John Paul Bida August 2015 bida.john@gmail.com

Lyra Data Science Interview Problem

Summary:
The objective was to analyze the Medicare Fee-For Service Provider Utilization & Payment Data Part D Prescriber to determine if their were any differences between how primary care providers (PCP's) and Psychiatrists prescribed medication.

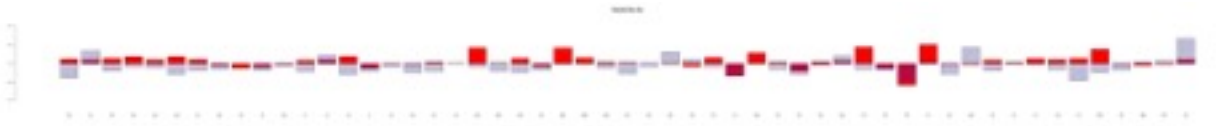
You will find all supporting code in the github repository here: https://github.com/jpbida/partD

I began by normalizing all data using z-scores for total_day_supply and cost. I then made plots of PCP z-scores and Specialist z-scores for each across all states for a given drug.

I noticed three types of drugs. First there was the ***specialist drugs***. These were over prescribed by specialists and under prescribed by PCP's across all geographies.

./results/bar_plot10.jpeg
Each bar represents a state. Red represents the z score for percent total day supply for PCP's and grey represents the zscore for Psychiatrists. We see red is always below zero and grey is always above zero.

Second there were ***common drugs*** that seemed to be over prescribed / under prescribed by both PCP's and Specialists.



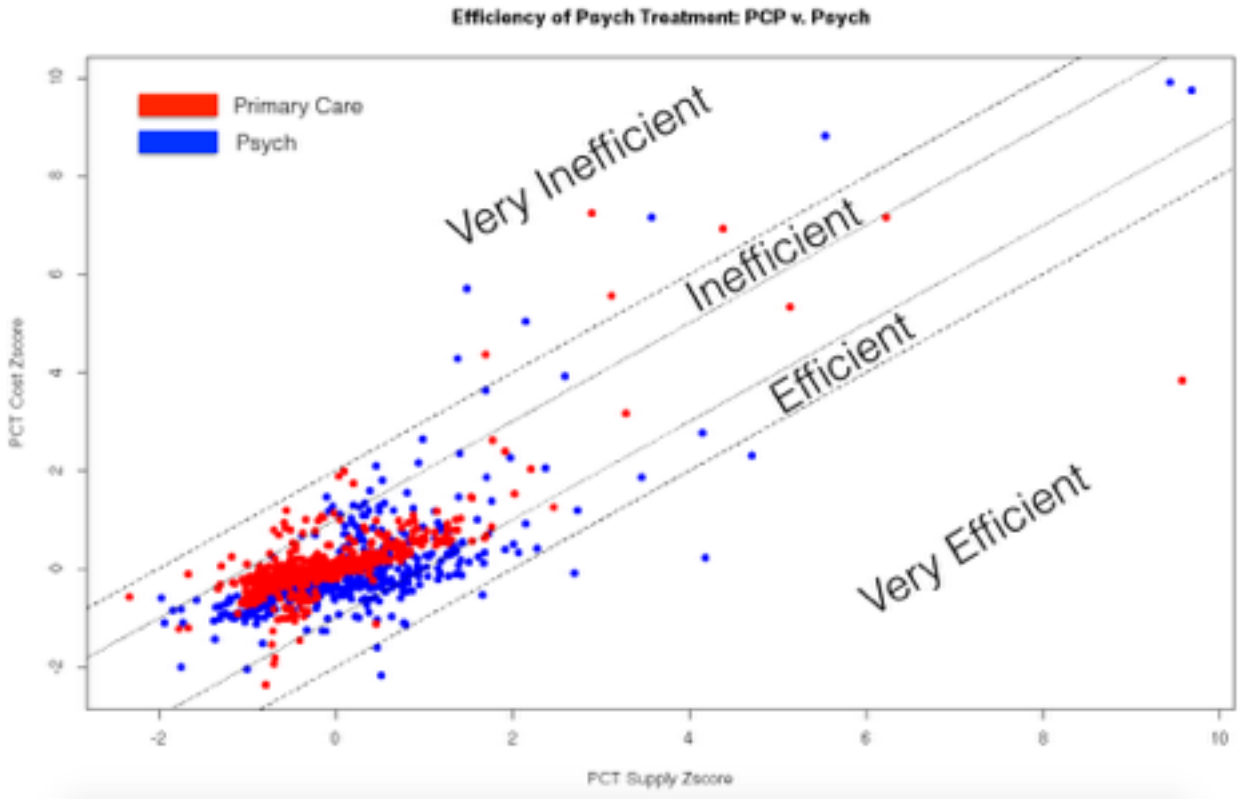
Third there were ***well defined*** drugs that had prescription rates equal across all geographies with low z scores for both PCP's and specialists.

Automatic classification into these three types would be an interesting next step. The "well defined" drugs could be used for normalizing against in future analysis. The common drugs is most likely where unneeded inefficiencies arise. The specialty drugs could be used to look at what happens when there are a low number of specialists around. Do the PCP's make similar choices?

John Paul Bida August 2015 bida.john@gmail.com

John Paul Bida August 2015 bida.john@gmail.com

I was interested in the inefficiency question because it was easier to look at. Made a plot of normalized cost v. supply. There are a lot of assumptions in this plot. First, I'm assuming the populations across geographies have similar relative prevalences of the diseases treated by the drugs being investigated. Ideally, you could use county level CDC data and stratify across similar populations. I thought it was a simple plot that could be made in the future with the above adjustments.



Drugs in efficient cases

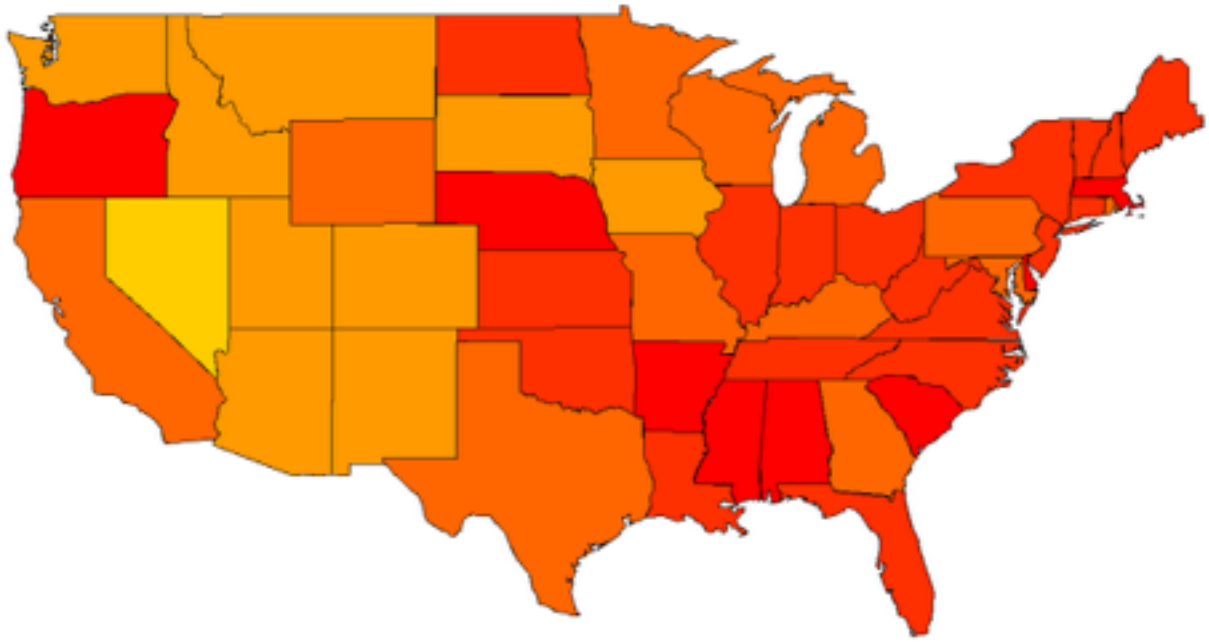
Drugs in inefficient cases

	pcp	psych
-2	0	2
-1	0	20
0	38	28
1	16	1
2	0	1

BUPROPION HCL is interestingly used in the efficient range by psychiatrists and the inefficient range by PCP's.
(-2 = Very Efficient Range, -1 = Efficient, etc..)

It was found that PCP's tended to inefficiently under prescribe psych and over prescribe under prescribe and overcharge were Psychiatrists were more efficient with their delivery overall.

When we plot back on the states with the efficiency of treatment we see the map below.



RED = efficient Treatment -> Yellow = Inefficient Treatment
*Maybe some correlation between remote areas and efficiency.

Future directions:

“House Keeping” Drugs (like house keeping genes in biology)
I think using prescription levels of “house keeping” drugs may be an interesting way to normalize this data set. House keeping drugs are those that are prescribed in very similar ways across all geographies . This gives you a measure of the prevalence of whatever was being treated and could be used to determine the similarity of geographies.

Missing BENE_COUNT data makes things difficult. You don’t know the size of the population you are serving. With 60% having missing data you could be off by a lot. You could probably model BENE_COUNT with house keeping drugs better.

“County Level Data”
Bringing in other data sources to give you a measure of the success of treatments at the county data might allow you to measure efficiency with quality. We don’t really know if any of the drugs are working. Something like cause of death data, special ed rates in schools.

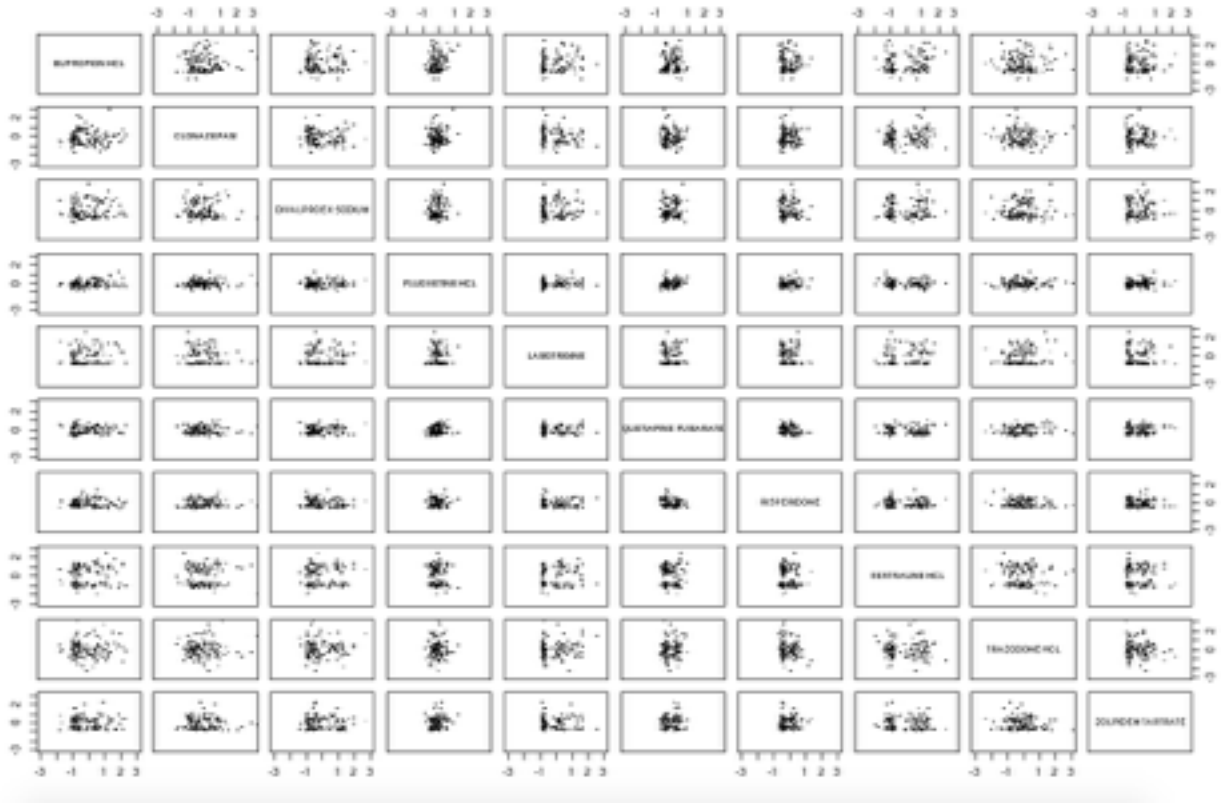
“Specialty Drug - Delivery in remote areas”

John Paul Bida August 2015 bida.john@gmail.com

Look at the drugs matching the specialty classifaciton then find remote locations with out specialists. Look at the PCP's in the are and see how they are doing delivering these particular drugs.

"Treatment Variations"
You could image two different treatments that use different sets of drugs. For example, you could preform surgery or just manage pain. One geography that prefers surgery may see lower dosages of pain meds and higher does of blood thinners. Can you pull out these treatment classes by looking at the relationships between prescription rates?

We have some pairwise correlations with CLONAZEPAM and LAMOTRIGINE having a high negative correlation suggesting possibility of two different treatments for the same symptoms.



John Paul Bida August 2015 bida.john@gmail.com

METHODS

Exploratory Analysis: explore.r

Dataset - First 10K rows of the PARTD full dataset

Observation 1: Missing data in BENE_COUNT

We have two data points that tell us approximately how much of a drug is being prescribed.

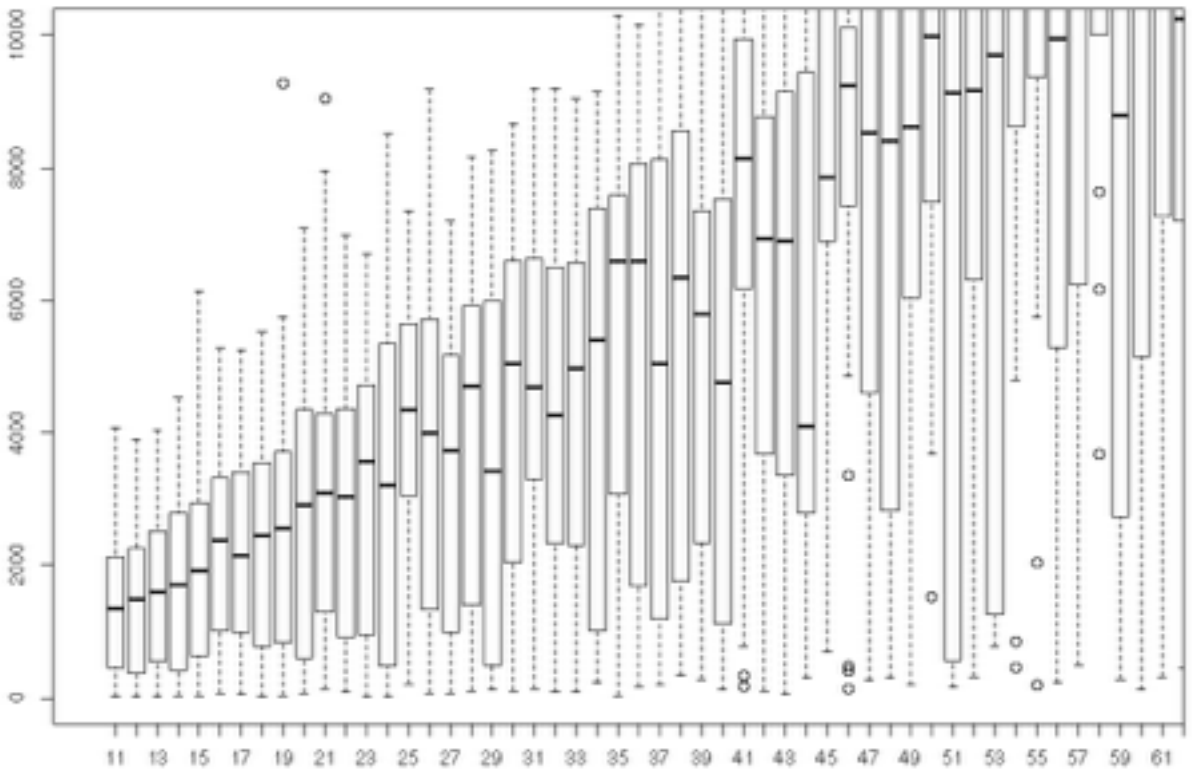
bene_count – The total number of unique Medicare Part D beneficiaries with at least one claim for the drug. Beneficiary counts fewer than 11 are not displayed.

total_day_supply – The aggregate number of days supply for which this drug was dispensed.

BENE_COUNT has a ton of missing data >60% due to the aggregation limits on the dataset.

We could build a model to impute the missing data points, but I think I'm going to just avoid it for now.

Plot below shows BENE_COUNT v. TOTAL_DAY_SUPPLY. Built a regression but didn't like the results. Moving on.



Observation 2: Quetiapine is different than the others…

One thought is that given a group of Pyschiatric drugs the ratios of the total_day_supply's would be consistent across NPI's with similar populations of people and similar treatment plans.

For example, if we had two Drugs A and B and two NPI's N1 and N2.

Similar Populations / Treatment Plans

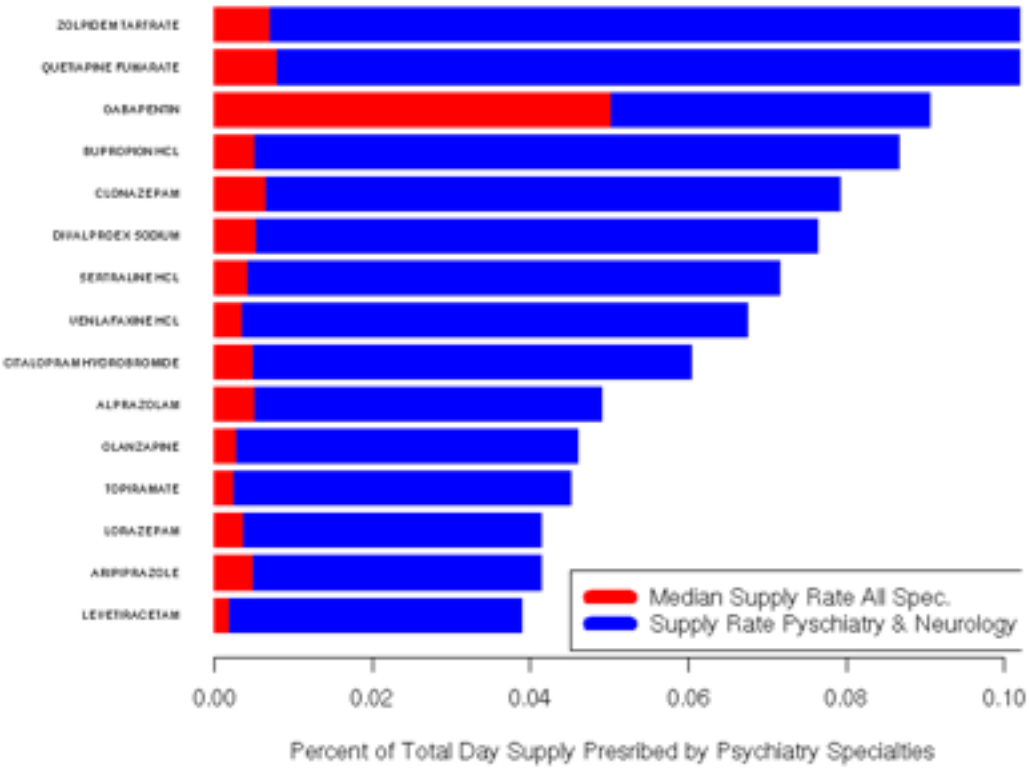
	A *Total_day_supply(%)*	**B *Total_day_supply(%)***
N1	100(50%)	100(50%)
N2	2000(50%)	2000(50%)

Different Populations / Treatment Plans

	A *Total_day_supply(%)*	**B *Total_day_supply(%)***
N1	100(50%)	100(50%)
N2	4000(66%)	2000(33%)

I started by looking at the percentage of total_day_supply for all drugs in the PSYCH specialties and compared it to the median of all other NPI's.

We see most drugs have a low percentage in the other NPI's except for GABAPENTIN.

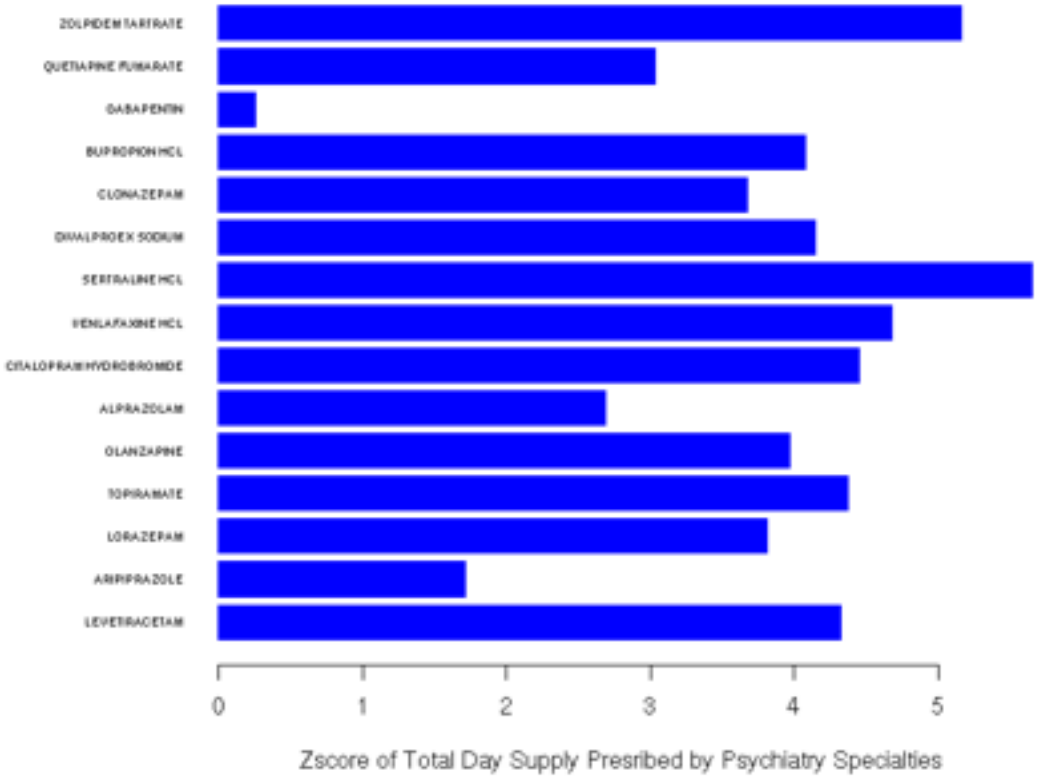


John Paul Bida August 2015 bida.john@gmail.com

PCT_SUPPLY = SUM (TOTAL_DAY_SUPPLY for a given GENERIC_NAME and SPECIALTY_DESC) / SUM (TOTAL_DAY_SUPPLY for a given SPECIALITY_DESC)

This tells us approximately how important this drug is to a given speciality. We can also normalize across all specialities using z-scores to determine how specific a drug is to a given speciality.

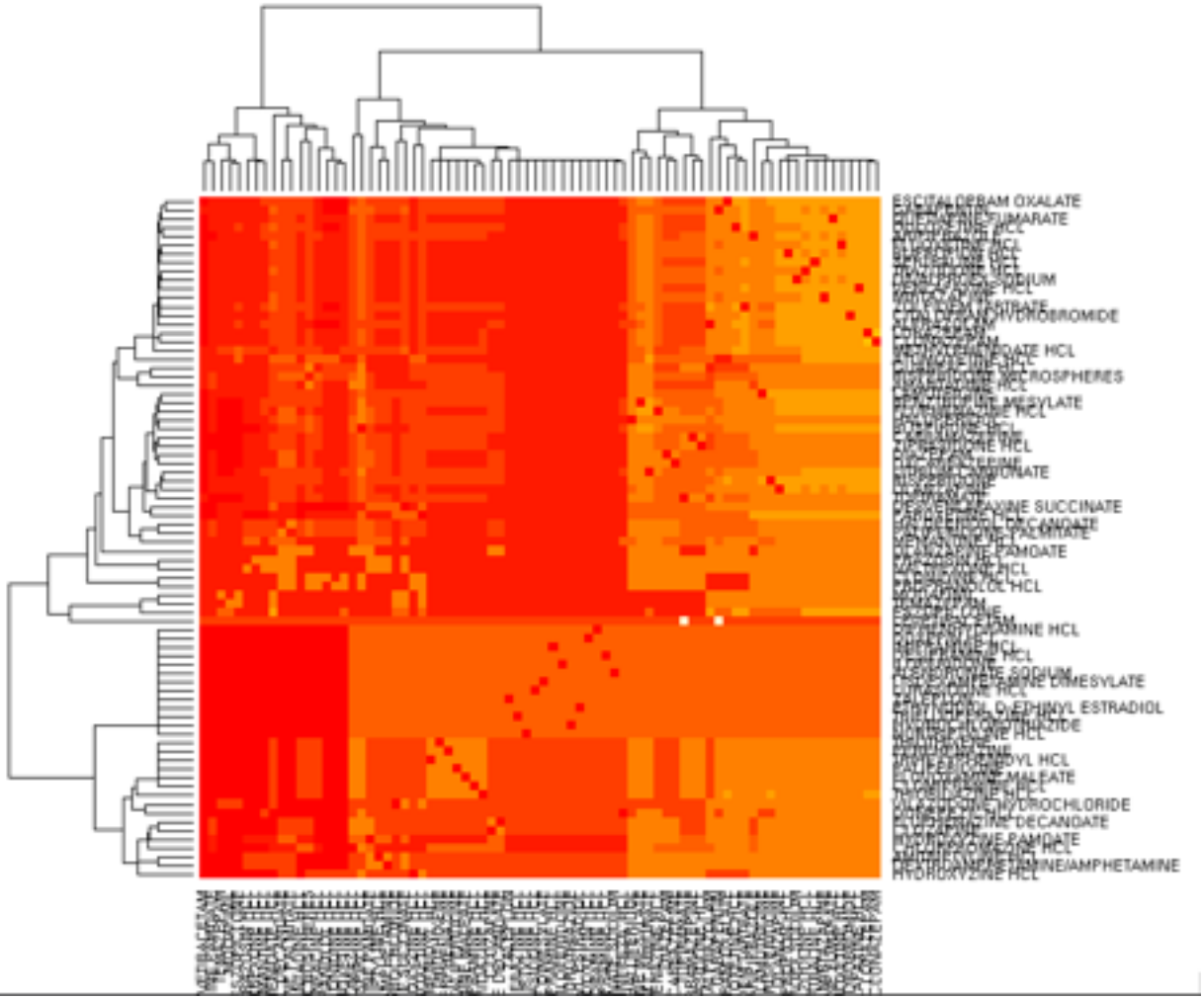
Z_SCORE = PCT_SUPPLY - MEAN_PCT_SUPPLY / SD PCT SUPPLY



We see GABAPENTIN has a relatively low z-score meaning it isn't uniquely prescribed by the Psychiatric specialty. Google search indicates it is an anti-seizure drug also used to treat symptoms of Herpes. The results make some sense.

Observation 3: NPI's co-prescription rates are messy

Another thought was that you could get a sense of a NPI's treatments looking at the co-occurrences of drug prescriptions. Below is a heat map showing the probability 2 drugs are prescribed by the same NPI. It was just too messy and hard to interpret, I decided to think more about z-scores for cost and total supply.



John Paul Bida August 2015 bida.john@gmail.com

Analysis Plan

After the exploratory analysis I decided to run with the z score method. We will be doing the following:

1) Find the top 10 Psych drugs by prescription rate among Psychiatrists.
2) Remove all other drugs from the dataset and calculate total_day_supply in each state for Psych and PCP groups. Use this total to calculate the percentage each of the top 10 drugs is of that total
3) Across all states calculate the z score for percent total_supply for each drug. Do the same for cost
4) Create classification of drugs based on their z-score for supplies between Psych and PCP providers across states.
    1) Specialist - (Across all states Psych have higher z-scores)
    2) Common - (Mix of z-scores for Psych's and PCP's at extremes - Life style drugs)
    3) Defined - (All scores are < 0.5 - everyone knows how to use this drug)

5) Plot cost v. supply and look for low-efficiency and high-efficiency deliveries by classification.

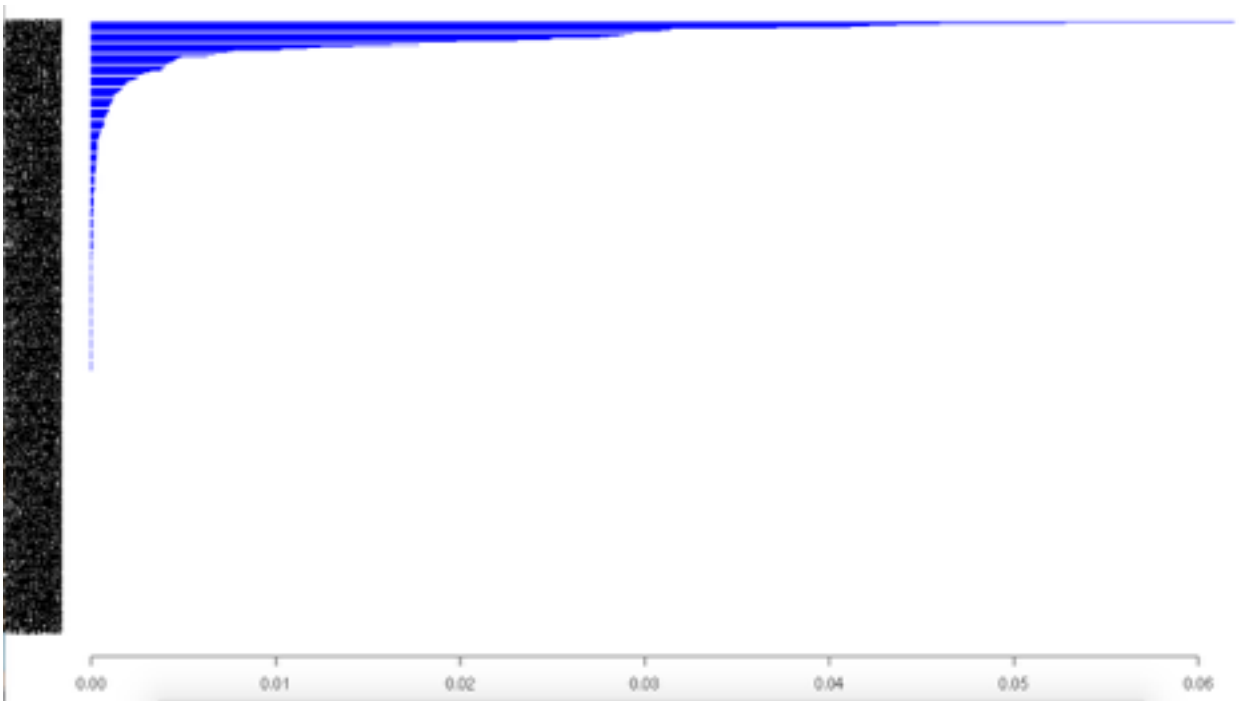
6) Plot efficiencies across geographic regions

1. Top 10 Psych Drugs - psych_drugs.r

Dataset - Created by grep'ing out the Pysch & Neuro entries from the full dataset

Plot below shows bar plot of percent of total day supply for all the drugs appearing in the psych data set. I ordered them by rate and outputted it to drugs.tab. The top 10 will be used for analysis.

All Top 50 - Red



2. Subsetting Data split.r (./hadoop/)
I originally thought the data set was small enough to do everything on a single machine. I was going to split by county. I couldn't find a way to split the dataset up fast enough. My attempt was captured in

split.r - Attempt at splitting data into different county's

Setup a Hadoop cluster on my laptop and wrote map reduce jobs. Look at the hadoop directory for more information.

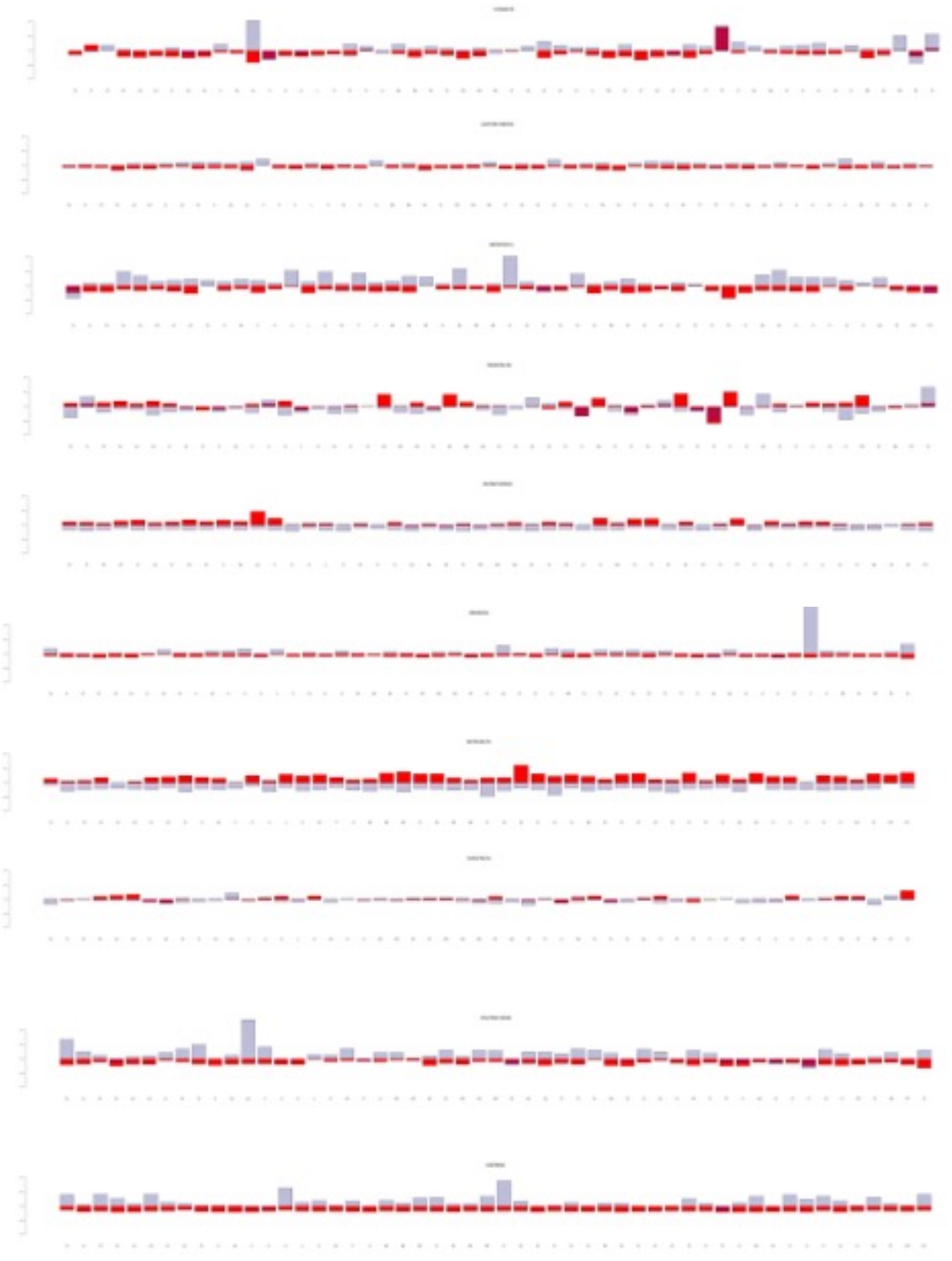
3. Rest of analysis - analysis.r

Really ins't much to it.

John Paul Bida August 2015 bida.john@gmail.com

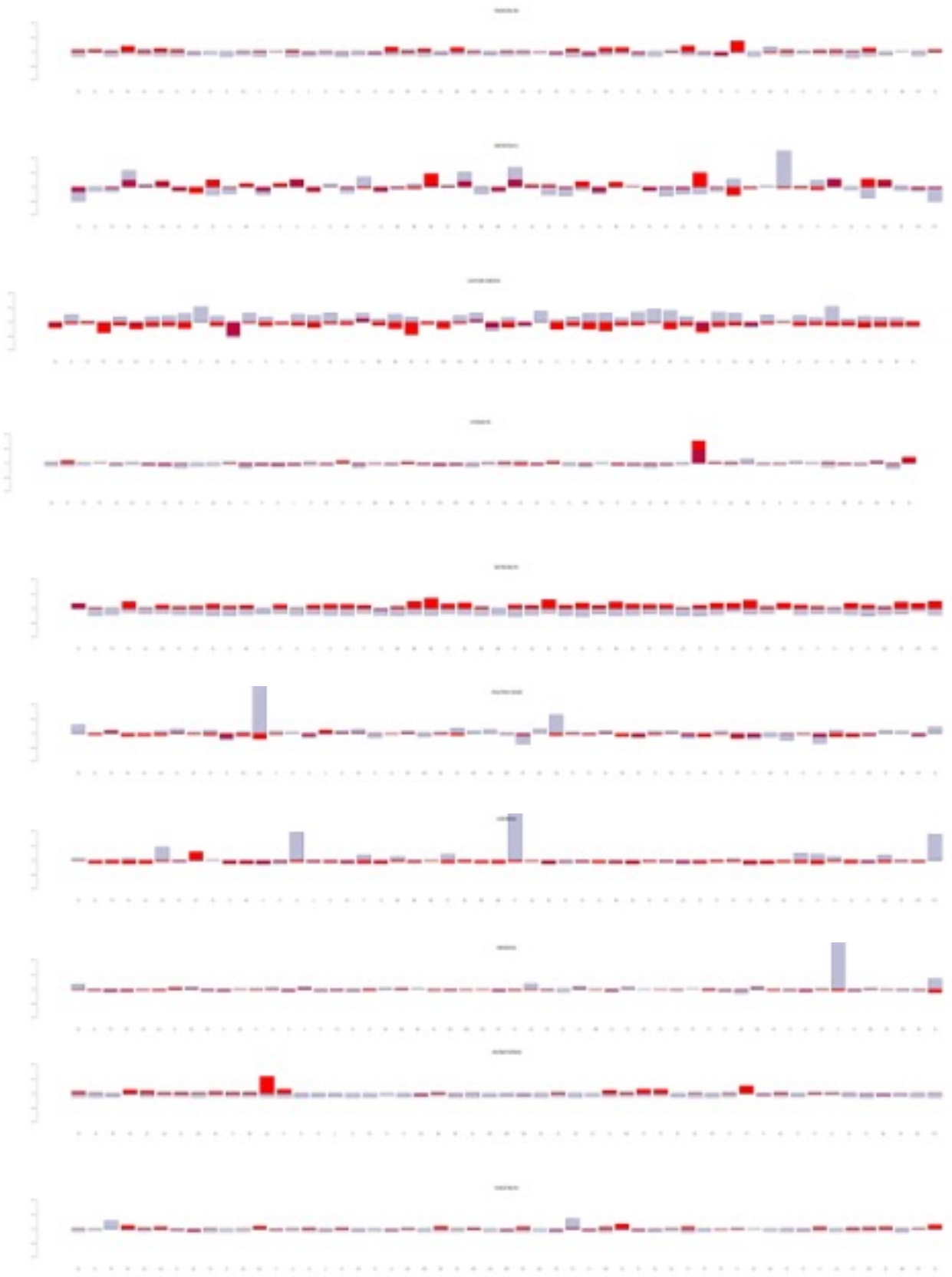
Barplots of Cost and Supply across states. (Look in results on github repo for random plots made by scripts)

Z-score Total Supply - Red = PCP , Grey = Psych



John Paul Bida August 2015 bida.john@gmail.com

Bar Plots of Cost



John Paul Bida August 2015 bida.john@gmail.com