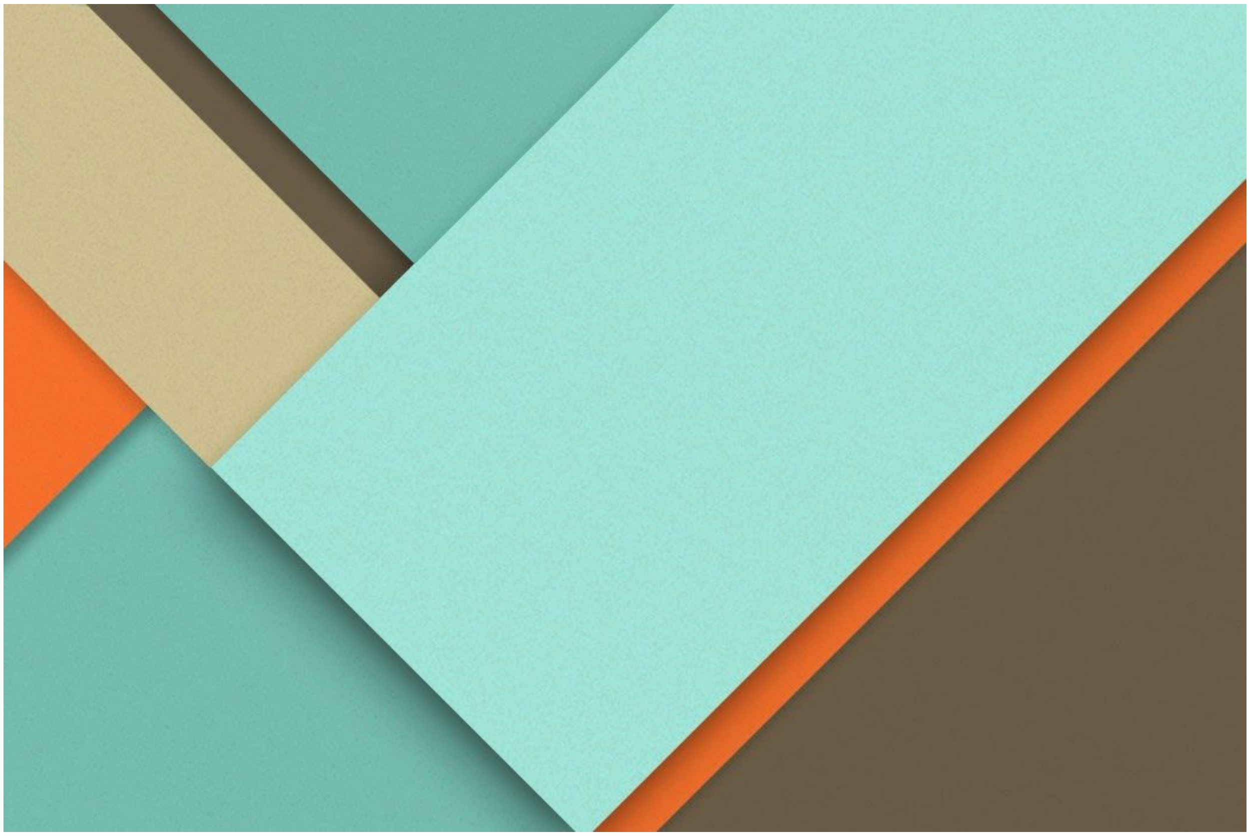


Visualizing prescription rate of pain-killers in America, in relation to Income, Housing, and Food

2020/01/15

Murali Behara
IBM Data Science Professional Certificate
Capstone Project

Table of Contents

1. Abstract

This project aims to visually explore the usage of pain-killer opioids in America. The intuition here is that those places in America with the availability of a richer variety & particularly of plant based food, combined with higher per-capita incomes & higher property values may be inversely related to the usage of prescription pain-killer opioids. Data about the prescription rates, individual incomes & home values are gathered by county in each state, & using the geolocation data, heat maps are created to analyze & validate the intuition.

2. Introduction

In a recent significant congressional hearings on an unrelated political matter, the number of times the congressional representatives mentioned opioid crisis as an urgent problem that deserves a joint congressional action sounded more significant. The use of pain-killer opioids as a crisis particularly related to the prescription rates has been in the news for a few years now, and it behooves one to ask why so many are in so much physical pain.

The aim of this project is to utilize location data to investigate and identify potential relationship(s) between the prevalent use of pain-killer opioid crisis, in relation to food, income & housing. It's assumed that the above essentials have some bearing on the opioid crisis, but a more thorough investigation to establish any causal links is outside the scope of this project.

The intended audience for this report are economists who are interested in understanding how one's quality of life in terms of physical pain & access to good food may impact one's earning potential & indirectly the property values in a geography.

Analysis - This project relies on historical data from reputable sources to answer any questions and hence machine learning to identify the patterns, relationships and trends in historical data that are otherwise understood mostly only as anecdotal stories. The goal here is to show relationships & hence the use of clustering, a descriptive model.

Diagnosis - Most problems in Data Science are classification problems. The analysis may help continue this line of investigation to perhaps predict how likely, a pain-killer will be prescribed based on access to better food, better income & better habitat.

Prediction - Predicting the probabilities or the likelihood of any of the direct or indirect variables like the increased access to better food, better paying jobs & better habitat is outside the scope of the project and therefore any speculation about the trends & future trajectories.

Prescription - Prescriptive solutions to the pain-killer opioid usage are also outside the scope of this project.

3. Literature Review

Do the basics of life, as in income, habitat and food have any bearing on one's physiological pain. I wanted to pursue this question and utilize data and analysis to answer and perhaps even discover new factors & lines of inquiry. During the recent congressional hearings about impeachment the number of times opioid crisis was mentioned by congress persons as an urgent matter to be prioritized and deserving time and resources was unmistakable. The abuse of pain-killers particularly, the high rates of prescribing the same has been in the news and under the microscope for a few years now. Pain is also cited to have a significant impact on the economy in Brookings Papers on Economic Activity, in the 2017 Fall edition, in an essay titled, Where have all the workers gone? An inquiry into the decline of the U.S. labor force participation rate, published by Alan B. Krueger, Bendheim Professor of Economics and Public Affairs of Princeton University. So this may be a worthy pursuit. We've all seen advertisements indicating that inflammatory pain particularly arthritic pains have something to do with life style including access to healthy food and the opportunity & effort to stay physically active. It is common knowledge that vegetables and fruits have lower inflammation rates compared to animal derived foods including meat and dairy. According to the physicians committee for responsible medicine, plant based food is recommended to ease arthritis pain, and this may particularly help those with chronic arthritic conditions such as gout. In a systematic review of effects of plant-based diets on body and brain using intervention research techniques published in nature science journal, it is observed that there's robust evidence for benefits of plant-based diets including weight loss, energy metabolism and lowering systemic inflammation. All of the above leads me to believe that access to vegetarian food must play a significant role in level of pain and management of the same. In my experience, fresh produce is also more expensive than most of the factory farm produced meat and dairy based food, artificially lowered, discounting the cost to our environment, fragile living ecosystems, health care and the economy itself. Thus in addition to the popularity and availability of vegetarian food, the ability to afford the same may also depend on economic well being. Income & property values in a particular geographic region are generally good indicators of economy & quality of life in that region. Thus in addition to the popularity and availability of vegetarian food, the ability to afford the same may also depend on economic well being. income & property values in a particular geographic region.

4. Methodology

It seems to be common knowledge that pain is physiological aspect of life can be attributed to lifestyle which in turn seems to largely depend on food & fiscal wellness. But that's an **ASSUMPTION**, and we'll make a number of other assumptions both qualitative & quantitative, and will take the help of **Machine Learning** to validate the assumptions without torturing the data, at the same time doing the necessary & due diligence to find out the **story hidden in the data**.

DataScience problems are also broadly categorized into classification or regression problems. Here in this project, we will utilize machine learning in an attempt to classify each county in the United States as either Painful or Painfree based a chosen threshold of prescription rate; all the counties that have a **pain-killer prescription rate of 50% or below will be categorized as painfree counties** & the rest as painful. To estimate the prescription rate is a larger undertaking but given the common knowledge about pain, it seems logical to investigate if there are parameters that could predict a geographical region as either painful or pain free.

First we will have to ensure that all the baseline features are collected and the data is complete; for instance there are **over 3000 counties** that CDC has data about painkiller prescription rate; The bureau of economic affairs per capita income data on almost all of the counties, however, when it comes to **median home values** we have data about less than 2000 counties. So we have to estimate the median home values for the remaining counties. The reason why we **consider income & property values** is to catch the two important economic criteria for fiscal well being. What we mean is that those who have inherited wealth may have mostly capital gains income which is reinvested & may under report the income but factoring in the home values may help the data to be less skewed. We will first use **folium & choropleth maps** to visualize the datasets with relevance to geolocation.

Secondly we will utilize a REST style Web services API to an external data source to retrieve geo location based information. **The department of health & human services of which CDC is part of, also recommends a healthy serving of vegetables & fruits. Vegetarian food is known to lower inflammation**. So using **FourSquare API** we will then retrieve the **availability of vegetarian food within 10 kilometer radius from the geographical center of a county(another assumption because some counties in the western states are too big & the geographical centers do not coincide with the business centers)**.

Third and perhaps the most important part of the project is where we will utilize a classical kernel method called **Support Vector Machine** to model the data and predict a county as either a pain free county or otherwise based on availability of vegetarian food, higher per capita income, & higher median home values, and evaluate the accuracy of the model using Training & Testing data sets.

Data requirements - To answer the questions about the use of **prescription pain-killer opioids**, the prescription rates along with geolocation data is required. The geolocation based prescription data helps join other relevant geolocation information such as **availability & type of food**, **personal income** & **home prices** to make any comparisons & potential associations.

Data collection - Multiple data sources will be utilized to find the associated information. For this particular project, most recent historical data about prescription rates & per-capita personal incomes, property prices will suffice for data analysis. All the data is available as downloadable files. Foremost the project recommends & stipulates use of data from FourSquare, accessed via a REST API.

Finally, all the data collected is staged in the IBM public cloud object storage (COS), that's also publicly available.

United States Health & Human Services provides information & navigation links to the **pain-killer opioid prescription data**, available from the **Centers for Disease Control (CDC)**. However the file is not downloadable but the table can be scraped from the Web page.

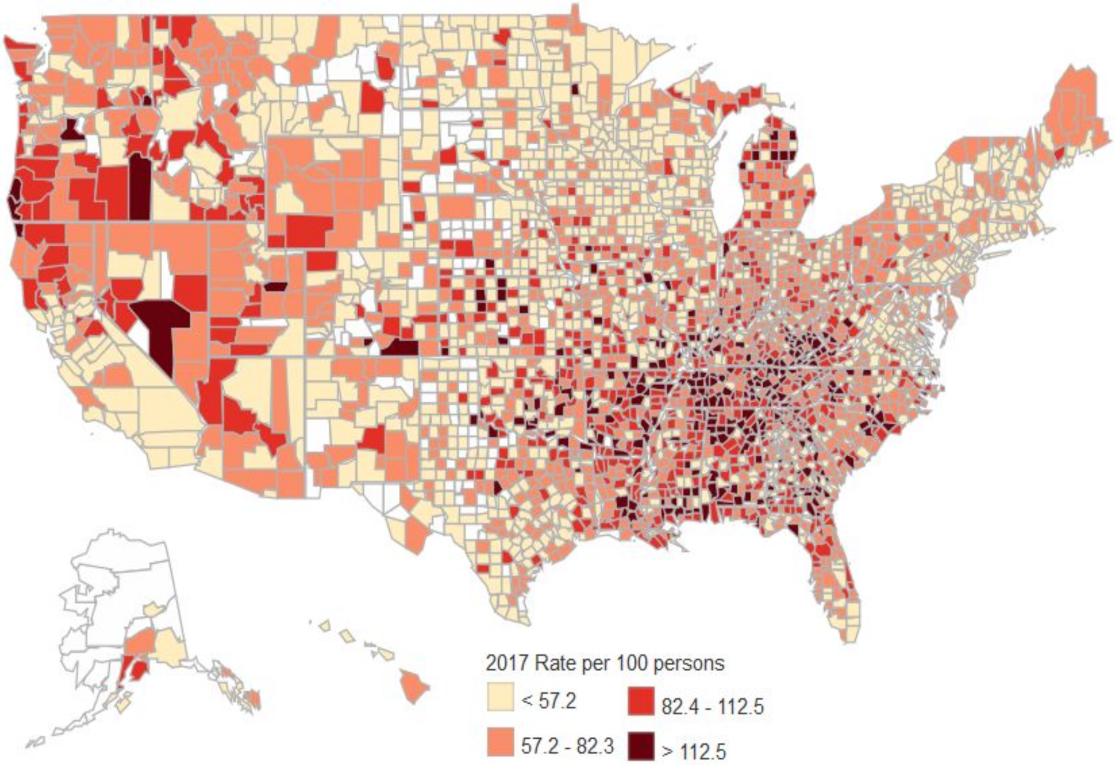
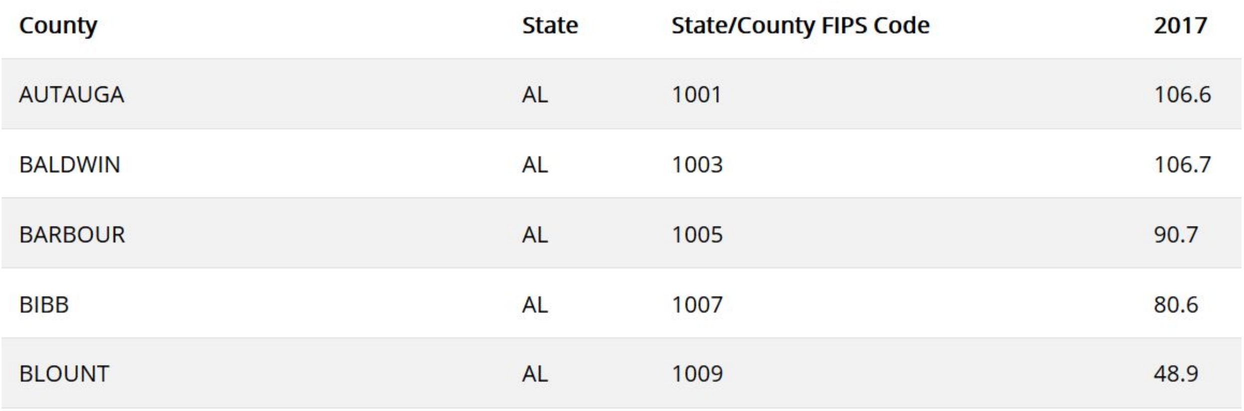


Figure 1 - US county wise pain-killer opioid prescription rate (percent) in 2017



County	State	State/County FIPS Code	2017
AUTAUGA	AL	1001	106.6
BALDWIN	AL	1003	106.7
BARBOUR	AL	1005	90.7
BIBB	AL	1007	80.6
BLOUNT	AL	1009	48.9

Figure 2 - Tabular format of data showing County, State, County Code & Prescription Rate

Thanks to the **US Bureau of Economic Affairs**, we can obtain **personal income data by county, & by metropolitan areas**. Below is the map of income data by US counties for fiscal year spanning 2017-2018.

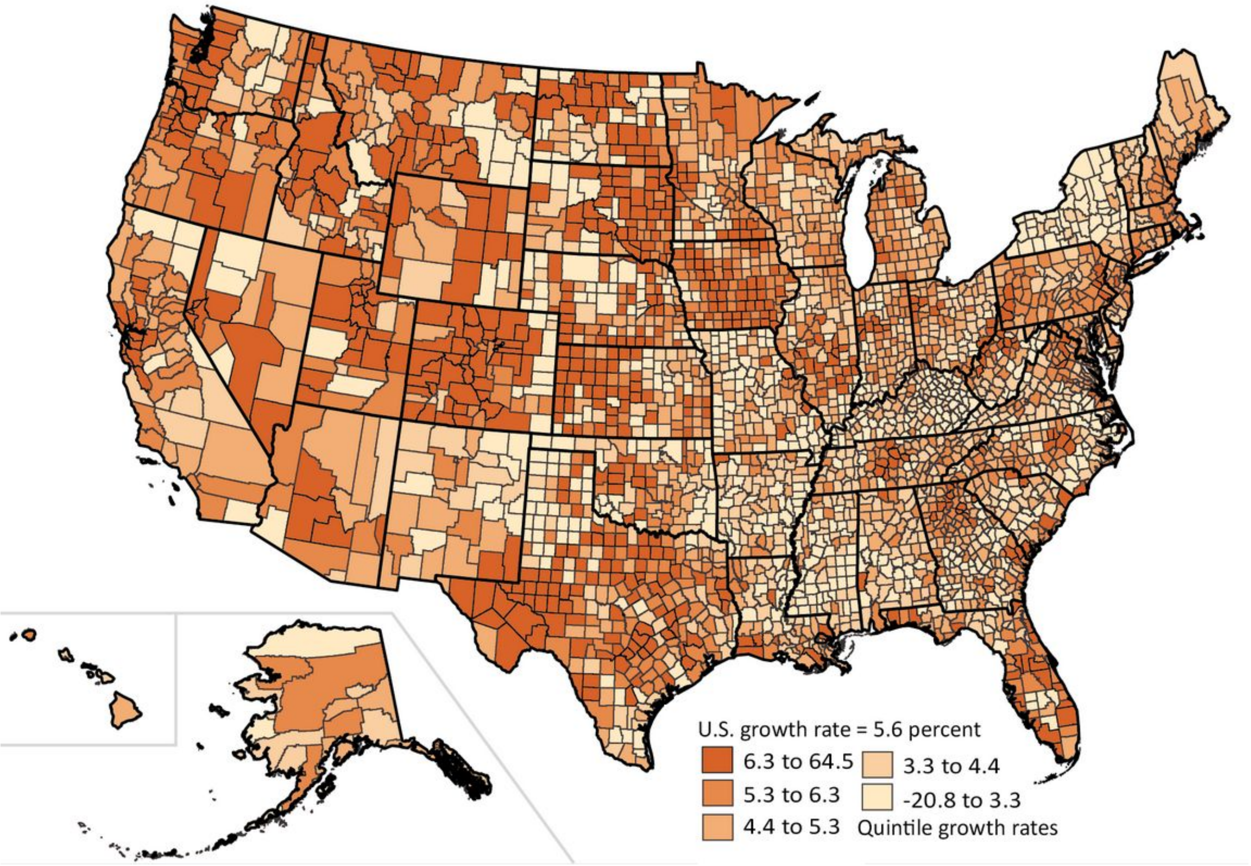


Figure 3 - US county wise personal income growth rate between 2017-2018

Zillow is a popular & a reputable private sector organization that publishes the popular annual **median home values**. The data is made available for download & analysis. The data that will be analysed as a part of this exercise is from the year 2017.

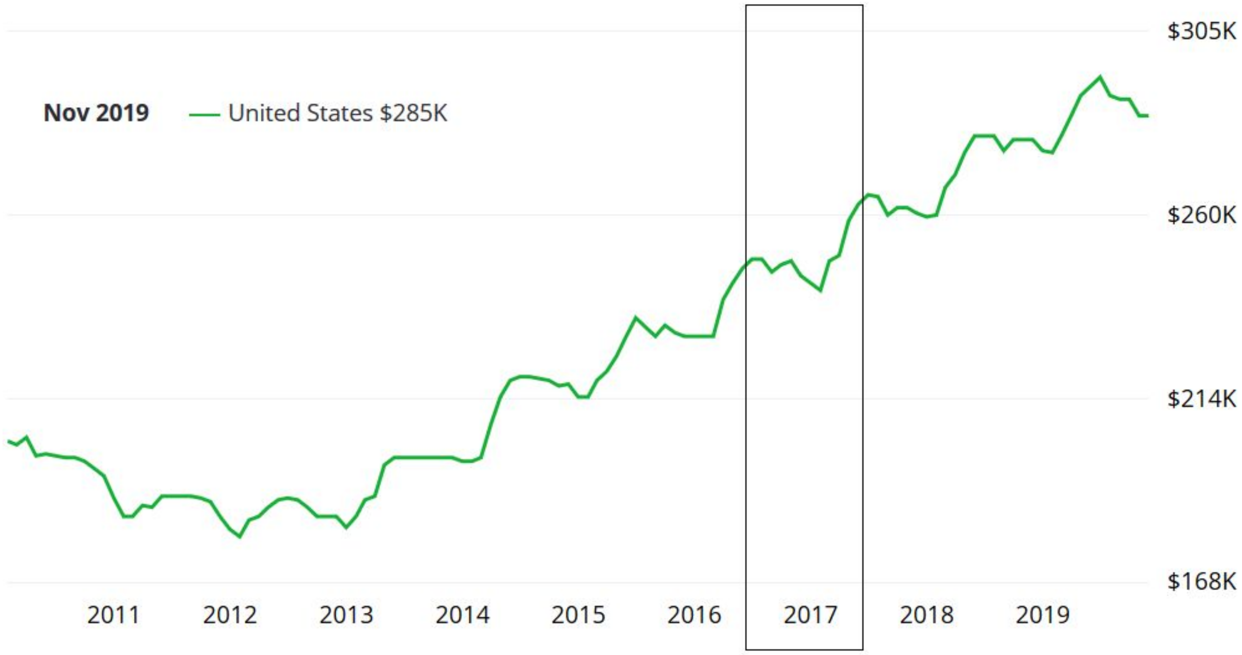


Figure 4 - US median home prices from the year 2011 through 2019

FourSquare offers an API to search venues across the United States using longitude and latitude. From the interactive map, a search for affordable plant based food is available only in select cities as shown below.

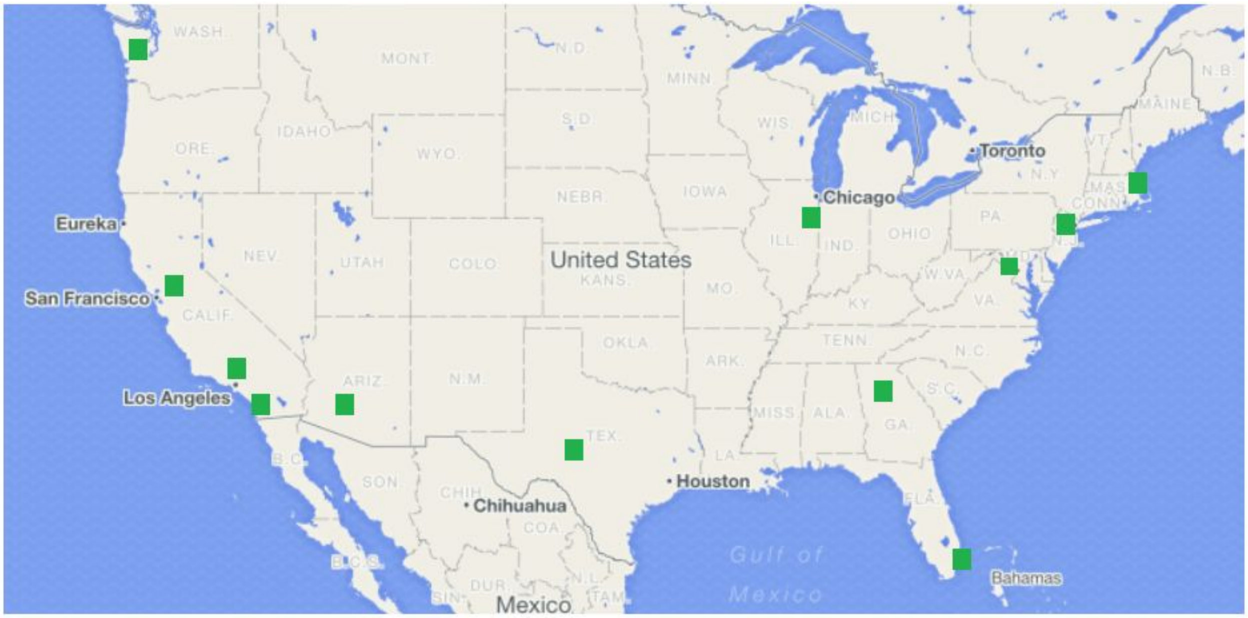


Figure 5 - FourSquare Map showing the cities where vegetarian food popular in the US

The idea is to use the main longitude & latitude intersection for each of the 3000+ counties in United States to visualize a heat map displaying the availability of affordable plant based food (vegetarian), ethnic, standard chain restaurant food & fast food color coded in just 4 categories shown below as an example.

Green	**Orange**	**Red**	**Dark Red**
Gastronomy Restaurant	Chinese Restaurant	Fish & Chips Shop	Burger Joint
Vegetarian Restaurant	Indian Restaurant	American Restaurant	Fast Food
Sushi Restaurant	Japanese Restaurant	Seafood Restaurant	Pizza Place
	Mediteranian Restaurant	Mexican Restaurant	Wings Joint
	Burrito Place	Italian Restaurant	
	Sandwich Place		

Data understanding - As a first step to conduct data science, data needs to be visualized & understood to ensure that it is of appropriate quality & quantity. Descriptive statistics is a quick way to get the standard measures of mean, median and mode, variance & standard deviation. Almost all data analysis tools from spreadsheets to python offer functions to execute descriptive statistics. Box plots are a great way to visualize the descriptive statistics. Seaborn package can be utilized to visualize the data. Then of course data can be sorted into bins to create histograms. There are at least 3 main data sets & a cross-reference dataset that will be utilized in this project. Let's visualize each of the datasets separately.

CDC's Prescription Rate data table by county

```
RangeIndex: 2969 entries, 0 to 2968

Data columns (total 4 columns):
CountyName          2969 non-null object
StateCode           2969 non-null object
CountyCode          2969 non-null int64
PrescriptionRate    2955 non-null float64
```

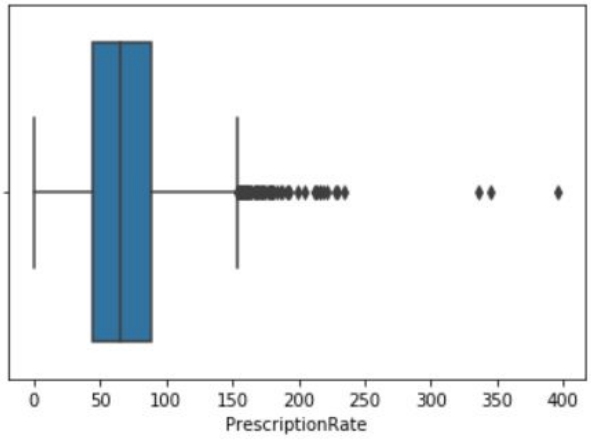
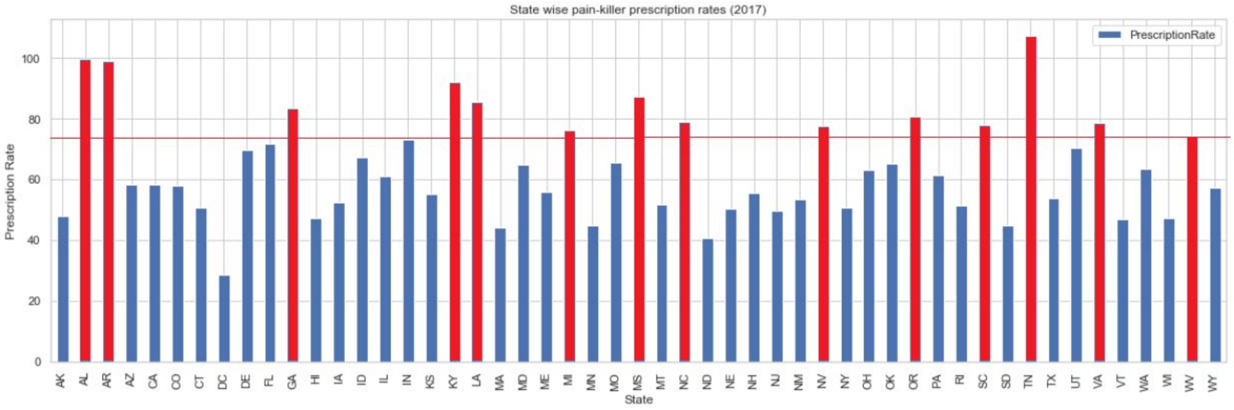
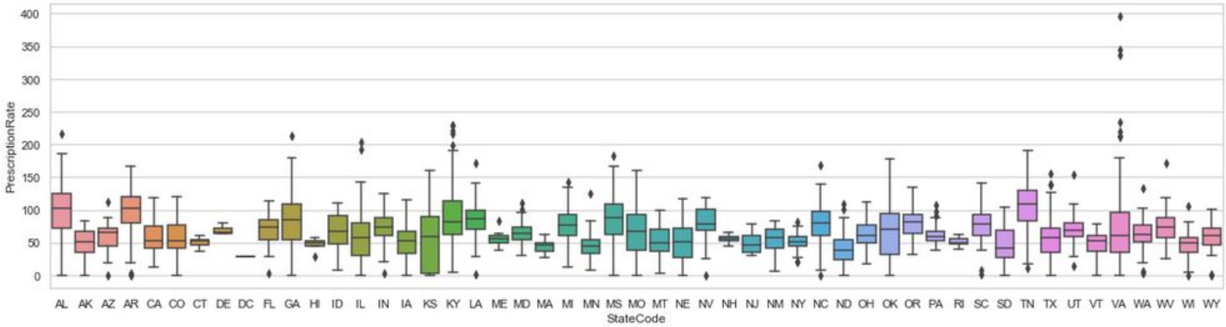


Figure 6 - Prescription Rates Table description & a boxplot summarizing prescription rates



Quick visualization of pain-killer prescription rate table shows that data is gathered for for 2969 out of 3000+ counties, and missing data for reporting counties. Box plot is a shows that the median prescription rate is close to 75%.



Aggregating the prescription rate by state, it appears that over a dozen states have higher than median prescription rates; Alabama, Arkansas, Georgia, Kentucky, Mississippi, Missouri, North Carolina, Oklohama, Tennessee, Texas & Virginia are interesting states with long whiskers and more than 50 prescriptions for every hundred people! So what's going on in these states?

US Bureau of Economic Affairs - Income Per Capita by County - The descriptive statistics output shows us the following.

```
count      3,080.00
mean      44,112.22
std       12,738.97
min       18,541.00
25%       36,557.75
50%       41,961.00
75%       48,748.50
max      251,728.00
```

It seems that the average per-capita income in the U.S., during 2017-2018 is roughly over $44,000 for all 3,080 counties listed in the table for all 50 states.

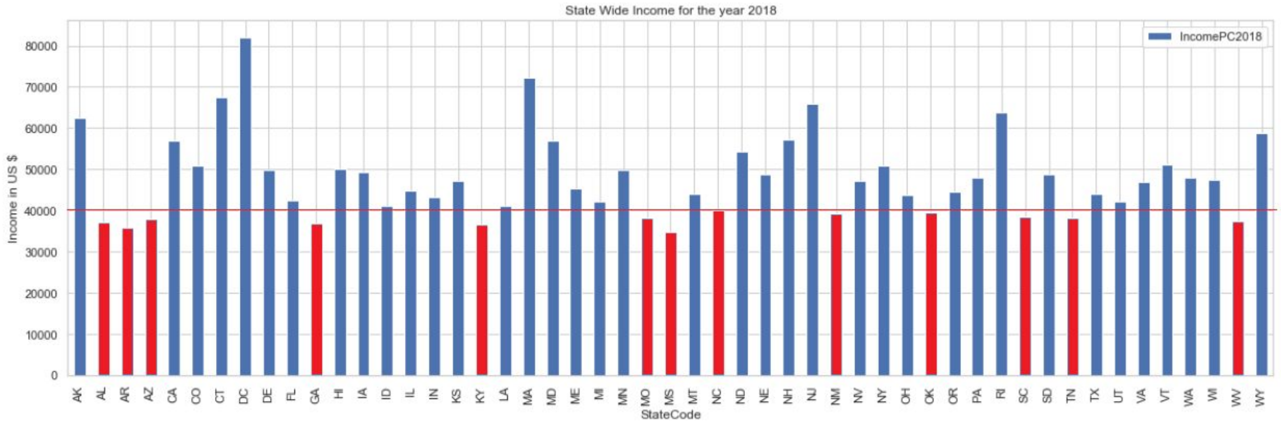


Figure 7 - Simple bar chart summarizing the average per capita income for each state

Contrasting with Figure 1, it is quickly apparent that this line of investigation may after all be fruitful to make qualitative statements, as one can see a few of the states with income below average, all have high pain-killer prescription rates, & the District of Columbia, the northeastern states all have significantly higher levels of income & visibly lower pain-killer prescription rates.

Zillow median home values - Quick descriptive statistics report of the home values table indicates that data for 1979 counties were recorded and average median home value is roughly $225K, in the year 2017.

```
count     1.979000e+03
mean      2.251729e+05
std       1.658781e+05
min       5.000000e+04
25%       1.399000e+05
50%       1.855385e+05
75%       2.640467e+05
max       2.772500e+06
Name: MedianHousePrice, dtype: float64
```

The barchart here below shows a summary of median home values, averaged for all the counties in a state. While it may not be surprising that although not proportionate the personal income roughly corresponds to home values; more interesting observation perhaps is one of qualitative and again one would see that roughly speaking the pain-killer prescription rates & median home values are inversely related.

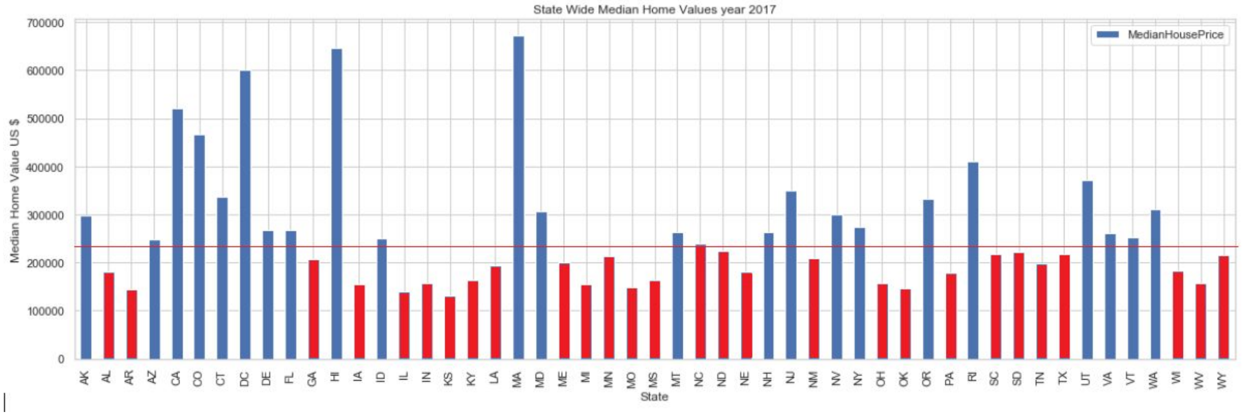
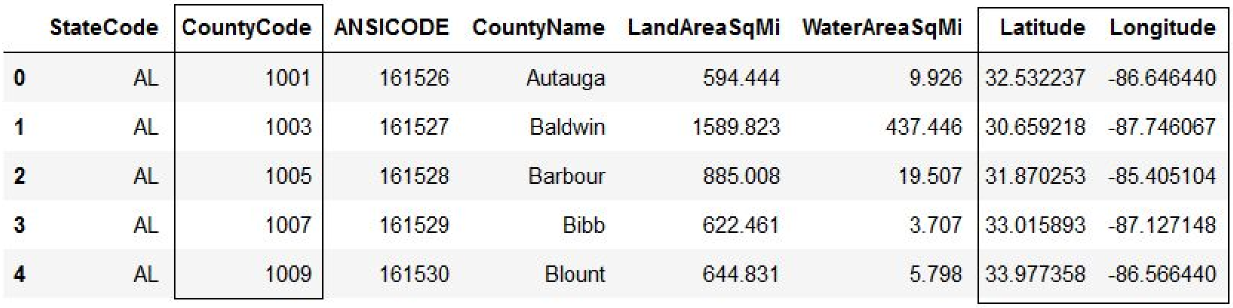


Figure 8 - Simple bar chart summarizing the median home value in each state

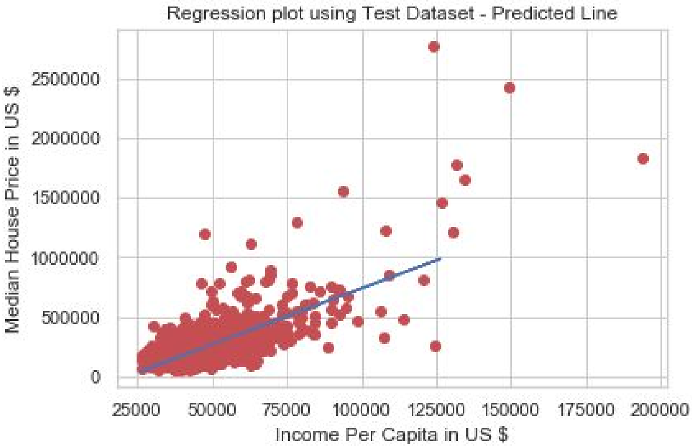
Data preparation - Data in the wild almost always contains missing or inconsistent data; there may be a need to join data using foreign keys as in this case & last but not least data needs to be normalized to get a sense of proper scale when comparing two quantitative entities on scales of very different magnitudes. A cross reference dataset will be also utilized to cross-reference Fed. Information Processing Standards (FIPS) county code, to obtain geographic latitude and longitude information. The cross-reference data is provided by the U.S. Census Bureau as a tables called the U.S. Gazetteer Files, providing geographical details including the land area, latitude and longitude for each county.



	StateCode	CountyCode	ANSICODE	CountyName	LandAreaSqMi	WaterAreaSqMi	Latitude	Longitude
0	AL	1001	161526	Autauga	594.444	9.926	32.532237	-86.646440
1	AL	1003	161527	Baldwin	1589.823	437.446	30.659218	-87.746067
2	AL	1005	161528	Barbour	885.008	19.507	31.870253	-85.405104
3	AL	1007	161529	Bibb	622.461	3.707	33.015893	-87.127148
4	AL	1009	161530	Blount	644.831	5.798	33.977358	-86.566440

Figure 9 - FIPS County Code can cross reference Latitude & Longitude

Let's first extrapolate the home values based on per-capita income using Linear Regression. Inner join the 1892 Zillow Home values with income per capita. Predict Home Values using Income Data. Split records into Training & Testing datasets (for linear regression) using SciKit Learn. Use Linear Regression model of (Sci-Kit Learn) estimate house values.



r2_score() of sklearn to compute the correlation coefficient

$$R^2 = \left(1 - \frac{Residual\,Errors\,Squared}{Mean\,Errors\,Squared}\right)$$

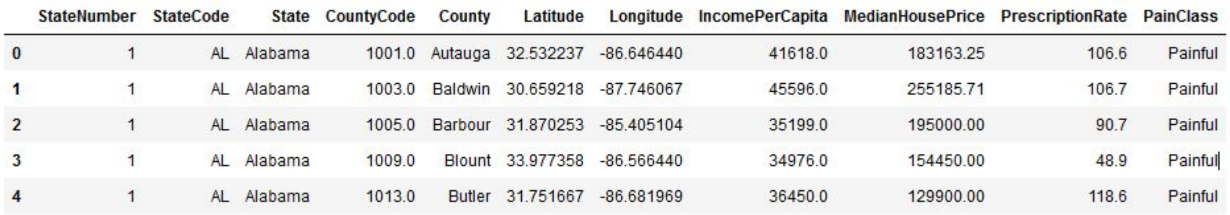
$$R^2 = 1 - \frac{(Y_i - Y_f)^2}{(Y_i - Y_m)^2}$$

Residual Errors Squared is $(Y_i - Y_f)^2$

Mean Errors Squared is $(Y_i - Y_m)^2$

Figure 10 - Correlating Per Capita Income to Median House Prices across US counties

Not the best but Median Home value is definitely correlated to the Income, which accounts for 72% of all the variation. Reproduce a heatmap using the data we've collected. We then proceed to create a complete feature dataset as shown below.



	StateNumber	StateCode	State	CountyCode	County	Latitude	Longitude	IncomePerCapita	MedianHousePrice	PrescriptionRate	PainClass
0	1	AL	Alabama	1001.0	Autauga	32.532237	-86.646440	41618.0	183163.25	106.6	Painful
1	1	AL	Alabama	1003.0	Baldwin	30.659218	-87.746067	45596.0	255185.71	106.7	Painful
2	1	AL	Alabama	1005.0	Barbour	31.870253	-85.405104	35199.0	195000.00	90.7	Painful
3	1	AL	Alabama	1009.0	Blount	33.977358	-86.566440	34976.0	154450.00	48.9	Painful
4	1	AL	Alabama	1013.0	Butler	31.751667	-86.681969	36450.0	129900.00	118.6	Painful

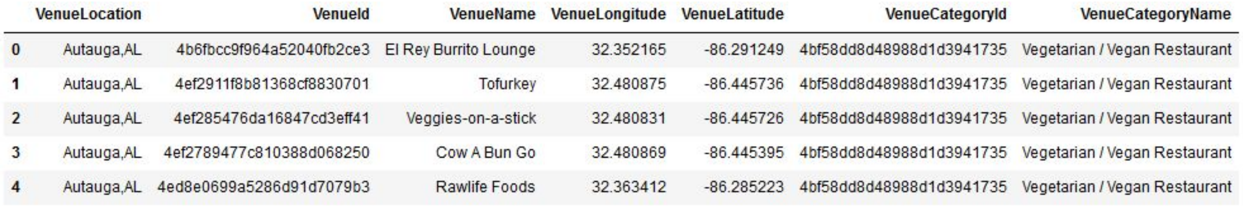
Figure 11 - Feature DataFrame displaying the features and predicted classes (2)

Use Four Square API to retrieve the type of food (restaurants) available in each of the counties. Set the variables for Four Square API credentials.

```
FoodCategory = '4d4b7105d754a06374d81259'
      FastFood = '4bf58dd8d48988d16e941735'
      Barbeque = '4bf58dd8d48988d1df931735'
```

```
American = '4bf58dd8d48988d14e941735'

FriedFood = '4d4ae6fc7a7b7dea34424761'

Burger = '4bf58dd8d48988d16c941735'

Pizza = '4bf58dd8d48988d1ca941735'

Wings = '4bf58dd8d48988d14c941735'

Steak = '4bf58dd8d48988d1cc941735'

Salad = '4bf58dd8d48988d1bd941735'

Vegetarian = '4bf58dd8d48988d1d3941735'
```

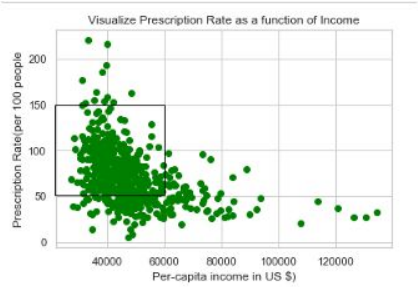
To simplify the problem we will focus on just the Vegetarian restaurants. Using Four Square API to explore locations, I then proceeded to retrieve the info about the availability of vegetarian food as in the number of vegetarian restaurants (category id = '**4bf58dd8d48988d1d3941735**'), within 10 kilometer radius from the center of each county (limiting to 10 restaurants); once again making an assumption. The API surely returned 2400 vegetarian restaurants, but also returned other types restaurants, with an overall accuracy of 88%.

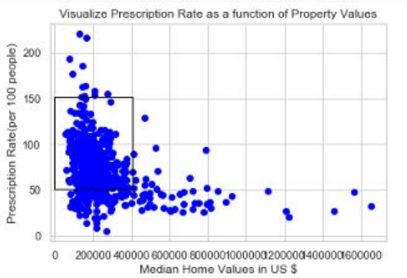


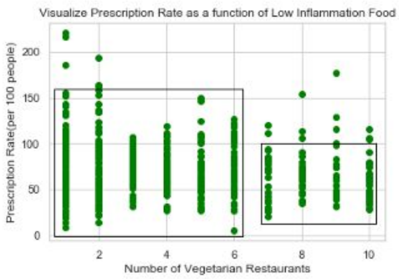
	VenueLocation	VenueId	VenueName	VenueLongitude	VenueLatitude	VenueCategoryId	VenueCategoryName
0	Autauga,AL	4b6fbcc9f964a52040fb2ce3	El Rey Burrito Lounge	32.352165	-86.291249	4bf58dd8d48988d1d3941735	Vegetarian / Vegan Restaurant
1	Autauga,AL	4ef2911f8b81368cf8830701	Tofurkey	32.480875	-86.445736	4bf58dd8d48988d1d3941735	Vegetarian / Vegan Restaurant
2	Autauga,AL	4ef285476da16847cd3eff41	Veggies-on-a-stick	32.480831	-86.445726	4bf58dd8d48988d1d3941735	Vegetarian / Vegan Restaurant
3	Autauga,AL	4ef2789477c810388d068250	Cow A Bun Go	32.480869	-86.445395	4bf58dd8d48988d1d3941735	Vegetarian / Vegan Restaurant
4	Autauga,AL	4ed8e0699a5286d91d7079b3	Rawlife Foods	32.363412	-86.285223	4bf58dd8d48988d1d3941735	Vegetarian / Vegan Restaurant

Figure 12 - DataFrame containing information about Vegetarian restaurants

Needless to say the results were cleansed. Also, the API did not retrieve all the restaurants but at this point I had enough data points (for 512 counties) to be able to predict pain classification, essentially as either Painfree or Painful based on an arbitrarily chosen but a reasonable threshold value for pain-killer prescription rate of 50 for every 100 persons.







Income, Habitat, Food See where most of the data points are & the number of vegetarian restaurants where prescription rate is above and below 100%. We could stop the analysis here and draw some broad conclusions without implying causal links but what I was interested in this capstone project was to see how much does the food factor into the prescription of pain-killers.

5. Results

Utilizing Support Vector Machines (SVM), a kernel method we will see if there's an optimal boundary between the sets of data points that collectively describe pain as a combination of income, property value, and non-inflammatory food (vegetarian). The reason why SVM is an appropriate model is the suitability of the problem, where the interest is to separate feature rich data manifold into two or more classes using a hyperplane. In our case we have just two classes. Painful and Painfree.

So, using the above 3 features in a training data set, I constructed a Support Vector Machine (SVM), a **kernel method** to see if there's an **optimal boundary** between the sets of data points that collectively described pain as a combination of income, property value, & non-inflammatory food, and trained the model to classify the data as either Painful or Painfree . Using the model and the test data (independent variables) that the model has not seen, I made predictions & compared with actual class; the model was able to predict a Painful county with 72% accuracy; however the model was unable to make any predictions about Painfree counties based on the features. As a sanity check, I then recreated the model with just 2 features, income and habitat and retested the model and surprisingly it performed the same suggesting that food may not after all factor into the high rate of prescription pain-killers, not to say much about pain & inflammation itself.

6. Discussion

The ability to digitally record data and the ubiquity of the same, combined with the availability of very sophisticated open source software allows more of us to apply complex scientific methods to investigate a great variety of real world issues. Before diving in and applying analytical techniques to data there is a tremendous amount of legwork to get all the data right in the desired format ready for analysis. This effort is referred to in our business as data wrangling. While it was surprising to learn that income and house values alone can classify if the pain killer prescription rate is above or below 50%, it also invites one to look closely and carefully at some of the assumptions made and how data is normalized. The Four Square API returned 2400 vegetarian restaurants, but also returned other types of restaurants, with an overall accuracy of 88%. The API did not retrieve all the restaurants, perhaps due to overloading or perhaps due to unavailability of data for some counties. This alone requires to return to this study and perhaps pick just one state, a state that is more interesting to investigate the problem. The state of Virginia seems to be one of the more interesting states that has long whiskers indicating the long range of minimum and maximum values and many outliers.

7. Conclusion

This has been a useful exercise that served as a warmup for a more rigorous work. We did validate that-

(1) Economically depressed counties have a high prescription rate of pain-killers (above the threshold of 50 prescriptions per 100 persons), and it's predictable

(2) Areas that have 6 or more vegetarian restaurants, all have a prescription rate of below 100%, and 5 or less indicated the opposite

(3) Perhaps the prescription of pain-killers after all have not much to do with treating pain itself

This case is far from settled and the next step for us is to gather more data, and investigate a more interesting state such as Virginia to find out more about the factors that may be contributing to high prescription rate of pain-killers.

8. References & Citations

1. Where have all the workers gone? An inquiry into the decline of the U.S. labor force participation rate. Alan B. Krueger, Bendheim Professor of Economics & Public Affairs. Princeton University. Brookings Papers on Economic Activity, Fall of 2017.

2. Tackling the Opioid Crisis: A Whole-of-Government Approach. US Senate Judiciary Committee. Tuesday, December 17, 2019.

3. Centers for Disease Control. Prescription Opioid Data. 2017. "Prescription opioids are often used to treat chronic and acute pain and, when used appropriately, can be an important component of treatment".

4. Focus on plant-based foods to ease arthritis pain. Physicians Committee for Responsible Medicine. Neal Barnard, MD, FACC, President, Physicians Committee.

5. The effects of plant-based diets on the body and the brain: a systematic review. Transl Psychiatry 9, 226 (2019) doi:10.1038/s41398-019-0552-0. Medawar, E., Huhn, S., Villringer, A. *et al.* 12 September 2019.

6. FourSquare APIs Documentation. Places API. Venues, Users, Photos, Tips

7. US Bureau of Economic Affairs(BEA), personal income data by county, metropolitan areas for fiscal year spanning 2017-2018.

8. Zillow is a popular & a reputable private sector organization that publishes the popular annual median home values.

9. Acknowledgements

Foremost, I'd like to acknowledge the support of my wife Jessica Manna, who understood my struggles working on weekends and after working hours sometimes late into night wrestling & wrangling with data. I'm grateful for her encouragement and useful critique of ideas, particularly this capstone project.

This capstone project would not have come to fruition without the outstanding guidance, lectures, laboratory examples and notes by the IBM Data Scientists and instructors of this course, Alex Akelson, Saeed Aghabozorgi, Polong Lin & Joseph Santarcangelo, and certainly thankful to the AI & Data Science Program Director Rav Ahuja also from IBM.

10. Appendix & Disclaimer

The intent for this report is to meet the requirements for Applied Data Science capstone project towards the award of a Data Science professional certificate. Although this report uses data science methodology and appears like one, this is not a formal research publication and hence the content is not formalized. However the document will be used as a baseline to conduct further investigations & perhaps formalized to submit for a peer review and subsequent publication.