CURES: Ventura County Prescription Data

Alona Kryshchenko, Cynthia Flores, Olivia Alexander, Andrew Nelson, Cynthia Sillas

Abstract: CURES is a database of Schedule II, III and IV controlled substance prescriptions committed to the reduction of prescription drug abuse without affecting legitimate medical practice or patient care. Given a subset of data from California's Prescription Drug Monitoring Program, we are focusing our efforts on Ventura County. With general patient information such as age, zip code of pharmacy where the prescription is filled as well as the prescription itself, we are analyzing our dataset in order to spot potential drug abuse. We are interested in finding a correlation, if it exists, between age, zip of pharmacy, or gender and how this relationship may affect prescriptions provided within our county. Through creating classification targets in order to simplify our given attributes of zip code of pharmacy, prescriptions provided, and drug strength, we aim to fit a set of attributes to a linear regression model.

1. *Introduction*

With our country currently in a state of panic over the Opioid Epidemic, we have set a goal of creating a "flag" system and designing a corresponding attribute for those provided with ample pills and/or days supplied to consider addition potential as well as zip codes that may be unethically facilitating abuse or distribution of controlled substances. Our research group has personal connections with this epidemic and are affected by these addictions in some personal manner. Seeing family members deteriorate over opioid abuse has shown us that the dependence can stem from legitimate mental dependence on the drug, whether it was initially necessary or not. Controlled substance prescriptions are often abused by the prescribed patients' family members or are stolen and sold on the black market or streets. Our research groups goal as well as CURES ultimate commitment is our citizens be able to obtain prescription drugs when necessary and they are used appropriately. Prescribers should be mindful in who they prescribe to and how often. Additionally, patients should be considerate that as this epidemic increases and physicians are being encouraged to supply less Schedule II, III, and IV drugs, there may be legitimate needs not being met because of the current level of abuse. When people are in pain, mentally induced or physically authentic, they may turn to street drugs of a higher level than are easier to get from a doctor such as heroine, which could increase the national rate of drug overdoses, arrests, crime, violence, and death.

We can increase the dimension of our data with feature engineered attributes. We aim to categorize information automatically through training analysis, creating models of linear regression, and machine learning classification by identifying target classes and engineering new parameters for our analysis. With supervised learning we test different combination of attributes, engineered or initially provided, and aim to fit linear regression models to demonstrate possible correlations. The ultimate aim is Ventura County Public Health to have the ability to use our model for future datasets. We hope that our efforts may assist CURES or other drug monitoring programs in helping prevent drug abuse by using our engineered "flag" attribute.

We consider two datasets, one consisting of 1,400 patients and the second with 14,000 patients. For our focus of Ventura County, we fix our data frames to consider only zip codes within our county. The included zip codes are assigned an integer value through a mapping dictionary in order to simplify our dataset. Additionally, we create a mapping dictionary to fit all prescriptions into classifications. We assume that the specific name of the prescription is less important than its class/purpose. Our roughly 150 prescriptions are classified into 8 groups in order to address our research question directly and explore if there are certain classes being overprescribed or abused. The classifications are assigned an integer value ranked by levels of addiction/abuse potential. Of the 8 drug classes, we are focusing our efforts and dataframes on the 3 most addictive classes of drugs; narcotics, sedatives, and amphetamines. Other classes are muscle relaxants, antihistamines, anticonvulsants, steroids, and weight loss. Of the roughly 12,000 patient files from Ventura, 85% are given Narcotic and Sedative prescriptions (4903 Narcotics, 4868 Sedatives) while Amphetamines account for 1726 cases, seen in Figure 1.

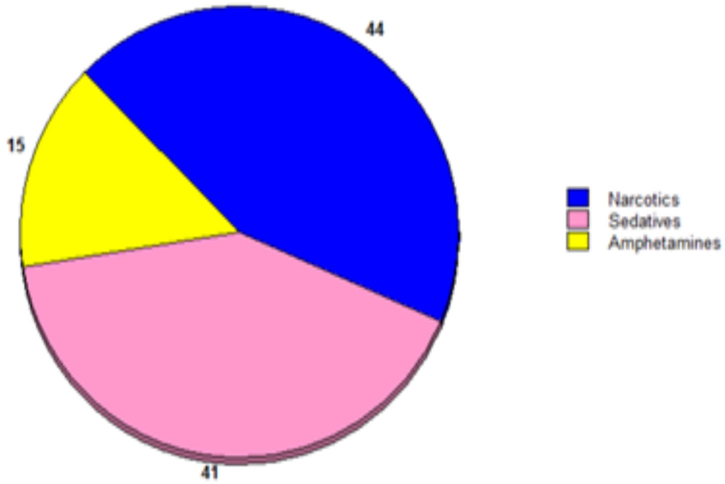
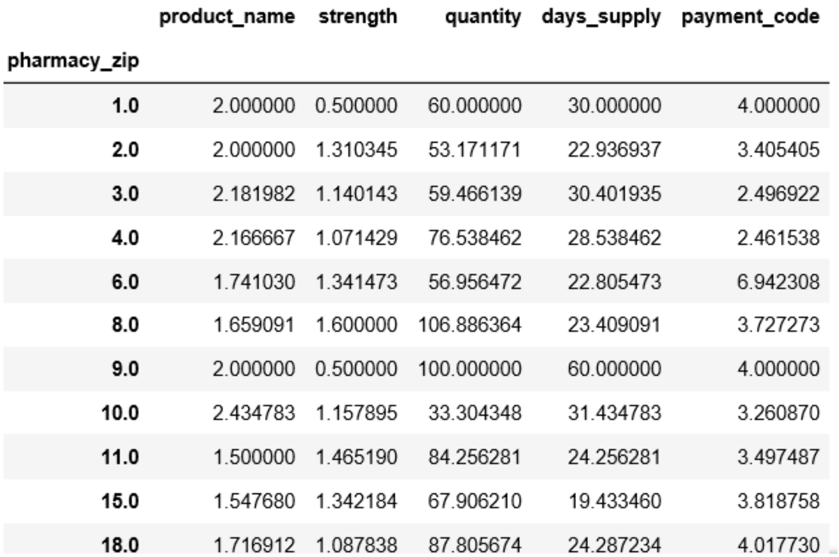


Figure 1 *Drug Class Percentages for Ventura County*

Explanatory variables are also known as independent variables. In experiments and research, this means a variable that is changed or controlled in order to test the effects on the dependent variable. We refer to our dependent variables as response variables which are being measured and tested. In the case of our study, we say that the zip code of the pharmacy is the explanatory variable and the prescription/drug class is the response variable. This response variable is also our target attribute. Our ultimate goal is to learn how to categorize based on the other attributes. We assume that the zip code is the “cause” of the problem and the drugs provided is the “effect”. We have the zip code attribute controlled and observe how other attributes react with this information such as prescription, quantity of pills, days supplied, and drug strength. We can categorize these variables as categorical (qualitative) or quantitative. As categorical variables are descriptions of things, much of our information may be classified as such. Our most clear categorical variable is the prescription attribute. Although this information is now sorted “numerically” by drug class, its data is being grouped, or *categorized*. Additionally, the gender of the patient qualifies at *categorical*, as we could not combine the two sexes to analyze mathematically; this information provides descriptions of the two classes. Similarly to the gender attribute, the drug form attribute only provides two values, but functions to sort and classify the prescription into *groups*. The payment code is a categorical variable because it is not a measurement but rather a *sorting* of how a patient paid for their treatment.

Quantitative variables are those which involve numerical values. It follows that we may sort the quantity of pills provided to a patient as a *numerical* attribute. Similarly, the number of days supplied for a prescription is a numerical value which can also be used mathematically to *calculate* values such as averages/means, standard deviations, etc. Drug strength is quantitative because it is a *measurement* of the prescription. Patient birth year is classified as qualitative. Each year is a group which tells us information about the patients; it is a *description* of the patient.

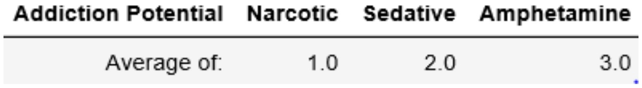
With an aim of discovering how age, zip code, and gender may affect prescriptions provided, we expect to find a strong correlation when testing these attributes with our Class 1 and 2 prescriptions of Narcotics and Sedatives. In order to find any suspicious behavior in an area, we compute the mean for quantitative attributes of each zip code, as seen in Figure 2, by using the python command: *df.groupby(['pharmacy_zip']).mean()*.



	product_name	strength	quantity	days_supply	payment_code
pharmacy_zip					
1.0	2.000000	0.500000	60.000000	30.000000	4.000000
2.0	2.000000	1.310345	53.171171	22.936937	3.405405
3.0	2.181982	1.140143	59.466139	30.401935	2.496922
4.0	2.166667	1.071429	76.538462	28.538462	2.461538
6.0	1.741030	1.341473	56.956472	22.805473	6.942308
8.0	1.659091	1.600000	106.886364	23.409091	3.727273
9.0	2.000000	0.500000	100.000000	60.000000	4.000000
10.0	2.434783	1.157895	33.304348	31.434783	3.260870
11.0	1.500000	1.465190	84.256281	24.256281	3.497487
15.0	1.547680	1.342184	67.906210	19.433460	3.818758
18.0	1.716912	1.087838	87.805674	24.287234	4.017730

Figure 2 *Mean of Attributes by Zip Code*

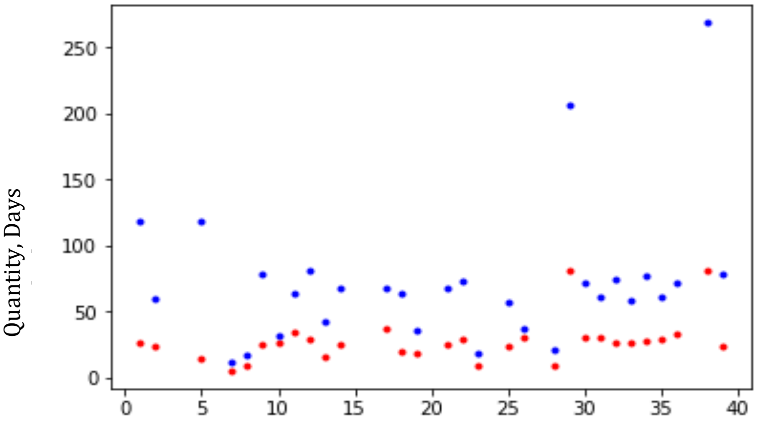
The product name mean reveals what class of drug the zip code is a high prescriber of, e.g. zip code #15 has a drug class average of 1.5, implying high number of cases for narcotics and sedatives whereas zip code #10 has an average of 2.4, not being a high prescriber of narcotics nor sedatives per se. We expect that the zip code of the pharmacy is the most significant attribute of our datasets and will test this information against a combination of different attributes. Using the data from this column, we create a new attribute of "addiction potential," sumarized in Figure 3. This feature engineered attribute will summarize the average drug class prescribed for a given zip code in order to consider areas of concern for overprescribing or abuse potential on the patient side.



Addiction Potential	Narcotic	Sedative	Amphetamine
Average of:	1.0	2.0	3.0

Figure 3 *Addiction Potential Classifications*

We also expect a relationship between the days supplied for a prescription and the number of pills given for each refill of that prescription. We hypothesize that comparing these two attributes can also assist us in our research question by providing additional information in forming addition potential. Prescribers providing low number of pills and/or days of prescription may be due in part from them being skeptical of a patient. Using a simple scatterplot, we assess a possible correlation for our first target plot between zip code and quantity and days supplied, shown in Figure 4. As the sine wave pattern is shown in Figure 5, we observe our plot in Figure 4 has some periodicity with similar patterns to that of sine waves. The curve shows configurations of periodic oscillations with a fairly consistent amplitude. We observe that the zip code shares a similar relation to quantity of pills than it does to days supplied. Both of the attributes graphed on the y-axis share similar oscillations to that of the sine wave. Note that quantity is shown as blue and the red points represent days supplied.



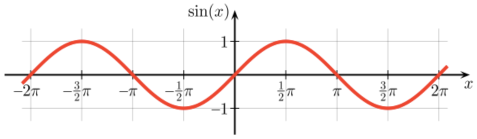


Figure 4 Zip Code vs. *Quantity, Days Supplied*

Figure 5 *Sine Wave Pattern*

With 9 attributes in total, we determine which are most important and find c ns for our linear regression model and our "flag" system for low, medium, and high risk patients. We also make assumptions about the patient birth year and the payment code from the transaction. Considering the patient birth year, we expect to see a high number of pills/days provided with prescribers not per se being wary of elderly addiction/abuse, however, they may need refills often because of relatives stealing their drugs, or they may have addiction that is overlooked by many physicians. We aim to consider this attribute in our "flag" system as well as the payment code of the prescription. It is unclear that this information can genuinely be considered without bias; however, we are unaware which payment code correlates to what kind of insurance/free health care. We believe this attributes may explain certain abuse based on zip code or class of prescriptions. In acknowledging that different payment types may correlate to different socioeconomic status, we may spot abuse in this form as well. To supplement this information using a larger scope, we research the average income of each zip code within our datasets for Ventura County. This additional attribute in our dataframe will allow us to test relationships between socioeconomic status and how it may affect prescription abuse in an area.

We examine this combination of attributes by plotting average income and addiction potential by attempting a regression to find an estimation of the parameter. Where x_i is our data, independent variable, of income, y is our response (dependent variable), of addiction potential.

We have $x_1, x_2, \dots, x_n$ as the mean income value of each zip code of our datasets within Ventura. Our y, the feature engineered addiction potential, is a numerical value relating to the average prescription drug class provided within a zip code. Since y is numeric, a quantitative variable rather than categorical, we use linear regression. If we test regressions of a categorical response variable, we may attempt logistic regression. For a linear regression model, we aim to fit our data closest to the line $y = f(x)$ but will of course have some variance. Our function (as $f: \mathbb{R}^n \rightarrow \mathbb{R}$), creates regression of $y = f(x) + noise$. As our $f(x)$ can be represented as $ax + b$, we aim to have our a and b close to 1 in order to closely configure our data to the line $y = f(x)$ as a linear function. These regression methods are utilized to test if there is a strong correlation between x and y or not. In this case, we aim to see if income relates to the given attribute of quantity, shown in Figure 6, or our target classification of addiction potential, shown in Figure 7. These target classes, used with other parameters, may be used for new data.

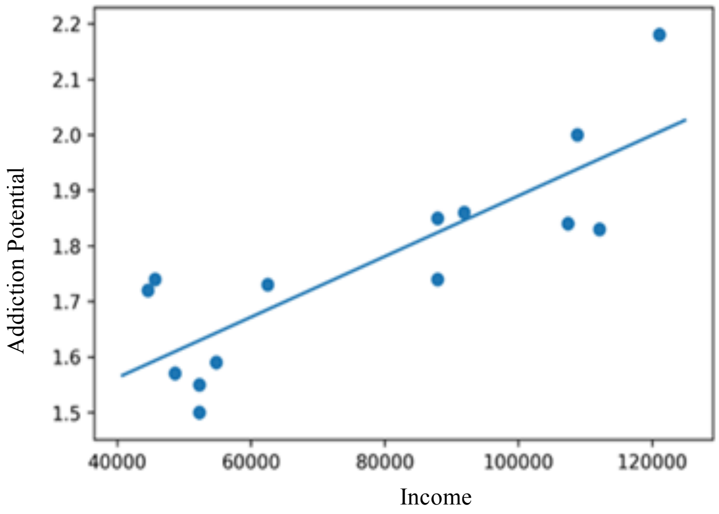


Figure 6 *Income vs. Addiction Potential*

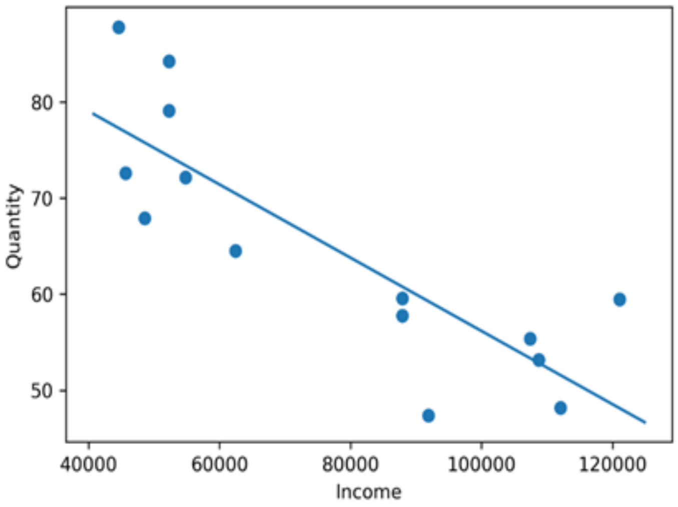
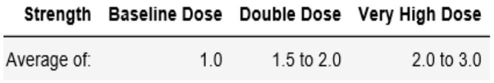


Figure 7 *Income vs. Quantity*

This plot demonstrates a strong correlation between these two pertinent feature engineered attributes. Note that low addiction potential values correlate to highly addictive classes of drug. We note that on average, lower income areas are prescribed the largest amount of narcotics and sedatives, the two most addictive drug classes in our sample. We also test the correlation between average income per zip code and the average quantity of pills provided for a prescription within that zip code, shown in Figure 7. We observe the highest quantities supplied are within the low-income areas. With further analysis, we can flag areas where drug addicts may be manipulating the pharmacies.

Also given in the second dataset is an additional attribute for the prescription drug strength. Most prescription strengths are defined in terms of milligrams and/or milliliters, however, for the prescription "Androgel" this attribute is listed in percentages. To remedy these differences, we research the baseline dosages for each prescription to be able to compare one drug dose to another and create levels of strength as a new attribute to use in our analysis. We define a baseline dose as prescription drug strength of 0.50-1.30, a double dose as 1.50-1.75, and 2.00-2.50 as a very high dose. The strength attribute is simplified as described in Figure 8. We analyze this new attribute in terms of numbers of cases, shown in Figure 9, for each drug strength dosage level to determine if physicians may be inadvertently contributing to prescription reliance, addition, and subsequent need for even stronger drugs on the streets.



Strength	Baseline Dose	Double Dose	Very High Dose
Average of:	1.0	1.5 to 2.0	2.0 to 3.0

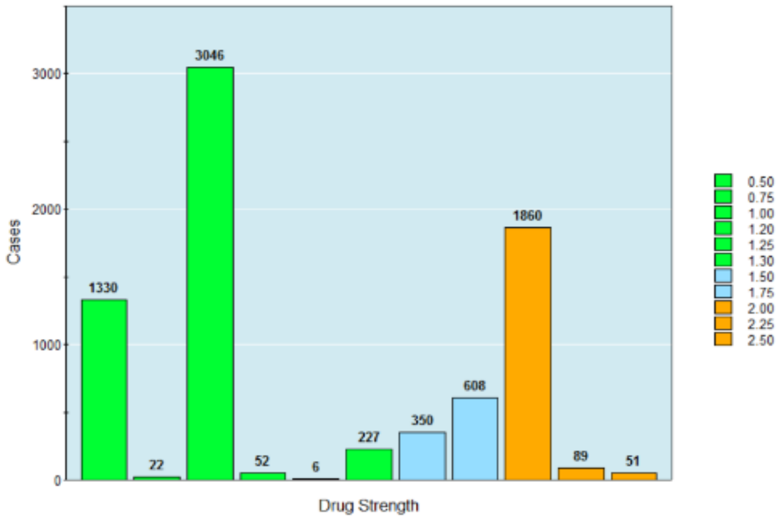


Figure 8 *Dose Strength Classes*

Figure 9 *Cases per Dose Class by Prescription Strength*

Seen in Figure 10, we are relieved to find that the biggest portion of drug strengths provided is within what we define to be a baseline dosage, accumulating 62% of the patient prescriptions we are considering. However, the second largest class is within what we describe to be a very high dosage, at 26% of cases in Ventura. Note that the cases considered are that of narcotics, sedatives, and amphetamines, which are strong prescriptions even at baseline dosages.

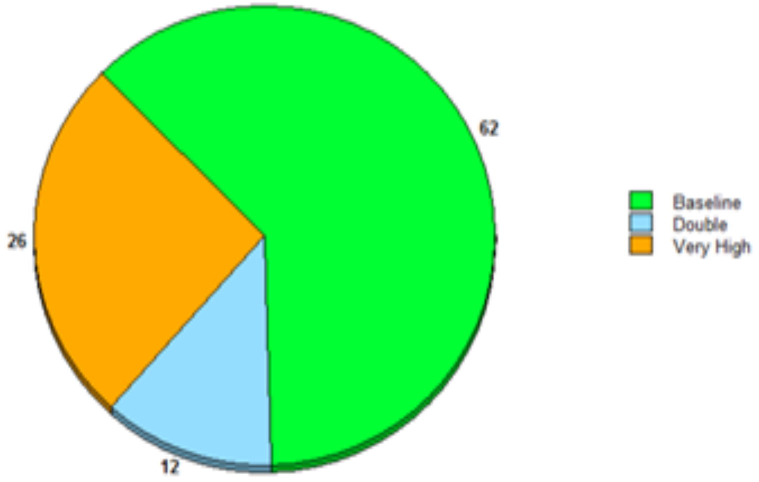
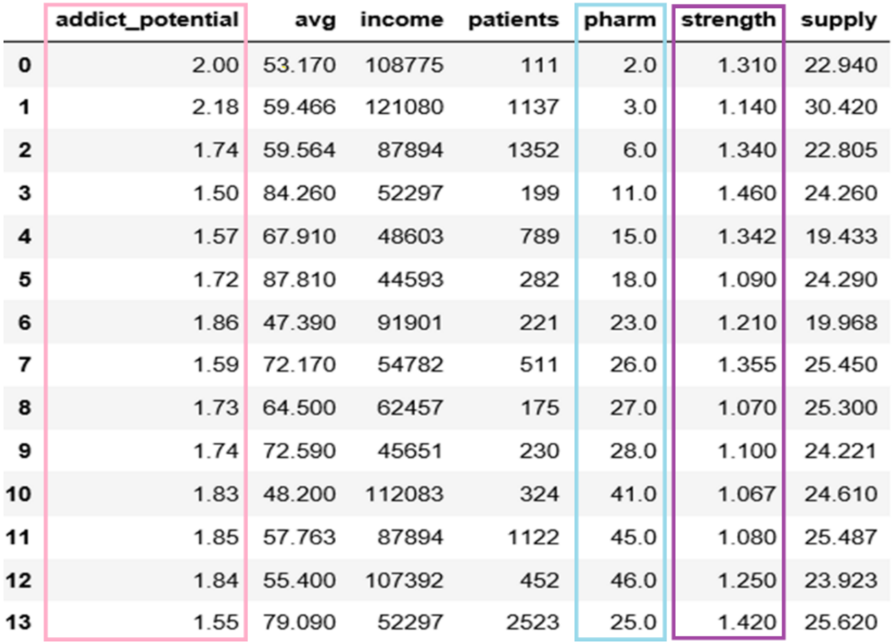


Figure 10 *Dose Class Percentages for Ventura County*

We compile the most important attributes from our datasets into a dataframe, shown in Figure 11, where the feature engineered attributes are highlighted. We include the feature engineered attribute of addict potential, created by the mean of the drug class, which is also feature engineered. We also consider the average quantity of pills, days supplied, income, drug strength, and pharmacy zip code to be pertinent to our investigations. Also included is the redefined zip codes related to these averages as well as the number of patients being supplied in each zip code. Using the information in this chart, we create a legend for color coding each zip code for areas of concern within Ventura County.



	addict_potential	avg	income	patients	pharm	strength	supply
0	2.00	53.170	108775	111	2.0	1.310	22.940
1	2.18	59.466	121080	1137	3.0	1.140	30.420
2	1.74	59.564	87894	1352	6.0	1.340	22.805
3	1.50	84.260	52297	199	11.0	1.460	24.260
4	1.57	67.910	48603	789	15.0	1.342	19.433
5	1.72	87.810	44593	282	18.0	1.090	24.290
6	1.86	47.390	91901	221	23.0	1.210	19.968
7	1.59	72.170	54782	511	26.0	1.355	25.450
8	1.73	64.500	62457	175	27.0	1.070	25.300
9	1.74	72.590	45651	230	28.0	1.100	24.221
10	1.83	48.200	112083	324	41.0	1.067	24.610
11	1.85	57.763	87894	1122	45.0	1.080	25.487
12	1.84	55.400	107392	452	46.0	1.250	23.923
13	1.55	79.090	52297	2523	25.0	1.420	25.620

Figure 11 *Pertinent Attributes*

(Author: Olivia)

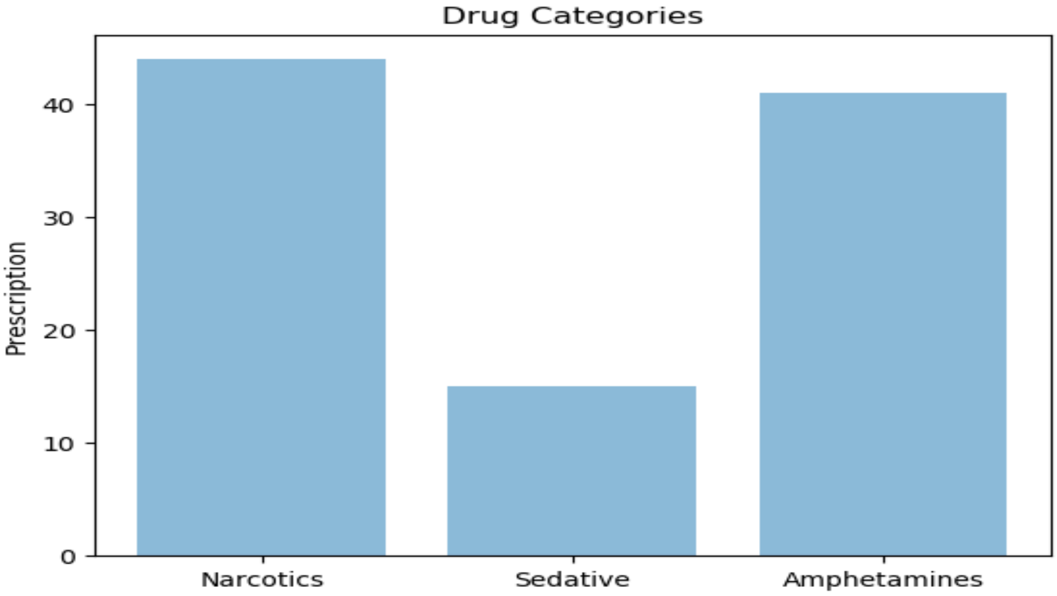
2. *Data Description*

The purpose of our data is to see if we can find any correlation between zip, gender, and age of patients who were prescribed drugs in Ventura county. The Ventura County Public Health, provided us with a data set containing 1600 patients, 55 drugs and 47 zip codes. However, after analyzing the data we notice a few discrepancies such as patient information from outside of Ventura County that eluded us to make cuts and classify the data for more accurate results. Considering some of the information given to us was either inconclusive or not in the area of interest that didn't help our research. We revised a new data set containing 1400 patients, 8 drug classifications and zip codes only in Ventura county. This new data set gave us more accuracy in finding the final results for our research and made it easier to analyze as a whole.

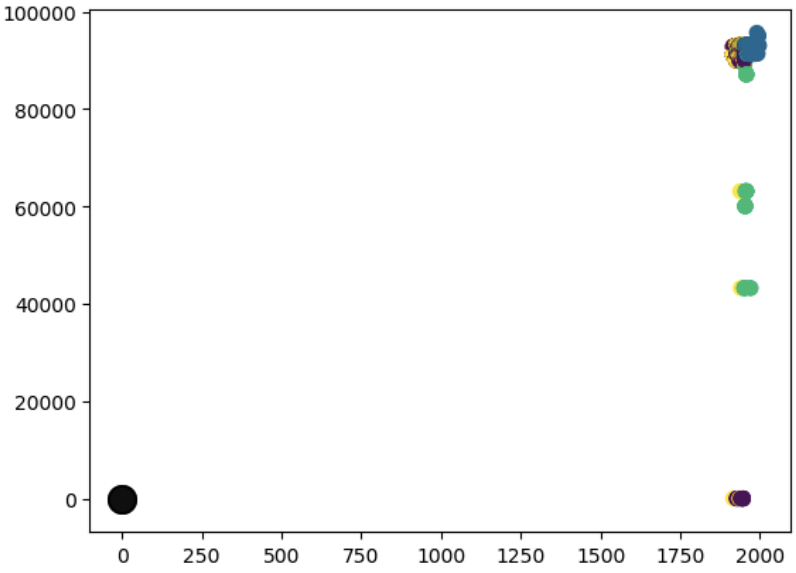
After making cuts and classifying our data, we had a well established data set for an observational study. Where we observed our observational units which were the patients prescriptions from the drug classifications, quantity given, average income in the area, and zip code. Which is where randomness plays in since all our patients are from the same area it's considered random sampling. Since the Ventura County Public Health provided us with the patients ultimately they were the ones who conducted the first random sampling to choose the patients in our data set. However we also conducted our own form by classifying the drugs and separating our patients by zip code to possibly flag areas. This was useful to measure the amount of drugs prescribed per patient in each part of the county.

3. *Descriptive Analysis*

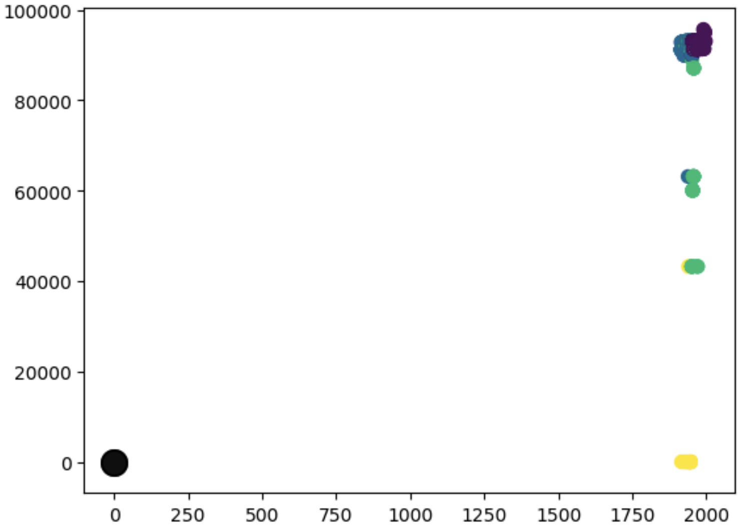
For further clarity, in measuring our variable and finding possible correlations we used plots and graphs to transform our dataframe. After revising our new data set we used python to create graphs to demonstrate the true impact of each drug classification prescribed. From this graphs we learned that three out of the eight there highly addictive drugs that were primarily prescribed which were narcotics, sedatives, and, amphetamines. We concluded, that narcotics were the most prescribed by doctors and are also the most addictive type of drugs, according to the FDA. This allowed us to focus on these three drug classifications instead of eight because these were the most prescribed and also the strongest addictive potential.



After looking at the Drug categories and seeing the amount of each drug type prescribed we realized that we could try to find more correlations to our data by using unsupervised clustering on python. This means we allow our data to run and form cluster on it's own. We began to develop this by using K-means clustering where we imputed our data to see if we could find four clusters to develop a map of Ventura County with four classifications to flag areas. However, we found little to no evidence to prove that there was four clusters in our data using K-means because two of the colors kept overlapping with one another and there wasn't a clear cluster forming that gave enough evidence for our data. Although it did provide us a good starting point to see that there was a possibility for our data to develop into four clusters.



Hence, we decided to try another method to see if we can find any other evidence for our research. Therefore we tried using Principal Component Analysis (PCA) before k-means clustering were we found significant amount of data to see that four clustering were indeed forming although there was still a little overlapping between some colors. Which gives us hope that we could keep on working and developing the research project and could possibly find the four cluster we are looking and prove that our results are in fact true. However using unsupervised clustering direct correlate with the information we wanted it. It does give us a good insight we were on the right track because some sort of clustering was forming in both K-means with PCA clustering.



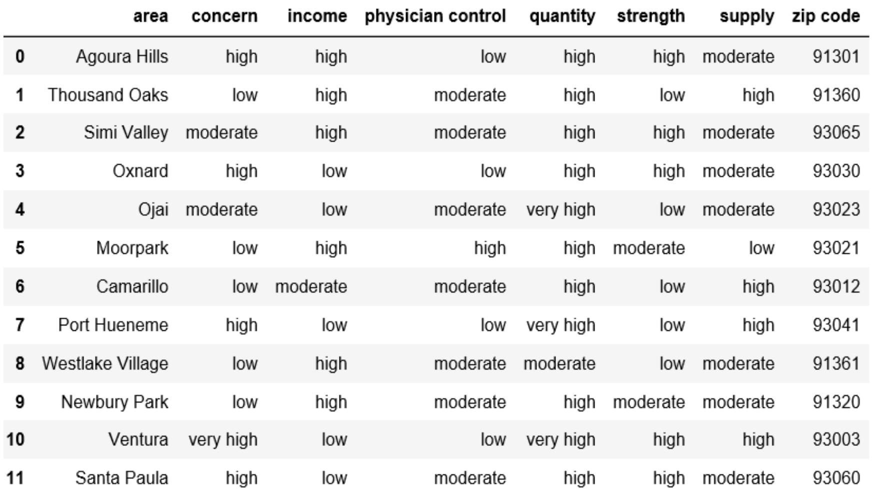
4. *Inferential Analysis*

After the data was loaded into the Dataframe, we began our analysis by plotting several pairs of attributes on scatter plots. Since our main focus is the exploration of behavior of the attributes with respect to each zip code, we reassagined each zip code with a smaller integer value (1-49) such that the scatter plots are easier to interpret. Furthermore, we explored the number of patients and the amount of drugs per type for each pharmacy. This gave us some initial understanding of the Dataframe, however, it was clear that new attributes were needed to generate conclusions. The mathematical methods we used to process the data were intended to create new attributes that we can then qualitatively interpret. Since our main emphasis was on creating attributes describing the data in new ways, with respect to the zip code in which each individual received their medication. In our effort to assess the possibiltiy of addiction each zip code, we created two new attributes, drug strength and addiction potential. We only used the patients in the data set that were prescribed either narcotics, sedatives, or amphetamines. Following findings from the FDA, we determined that opiate narcotics were the most addictive, followed by benazodaprine or similar sedatives, and amphetamines. We assigned a value the value 1 for narcotics, 2 for sedatives, and 3 for amphetamines. Analyzing the data we found that the majority of the patients received either narcots or sedatives. We averaged the previously mentioned values for each pharmacy, to create the addiction potential attribute. This attribute attempts to state the overall addictive potential of all drugs, as well as inherently the amount of narcotics and sedatives. Narcotics are our main focus, and the closer the addiction potential number is to 1, the zip code data set indicates a high percentage of narcotics.

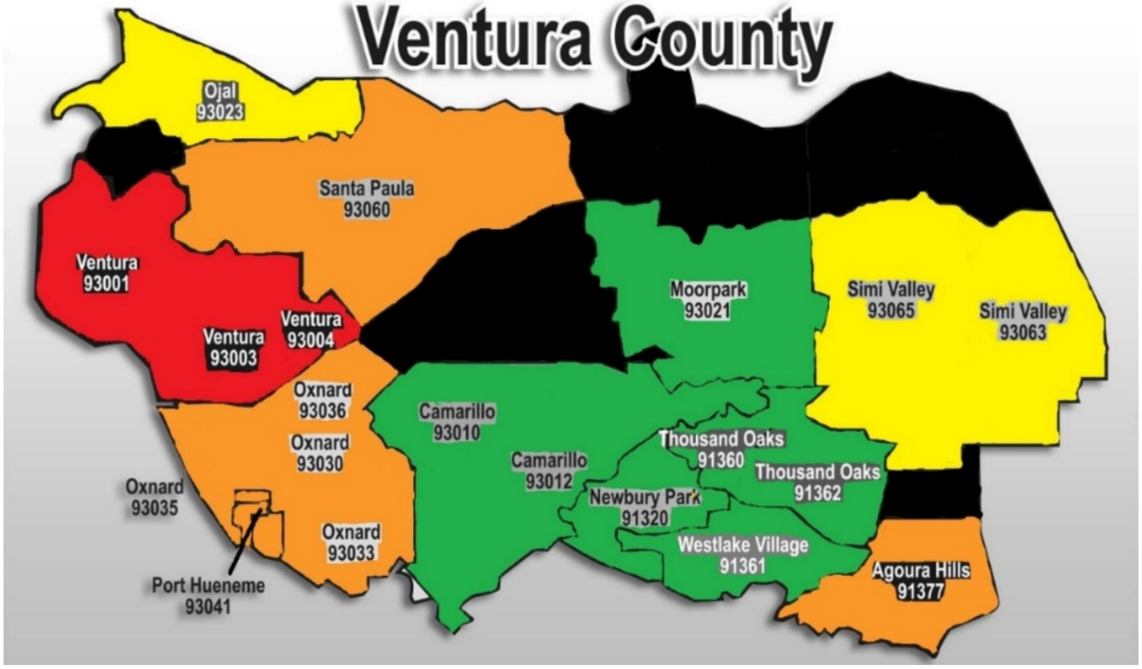
Our other attribute we created was drug strength based of the dosage, considering the nature of each medication. For example, 1mg of benzodaprines and 5mg of opiates are the most prescribed medications. We assigned numerical values to every drug's strength. Some doses far exceeded the average, such as 20mg of vicodin, or very large doses of fentanyl. The biggest challenge was cross referencing what would be considered baseline, or ”double” doses for every brand of medication in each category. We assigned the double dose a value of 1.5-2 and a high does as 2-3, with the baseline dose being given a 1.0. Averaging each patients medicine strength per zip code we were able to create the drug strength attribute.

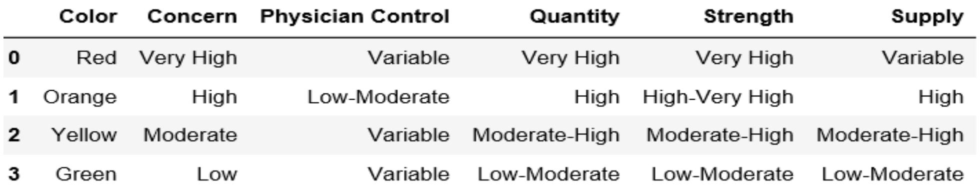
Utilizing these two attributes we found possible correlations between the income of each area and addiction potential, as well as income versus drug strength. This enabled us to take all of our attributes, (drug strength, income, quantity, days supply, addiction potential) for each pharmacy and assign qualitative values (low,medium, high, very high) for each attribute per pharmacy. We then set a certain range of values for each attribute that would ultimately decide the overall concern level per pharmacy. We noticed very strong indications that addiction potential is high in lower income areas. Ventura in particular, has been giving very strong narcotics to nearly every patient in good quantity. However, the days supply attribute indicated tht Ventura on average gave less than half the days of supply of every other zip code. We interpreted this as possibly physicians trying to control the problem, making it more difficult for addicts to find narcotics as their supply ran out. This observation in particular, led us to the creation of the physician control attribute. All in all, we color coded every zip code based off of

concern, with red being very high and green being low. This visual can ”direct” if you will, the researcher to areas of most concern, and guide them in creating new objectives, such as idenetifying individual patients. While, of course, there is an amount of bias in every conclusion, we feel strongly that our final product can indicate areas of most need.
(Author: Andrew)



	area	concern	income	physician control	quantity	strength	supply	zip code
0	Agoura Hills	high	high	low	high	high	moderate	91301
1	Thousand Oaks	low	high	moderate	high	low	high	91360
2	Simi Valley	moderate	high	moderate	high	high	moderate	93065
3	Oxnard	high	low	low	high	high	moderate	93030
4	Ojai	moderate	low	moderate	very high	low	moderate	93023
5	Moorpark	low	high	high	high	moderate	low	93021
6	Camarillo	low	moderate	moderate	high	low	high	93012
7	Port Hueneme	high	low	low	very high	low	high	93041
8	Westlake Village	low	high	moderate	moderate	low	moderate	91361
9	Newbury Park	low	high	moderate	high	moderate	moderate	91320
10	Ventura	very high	low	low	very high	high	high	93003
11	Santa Paula	high	low	moderate	high	high	moderate	93060





	Color	Concern	Physician Control	Quantity	Strength	Supply
0	Red	Very High	Variable	Very High	Very High	Variable
1	Orange	High	Low-Moderate	High	High-Very High	High
2	Yellow	Moderate	Variable	Moderate-High	Moderate-High	Moderate-High
3	Green	Low	Variable	Low-Moderate	Low-Moderate	Low-Moderate

Code Implementation

We experienced challenges in code implementation when attempting to rename attributes. To redefine zip codes into a numerical list, we was able to use a dictionary which took the values from the relevant column and reassigned its values. However, when attempting to use this same procedure for the prescription names, we was unsuccessful. When attempting to define several different prescriptions to the same value, in our case to a specific drug class, it was unable to run this information. We instead used a mapping code which not only classified the prescriptions into our desired classes but replaced this attribute column of our dataframe instead of creating a new attribute like the dictionary for zip codes did in our initial coding method.

In our second dataset, we are given general information for 14,000 patients. A large dataset is helpful in testing the accuracy of our classification models, target classes, and linear regression models. As we are focusing our efforts on patients in Ventura County, we built our dataframes to consider only zip codes within our county. Additionally, we are not considering incomplete patient files i.e. cases that were missing the prescription name or another pertinent attribute. These incomplete files are not appropriate for our analysis, therefore we are not focusing on these results. Even after focusing our dataframes, we are working with roughly 12,000 patients within the county from our second dataset as well as our original dataset of 1,600 which was constricted to 1,400 patients for similar reasons. There were an additional 50 prescriptions listed in the second data set that were not listed in the original. We added these prescriptions to our mapping dictionary after classifying their purpose.

5. *Conclusion*

Our Analysis revealed that our research question involves many more factors than initially predicted. Creation of new attributes is a difficult process, and much more can be done to solidfy these algorithms, as well as analyze them with the rest of the data set. Our findings indicated that certain attributes of an area could possibly dictate how and why drugs are prescribed the way they are. Income was shown to have a possible correlation with quantity of drugs administered, drug strength, and the feature engineered addiction potential in- dicator. However, there are many ways to represent a zip code quantitatively, and median income is just one. We cannot say we absolute certainty that we can derive a cause-effect conclusion, however, our findings seem to open the door allowing new questsions and methods to be implemented, in the vein of the work we have completed. Analying Ventura County is a prudent idea, however our algorithms can and should be expanded to many zip codes, allowing us to see if our findings in Ventura County are consistent with that of other counties in California. Our investigation went in a direction that we believe should be followed to legitimize our conclusions and answer new questions arising from the results, and processes of our findings. If we had more data, the legitimacy of our data may objectively improve. Furthermore, if we had data expanding our search to much larger parts of California, we could take our work in entirely new directions. Overall, we learned that analyzing prescription data opens the door to many questions, and possible methods of generating conclusions. It is an ever continuing process, and we hope more than anything that our conclusions can f the objective confidence in our variables and legitimize further analysis of the variables that we perceived could exhibit possible correlations.

There is bias in the subjective assignment of strength to each type of drug chosen. However, since there are generally baseline amounts prescribed, along with doses of double or even triple the most common dose, we were able to generalize this process for all the drugs chosen.

Furthermore, the attempts to derive conclusions using PCA and K-means was difficult. While there was significant clustering, our goal of finding four clusters based off our flagging attributes failed. However, if we are able to find methods of better categorizing data, we could possibly derive very solid results from whichever PCA method we find best suited for the data.

Unfortunately, our sample size is not large enough to generate very solid conclusions. Many of the zip codes given had less than 40 patients, which isn't enough to generate comparison to other larger zip codes. Second, we were not able to identify the amount of refills each patient was given, or sought. This is a good indicator of addiction. Finally, We were not given the condition of each patient. More serious conditions can justify a large amount of pills being prescribed. We were only able to flag pharmacies but with more information, we can flag individual patients.