U.S. Pain Pills Analysis using HIVE and PIG

Authors: Adrian Marroquin, Aleksander Sekowski, Fereshteh Mamaghani, Jinhui Liu, Siying Chen
California State University, Los Angeles
CIS 5200 - SYSTEMS ANALYSIS AND DESIGN
amarro15@calstatela.edu, fmamagh@calstatela.edu, schen112@calstatela.edu, jliu2@calstatela.edu
asekows@calstatela.edu

Abstract: Pain medication prescription sales data can reveal trends and insights into one of the most popular types of medication sold in the United States of America. In this paper we work to analyze a comprehensive data set of all prescription sales of hydrocodone and oxycodone-based medications in all US states between 2006 and 2012. Among others, we reveal which configurations of drugs have the highest market share, which states have the highest and the lowest growth in sales in the time period, which institutions are the biggest buyers of pain medication and which urban centers are responsible for the most sales volume. We use big data tools such as Hive and Pig to create data visualization artifacts such as spatial temporal diagrams, market share pie and column charts and heatmaps with Tableau and MS Excel. The resulting analysis and trends found in the prescription data could interest researchers trying to find a correlation between increase in prescription pain medication and opiates sourced from illegitimate sources.

1. INTRODUCTION

Our group project is based on the data coming from the Automation of Reports and Consolidated Orders System (ARCOS).As an automated system created by the Drug Enforcement Administration (DEA), the Automation of Reports and Consolidated Orders System helps DEA monitor selected controlled substances by capturing current and historical records of selected controlled substance inventories and relative transaction dates. According to the Code of Federal Regulations (CFR) 21 CFR1304, manufacturers and distributors must report their inventories of selected controlled substances to DEA periodically. Any changes to their inventories of these substances must be recorded and submitted through the ARCOS system. In addition, the manufacturing and procurement quotas for Schedules I and II controlled substances are restricted by the United States Controlled Substances Act. No manufacturers or distributors allow exceeding the yearly controlled substances quotas set by the government.

Our dataset relies on the source name "Pain Pills in the USA," created by Paul Mooney. The original date Paul applied comes from the Washington Post report, which we know it supports by the ARCOS system fundamentally.

2. RELATED WORK

The ARCOS dataset became publicly available after a year-long legal battle between the DEA and the Washington Post/HD Media. With the dataset finally released in 2019, many drug companies are now being sued in federal court “by nearly 2,000 cities, towns, and counties alleging that the drug companies have conspired to flood the nation with opioids.”[1] The article also details the major players in the distribution of “Oxycodone” and “Hydrocodone” and the continuous settlements that have occurred since the opioid epidemic.

Drug Enforcement Administration (DEA) publishes ARCOS reporting annually[2]. “ARCOS provides an acquisition/distribution transactional records of ... certain controlled substances ”for each year[2].The reports include “total drug amounts (in grams) distributed to retail registrants in each state, by zip code”[3] and “by quarter”[3]. The reports also reflect the total consumptions of the components per 100K population ,grouped by each state, quarterly and annual total in descending order[3]. The reports also provide the average amount of transactions and number of buyers, grouped by business activities, in each state and within the US[3].

In contrast, we have demonstrated the trend of total purchases across the United States within 6 years in a 3D map. Also Top 10 cities with the highest amount of pain pills’ components distributed to retailers are illustrated by the heat maps. With regards to population, 15 most popular products among people are identified and the top 10 market shares of products are investigated. Focusing on buyers, we demonstrate the distribution of significant buyers and their highest amount of purchase.

3. PROPOSED PROJECT

We divide our project into several steps. Our proposed procedure for this project starts from the data collection by downloading the relative date from ARCOS. The data is under tsv files, which need to be uploaded into the HDFS for further application. Inside the HDFS, we use the Hive and Pig command to create tables and relations to analyze the data. Next step, we will transfer the data into “CSV” or “TSV” format in order to apply them into Excel. Finally, we will provide a detailed analysis to support our project by giving proper diagrams, pie charts, and other relative graphs through excel and PowerBI.

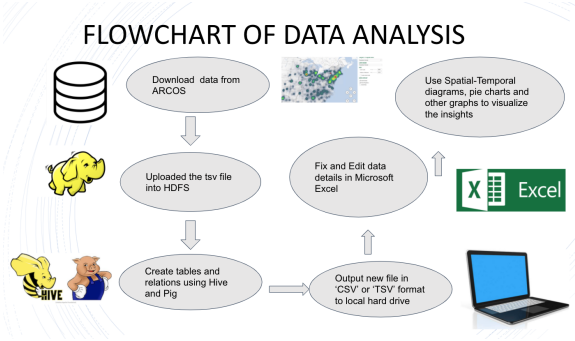

Figure 1. Data Analysis Flowchart

3.1 Query 1

Query 1 is focusing on the market share for pain pills. Because market share is a key indicator to help the company to understand consumers' behavior, and also provide a reference when the company makes its decisions. The first part of Query 1 is talking about the top 15 pain pill products' market share from 2006 to 2012. And will present the result in both the pie chart and bar chart to indicate the percentage/values of each products' market share. The second part of Query 1 is focusing on the top 10 products' market share each year. Instead of using excel to visualize the top 15 products' market share, Power BI will be the tool to visualize the top 10 products' market share and animate the result as a bar chart race by using the "Animated bar chart race" add-ins. In this case, the animation will show clear results to the audience of how the top 10 market share changes each year.

3.2 Query 2

In order to visualize the growth of opioid sales Query 2 provides data to create a spatial temporal diagram of drug sales per state per year between 2006 and 2012. All prescription records will be grouped by main ingredient name (OXYCODONE, HYDROCODONE), location (state) and date (year). The resulting table will be visualized as a 3D map with column charts and a timeline animation playback. The timeline playback will make it easier for the audience to examine where growth in opioid sales has grown at the fastest pace. Breaking up the data by geographical location will let us find out about states that register low volumes of sales of pain medication.

3.3 Query 3

The method used for the third query will be utilizing Hadoop and the PIG platform to analyze the dataset in order to answer "Top 10 Cities with the highest quantity distribution by drug and by state". The code that was written was focused on being reusable in order to filter the massive dataset TSV file by state and also by drug name. By doing so, it allows for smaller data files to be extracted in order to drill down and visualize specific data. The states of California and Virginia have been selected to show each state's distribution of the drug Oxycodone.

3.4 Query 4

The purpose of query 4 is to visualize the distribution of major buyers of Oxycodone, one of the key components of pain pills, and the buyers' significant transactions across the country between 2006 and 2012. For making the transactions relevant in the spatial-temporal animated playback, the dataset is prepared in two phases. In the first phase, the top N buyers are determined based on their total transactions' quantity. In this query, the N is equal to 100. In the next phase, the Top N transactions based on quantity per each buyer are extracted. The N is selected as 2. To conduct the query, the main table is built in Hive and exported to Pig by HCatalog and Joins and Nested Foreach queries were used in Pig to extract Top 2 transactions per each buyer in top 100 buyers. The data type of Date is text and the original format is modified by substring function for compatibility with 3D Map in Excel.

4. Analysis and Visualization

4.1 Query 1

As we can see in figure 2. Query1-1, from 2006 to 2012, Oxycodone HCL/Acetaminophen 5mg/325m have the largest market share with 5,482,988, which is almost 1.5 times greater than the Hydrocodone BIT. 10mg/Acetaminophen (3,726,823). (figure 2. Query 1-1) is a bar chart to present the top 15 total market share from 2006 to 2012.

Figure 2. Query 1-1 Top 15 Bar

And according to figure 3. Query 1-2, Oxycodone HCL/Acetaminophen 5mg/325m also takes 8% of the total mark. The other 14 products take from 2% to 5% of the market. And all the rest products take 53% of the market, which means the top 15 products share 47% of the market. The second chart

(figure 3. Query 1-2) is a pie chart based on the same data as the first chart. This pie chart will point out the percentage for the top 15 product market share.

Figure 3. Query 1-2 Top 15 Pie

Based on the animate bar chart (figure 4. Query 1-3), The Oxycodone HCL/Acetaminophen 5mg/325m always taking first place in the market from 2006 to 2012. And Hydrocodone. BIT./Acet.,10mg & 352mg/ taking over Hydrocodone BIT. 10 mg/Acetaminophen's place as the second large product in the market in 2012. Hence, Oxycodone with 5mg/325m units is the most popular pain pill in the market. And the last chart (figure 3. Query 1-3) is an animate bar chart. However, the word file cannot present a gif file. Therefore, the screenshot was provided for last year's (2012) top 10 market share.

Figure 4. Query 1-3 Top 10 Bar

4.2 Query 2

As seen in Figure 5. Query 2-1 there is a significant growth in sales of opioids in multiple states. The highest absolute growth numbers are in the states of California, the North East and Florida. The map allows us to analyze the difference in popularity of specific ingredients per state. Hydrocodone corresponds to a significant majority of sales volume in states such as California, Texas and Indiana. Oxycodone registers a much faster growth in sales in the state of Washington, New Jersey and North Carolina. Several states register very low or non-existing sales volumes. Among them are the states of Wyoming, Montana, Idaho and Iowa. These states have recorded almost no growth of sales over the examined period. Query 2 is a timeline playback. Below are screenshots of each step-in progression.

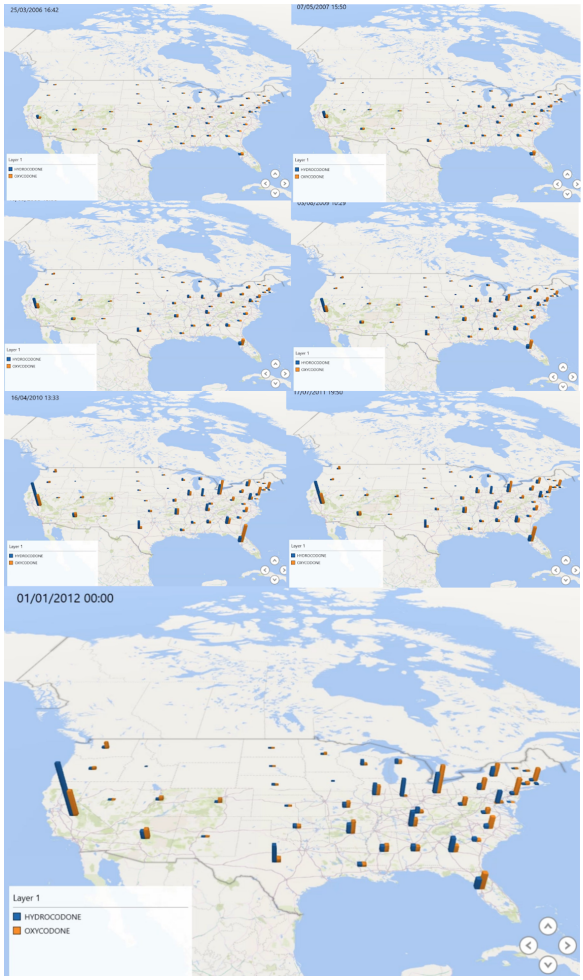

Figure 5. Query 2-1 step-by-step.

4.3 Query 3

The following results show heat maps visualizing which cities in Virginia and California had the highest distribution of Oxycodone. The same Pig Latin code can be reused to create datasets for any state or drug. This can help analysts pinpoint which cities distributed the highest amount of which drug.

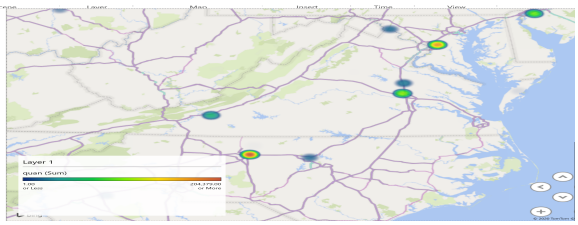

Figure 6. Query 3-1 Virginia Heatmap

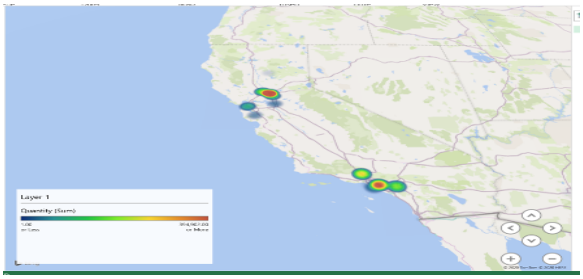

Figure 7. Query 3-2 California Heatmap

4.4 Query 4

The below chart shows Top 100 buyers of Oxycodone and their business activities are either retail or chain pharmacies with only one practitioner, who is at the 43rd row. Also, there are 12% of buyers who made much higher purchases than others in the top 100. The result of the top 100 is the basis for researching the distribution of significant transactions across the US.

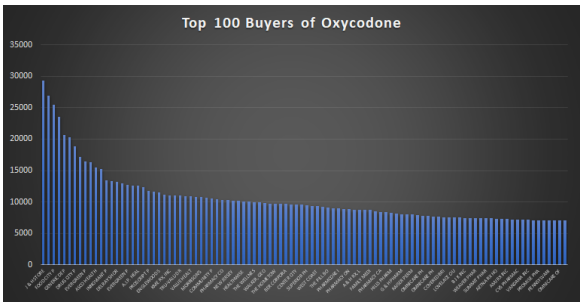

Figure 9. 4-1 Top 100 Buyers Bar chart

The spatial-temporal diagram shows the distribution on Top 100 Buyers and the bar charts show the sum of two highest purchases per buyer per city. Florida encompasses the highest number of large transactions. Oregon is the next, that consists mostly of Kaiser related pharmacies. There is only one buyer in Louisiana, whose business activity is practitioner, and is responsible for part of major transactions. Next spot is North East with an emphasis on Pennsylvania.

Figure 10. 4-2 Spatial-Temporal Diagram of transactions

5. CONCLUSION

Above all, based on our research and queries, the conclusions can be drawn as follows:

I. Oxycodone HCL/Acetaminophen 5mg/325m taking the largest market share.
II. The following states/areas have the highest demand for pain pill products: California, Florida, and the North East of the USA.
III. Bay area and Greater Los Angeles have the most distribution of pain pill products. Roanoke has the largest in Virginia.
IV. The biggest buyer of Oxycodone is located in Florida and most of the distribution of large transactions are located in Florida, on the north east and north west.

However, the dataset we are using is a sub-dataset from the DEA's database. The year of the data is from 2006 to 2012. Therefore, the results and conclusion may be different compared to the whole database. And the same queries can be done with the bigger dataset by using our framework. For more information and details, please refer to the project's GitHub (https://github.com/aleksUIX/CIS5200-team5#readme).

6. REFERENCES

[1] Higham et al., 2019. (2019, July 16). 76 billion opioid pills: Newly released federal data unmasks the epidemic. Washington Post. Retrieved from: https://www.washingtonpost.com/investigations/76-billion-opioid-pills-newly-released-federal-data-unmasks-the-epidemic/2019/07/16/5f29fd62-a73e-11e9-86dd-d7f0e60391e9_story.html?itid=lk_inline_manual_2

[2] ARCOS Retail Drug Summary Reports https://www.deadiversion.usdoj.gov/arcos/retail_drug_summary/

[3] REPORTING PERIOD - 2009 https://www.deadiversion.usdoj.gov/arcos/retail_drug_summary/2009/index.html