Mohammed Khan
Due 8/28/2020
DSA 5010: Foundations of Data Management
Final Project
Dr. Peter Salemi

EXECUTIVE SUMMARY

Overview

The ABCPharmacy business model contains reports for customer and franchise metrics. Various entities will be used to provide data and context as a data model to help increase performance and profit, while decreasing cost and labor. These insights will be presented in conjunction with recommendations of how to fulfill an increase in business efficiency.

Business Goals

This analysis will focus on three general aspects. The first aspect is to identify products that are most beneficial for ABCPharmacy, and also those that do not entice the customers. The second aspect is to provide location and time sensitive information about the customers, because valuable insight exists in predictive behaviors which ABCPharmacy can use to stay proactive with customer demand. The last aspect is for benefit on the store level, having to do with employees and store layout.

Data Description

ABCPharmacy is data of relative high quality and fitness for this study. It provides the requisite data and relationships to be relevant to how ABCPharmacy functions. The drawback of using this data is how outdated it is, with the data focusing on the first half of 2016. Otherwise, the lack of missing values and low instances of outlier data deem this dataset easy to work with, requiring minimal clean up.

Methodological Summary

RStudio and R are used in combination for this analysis. The package used for filtering and grouping the data is dplyr. There is a metadata provided with the dataset, providing context and a complete set of definitions and notes which aided in forming relationships between entities.

Results

Example output of the first query:

```
> #-------------------------------------------------------------------------------------------------------------------------------
> # 1. Which products were most purchased in bundles of 2, 3, and 4 in the same cart?
> #-------------------------------------------------------------------------------------------------------------------------------

# A tibble: 50 x 2
  PROD_NBR          n
  <chr>         <int>
 1 90400000000   4398
```

```
 2 1820025008              3036
 3 90800000000             2855
 4 90300000000             2088
 5 90600000000             1869
 6 91100000000             1692
 7 92000000000             1298
 8 98650000000000273630700  1095
 9 99100000770330000000003   982
10 90900000000              971
# ... with 40 more rows
```

Recommendations

The analysis concluded that there is room for improvement in store efficiency and customer satisfaction with the ABCPharmacy business model. However, without knowing the current state of the business or the financial flow of data, it is impossible to tell if these recommendations are already being implemented.

CORE REPORT

Overview

For the ABC Pharmacy business model, this analysis will be focusing on geographic, store and product metrics. Having context through all three of these metrics will help give us the most clear understanding in which areas are in need of enhanced performance and how we can implement those recommendations. The information given here consists of data from all three of those focus areas, however, most of the entities serve to define and categorize products. PHRMCY_MASTER and POS_TRANS provide data on the geographical and store metrics, respectively. The POS_TRANS table was especially useful in formulating questions to assist in analysis because it contains two foreign keys as part of its primary key. This entity also contains data on pricing and sales, which gives us information to improve the bottom line.

Business Goals

This analysis should be able to give us an idea of which products are improving performance and which products could be done away with (7-10). There is also an analysis done on which products might be most lucrative when assigned a ‘buy x, get 1 free’ promotion (1), and there is also a query done to assign another attribute to store price values for single products derived from extended sales (2). There is a question for the geographic metrics (3), which might save costs and help increase net profit based on location. The rest (4-6) provide value in the store setting, ranging from layout improvements to employee coverage predictions and potentially rewarding stores that are among the best assets in the company. These recommendations would serve to streamline a model that is already lucrative, making way to improve efficiency and potentially grow in other aspects of the company.

Business Questions

Questions:

1. Which products were most purchased in bundles or 2, 3, and 4 in the same cart?
2. What is the individual sales amount for each distinct product in ascending order?
3. Which zip codes contain the most pharmacies?
4. Count the number of category codes.
5. Which dates resulted in the most basket sales?
6. Which pharmacies, on average, sell the most baskets per day?
7. Which products sold the most in terms of cumulative sales?

8. Which products sold the least in terms of cumulative sales?
9. Which products sold the least number of units?
10. Which products sold the most number of units?

Data Description

Data Quality

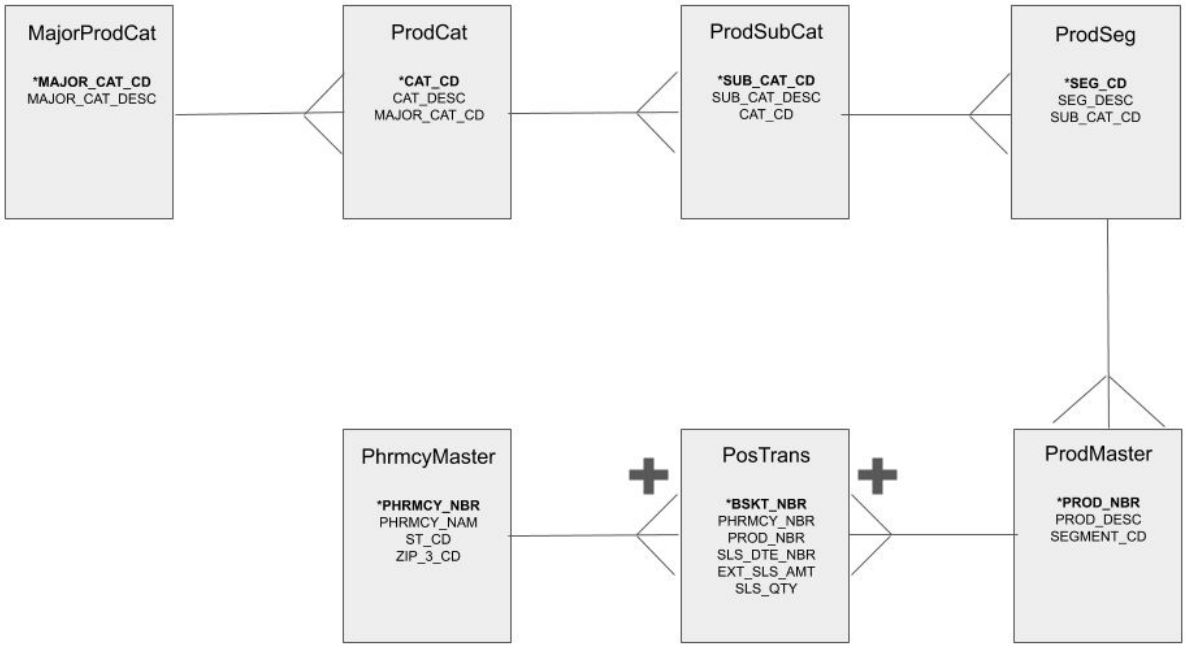
We can define data quality as “keeping data accurate, complete and current.” There are many measures that can be used to vet and confirm the quality of a particular set of data. We can choose to focus on completeness, timeliness, and fitness as measures of data quality.

The ABCPharmacy data happens to run very low on missing values and misplaced values. It consistently and meticulously has the data at the correct place in the correct form and type. Manipulation and readability of the data is easily accessible. However, there are issues present when sorting by the extremes of the dataset. There appear negative and infinite values when the dataset is imported into an IDE, which might speak to clerical errors when the dataset was being constructed. Otherwise, in these instances it would be highly impractical to observe a negative value (such as a negative product number or a negative sales price). Once those outliers are dealt with properly and removed from the imported dataset, the data model becomes highly workable. In the context of the data, it would be justifiable to outright remove the negative valued data since it makes no sense within the bounds of the data being measured.

The timeliness of the data is lacking, as the data is a record of values from the former half of the year 2016. There are many potential issues with this, such as the discontinuation of products, closing of pharmacies, and even a rehaul of the inventory system used in ABCPharmacy. It is impossible to glean any updates from this data that might be active in the year 2020, so we have no choice but to work with and make conclusions off of outdated information. The data fails in this quality as a four year difference is significant in the realm of running a franchise.

In terms of fitness, the data is clearly labelled and the metadata provided gives a complete picture of what each table, column and row are trying to measure. As the questions in this report were formulated using the metadata and data model itself, the fitness is likely very high. There are no assumptions made in the questions and all of the queries are taken directly from the imported data. The data was formatted and presented in a way that made the queries simple to write and extract information from. This is an encouraging sign of fitness.

Data Model



Methodological Summary

This study is based on the data provided within the ABCPharmacy data model. The metadata provided is clearly divided into the separate entities within the data model, along with the data definition and notes for each attribute. Studying the metadata helps identify the relationships between the separate entities, as shown in the data model above. The metadata also offers the full word alternatives of the abbreviations used in the naming of the entities and attributes, which are helpful in communicating the data and drawing accurate conclusions.

RStudio is the IDE selected for this study, as it provides ample support and functionality to perform the queries and make notes within the code itself. The use of dplyr is necessary for this analysis, along with the tibble library for more versatile interaction of the datasets. The readr library is utilized to read in the .csv files containing the datasets within the local work directory. Readr also allows us to assign column types in a simple manner. Any issues within the importing of the data model was also taken into account, to make sure the entities and the data were brought into the IDE successfully and completely. One of the queries required a conversion of the date from a character type to a datetime type, so this was done next, along with formatting the date for easier readability. The queries were done with dplyr, using piping to filter and group rows and columns until the query is appropriately answered. Comments are added before each block of code, providing context for which code was chosen and why. The console served as the output display for the R script, and the global environment served as easy access to the datasets and finished queries.

Results

```
> #-------------------------------------------------------------------------------------------------------------------------------------------
> # 1. Which products were most purchased in bundles of 2, 3, and 4 in the same cart?
> #-------------------------------------------------------------------------------------------------------------------------------------------

# A tibble: 50 x 2
   PROD_NBR                    n
   <chr>                   <int>
 1 90400000000              4398
 2 1820025008               3036
 3 90800000000              2855
 4 90300000000              2088
 5 90600000000              1869
 6 91100000000              1692
 7 92000000000              1298
 8 98650000000000273630700  1095
 9 99100000770330000000003   982
10 90900000000               971
# ... with 40 more rows


> #-------------------------------------------------------------------------------------------------------------------------------------------
> # 2. What is the individual sales amount for each distinct product in ascending order?
> #-------------------------------------------------------------------------------------------------------------------------------------------

# A tibble: 178,583 x 2
   PROD_NBR                 IDV_SLS_AMT
   <chr>                          <dbl>
 1 8770140726                      0.01
 2 66597306412                     0.01
 3 8770140818                      0.01
 4 98650000000000273630700         0.01
 5 8770190003                      0.01
 6 8770189806                      0.01
 7 8770154970                      0.01
 8 8770198111                      0.01
 9 8770140510                      0.01
10 8770140766                      0.01
# ... with 178,573 more rows


> #-------------------------------------------------------------------------------------------------------------------------------------------
> # 3. Which zip codes contain the most pharmacies?
> #-------------------------------------------------------------------------------------------------------------------------------------------

# A tibble: 10 x 2
   ZIP_3_CD     n
   <chr>    <int>
 1 112         91
 2 070         57
 3 100         57
 4 113         48
 5 191         45
 6 104         44
 7 088         41
```

```
 8 009         40
 9 007         29
10 006         28


> #-----------------------------------------------------------------------------------------------------------------------------------------------
> # 4. Count the number of category codes.
> #-----------------------------------------------------------------------------------------------------------------------------------------------

# A tibble: 1 x 1
  `sum(n)`
     <int>
1       62


> #-----------------------------------------------------------------------------------------------------------------------------------------------
> # 5. Which dates resulted in the most basket sales?
> #-----------------------------------------------------------------------------------------------------------------------------------------------

# A tibble: 10 x 2
   SLS_DTE_NBR             n
   <dttm>              <int>
 1 2016-02-12 00:00:00  4725
 2 2016-06-17 00:00:00  4670
 3 2016-03-03 00:00:00  4610
 4 2016-06-03 00:00:00  4564
 5 2016-03-07 00:00:00  4505
 6 2016-05-06 00:00:00  4467
 7 2016-05-07 00:00:00  4441
 8 2016-06-16 00:00:00  4400
 9 2016-03-11 00:00:00  4397
10 2016-05-31 00:00:00  4394


> #-----------------------------------------------------------------------------------------------------------------------------------------------
> # 6. Which pharmacies, on average, sell the most baskets per day?
> #-----------------------------------------------------------------------------------------------------------------------------------------------

# A tibble: 10 x 2
   PHRMCY_NBR          perDayBSKT
   <chr>                    <dbl>
 1 1174450154022548624       350.
 2 6991356705459241502       265.
 3 2759406693434064370       224.
 4 61520549788616420         209.
 5 8506230257184703229       207.
 6 3009693108150153253       194.
 7 4416100399456673861       188.
 8 9201518331233084207       177.
 9 3241946627864485090       166.
10 3216540913770909647       150.
```

```
> #-----------------------------------------------------------------------------------------------------------------------------------------------------
> # 7. Which products sold the most in terms of cumulative sales?
> #-----------------------------------------------------------------------------------------------------------------------------------------------------

# A tibble: 20 x 2
   PROD_NBR              totalEXT_SLS_AMT
   <chr>                            <dbl>
 1 99100000770330000000003         783415.
 2 99100000234116000000009         256944.
 3 90400000000                     234125.
 4 91000000000                     221579.
 5 90800000000                     182751.
 6 90300000000                     171306.
 7 90600000000                     136317.
 8 98650000000000944904150         135200.
 9 99100000066590000000006          81239.
10 98650000000000185441572          59452.
11 92000000000                      53405.
12 98650000000000273630700          49044.
13 91100000000                      43702.
14 90900000000                      36012.
15 2610000573                       34250.
16 2820000357                       33979.
17 2820000384                       28093.
18 35310020010                      22971.
19 91800000000                      22578.
20 81460502007                      20680


> #-----------------------------------------------------------------------------------------------------------------------------------------------------
> # 8. Which products sold the least in terms of cumulative sales?
> #-----------------------------------------------------------------------------------------------------------------------------------------------------

# A tibble: 169 x 2
   PROD_NBR    totalEXT_SLS_AMT
   <chr>                  <dbl>
 1 1542102427                 0
 2 1542197883                 0
 3 1542197893                 0
 4 18436900005                0
 5 2318514515                 0
 6 2360101336                 0
 7 2360112030                 0
 8 2529487400                 0
 9 2571523285                 0
10 2571566711                 0
# ... with 159 more rows
>
```

```
> #-----------------------------------------------------------------------------------------------------------------------------------------
> # 9. Which products sold the least number of units?
> #-----------------------------------------------------------------------------------------------------------------------------------------

# A tibble: 118 x 2
   PROD_NBR    totalSLS_QTY
   <chr>              <int>
 1 1542102427             0
 2 1542197883             0
 3 1542197893             0
 4 2318514515             0
 5 2360112030             0
 6 2571523285             0
 7 2571566711             0
 8 2571566999             0
 9 2930000002             0
10 30005556719            0
# ... with 108 more rows


> #-----------------------------------------------------------------------------------------------------------------------------------------
> # 10. Which products sold the most number of units?
> #-----------------------------------------------------------------------------------------------------------------------------------------

# A tibble: 50 x 3
# Groups:   PROD_NBR [27]
   PROD_NBR                 SLS_QTY totalSLS_QTY
   <chr>                      <int>       <int>
 1 90400000000                    1       16954
 2 99100000066590000000006        1        9956
 3 90900000000                    1        6549
 4 90300000000                    1        6306
 5 90400000000                    2        6004
 6 90600000000                    1        5448
 7 90800000000                    1        4700
 8 1820025008                     2        4166
 9 98650000000000273630700        1        3498
10 90800000000                    2        3408
# ... with 40 more rows
```

Conclusions and Recommendations

Based on the results of the queries which were analyzed in this report, we can take away conclusions regarding products identified by their number. This information would be useful to plug back into the PROD_MASTER table to be able to identify the name of each product that was given as a result for each query. Taking the names of the products and comparing them to each other using their category and subcategory codes will give us a clear picture of how products are seen by the customers. An actionable insight would be to perform an analysis on the products that are best sellers in terms of both revenue and quantity. Finding an overlap of products within these two measures would be instrumental in understanding which products the customers keep coming back to ABCPharmacy for. These insights can also be leveraged to

inspect the layout of the store itself, so ABCPharmacy can make sure to place the best selling products in a location within the store that attracts the most attention, boosting the revenue and stock sold even more. Conversely, finding the worst sellers in the store would aid in keeping that type of stock away from the public eye, in favor of better performing products to encourage more guaranteed sales without the need of advertisement. The worst performing products might be discontinued outright, saving on cost of transportation and stocking these products.

The products ABCPharmacy sells are divided into 62 category codes. It is feasible to do an analysis on which categories are best sellers to be able to place the best performing categories front and center in the store. These products would most likely be necessities that are essential to all customers, regardless of demographic. ABCPharmacy carries an ample stock of general drugstore products, which should create a facile and efficient shopping trip for the customer. This would also cut down on the time employees spend helping customers that have questions about where and when a product will be in stock. The time used in answering customer inquiries can instead be spent in the general maintenance of the store, which would also increase the performance of the employees themselves.

Taking into account the geographical and time sensitive measures of ABCPharmacy stores, we can make use of these metrics to find which dates and locations are the most traffic heavy in customer volume. For example, six out of the top ten dates in basket sales landed in May and June. This might suggest that ABCPharmacy carries products that are especially in demand during these months. This pattern can also be attributed to natural factors, such as weather conditions and season. Holidays are also worth taking into account. Even though cutting off a query at just ten can be arbitrary, there is a potentially significant insight here, considering May and June only make up two of the months recorded in the data, yet they make up six of the top ten busiest days in terms of basket sales. The results of location by zip code can also be actionable information, but more data is needed to make a sound insight such as geography of the locations and their proximity to consumer heavy locales. However, the zip code 112 contains 91 pharmacies, the most by far of any other zip code. This might suggest that this particular pharmacy is located in a very densely populated consumer base. It might also be a signal that the concentration of too many pharmacies in the same general location leads to inefficiencies in customer flow when other demanding locations might need the support of additional stores more.