696, 34368, 87619, 15940, 28233, 28236, 18006, 87643, 24157, 24158, 36447, 87646, 87647, 42599, 65127, 15982, 30318, 5744, 44658, 34422, 34423, 28279, 34433, 22156, 22157, 34449, 40593, 40594, 40599, 34456, 34457, 34458, 59037, 77487, 65200, 89786, 89787, 22207, 67267, 89796, 22215, 42703, 22228, 57051, 42716, 42717, 42718, 65254, 65256, 65257, 65258, 89836, 89837, 1799, 36618, 59154, 34579, 34578, 20248, 10008, 57125, 87849, 57148, 77632, 14192, 77683, 87937, 77699, 77709, 20366, 77714, 65426, 18348, 77741, 77740, 18352, 34746, 34747, 49086, 34753, 67526, 67527, 77776, 28625, 88018, 28627, 26586, 53213, 53214, 53216, 34787}

**Shop ‘4164’ unique products**

Unique products in shop number ‘4164’

setB = {11776, 11777, 25604, 34820, 25606, 36873, 36875, 11788, 36878, 43025, 43026, 43028, 19476, 11290, 11294, 11296, 36904, 43050, 43051, 10802, 10803, 10805, 10807, 39492, 77903, 36965, 89191, 36968, 36969, 37991, 36971, 37996, 37994, 37998, 89194, 89196, 43121, 89199, 28275, 12404, 19061, 43126, 34423, 19064, 28279, 12408, 19066, 43124, 38008, 43125, 43136, 34433, 27780, 77956, 40070, 27783, 40592, 40593, 40594, 37012, 86691, 86692, 38056, 100015, 39419, 17086, 26820, 26821, 26824, 26826, 84172, 34001, 34004, 34006, 42716, 41693, 45276, 72928, 65251, 88294, 65254, 65256, 88296, 26858, 31474, 31475, 78078, 78079, 40192, 43777, 1799, 15626, 46350, 43285, 38174, 38176, 38178, 38180, 38194, 81204, 10550, 81206, 17206, 43331, 43334, 43336, 43338, 88406, 64858, 73055, 64864, 64863, 64866, 52579, 86884, 86885, 86886, 64870, 64865, 87403, 87408, 52594, 52596, 77714, 27544, 68009, 77740, 75183, 28081, 86456, 75208, 36301, 36304, 36306, 36308, 68052, 58838, 66518, 26585, 37338, 53214, 27102, 48101, 32231, 32232, 48105, 32234, 32235, 32236, 39917, 39918, 27629, 87026, 58868, 65013, 27125, 58872, 11771, 11774}

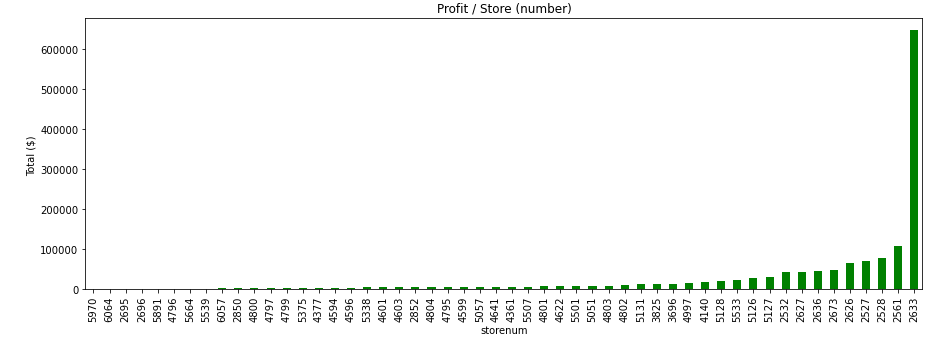
**Product recommendations for shop number ‘4164’**

Whichever products appeared in the top 500 profitable list but aren't in shop number 4164’s product stack.

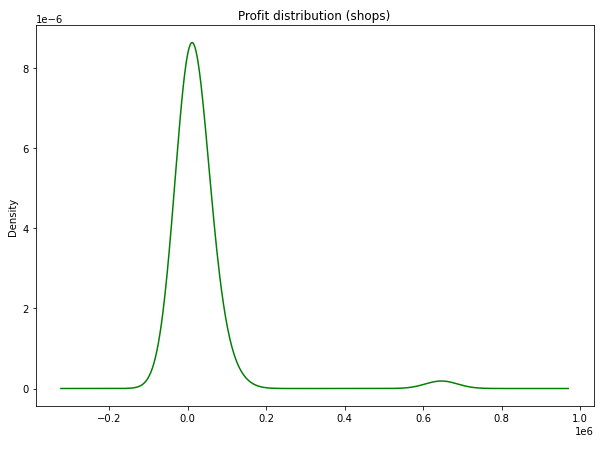
set A - setB = {34819, 34821, 69637, 36874, 77842, 77843, 36886, 36887, 69657, 43036, 77852, 43038, 49185, 49186, 36900, 49189, 36903, 36908, 88116, 34871, 100413, 34881, 100423, 26710, 43095, 86112, 36967, 36970, 36974, 36976, 43120, 36978, 41077, 43127, 43128, 88186, 36989, 43137, 34972, 34995, 45248, 28866, 28867, 26823, 84171, 26828, 26827, 45277, 45278, 82147, 88298, 86251, 43244, 22784, 22786, 4356, 22788, 266, 16676, 10548, 43316, 69946, 69947, 43328, 43333, 43337, 74086, 45419, 43387, 10627, 10628, 37258, 27025, 27026, 68022, 29119, 68036, 68037, 989646, 16850, 989653, 37336, 88536, 37346, 37347, 88548, 37348, 88556, 35314, 35316, 35317, 35318, 918010, 4626, 35354, 10784, 10791, 10792, 37418, 10804, 10808, 12856, 10809, 80456, 35416, 35418, 29287, 19063, 19067, 8828, 19068, 41604, 64136, 86670, 86672, 64147, 86693, 47786, 82607, 33467, 76478, 17087, 82627, 76487, 82636, 82637, 41692, 41694, 17127, 41705, 68340, 4866, 4867, 86796, 27410, 19226, 37663, 37665, 27433, 72504, 27453, 45886, 101187, 82787, 86887, 41846, 33658, 33663, 5006, 15248, 82846, 82847, 64418, 31658, 5036, 5037, 82867, 27605, 48099, 48102, 31719, 48106, 48107, 48108, 48122, 37886, 64513, 25607, 25608, 19477, 19486, 27680, 11297, 89121, 11299, 11298, 48164, 74801, 37938, 64573, 89154, 89164, 66636, 35918, 89175, 64601, 52314, 21596, 52316, 52318, 21598, 21597, 37993, 89193, 35947, 89197, 87150, 89198, 89200, 37997, 40053, 38006, 89207, 64636, 89215, 40071, 33937, 89242, 64676, 38058, 44217, 38088, 64715, 5326, 5327, 5329, 34003, 34007, 34008, 15582, 34014, 64736, 5346, 5347, 81124, 34029, 34030, 87280, 64752, 34036, 40186, 89339, 40193, 64776, 15627, 15628, 23828, 66836, 15644, 38177, 38179, 89387, 64816, 15667, 5430, 81208, 11586, 11588, 34116, 5446, 34117, 87380, 73053, 64867, 64868, 89445, 17766, 89447, 5486, 52593, 916850, 34162, 87409, 66936, 44419, 40327, 64904, 28043, 64914, 75165, 46504, 75181, 28086, 28087, 87485, 67006, 75210, 36305, 36307, 87510, 87511, 44516, 64996, 32233, 89577, 32238, 58875, 17916, 11773, 5635, 89610, 11786, 85526, 89626, 17956, 87589, 17958, 89646, 34359, 34366, 5696, 34368, 87619, 15940, 28233, 28236, 18006, 87643, 24157, 24158, 36447, 87646, 87647, 42599, 65127, 15982, 30318, 5744, 44658, 34422, 22156, 22157, 34449, 40599, 34456, 34457, 34458, 59037, 77487, 65200, 89786, 89787, 22207, 67267, 89796, 22215, 42703, 22228, 57051, 42717, 42718, 65257, 65258, 89836, 89837, 36618, 59154, 34579, 34578, 20248, 10008, 57125, 87849, 57148, 77632, 14192, 77683, 87937, 77699, 77709, 20366, 65426, 18348, 77741, 18352, 34746, 34747, 49086, 34753, 67526, 67527, 77776, 28625, 88018, 28627, 26586, 53213, 53216, 34787}

**Most profitable shops**

**Bar chart of profits by store**

****

**Distribution of profits**

****

**Vendor recommendations 4641**

{259.0, 260.0, 521.0, 266.0, 267.0, 277.0, 285.0, 287.0, 35.0, 297.0, 300.0, 557.0, 301.0, 305.0, 306.0, 308.0, 564.0, 566.0, 55.0, 65.0, 322.0, 330.0, 85.0, 342.0, 346.0, 91.0, 357.0, 619.0, 368.0, 626.0, 115.0, 116.0, 370.0, 380.0, 125.0, 384.0, 130.0, 389.0, 391.0, 395.0, 402.0, 154.0, 410.0, 420.0, 421.0, 434.0, 184.0, 192.0, 451.0, 195.0, 461.0, 205.0, 978.0, 469.0, 214.0, 216.0, 479.0, 482.0, 229.0, 492.0, 239.0, 240.0, 497.0, 255.0}

**Vendor recommendations buying a pharmacy**

{479., 357., 342., 216., 497., 285., 308., 482., 564., 239., 557.,

287., 267., 451., 391., 978., 116., 91., 402., 384., 306., 214.,

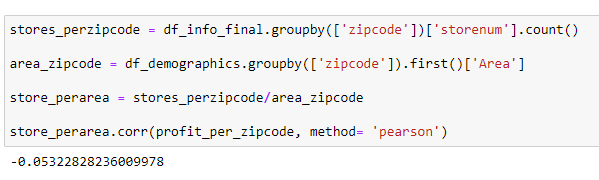
469., 346., 521., 130., 566., 184., 368., 492., 195., 154., 266.,

297., 322., 277., 125., 619., 461., 330., 305., 626., 229., 389.,

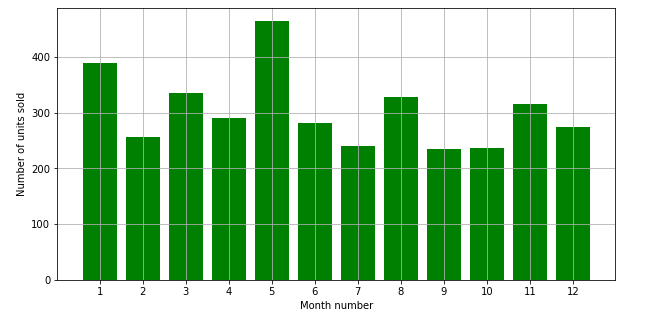
240., 255., 192., 300., 410., 380., 205., 259., 55., 115., 420.,

395., 35., 85., 434., 301., 65., 370., 421., 260.}

**Zipcode profitability correlation with store density**

****

**Appendix, Sales of 4641 by month**

****

**Decision comparison plot**

