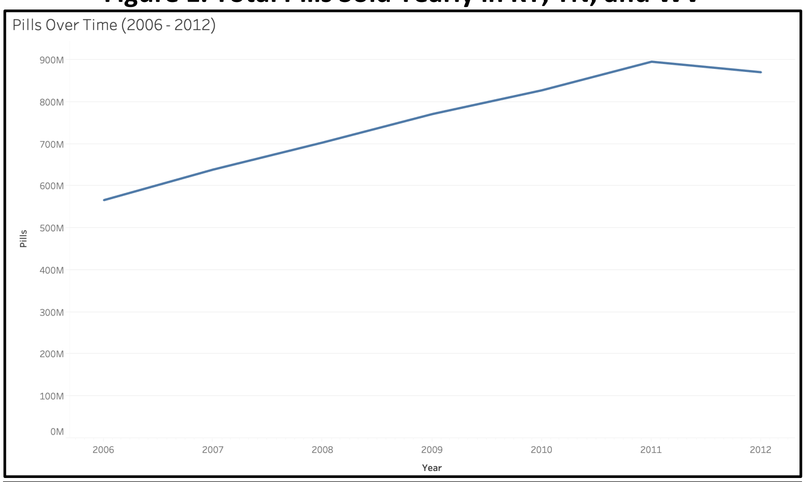
Joey Aramouni || Sofiane Hadji || Girish K Govindarajan || Rihab Rebai

WHO DUNNIT? US OPIOID EPIDEMIC

1) Problem Description

The US suffers from an opioid epidemic, in which there is an extensive overuse of opioid drugs (10.8M people misused opioid prescriptions in 2018[1]). There are large costs of this including loss of lives, health care, criminal justice, and lost economic productivity. Our objective will be to **identify pharmacies that have filled an excessive number of opioid prescriptions in the states of Kentucky, Tennessee, and West Virginia between the years 2006 and 2012**. This will point to pharmacies who may be illegally distributing opioids or neighboring doctors who overprescribe opioids. It is also a proxy for understanding which areas are suffering the most opioid abuse. Thus, we aim to inform senior decision-makers (1) which pharmacies should be investigated for potential malpractice and (2) which counties to focus interventions on.

Figure 1. Total Pills Sold Yearly in KY, TN, and WV



2) Data

The main data source is the Automation of Reports and Consolidated Orders System (ARCOS)[2] which was established by the Drug Enforcement Agency (DEA)[1]. The dataset tracks the path of every opioid pain pill (oxycodone and hydrocodone), from manufacturer to pharmacy, in the United States between 2006 and 2012. Some of the fields available include manufacturer name, manufacturer address, pharmacy name, pharmacy address, pharmacy type (retail vs. chain vs. physician), drug name, drug quantity purchased, and date of purchase. **Appendix 1** contains a sample screenshot.

Secondary sources of data used are US census data such as population, age, income, and population.[2] We also leveraged health indicators such as percentage insured and number of physicians per county.[3]
We also did some feature engineering. In particular, we found the concentration of pharmacies near a pharmacy by summing all pharmacies in a zip code.

[1] Opioid data can be found at the following link: https://www.washingtonpost.com/national/2019/07/18/how-download-use-dea-pain-pills-database/?arc404=true

[2] US census data can be found at the following link: https://factfinder.census.gov/

[3] Health indicators are available at the county level here: www.countyhealthrankings.org

3) Methodology

To narrow the scope, we did a high-level summary on the entire ARCOS dataset and analyzed which states had the highest pills sold per capita. We chose to focus on Kentucky, Tennessee, and West Virginia which are in the top 5 states with the highest pills per capita (**Appendix 2**).

3.1 Visualizations

To get an initial idea for which counties and pharmacies may be problematic, we created a series of visualizations which mapped pills sold per capita for each pharmacy between 2006 – 2012.

$$Pills\ sold\ per\ capita\ = \frac{Total\ Pills\ Sold\ by\ Pharmacy}{7\ \times\ Population\ (zip\ code\ level)\ in\ 2012}$$

3.2 Time Series

Analyzing pills per capita highlights pharmacies who sold the most opioids (relative to neighboring population) during all seven years, but it does not give information about time evolution. By studying the rate of opioid purchases by individual pharmacies, we can highlight worrisome changes and more easily compare pharmacies which do not have the same volume of sales. We benchmark each pharmacy's purchasing behavior against itself. However, this method does have limitations as the ARCOS dataset starts in 2006 which is at the height of the USA opioid epidemic. Thus, if the number of opioids sold by a pharmacy in 2006 is already abnormal, we would not be able to differentiate.

3.3 Clustering

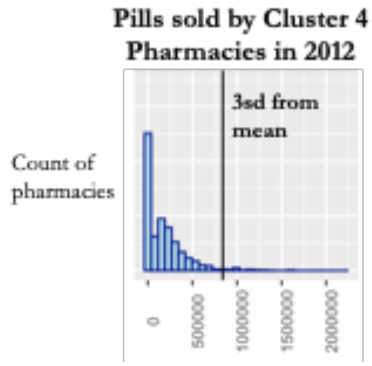
Through clustering, we can find 'peer groups' among pharmacies, i.e. pharmacies who are expected to sell a similar level of drugs. Through qualitative research, we found factors which may influence pharmacy sales. We then transformed the data such that each row represents a unique pharmacy and clustered on the following features: population size, income, concentration of pharmacies, percentage population above 65%, pharmacy type (chain, retail, or physician), number of neighboring doctors, and insurance coverage.

We first standardized our data - one-hot encoding was leveraged to handle categorical variables, and the continuous variables were normalized to have a mean of zero and a standard deviation of one. We then employed hierarchical clustering with the Ward-linkage dissimilarity to segment our pharmacies. This allows for iterative decision making based on the granularity at which we want to characterize peer groups of pharmacies. We settled on seven representative groups based on the Scree plot **(Appendix 3)**.

In order to qualitatively characterize the clusters, we plotted the distribution of each feature across clusters and compared each cluster to the distribution of the feature across all pharmacies **(Appendix 3)**.

3.4 Identifying Anomalies

Within each peer group, we computed the mean and the standard deviation of the total number of pills sold. The goal of the analysis was to get a sense of what constituted an "anomalous" quantity. We flagged pharmacies which sold pills several standard deviations away from the mean. As a result, we were able to identify pharmacies deviating from common sales levels in each cluster. In addition, we can now rank the pharmacies according to their degree of violation.

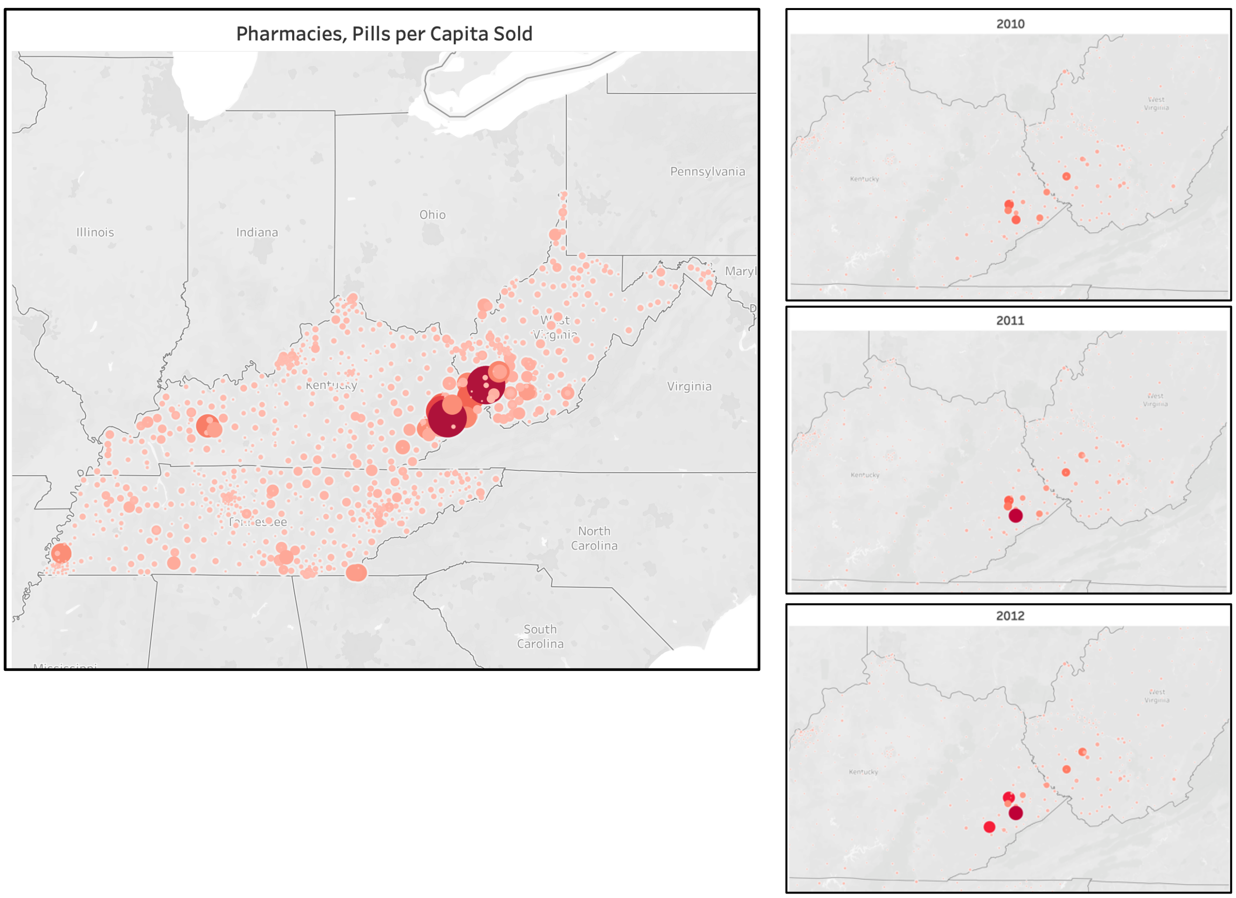


4) Findings

4.1 Visualizations Findings

Through visualization, we found that the highest concentrations of pills per capita sold were near the West Virginia – Kentucky border (Figure 2). Particularly, pharmacies in Floyd, Mingo, Pike, and Logan counties appear suspect. Moreover, more and more pharmacies appear suspect in recent years.

Figure 2. Evolution of pills per capita between 2010 and 2012 in the WV - KY border



4.2 Time Series Findings

There were two interesting series of patterns we observed when plotting pharmacy sales against time. One of these was a sudden sharp increase in sales and the other was a sudden sharp decreases in sales. Key illustrative examples can be found in **Appendix 4.** The sharp decrease may correspond to pharmacies reducing their purchases or big suppliers abruptly cutting off purchases to pharmacies after falling under the radar of the authorities. An example is the case of Tug Valley Pharmacy, they have a sudden drop in purchases in 2010. Our preliminary analysis would flag this pharmacy and recommend close-surveillance from authorities. The fact that this pharmacy went on to be shut down in 2018 about a decade later, further validates the approach. The idea is that pharmacies that were identified for unlawful purchasing or prescription must have indulged in these practices for a while before they were "found out". On the other hand, the sharp increase might be present because they secured a contract from a big distributor. In any case, there is a clear indication of unusual activity that has to be investigated closely.

This procedure allows us to isolate individual pharmacies based on abnormal buying patterns. However, it might be infeasible to extend this to all of the pharmacies in the Midwest. We outline a possible method to tackle this problem when we discuss limitations.

4.3 Clustering Findings

The clusters can be characterized as follows:

Cluster #	Population	# Nearby Pharmacies	# Nearby Doctors	Income	65+	Insured	Practitioner	Chain	Retail	# Pharmacies in Cluster	Mean Pills Sold
1	Avg	-	Avg	-	--	++	Avg	Avg	Avg	1117	803,630
2	++	+	++	-	--	-	Avg	+	-	686	729,984
3	++	++	-	Avg	+	Avg	+	Avg	-	702	579,005
4	--	--	--	--	++	--	-	Avg	++	1624	1,084,841
5	-	-	--	--	++	Avg	+	Avg	Avg	1516	784,874
6	Avg	+	++	++	--	Avg	++	Avg	-	554	296,926
7	++	+	-	+	--	++	Avg	+	-	494	715,765

The clustering results are counterintuitive to what we initially expected. The clusters with features that should indicate low sales, actually have higher number of sales and vice versa. However, the surprising results may indicate something suspect.

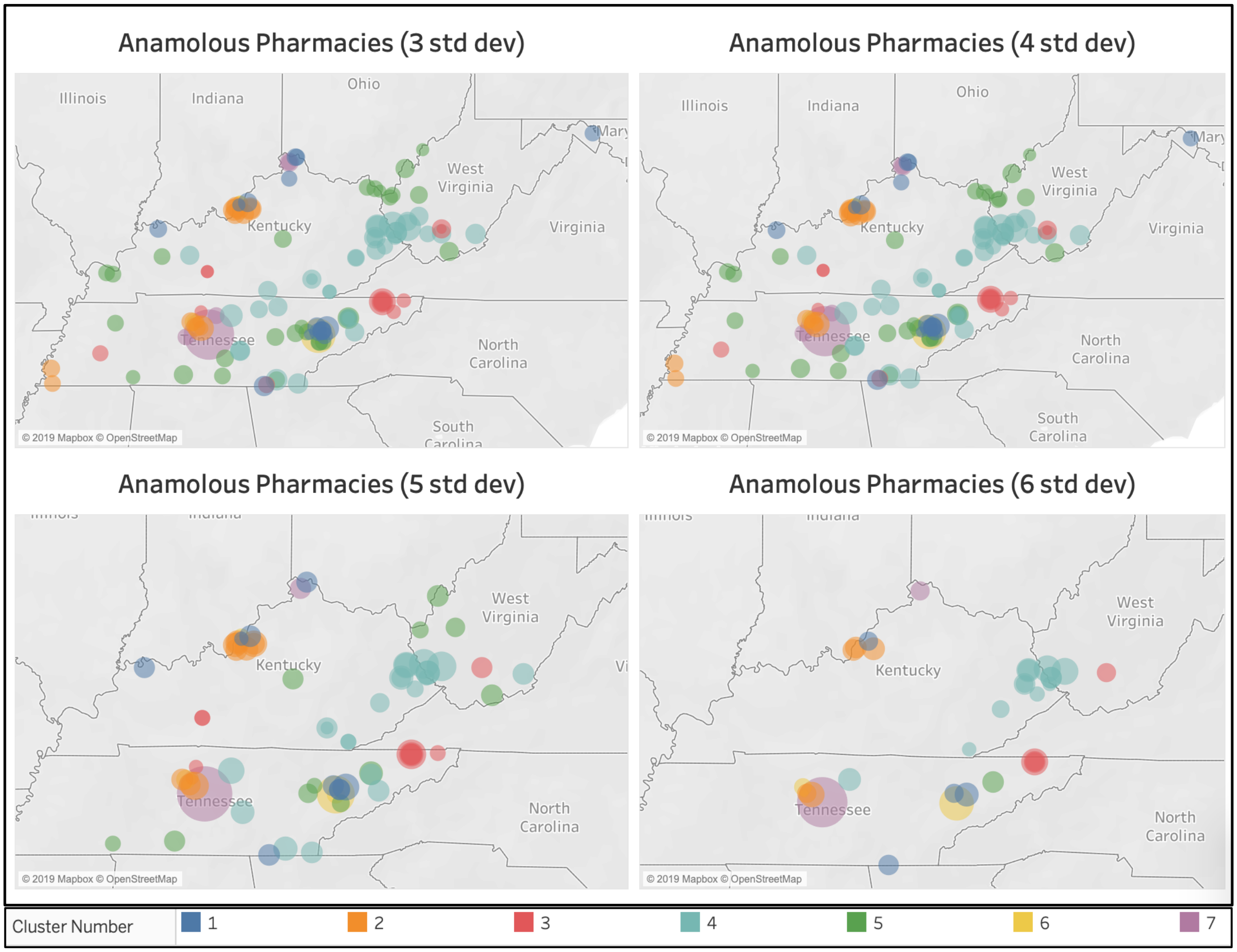
For example, note that cluster 4 is particularly interesting – although we would expect this cluster to have low number of pills (small population, few pharmacies, few doctors, low income, and low insured rate), the cluster actually has the highest average number of pills sold. Based on the cluster features, these pharmacies are likely located in rural areas. Indeed, this is the cluster which contains all the pharmacies which are at the West Virginia-Kentucky border.

On the opposite end, Cluster 6 which has the fewest average number of pills sold, has a high number of pharmacies, doctors, and income. There is also a larger proportion of practitioners.

4.4 Anomaly Detection Findings

The picture below highlights pharmacies which had sales greater than X standard deviations from the mean of its cluster in any year between 2006 – 2012. The size of the circle represents the total number of pills sold by a particular pharmacy. The color of the circle represents the cluster. As we increase the number of standard deviations, we zoom in on the most anomalous pharmacies.

Figure 3. Anomalous Pharmacies by Standard Deviations Away from Peer Pharmacies



The table below summarizes the number of pharmacies flagged at each standard deviation. A lower standard deviations is more conservative, and will capture more anomalous pharmacies but also more false positives. A higher standard deviation risks not capturing all anomalous pharmacies. There is also a cost associated with investigating each pharmacy. Thus, there needs to be some cost-benefit analysis when considering the threshold for flagging suspect pharmacies (more to follow in managerial implications section).

	Number of pharmacies	Pills sold
All pharmacies	6,693	5,274,625,276
3 standard dev	226	1,195,001,248
4 standard dev	121	785,883,198
5 standard dev	69	521,305,535
6 standard dev	32	307,983,070

It is interesting to note that pharmacies within 3 standard deviations represent 3% of total pharmacies, but sold 20% of all total pills.

5) Managerial Implications

The analysis indicates a number of regions where support services for opioid addicts should be targeted. These include the West Virginia-Kentucky border, Northeastern TN, and Nashville, TN (as highlighted in visuals above).

The analysis also supports decision-makers in determining how to best allocate limited resources to investigating suspect pharmacies. This may be resource-intensive, particularly considering that it may require an investigation of neighboring doctors who are responsible for the overprescribing. Based on the budget available and the level of suspicion (i.e. standard deviations away from the mean of peer pharmacies), there are two proposed levels of interventions:

(1) Low cost: targeted notifications warning pharmacies and nearby doctors that authorities
(2) High cost: conduct audits of suspect pharmacies and nearby doctors

This tiered approach not only ensures efficient allocation of resources by targeting pharmacies where there is likely to be a problem, but also is expected to reduce opioid sales compared to a baseline in which random pharmacies are audited or sent warnings.

We believe that the overall methodology can be leveraged by agencies such as the DEA (who aims to combat illegitimate drug distribution) to develop an anomaly detection system which proactively leverages their recent data to flag suspect purchases made by pharmacies. Some methods for scaling the current approach to more pharmacies are discussed in **Appendix 6.**

Concluding Remarks

In summary, the opioid crisis is complex and involves many stakeholders, namely patients, pharmacies, physicians, manufacturers, enforcement agencies, and other government bodies. Leveraging our analysis and findings as a proof of concept, we strongly believe that there is potential for data science to inform a more effective response to the crisis. We urge these key stakeholders to expand on this analysis, in both scale and methodology, in order to help develop a concerted approach to eradicate the epidemic.

APPENDIX 1 – ARCOS Data Sample

```
Classes ‘data.table’ and 'data.frame':  11737630 obs. of  42 variables:
 $ REPORTER_DEA_NO      : chr  "PB0034861" "PB0034861" "PB0034861" "PB0034861" ...
 $ REPORTER_BUS_ACT     : chr  "DISTRIBUTOR" "DISTRIBUTOR" "DISTRIBUTOR" "DISTRIBUTOR" ...
 $ REPORTER_NAME        : chr  "BELLCO DRUG CORP" "BELLCO DRUG CORP" "BELLCO DRUG CORP" "BELLCO DRUG CORP" ...
 $ REPORTER_ADDL_CO_INFO: chr  "null" "null" "null" "null" ...
 $ REPORTER_ADDRESS1    : chr  "5500 NEW HORIZONS BLVD" "5500 NEW HORIZONS BLVD" "5500 NEW HORIZONS BLVD" "5500 NEW HORIZONS BLVD" ...
 $ REPORTER_ADDRESS2    : chr  "null" "null" "null" "null" ...
 $ REPORTER_CITY        : chr  "NORTH AMITYVILLE" "NORTH AMITYVILLE" "NORTH AMITYVILLE" "NORTH AMITYVILLE" ...
 $ REPORTER_STATE       : chr  "NY" "NY" "NY" "NY" ...
 $ REPORTER_ZIP         : int  11701 11701 11701 11701 11701 11701 11701 11701 31793 31793 ...
 $ REPORTER_COUNTY      : chr  "SUFFOLK" "SUFFOLK" "SUFFOLK" "SUFFOLK" ...
 $ BUYER_DEA_NO         : chr  "AT8334447" "BA9132806" "BA9132806" "BC3865017" ...
 $ BUYER_BUS_ACT        : chr  "RETAIL PHARMACY" "RETAIL PHARMACY" "RETAIL PHARMACY" "RETAIL PHARMACY" ...
 $ BUYER_NAME           : chr  "TRIVILLIAN'S PHAR OF KANAWHA" "ADVANCED PHARMACY SERVICES" "ADVANCED PHARMACY SERVICES" "CLENS PHARMACY II" ...
 $ BUYER_ADDL_CO_INFO   : chr  "null" "null" "null" "null" ...
 $ BUYER_ADDRESS1       : chr  "215 35TH STREET, SE" "6620 LASCASSAS PIKE" "6620 LASCASSAS PIKE" "8622 ASHEVILLE HIGHWAY" ...
 $ BUYER_ADDRESS2       : chr  "null" "null" "null" "null" ...
 $ BUYER_CITY           : chr  "CHARLESTON" "LASCASSAS" "LASCASSAS" "KNOXVILLE" ...
 $ BUYER_STATE          : chr  "WV" "TN" "TN" "TN" ...
 $ BUYER_ZIP            : int  25304 37085 37085 37924 37821 37821 42101 42101 41501 38478 ...
 $ BUYER_COUNTY         : chr  "KANAWHA" "RUTHERFORD" "RUTHERFORD" "KNOX" ...
 $ TRANSACTION_CODE     : chr  "S" "S" "S" "S" ...
 $ DRUG_CODE            : int  9193 9193 9193 9143 9193 9193 9193 9193 9193 9193 ...
 $ NDC_NO               : chr  "00603388628" "00603388821" "00591038505" "60951079770" ...
 $ DRUG_NAME            : chr  "HYDROCODONE" "HYDROCODONE" "HYDROCODONE" "OXYCODONE" ...
 $ QUANTITY             : int  2 3 1 2 4 1 2 4 2 2 ...
 $ UNIT                 : chr  "null" "null" "null" "null" ...
 $ ACTION_INDICATOR     : chr  "null" "null" "null" "I" ...
 $ ORDER_FORM_NO        : chr  "null" "null" "null" "065309082" ...
 $ CORRECTION_NO        : chr  "null" "null" "null" "null" ...
 $ STRENGTH             : chr  "0000" "0000" "0000" "0000" ...
 $ TRANSACTION_DATE     : int  6022006 1042006 6262006 3022007 10102006 4302007 5032006 3192007 7242012 2092009 ...
 $ CALC_BASE_WT_IN_GM   : num  6.05 1.82 2.27 1.79 2.42 ...
 $ DOSAGE_UNIT          : chr  "1000.0" "300.0" "500.0" "200.0" ...
 $ TRANSACTION_ID       : int  696 729 6798 1010 3069 10126 907 6423 1099 1037 ...
 $ Product_Name         : chr  "HYDROCODONE BITARTRATE AND ACETA 10M" "HYDROCODONE.BIT & ACETA  10MG & 500M" "HYDROCODONE BIT.7.5MG/ACETAMINOPHEN" "ENDOCET TABS -
 10MG OXYCODONE.HCL &" ...
 $ Ingredient_Name      : chr  "HYDROCODONE BITARTRATE HEMIPENTAHYDRATE" "HYDROCODONE BITARTRATE HEMIPENTAHYDRATE" "HYDROCODONE BITARTRATE HEMIPENTAHYDRATE" "OXYCO
DONE HYDROCHLORIDE" ...
 $ Measure              : chr  "TAB" "TAB" "TAB" "TAB" ...
 $ MME_Conversion_Factor: num  1 1 1 1.5 1 1 1 1 1 1 ...
 $ Combined_Labeler_Name: chr  "Par Pharmaceutical" "Par Pharmaceutical" "Actavis Pharma, Inc." "Par Pharmaceutical" ...
 $ Revised_Company_Name : chr  "Endo Pharmaceuticals, Inc." "Endo Pharmaceuticals, Inc." "Allergan, Inc." "Endo Pharmaceuticals, Inc." ...
 $ Reporter_family      : chr  "Bellco Drug Corp" "Bellco Drug Corp" "Bellco Drug Corp" "Bellco Drug Corp" ...
 $ dos_str              : chr  "10.0" "10.0" "7.5" "10.0" ...
```

APPENDIX 2 – Pills per Capita for US States

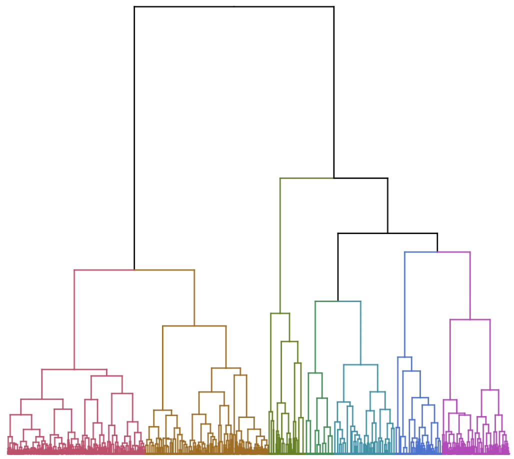
The table below shows the top five US states with the highest pills per capita in 2012, as well as the bottom five. Our analysis is focused on West Virginia, Kentucky, and Tennessee.

Rank	State	Count of Rows	Total Pills	Population	Pills per Capita
1	SC	544,937	357,392,065	4,896,146	73
2	WV	295,481	133,253,210	1,844,128	72
3	KY	673,543	313,335,497	4,425,092	71
4	AL	686,492	313,139,395	4,858,979	64
5	TN	963,781	424,173,794	6,600,299	64
45	NJ	627,867	240,111,942	8,958,013	27
46	NY	1,355,370	527,614,629	19,795,791	27
47	MN	396,706	143,921,404	5,489,594	26
48	SD	67,345	21,405,550	858,469	25
49	ND	53,787	16,998,650	756,927	22
50	DC	21,810	9,202,690	633,427	15

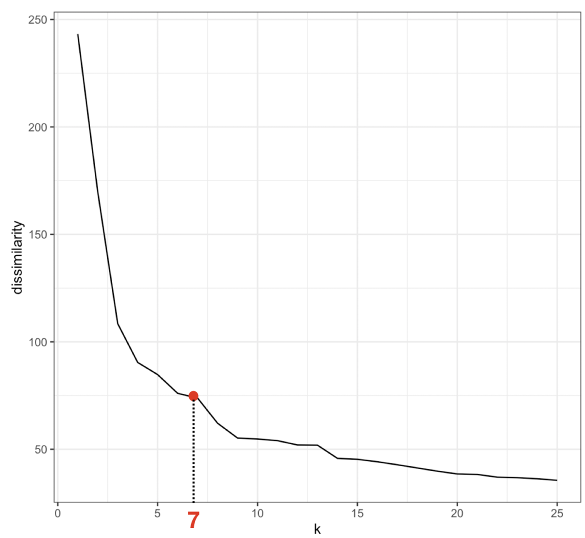
]

APPENDIX 3 – Clustering Details

Hierarchical Clustering Dendrogram



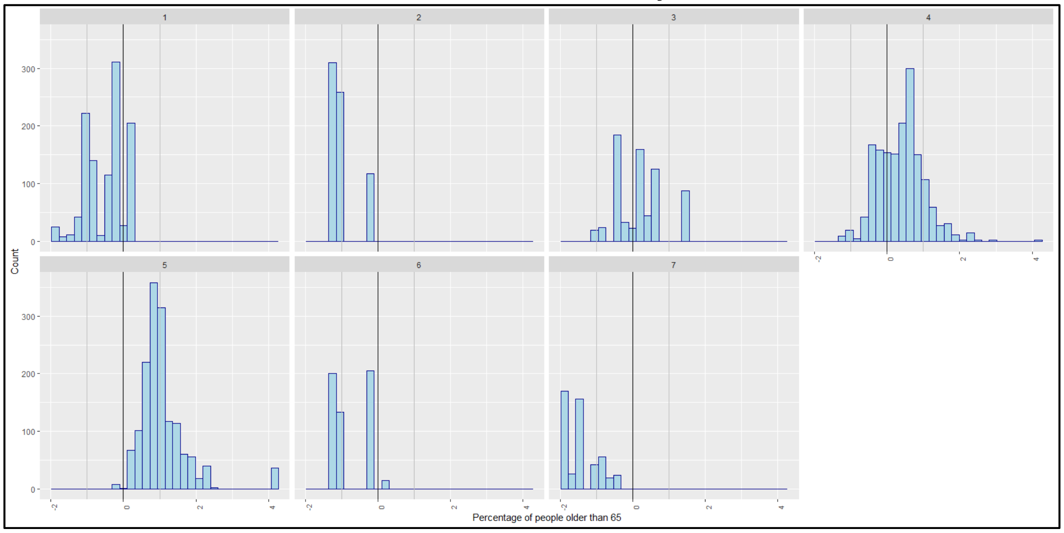
Scree plot



Characterization of Clusters

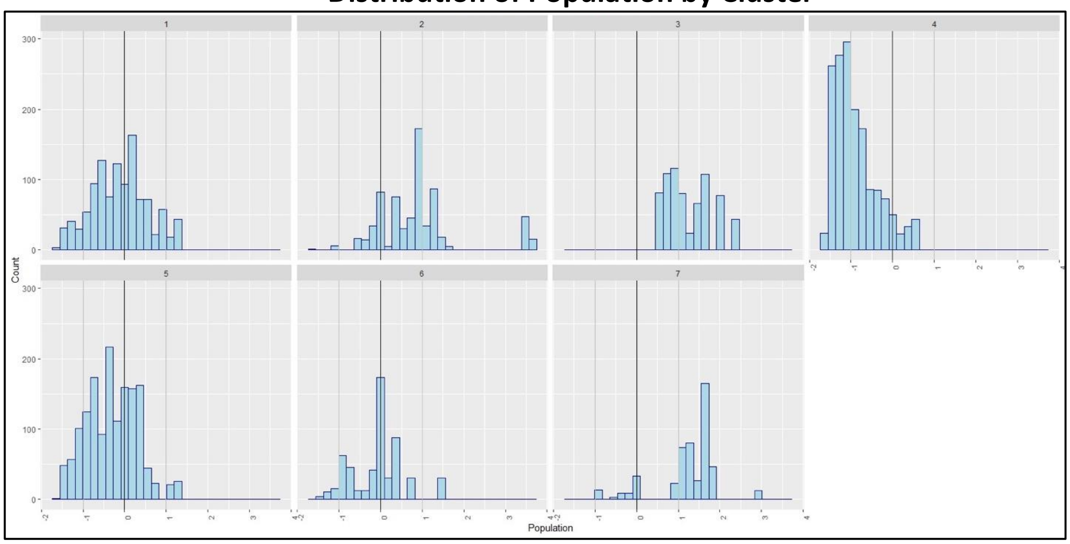
Each panel corresponds to the distribution of the feature for a particular cluster. The black line represents the mean of the feature for all pharmacies and the gray lines represent one standard deviation from the mean.

Distribution of % Over 65 by Cluster



Clusters 1, 2, 6, and 7 are younger than the average. Clusters 3, 4, and 5 are older than the average.

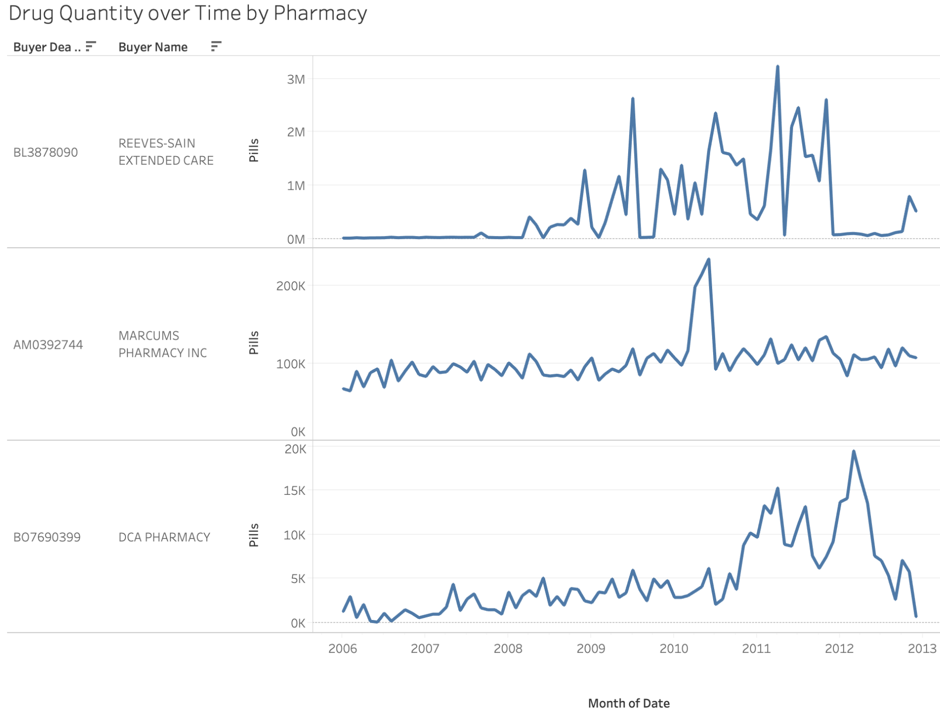
Distribution of Population by Cluster



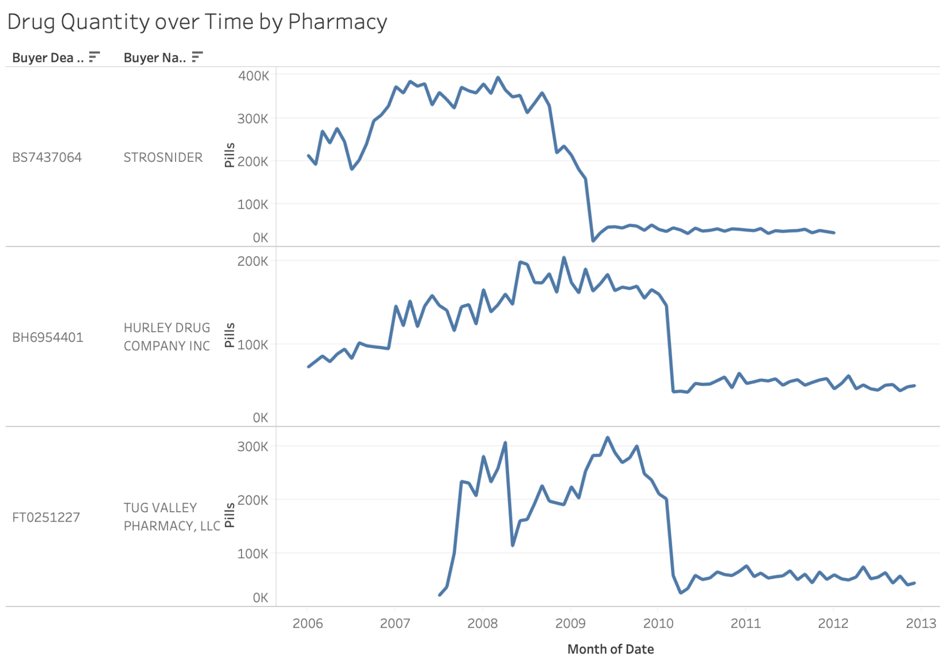
Clusters 4 and 5 are less populated than the average. Clusters 2, 3, and 7 are more populated than the average. The remaining clusters are about average.

APPENDIX 4 – Temporal trends in Pharmacies

Increasing trend



Decreasing trend



APPENDIX 5 – Top 15 Pharmacies by Pills Sold

Top 15 Pharmacies by Pills Sold per Capita

Pharmacy's - Pills Sold per Capita

Buyer Dea No	Buyer Name	Buyer County	Buyer State	Pills	Pills per capita
AF1972656	FAMILY DRUG WHEELWRIGHT	FLOYD	KY	1,026,600	2,194
BF0660565	FAMILY DISCOUNT PHARMACY INC	LOGAN	WV	12,849,040	2,181
BL4844329	LACKEY PHARMACY	FLOYD	KY	1,654,460	1,421
BF1363441	FAMILY PHARMACY	PIKE	KY	4,257,710	1,080
FH1454999	HOWARD FAMILY PHARMACY, INC.	FLOYD	KY	2,380,620	1,047
BF6608698	FOOD CITY PHARMACY #425	PIKE	KY	3,708,800	941
BM7100427	MEDICAL CENTER PHARMACY	MUHLENBERG	KY	3,196,370	799
AM1175024	MEDICINE STOP PHARMACY	BOONE	WV	1,689,590	720
BS7437064	STROSNIDER	MINGO	WV	13,168,350	685
AM9213935	MARROWBONE CLINIC PHARMACY	PIKE	KY	1,861,120	677
BH0695001	KENTUCKY CVS PHARMACY, L.L.C.	PIKE	KY	2,634,860	668
AN6397930	USN - NAVY EXCHANGE PHARMACY	SHELBY	TN	1,027,660	603
FS2880070	SPECIALTY CARE CENTERS OF EASTERN K..	PERRY	KY	86,700	602
AB9715674	BETSY LAYNE PHARMACY INC	FLOYD	KY	2,286,450	598
BK7519828	KROGER PHARMACY	LOGAN	WV	3,441,400	584

Top 15 Pharmacies by Total Pills Sold

Sum of Pills Sold by Pharmacy

Buyer Dea No	Buyer Name	Buyer County	Buyer State	Pills	Pills per capita
BL3878090	REEVES-SAIN EXTENDED CARE	RUTHERFORD	TN	45,925,980	153
BF7000526	FOOD CITY PHARMACY #674	KNOX	TN	21,555,400	124
BS7437064	STROSNIDER	MINGO	WV	13,168,350	685
BF0660565	FAMILY DISCOUNT PHARMACY INC	LOGAN	WV	12,849,040	2,181
AG0388238	P & S PHARMACY	SULLIVAN	TN	12,662,200	52
BT7485166	THE WELLNESS PHARMACY INC	DAVIDSON	TN	11,849,020	25
BF4584024	FOUR WAY PRESCRIPTION SHOP	KNOX	TN	10,784,070	160
BV7526176	VALUE-MED INC	JOHNSON	KY	10,449,480	242
BP4202610	P C A CORRECTIONS	JEFFERSON	KY	9,611,380	41
AS2660365	SMITH COUNTY DRUG CENTER INC	SMITH	TN	9,508,190	217
BL6079948	LITTLE & WADDELL INC	FLOYD	KY	9,274,090	123
BD8968173	DBA MEDICINE CABINET PHARMACY	JOHNSON	KY	9,183,940	213
BH6954401	HURLEY DRUG COMPANY INC	MINGO	WV	8,890,370	217
FT0251227	TUG VALLEY PHARMACY, LLC	MINGO	WV	8,827,860	215
AW2992863	WALGREEN CO.	JEFFERSON	KY	8,725,790	35

APPENDIX 6 - Limitations & Further Work

We found it difficult to validate the legitimacy of the clusters, i.e. how do we know these are actually 'peer groups'. It is not clear how we can benchmark the results of our model against the truth. A naive approach could be to track the activity of some of the pharmacies we flagged as overselling, over a period, and seeing if they eventually get involved in legal issues or even shut down - e.g. in the case of Tug Valley pharmacy, where they exhibited a long period of overselling and went on to be taken down by the authorities for violating prescription laws, only about a decade later. We need more rigorous characterizations of our results to utilize the model with certainty, this might also make it easier to convince policy makers or stakeholders of its value.

Whenever a new pharmacy or a group of pharmacies are added to the database, or we obtain more data for later periods for the existing pharmacies, we may have to perform clustering again. One can assign new pharmacies to existing clusters based on their pairwise dissimilarities with the existing pharmacies but this would not make use of the new information we acquire. This also does not address the fact that the behavior of pharmacies can vastly vary with time, as we have seen in our analysis. An approach to handle this can be to cluster pharmacies based on their sales in each month over the 7 year period. We can then identify peer groups of buying patterns, similar to the analysis carried out without the temporal variations earlier. We flag pharmacies based on whether the net percentage increase is greater than that of the cluster centroids by a number of standard deviations away.

Another limitation is that we characterize a region to be a problem-area based on whether pharmacies in that area overprescribe or not. This need not always be the case, as it is widely known that people living in suburban areas sometimes travel up to a few hours to meet with a practitioner or visit a pharmacy hundreds of miles away to pick up their prescriptions. We can augment our analysis only if we have data at the next level of the supply chain - i.e. from buyer pharmacies/practitioners to customers.