SI 618 Project Report I

1. Motivation

The nature of my project is to see if there is a correlation or some kind of relationship between homelessness and energy usage. The three energy usages that I am examining are energy consumption, energy production, and energy expenditure in all states in the US. Knowing that homelessness is strongly related to income level, I wanted to see if there was a relationship between energy usage and income first. After examining that relationship, I want to see the relationship between homelessness, income, and the three energy usages. The idea came to me after looking at my electricity bill and walking around the city of Ann Arbor. I noticed that my electricity bill is much higher in Ann Arbor than when I was living in Oklahoma. I also noticed an increasing number of homeless people that are living in Ann Arbor. I was just curious to see if there might be some kind of explanation or link between these observations.

2. Data Sources

US Homelessness Dataset: https://www.kaggle.com/adamschroeder/homelessness

The US Homelessness Dataset was easy to download off the Kaggle website. It returned about 80,000 rows of data on a csv format. I was able to download this on my laptop and move it to cavium for quicker processing. This dataset lists years from 2007-2016 and all 50 states, D.C., and U.S. territories Guam and Puerto Rico in the first two columns. The dataset also shows CoC numbers and the exact CoC where the data was taken. The last two columns show the count of homeless people in that exact CoC and the description or level of homelessness when considered. For my project purposes, I only used data within the year of 2014 to keep everything uniform. Also, I excluded data from Guam and Puerto Rico because my other datasets did not have data on them. I also only found the count of homeless people in the CoC useful because I was only interested in the quantitative data rather than the qualitative. So, I used the State, Year, and Count for this dataset specifically.

Energy Census and Economic Dataset:
https://www.kaggle.com/lislejoem/us_energy_census_gdp_10-14

The energy census dataset only displays about 52 rows of data because it is limited to the fifty U.S. states and D.C. Downloading this off Kaggle was easy because it was a csv file and it had time references in the column spanning from 2010-2014. The data had multiple columns but the only ones I found useful were the states column (State), state codes column (StateCodes), the total energy consumption for every U.S. state in 2014 (TotalC2014), the total energy production for every U.S. state in 2014 (TotalP2014), and the total energy expenditure for every U.S. state in 2014 (TotalE2014).

Income and GDP per capita Dataset:
https://apps.bea.gov/itable/iTable.cfm?ReqID=70&step=1

My income dataset only shows data from all 50 U.S. states and D.C. I specifically grabbed the year of 2014 because my other two datasets shared the year 2014 and it was the most recent year in combination. This dataset was downloaded as a csv file, but it had bad formatting. Because of this, I removed empty space columns and made the headers clearer through excel and exported it back as a new csv. The columns I used for this dataset were states and 2014 GDP per capita.

3. Data Manipulation

For my homelessness dataset, I realized I needed to only grab the data pertaining to the year 2014. After reading in the csv file into pyspark using spark.read, I used a where condition in my query to grab only data pertaining to the year 2014 and disregard data from Guam and Puerto Rico. I registered this into a temporary table and used another pyspark.sql query to select data from the state and count columns within my new 2014 data table. This left me with the count of homeless people in every U.S. state and D.C. for the year 2014. I made an RDD by making that data into a tuple, but I quickly realized that some of the data in count were displayed as ‘1,539’ rather than ‘1539’. Because I needed to sum up all counts in each state later on in my manipulation, I mapped all the values with a lambda function to remove all commas in my RDD data and set that equal to a new RDD variable. Next, I mapped all the values again and turned my counts into integers to convert the values out of unicode. After, I reduced them by key to get all the states homelessness count and collected it all together. Using this data, I created a data frame to register it to another temporary table for later manipulation.

As I briefly mentioned for my income/GDP per capita data, there were a lot of empty columns and rows in the initial csv. To make things easier for my table joining, I went into excel and removed all the empty space and re-exported it as csv. I then used spark.read to load the csv file and register it as a temporary table. I then loaded in my energy csv file using spark.read. For my energy table, I used pyspark.sql queries to grab only state codes, states and whichever energy usage I wanted. I did a total of three queries, one for consumption, production and expenditure and registered each of those queries to their own temporary table. I then joined my income table to each of the three energy usage tables because they both had state as a column. After joining income to each of the three energy usage tables, I joined my homelessness table to each of the three income/energy usage tables because they had state code as a common column. This was the extent of my data manipulation and data parsing.

Also, in order to make everything uniform, I edited the headers each dataset column that were similar to they would be easily joined later on. For my first two datasets, handling missing data or ignoring data was pretty easy because I was able to just use queries to grab the data that I found important for this project. However, for my homelessness dataset, it was a little challenging because I had to figure out mapping functions in order to solve each problem I ran into. Because I knew that I wanted only specific years in my data and that I wanted to merge income per capita in every state, homelessness count in every state with total energy consumption, total energy production, and total energy expenditure in every state, it was not that hard to figure out the steps to manipulate the data. Running into the bugs because the dataset returned different types of data and figuring out a way to solve it was the time-consuming part of my project.

4. Analysis and Visualization

I initially wanted to see if there was a correlation of some sort between income per capita in each state and total energy consumption for every state. I also did this with total energy production and total energy expenditure for every state to see if any of them were somehow linked in any way. I created a scatterplot using the two variables after loading in the data. This was my initial analysis.

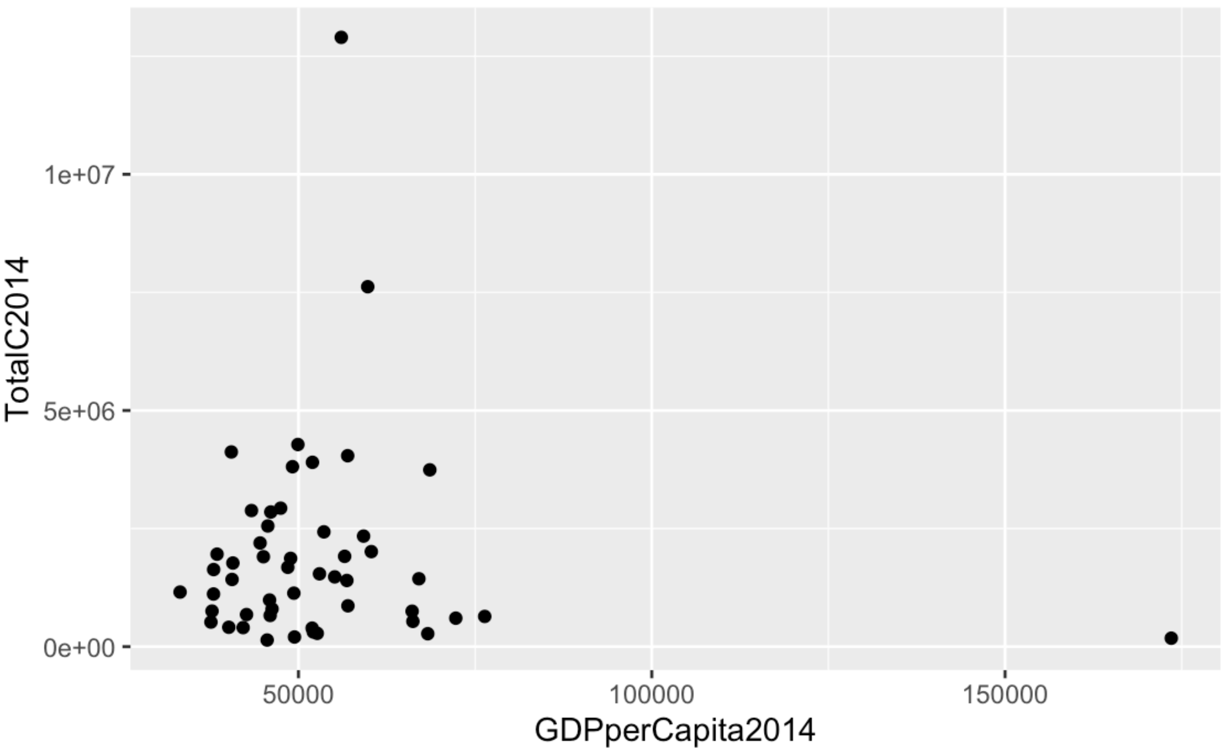

Figure 1. Scatterplot of Total Energy Consumption to GDP per capita for every U.S state and D.C.

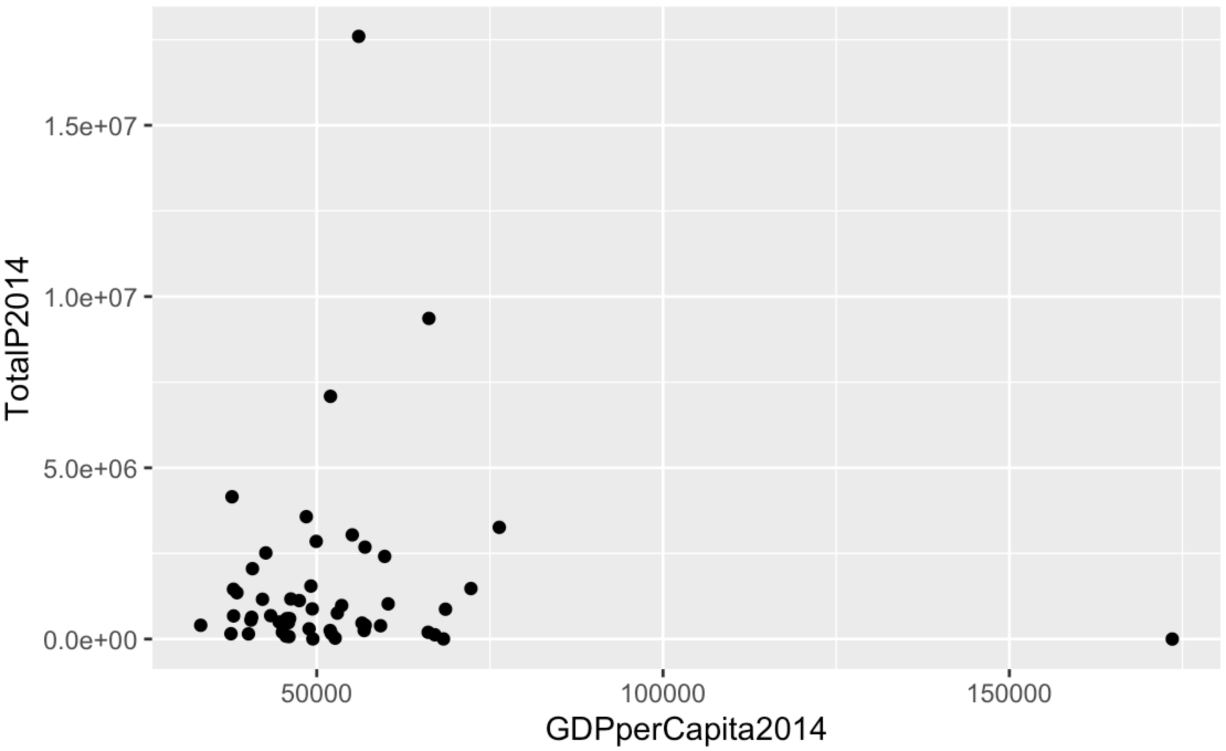

Figure 2. Scatterplot of Total Energy Production to GDP per capita for every U.S state and D.C.

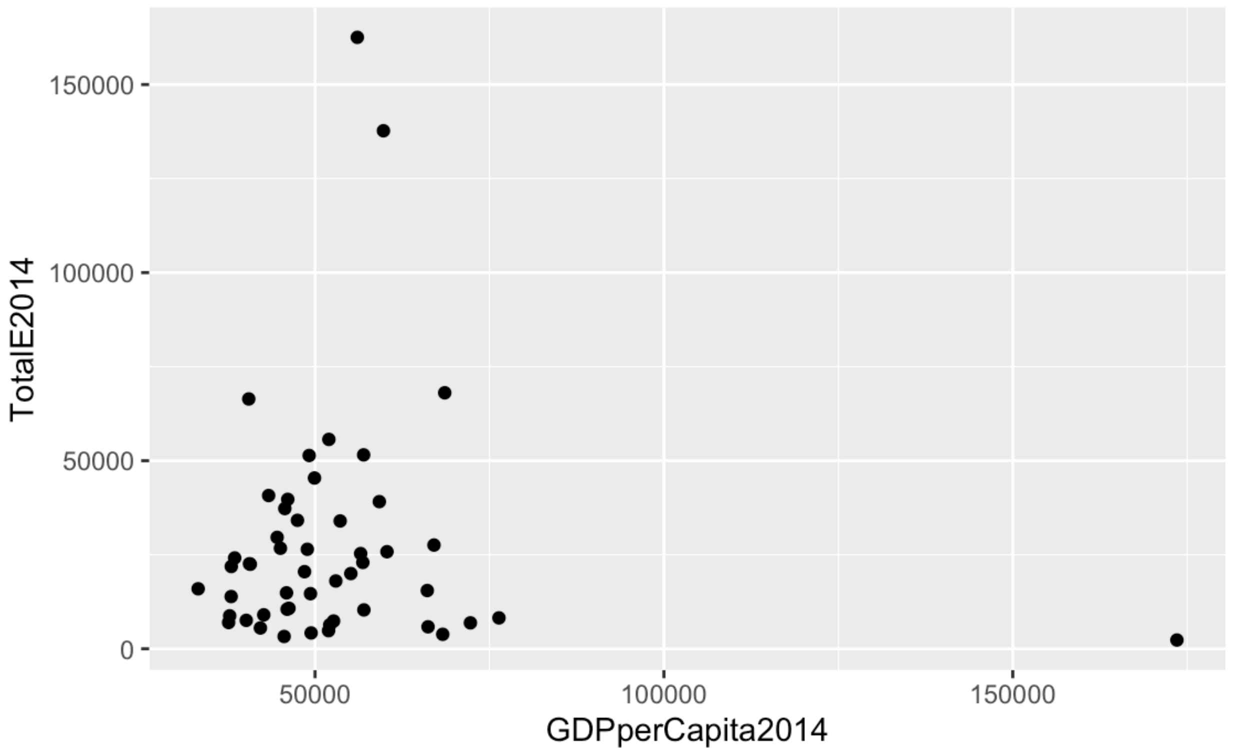

Figure 3. Scatterplot of Total Energy Expenditure to GDP per capita for every U.S state and D.C.

Looking at figures 1, 2 or 3 there does not seem to be a strong link to say that there is a correlation between the variables. Because there was not a correlation link, I moved onto how homelessness may relate to these three scatterplots. Using R studio again, I created a 3D scatterplot using my joined data of homelessness, income, and the three energy usage variables.

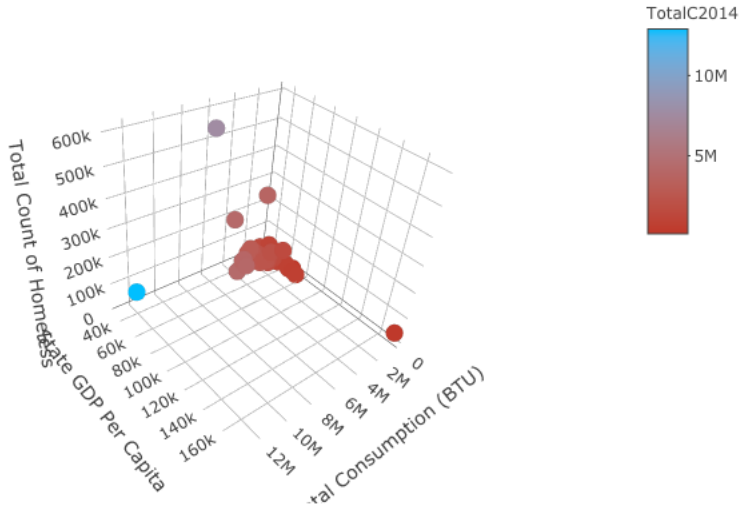

Figure 4. 3D scatterplot of total homeless people counts, total energy consumption, and income per capita in all U.S. states and D.C.

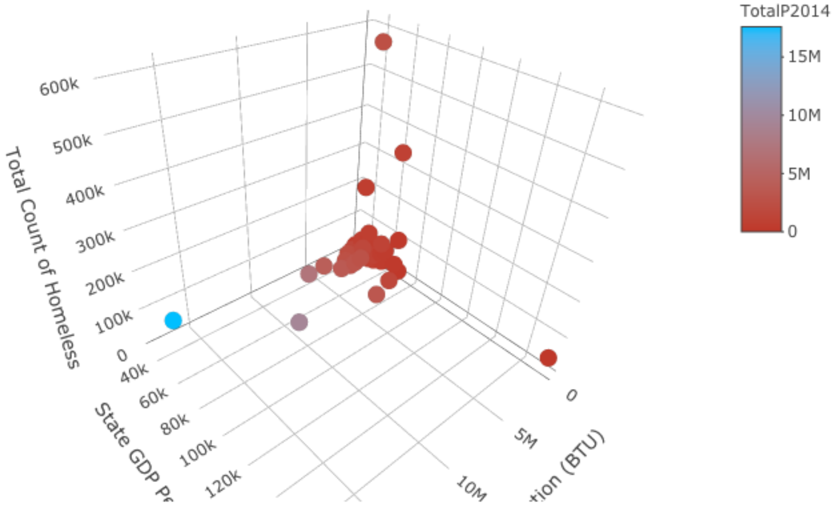

Figure 5. 3D scatterplot of total homeless people counts, total energy production, and income per capita in all U.S. states and D.C.

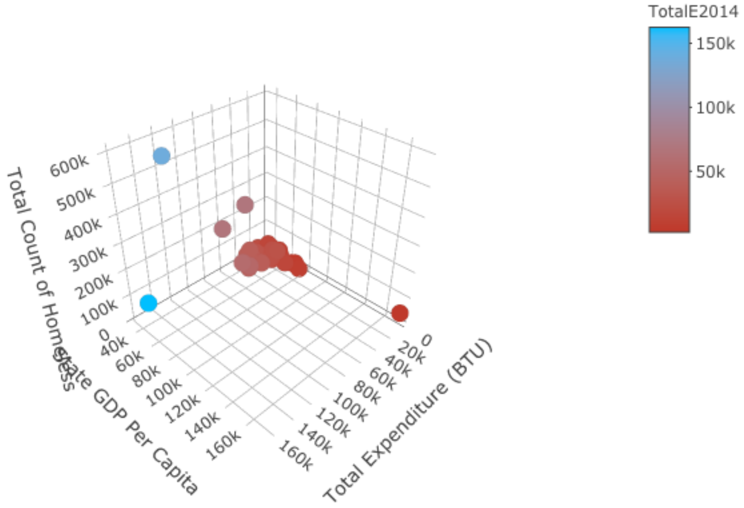

Figure 6. 3D scatterplot of total homeless people counts, total energy expenditure, and income per capita in all U.S. states and D.C.

Looking at my scatterplot visualizations, I noticed that there still was not a relationship between any of these things. It would make sense that areas that use more energy spend more money and therefore cost of living would be higher making it easier for more people to end up homeless. However, with my analysis, it does not seem to show that kind of correlation. I guess it did not work because there are a lot of other variables that account for homelessness such that energy affecting cost of living was not high enough to be much of a variability in this analysis.

Now that I have done my initial analyses, it is time to answer my main research questions.

Are there more homeless people in states that consume more energy? Are there more homeless people in states that produce more energy? Are there more homeless people in states that expend more energy?

Since the data analysis coding was similar for all energy usages, I will explain the process using consumption. In order to find out whether or not a there are more homeless people in states that have high energy consumption, I realized I had to bin consumption into three categories: low consumption (0 – 500,000 billions BTU), medium consumption (500,000-2,000,000 billions BTU) and high consumption (2,000,000 - Infinity billions BTU). So, I had to use another pyspark.sql query to select total consumption for 2014, total homeless people count, and the state codes. After doing so, I used another mapping lambda function to convert my consumption into an integer. I then created a bucketing variable that splits the thresholds of each of my bins. I created another data frame that incorporates my bins and used another lambda function to map my labels with my bins. After I created this, I used another pyspark.sql query to get consumption_bucket and Count together. I needed these to create a new crdd2 which I used another map and reduceByKey to get them all together. I made another query to get the average homelessness count for each consumption_bucket. The data for consumption, production, and expenditure are in table 1. I also wrote this data frame to a csv to put into R to create a bar graph to show the counts to energy consumption levels of states. I did this with production and expenditure as well. My expenditure bins had different thresholds (0- $30000 | $30,000 - $100,000 | $100,000 - $Infinity) because the numbers were much lower than consumption and production. You can see that in my bar graph (figure 7), the count of homeless people is much larger in high consumption than in low consumption and medium consumption. We can see that the average number of homeless people in high consumption tier states is higher than every other tiers' average homeless people count. This is true for production and expenditure.

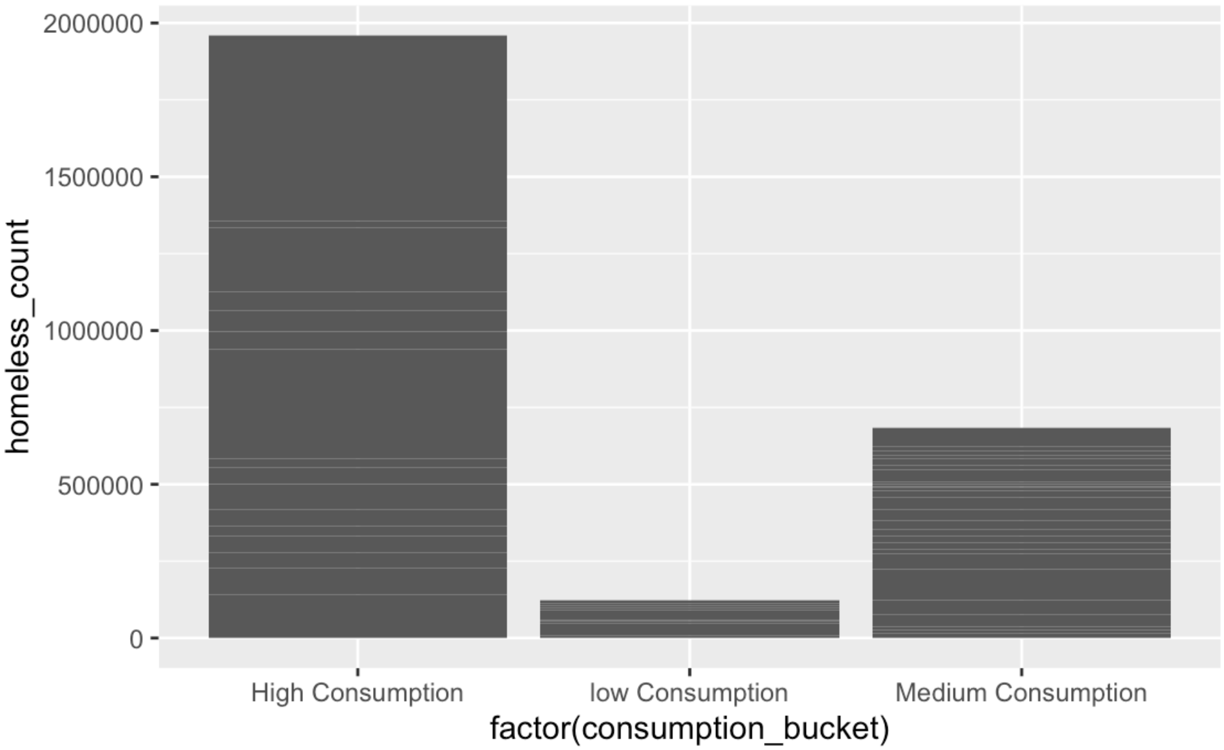

Figure 7. Bar graph of consumption tiers and homeless count in U.S. states and D.C.

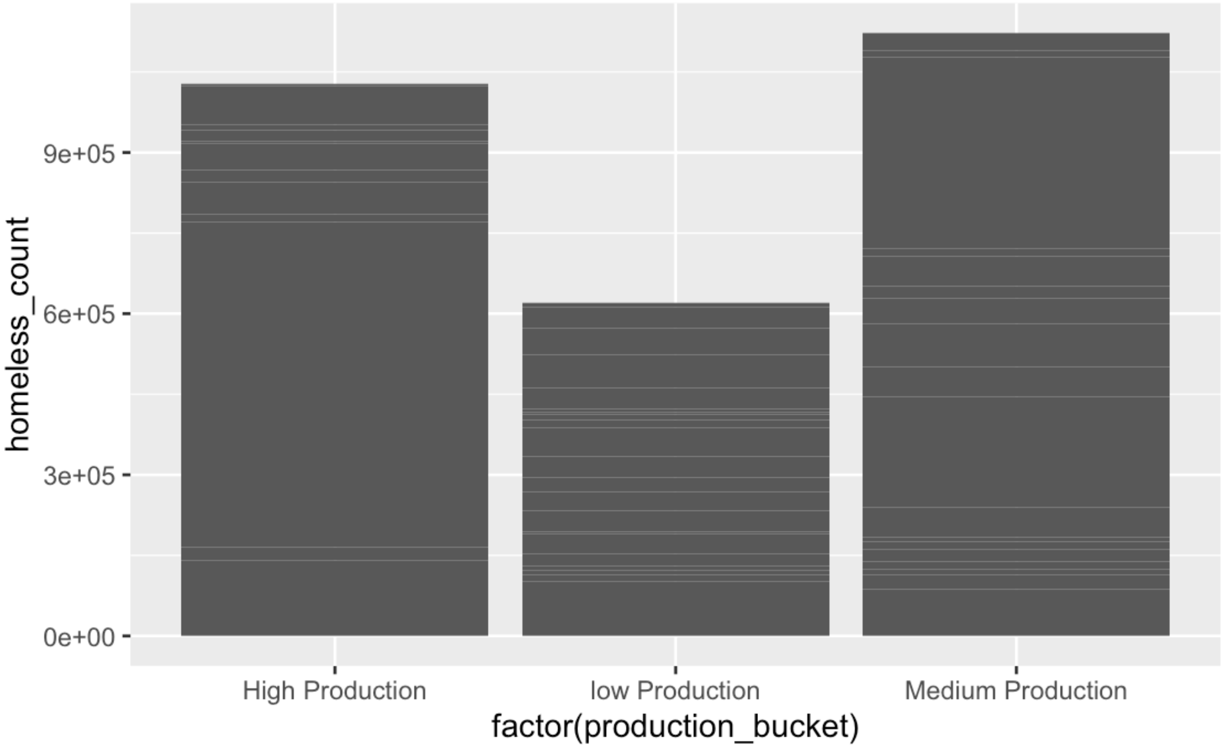

Figure 8. Bar graph of production tiers and homeless count in U.S. states and D.C.

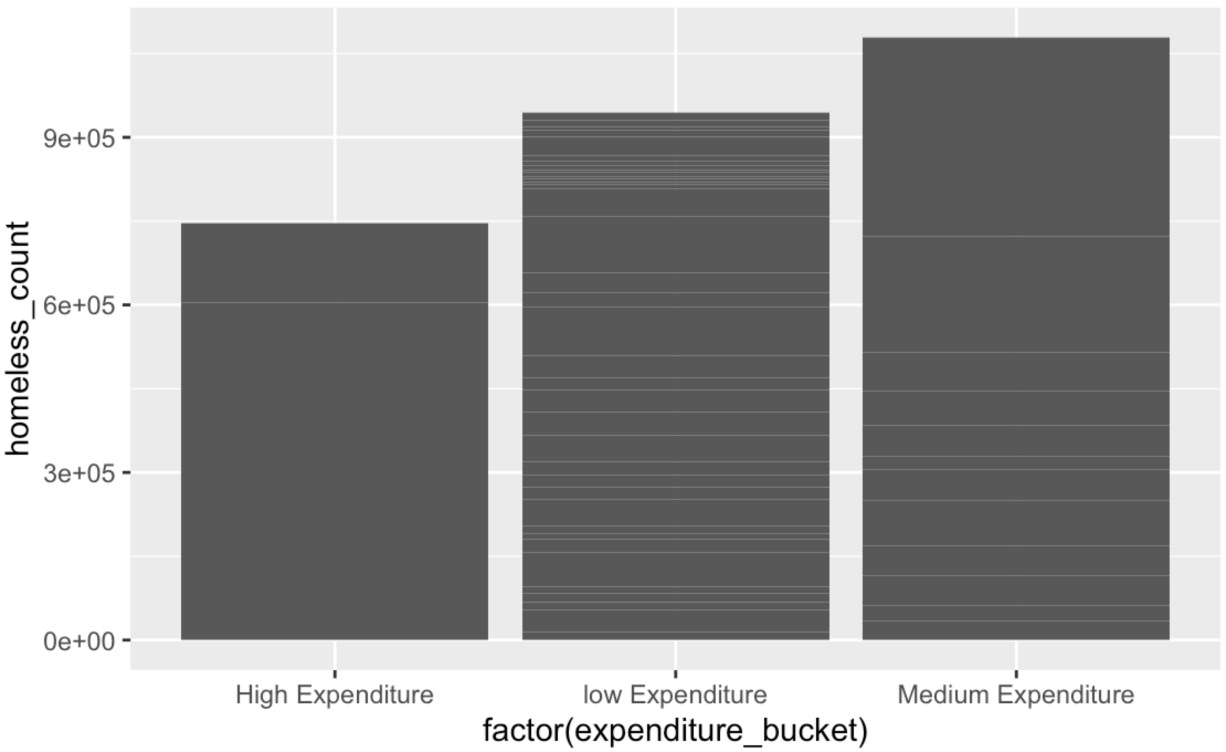

Figure 9. Bar graph of expenditure tiers and homeless count in U.S. states and D.C.

	High Usage (average)	Medium Usage (average)	Low Usage (average)
Consumption / Homeless Count	122495	27404	13752
Production / Homeless Count	85566	62373	29490
Expenditure / Homeless count	372696	89919	25523

Table 1. Average homeless count for high, medium, and low usage for consumption, production and expenditure.

5. Challenges

A big challenge I faced was contemplating whether or not to create actual correlations and averages for my analysis. After going to office hours, I realized that this project is not so much about specific analysis but more aggregation using data manipulation techniques. In order for me to get around this problem, I had to think of exactly how I wanted to join the datasets and what questions I could potentially answer using my combined data. Another big challenge for me was figuring out how to parse through my data using sql queries because I did not want to create data frames without grabbing only the information I wanted out of the dataset.