



Visualization of the use of pain-killer opioids in America, in relation to Food, Income & Housing

2020/01/05

Murali Behara

IBM Data Science Professional Certificate

Capstone Project



Table of Contents

Abstract	2
Introduction	3
Literature Review	5
Methodology	6
Results	14
Discussion	15
Conclusion	16
References	17
Acknowledgements	18
Appendix	19


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 indirectly proportional to the usage of prescription pain-killer opioids. Data about the prescription rates, individual income & home values are gathered by county in each state, and 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.



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

-- TBD --

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat.

4. Methodology

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.

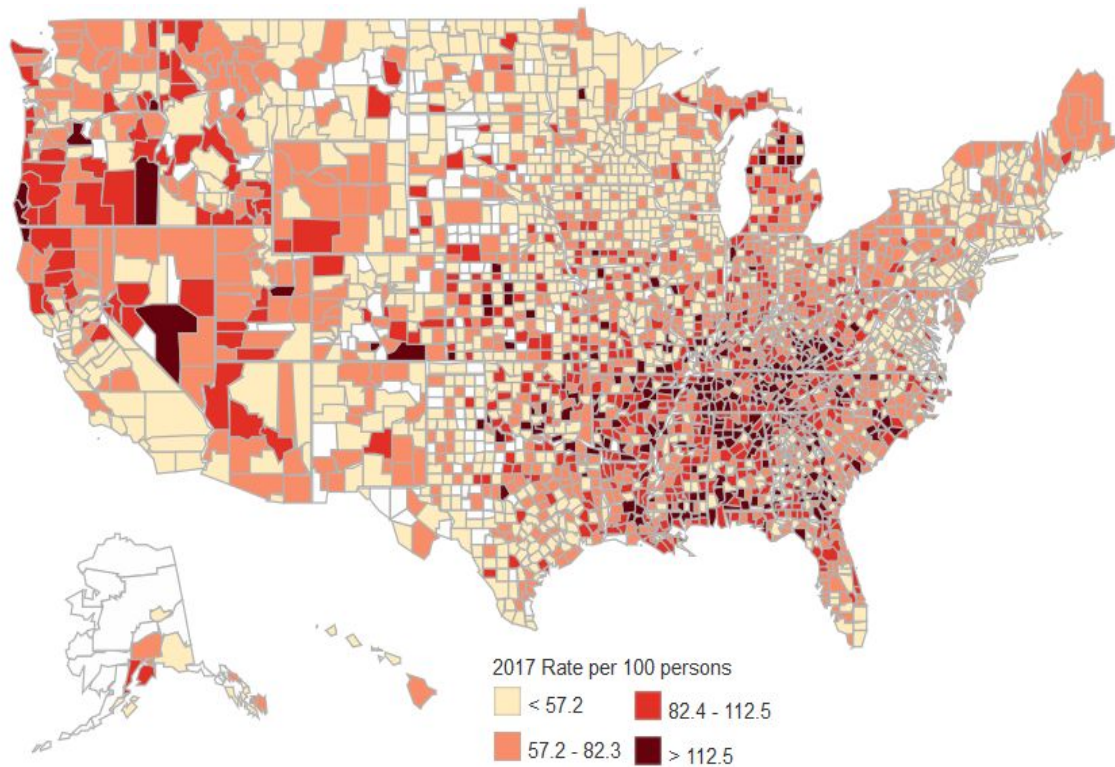


Figure 1 - US county wise pain-killer opioid prescription rate (percent) in 2017

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.

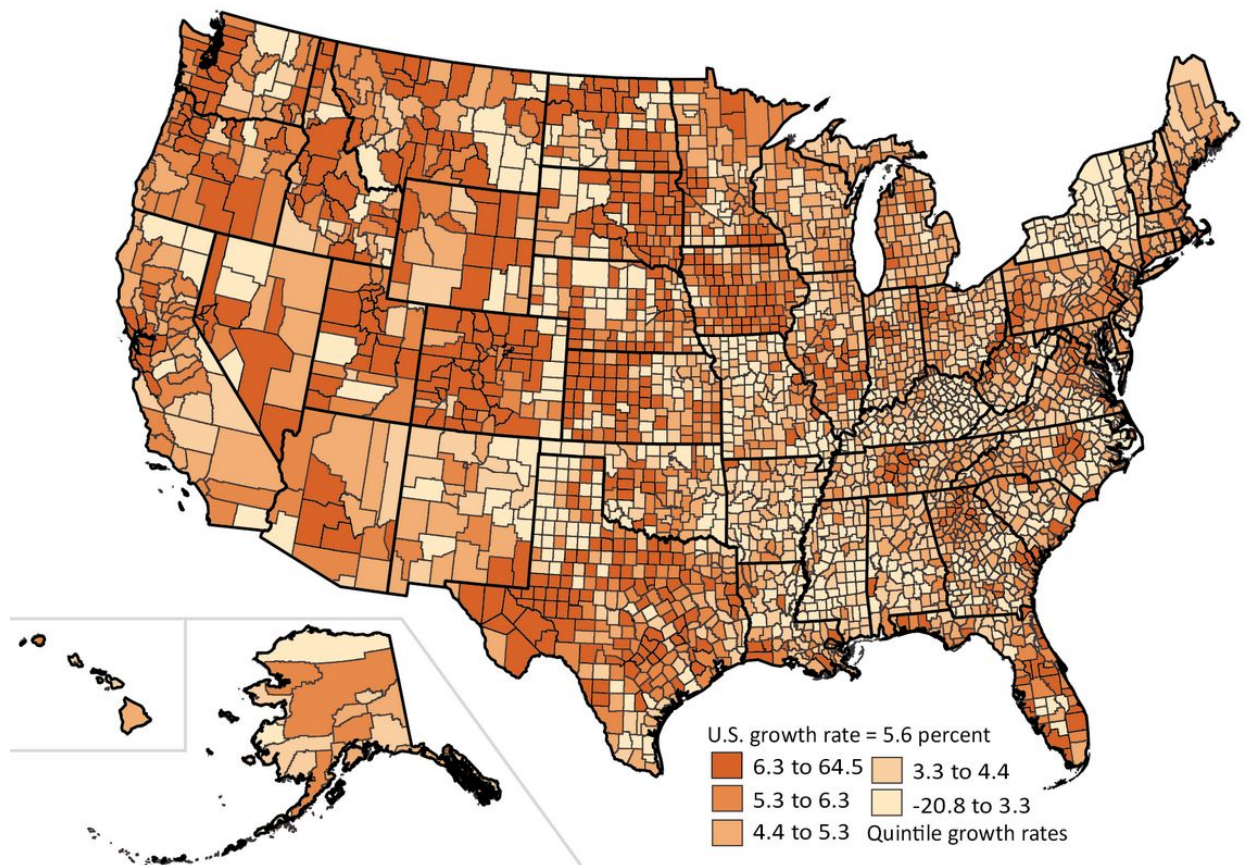


Figure 3 - US county wise personal income growth rate between 2017-2018

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.

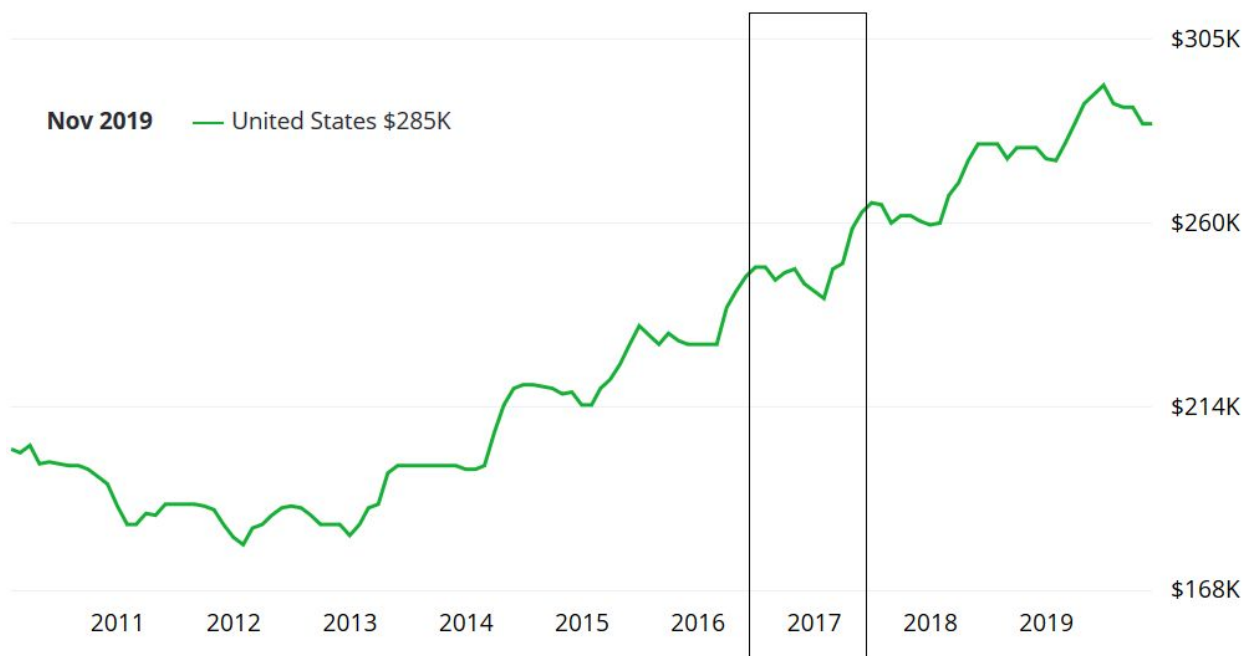


Figure 4 - US median home prices from the year 2011 through 2019

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.

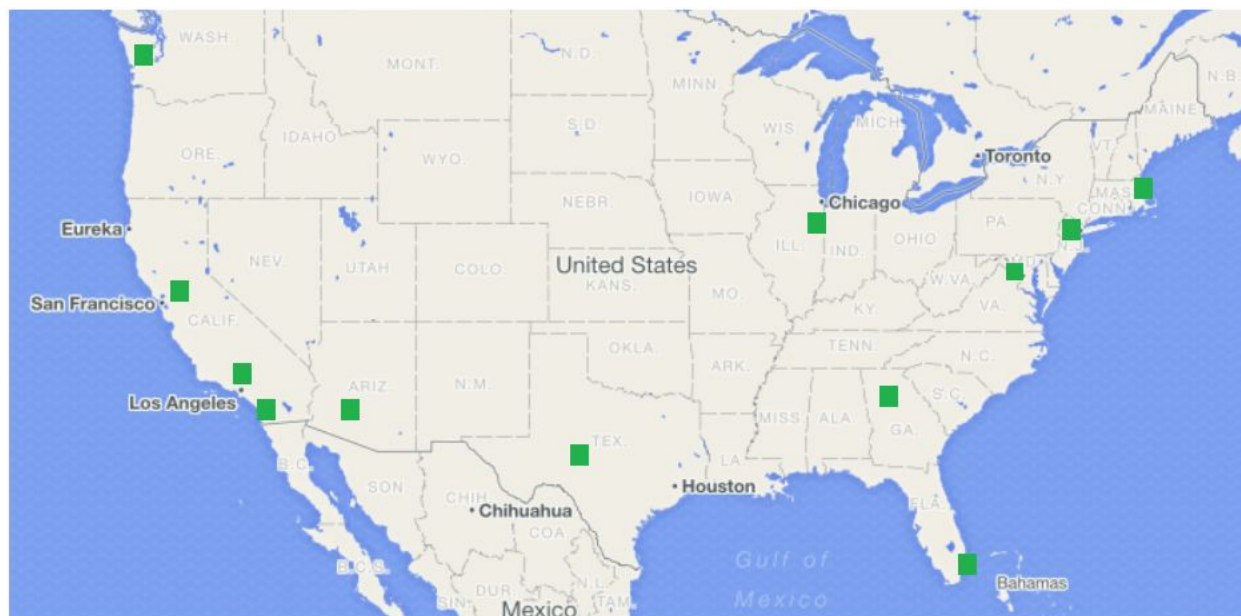



Figure 5 - FourSquare Map showing the cities where vegetarian food popular in the US

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
	Mediterranean 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

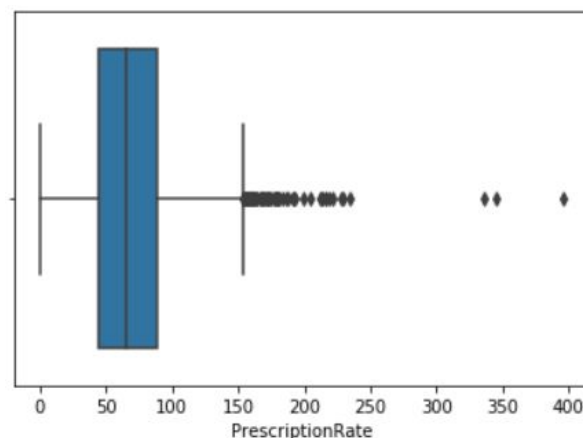


Figure 6 - Prescription Rates Table description & a boxplot summarizing prescription rates

Quick visualization of pain-killer prescription rate table shows that data is gathered for for 2969 out of 3000+ counties, and missing data for about 14 reporting counties. Box plot is a shows that the median prescription rate is close to 75%, equivalent to stating that 3 in every 4 people have been prescribed pain-killers.

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.

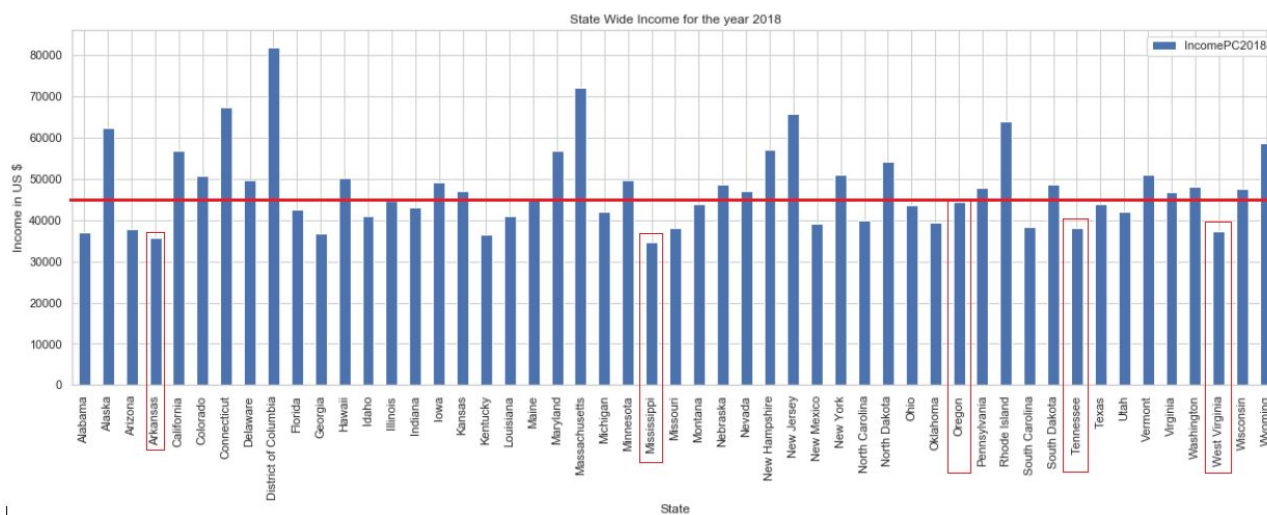


Figure 7 - Simple bar chart summarizing the average per capita income for each state

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.

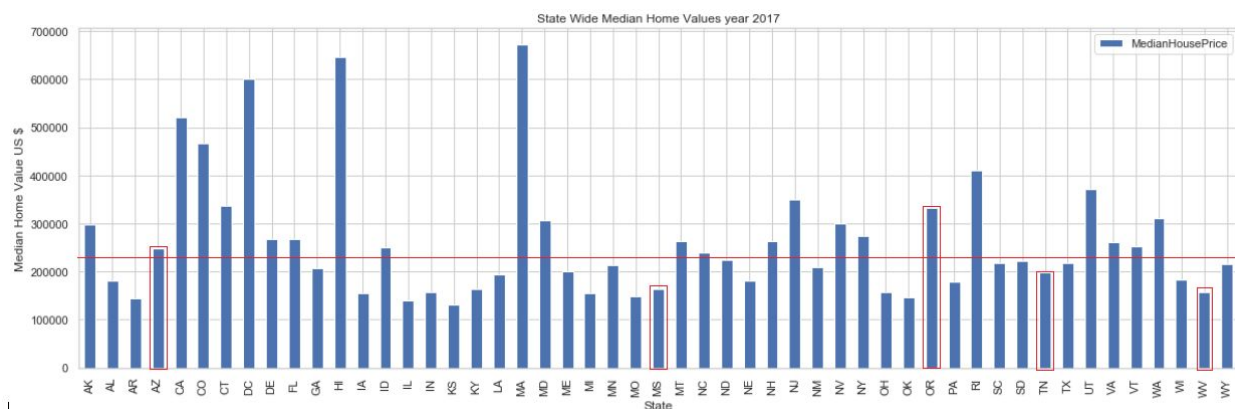


Figure 8 - Simple bar chart summarizing the median home value in each state

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 almost always 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

5. Results

Model building - Guideline question: How can I best model the data, visualize the data to steer my investigation.

Model evaluation - Guideline question: What kind of model is appropriate to analyze the data & does the model answer the question

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat.

Model deployment - Guideline question: Can the model be put into practice.

Incorporating feedback - Guideline question: Can we get constructive feedback to answer the question.

6. Discussion

-- TBD --

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat.

7. Conclusion

-- TBD --

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat.

8. References

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.



9. Acknowledgements

10. Appendix