**Exploring the Relationship Between Walkability and the Prevalence of Chronic Illnesses**

Renzo Espinoza

College of Information Technology, Western Governors University

D214: Data Analytics Graduate Capstone Task II

April 30th, 2022

**A: Research Question**

Homes in walkable areas have become increasingly sought-after in recent years, according to Corthright (2020). There are many reasons as to why pedestrian-friendly neighborhoods are becoming so desirable, one potential benefit is that walkable neighborhoods encourage a healthier quality of life. The objective of this study is to attempt to provide further evidence of the theory that living in walkable neighborhoods leads to better health outcomes

There are several studies pointing to the possible health benefits of walkability, as summarized by Casella (2022), one large study of 1.1 million adults found that those living in less walkable areas had a 20 percent higher chance of developing pre-diabetes over 8 years. Other studies found residents in walkable neighborhoods had a lower rate of obesity, had a reduced rate of high blood pressure, and got more exercise than their counterparts living in pedestrian-unfriendly neighborhoods. So what exactly is ‘walkability’? Companies and organizations have developed their own scores to answer this question, these scores take into account factors such as population density, pedestrian-friendly streets, and the amount of schools, workplaces, parks, and public spaces in the area. There are even more factors that vary between walkability scores from different organizations, the metric that I will be using is the Walkability Index developed by the U.S. Environmental Protection Agency.

The research question of this analysis is: is there a statistically significant relationship between the walkability of a county in the United States and the prevalence of chronic illnesses in that area? Finding underlying causes of chronic conditions is an important business need for hospitals all across the country, people with five or more chronic conditions make up only 12 percent of the population but account for 41 percent of all healthcare spending (Buttorff et al., 2017, p. 15). My hypothesis is that since it has been theorized that walkable neighborhoods could encourage healthier lifestyles, there might be a link between the prevalence of chronic conditions and the walkability of the area where a patient resides in.

**B: Data Collection**

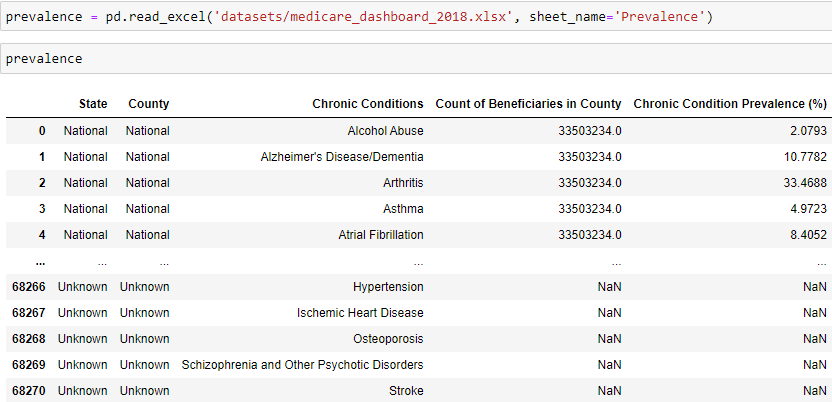
For this analysis, data was gathered from three different sources. The first data set is the Medicare Chronic Conditions Dashboard (Centers for Medicare & Medicaid Services, 2022), this dataset contains two data tables showing the prevalence of different types of chronic conditions in each county in the U.S. as well as the prevalence of multiple concurrent chronic conditions in 2018. To quantify how walkable each county is, I used the Walkability Index data set (U.S. Environmental Protection Agency, 2021). This data set has over 220,000 rows of data and 117 columns, the feature that I will center my analysis around is the ‘NatWalkInd’ feature. This metric acts as a score for how walkable an area is, this index formulized by the EPA takes into account the density of intersections in an area, the proximity to transit stops, and the diversity of land use.

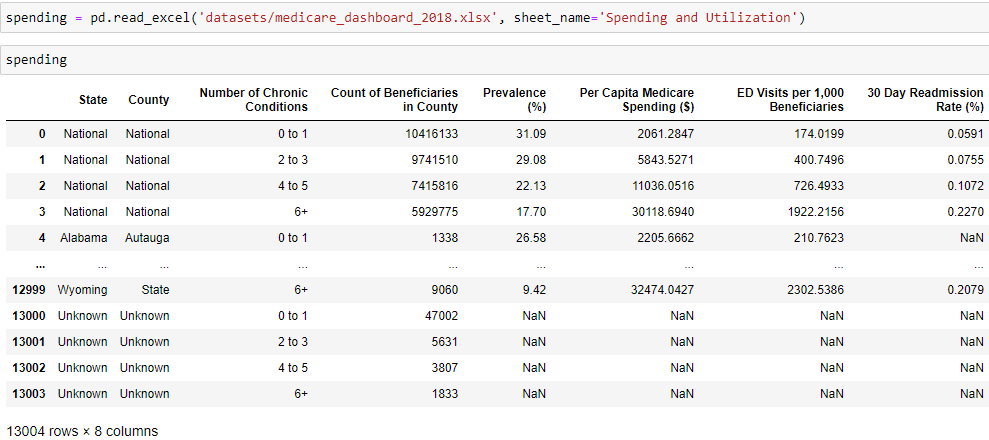
Each entry in the chronic conditions data set is identified by the state and name of the county, whereas the walkability index dataset uses three columns, in those three columns are codes that represent a state, county, and area within the county. The final dataset is a .csv file containing the Federal Information Processing System (FIPS) codes used to identify states and counties in the United States (Healy, 2018). I will use this data convert the FIPS codes found in the walkability index data set into their respective counties, allowing me to link the data with the chronic conditions data tables.

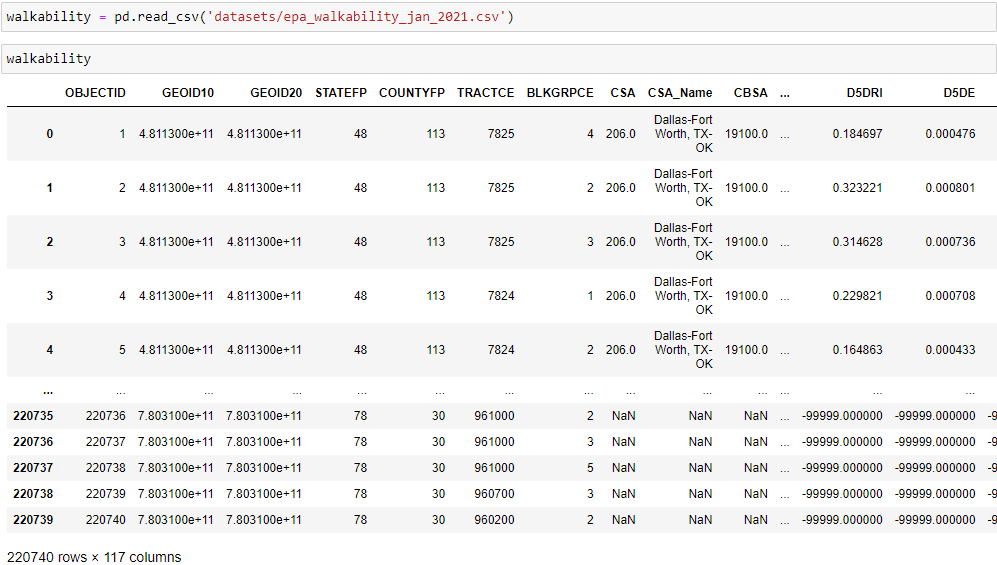
In total there is around 305,000 rows of data to process. The procedure I used to gather these three data sets is manually downloading them. The advantage to this data-gathering methodology is that it is fast and simple. A disadvantage of this data collection methodology is reproducibility. For someone to run this analysis on their machine, they would have to download the necessary data sets from their respective webpages themselves. It is not guaranteed that they will understand how to download the data. For example, the page for downloading the chronic conditions data contains links to two other data sets, so I would have to write out instructions for installing the required data. A Python script that downloads the data sets automatically would take the guesswork out of setting up the project. To overcome this disadvantage, I can package the data sets together along with all the other source files of the project when sending it to anyone interested, thus avoiding the need to manually download the data files.

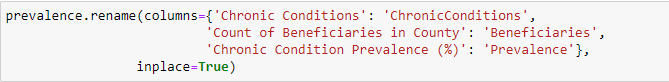
**Data Extraction and Preparation**

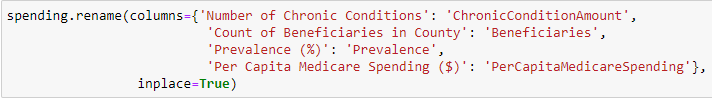
After collecting the data, the next step was to import the .csv and .xlsx files into a programming environment for cleaning and analysis. I chose Python as my programming language with Conda as my environment management system and my programming environment was a Jupyter Notebook. I chose these tools because the Python ecosystem has many useful libraries for performing data cleaning and analysis. One such library is Pandas, I used the ‘read\_excel’ and ‘read\_csv’ functions of the library to import the data into a Pandas DataFrame. An advantage of DataFrames over a two-dimensional array is that it is a more efficient data structure and comes with several built-in functions for fast data manipulation. After importing the data, I then display the DataFrame to get a peek into the data.



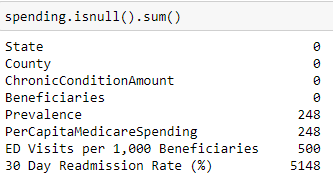
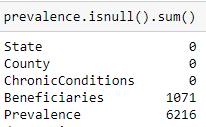


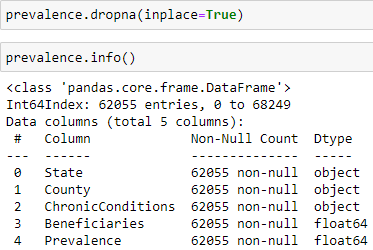


 Next, I renamed certain columns in the two data tables imported from the Medicare dashboard. I did this to shorten the column names and make it easier and faster to reference the columns in Python.

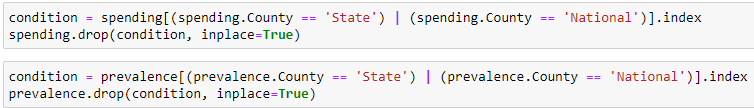


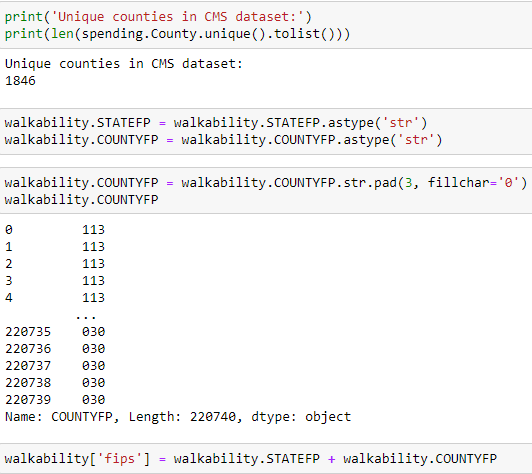
To find missing data, I used the ‘isnull’ function to get a count of how many entries appear in each column containing null values.

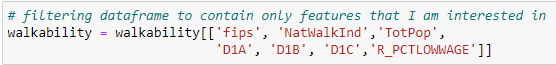


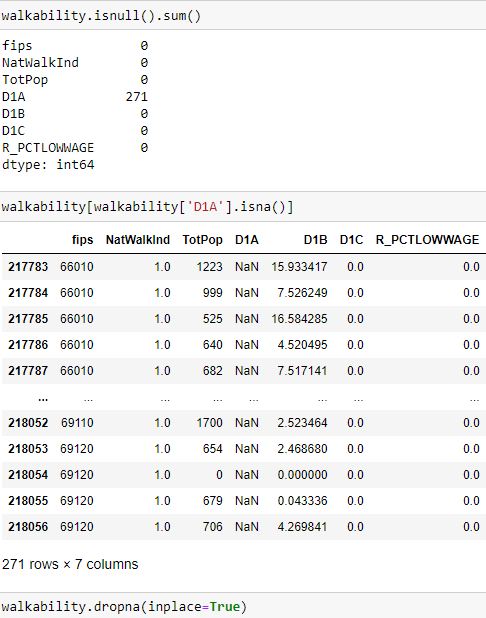
The ‘prevalence’ DataFrame contains the individual prevalence of multiple different chronic conditions. According to the summary above, there are 6,216 of the 68,271 total rows contain a missing ‘Prevalence’ value. For the ‘spending’ DataFrame, only 248 of the 13,004 rows contain a missing value for ‘Prevalence’. My solution to this issue is to use the ‘dropna’ method to simply remove the rows containing missing values. The downside to this approach is that you lose information when dropping rows. The missing rows in ‘spending’ accounts for only 1.9% of the data so dropping those entries would not result in a significant effect on the data. For ‘prevalence’, the missing entries are around 11% of the data, while dropping that much data is a considerable loss of information, my analysis is mainly centered on the ‘spending’ and ‘walkability’ DataFrames so I chose this approach to save time. Also I dropped the two columns, ‘ED Visits per 1,000 Beneficiaries’ and ‘30 Day Readmission Rate (%)’ as they are not important for my analysis or contain too much missing information.



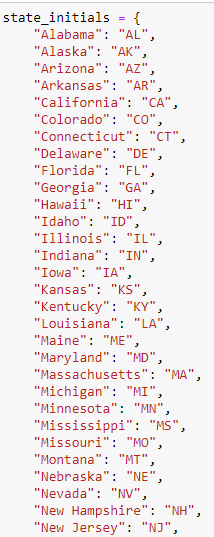
 The ‘spending’ and ‘prevalence’ data tables have state and nation level rows, I dropped these rows because I’m only interested in the county-level statistics.

 The ‘walkability’ data set has two columns, ‘STATEFP’ and ‘COUNTYFP’, that are used to identify what state and county they belong to. These columns are being imported as integer columns, so using the ‘astype’ function I converted the two columns to strings. I then left-padded the county column with zeros if the string is less than 3 characters so it can match the same format as the county codes in the ‘fips’ data set. Finally, I combined those two columns into one new column titled ‘fips’ by concatenating their strings together in each row.

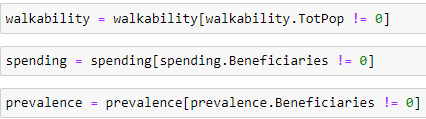
 Next, I filtered the DataFrame to contain only the features I am interested in analyzing to save memory and reduce complexity since there are 117 columns in the data set.

 From there I check to see how many nulls appear in each column. The only column will nulls is the ‘D1A’ column, containing only 271. I inspected these rows containing nulls and found that they also contain blank values for the rest of the columns, so I deem them worthy of removal so they do not affect the rest of the data when I perform aggregation later.

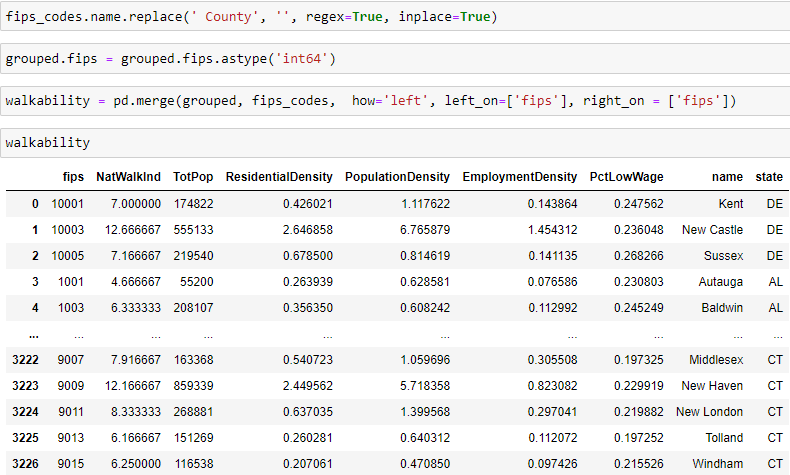
To eventually join the data tables together I will need them to have matching columns. I created a Python dictionary containing each state in the United States as a key, with their value being their state initials. I then use that dictionary replace each entry in the ‘State’ column of the ‘spending’ and ‘prevalence’ DataFrames with their respective initials.

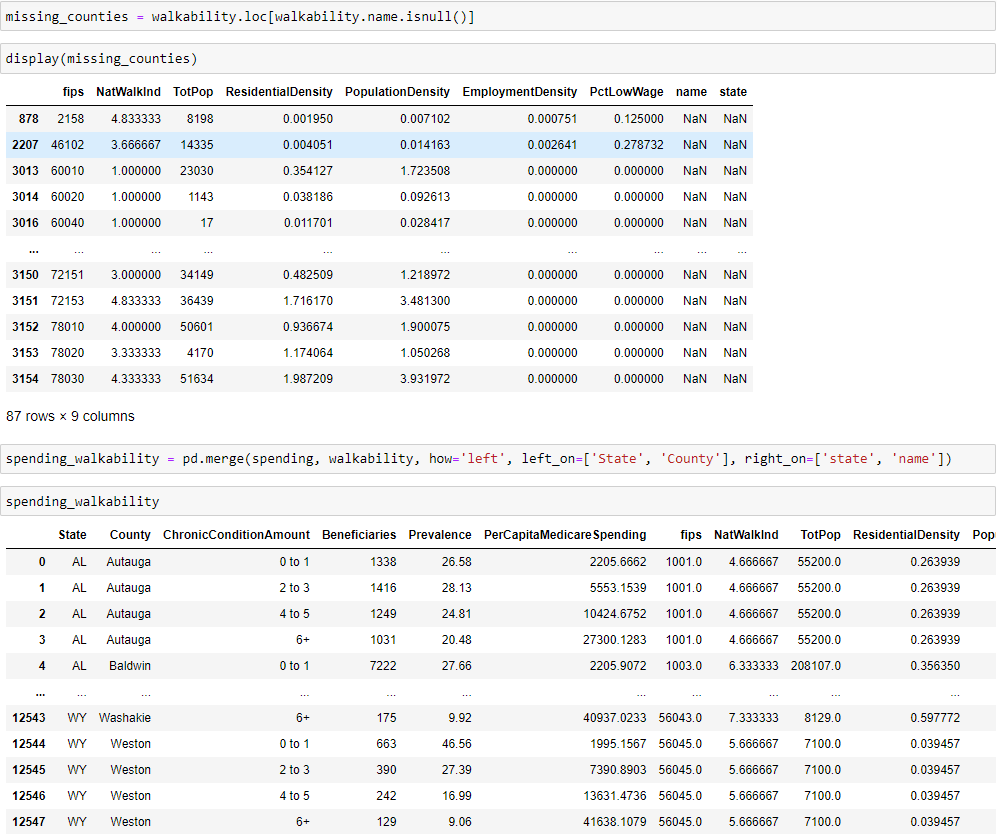


Then I dropped any county or area in the three datasets that has 0 people or beneficiaries in them as I believe they are input errors or skew the data.

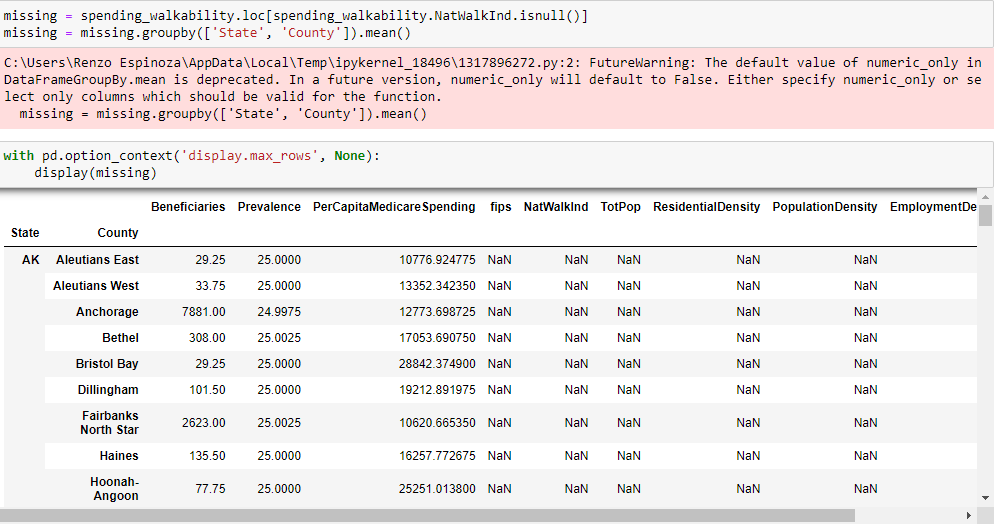


 Each entry in the ‘walkability’ represent a Census block group, there can be multiple block groups in a county, so to match up with the entries in the ‘prevalence’ and ‘spending’ data sets I need to aggregate the block groups into counties. To do this I used the ‘groupby’ function of the Pandas library to group each entry that has matching values in the‘fips’ column that I created earlier. After grouping the entries I renamed certain columns so a data dictionary would not be needed to understand them.

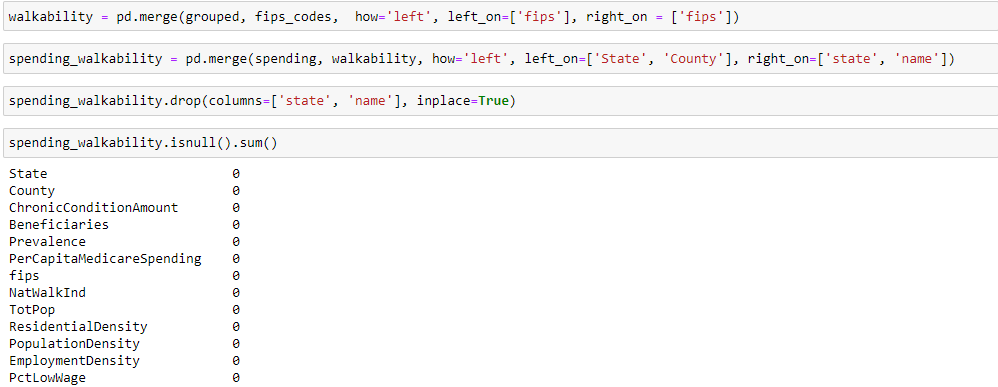
 The next step was to join the data set containing the FIPS codes with the new ‘grouped. I used the ‘merge’ function from the Pandas library to perform this join operation. Before that I converted the ‘fips’ column in the ‘grouped’ DataFrame to an integer so the ‘fips’ columns in the two datasets have a matching data type. I also removed the word ‘County’ from the strings in the ‘name’ column of the FIPS data set so it matches the format of the ‘spending’ and prevalence’ DataFrames.

I then checked to see how many rows did not match between the two tables and found that only 87 rows failed to have a matching FIPS code. As this is only around 3% of the rows in the DataFrame I chose to avoid searching on the internet for the county that their FIPS codes belong to, to save time. I use the same ‘merge’ function from before to perform this left join operation. Since the ‘walkability’ table is being joined onto the ‘spending’ table, the previously mentioned 87 rows with missing state and county name values will be left out of the resulting DataFrame.

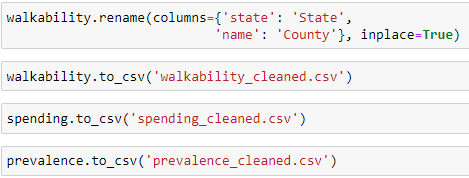
I then check to see which counties did not match between the two tables by creating a DataFrame containing missing values after the join operation.



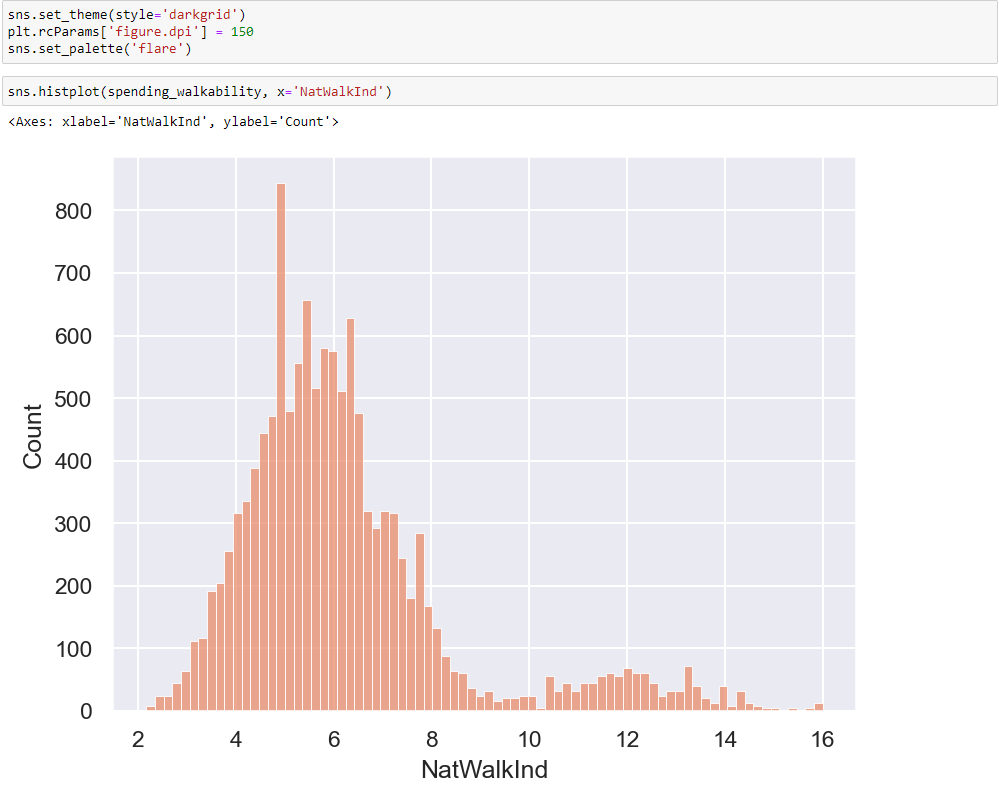
 After looking through the unmatched counties I found that they all had minor differences between the names that appear in both tables. Using regular expressions and the ‘replace’ Python function, I filtered out these minor differences

 Now with the names of the counties in both tables matching, I rerun the joining operation and this results in zero rows with missing values.

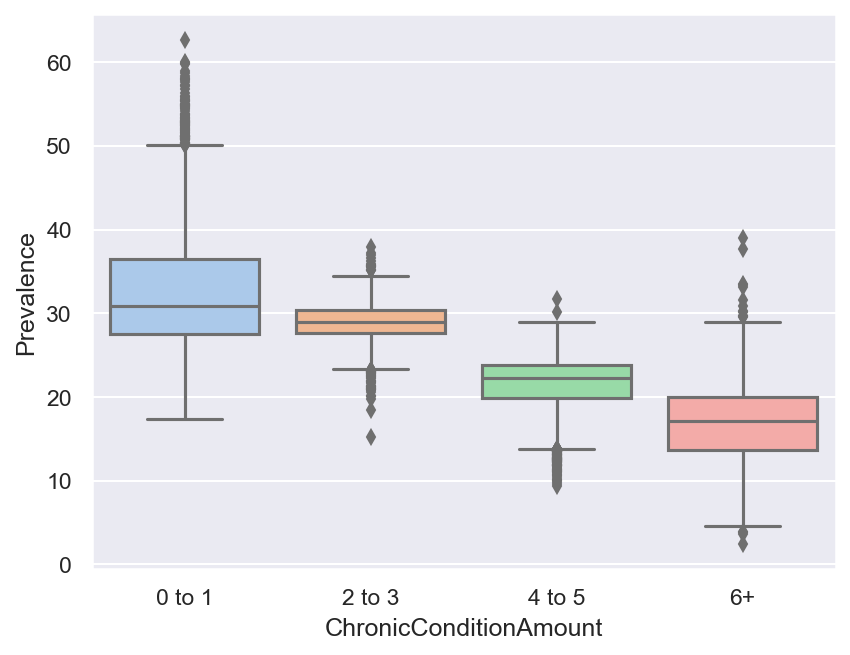
I renamed the newly added ‘name’ and ‘state’ columns in the ‘walkability’ DataFrame and then exported the cleaned datasets to separate .csv files using the ‘to\_csv’ function from the Pandas library.



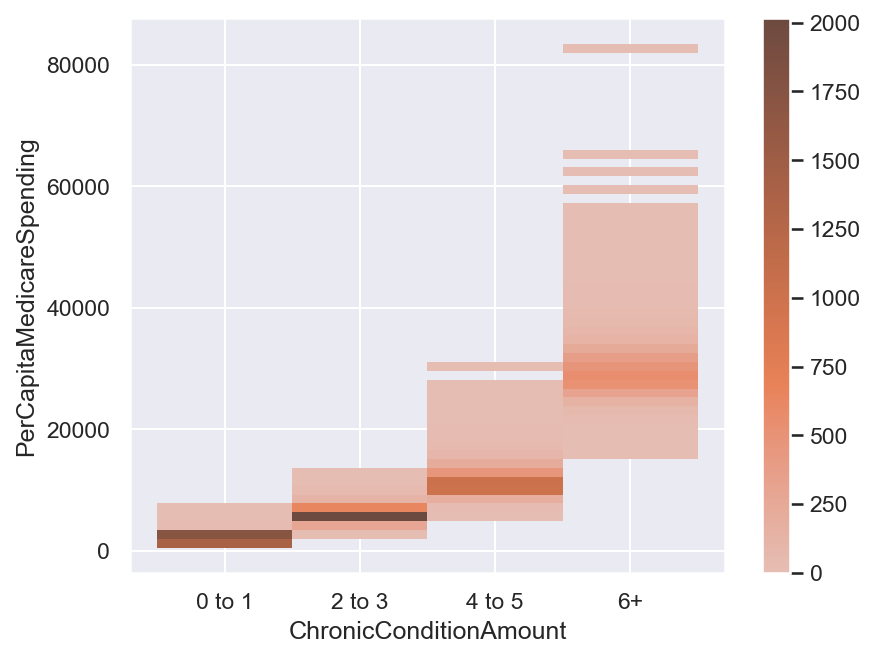
**D: Analysis**

 To perform exploratory data analysis, I will use the Python libraries Seaborn and Matplotlib as my tools for performing visualizations. I set the aesthetics of the graphs and then create my first plot of the distribution of the walkability index variable. From this histogram it is apparent that the data is positively-skewed. This makes sense as most places in the United States are not designed to be walkable, with only a few large metro areas that can be considered pedestrian-friendly.

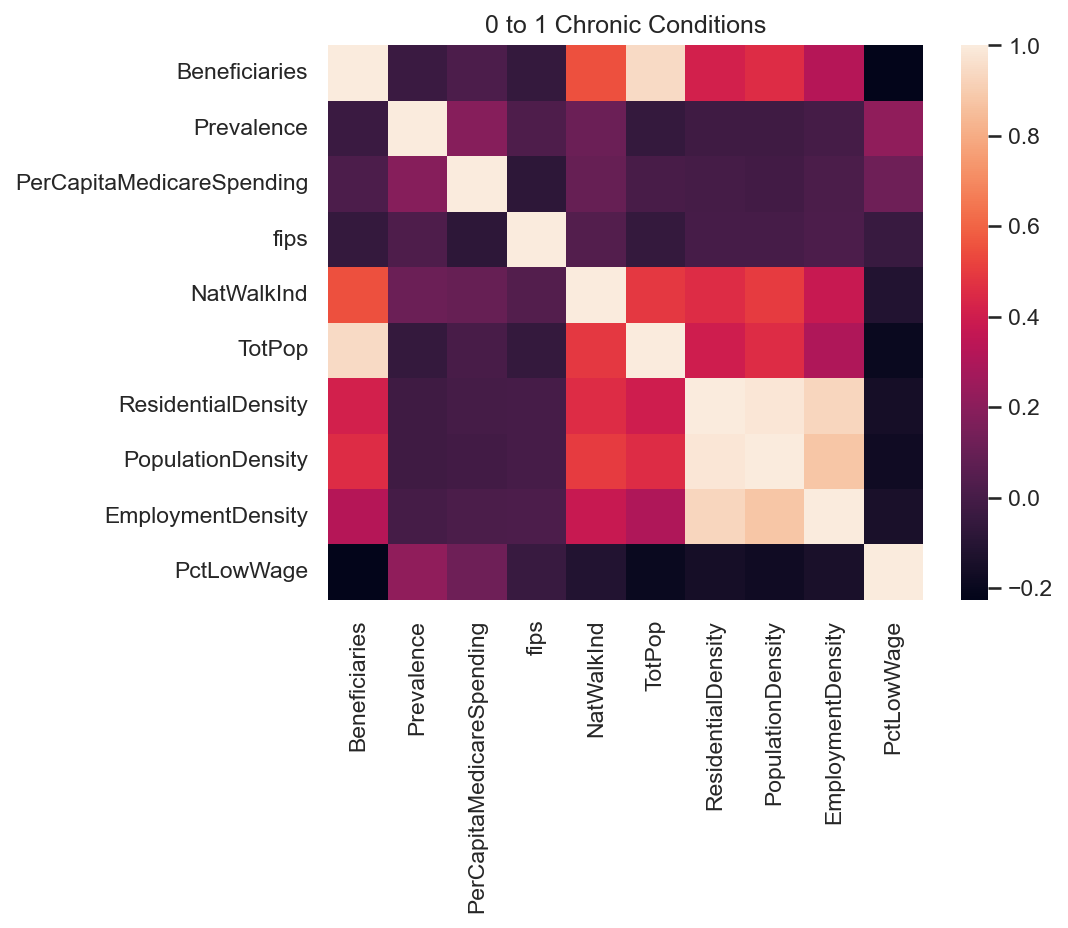
I then created a boxplot of the prevalence of different amounts of concurrent chronic conditions in every county. From this plot we can gather that, in general, Medicare and Medicaid beneficiaries with 0 to 1 chronic conditions are the most prevalent group, but there is an alarmingly high amount of counties with a 20% or greater prevalence of 6+ chronic conditions.

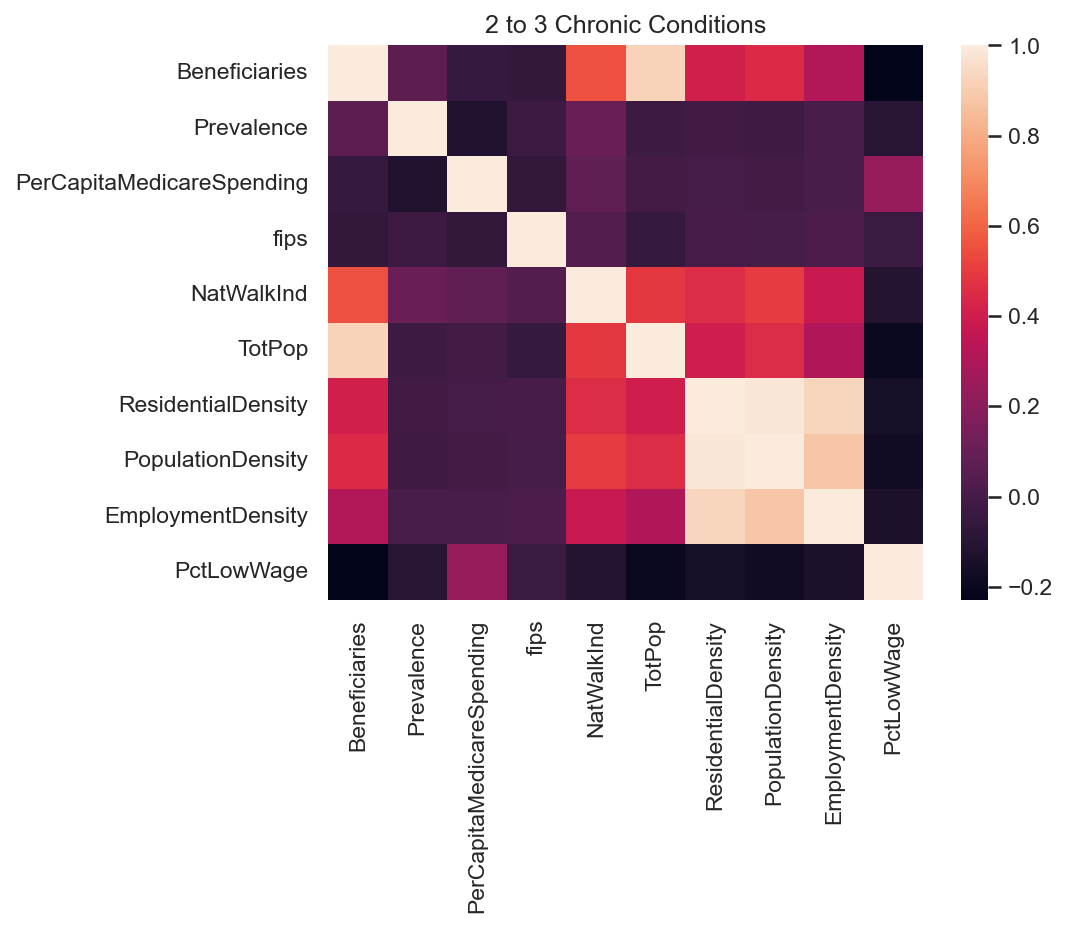


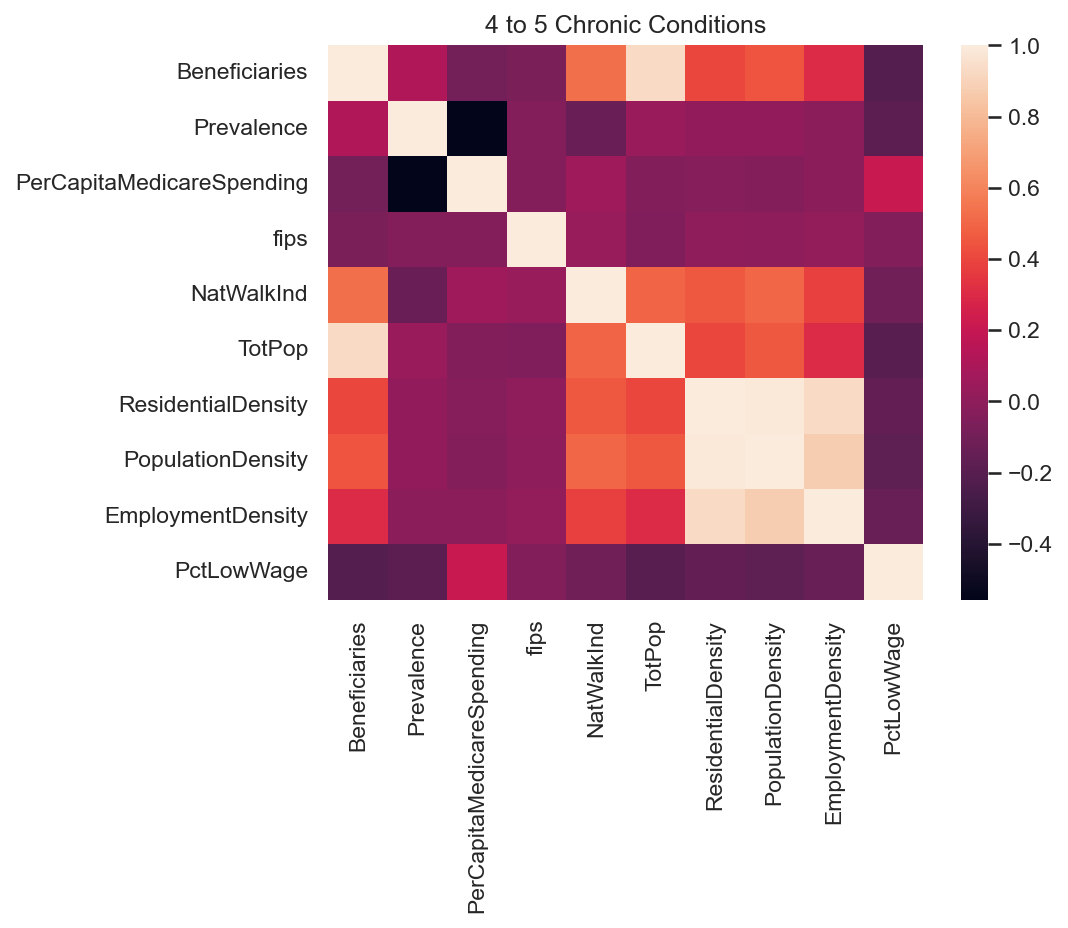
The economical reason why a high prevalence of 6+ chronic conditions is a concern is illustrated by the two-dimensional histogram below. As you can see, per capita Medicare spending of a county greatly increases as the amount of concurrent chronic conditions increases.

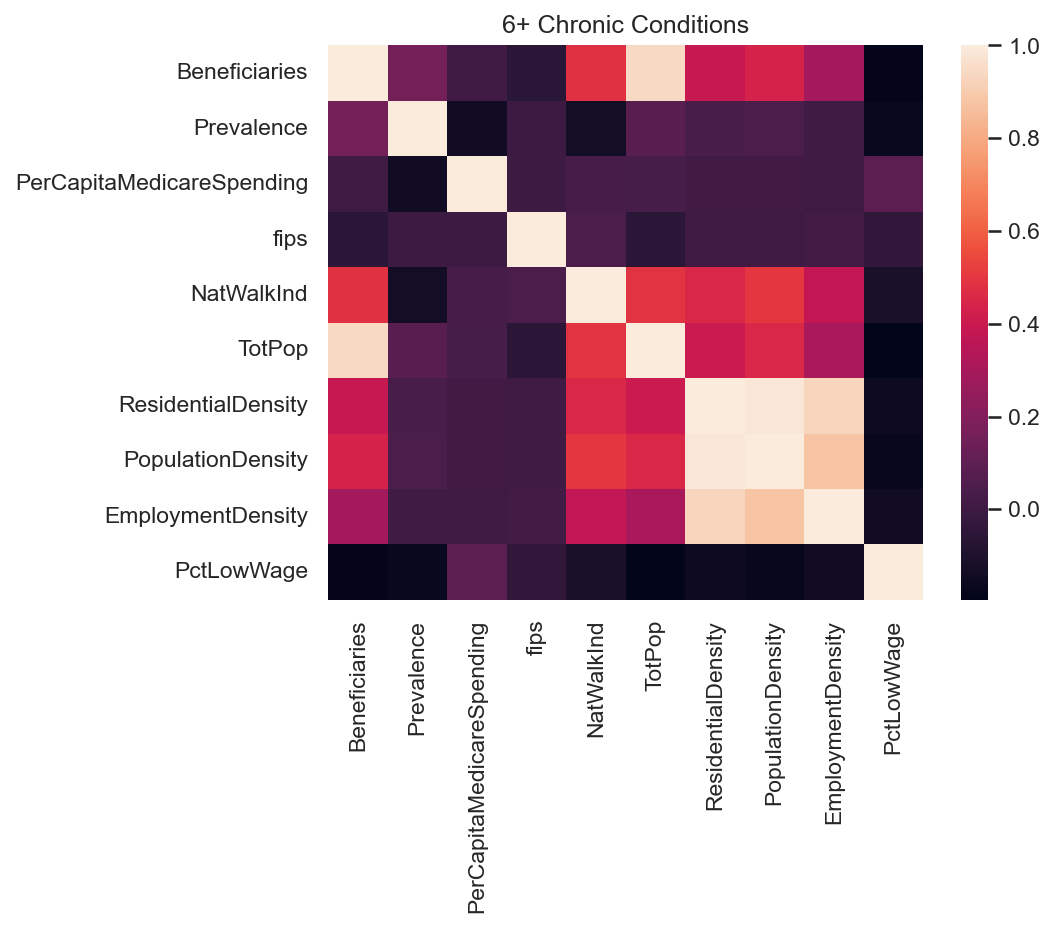


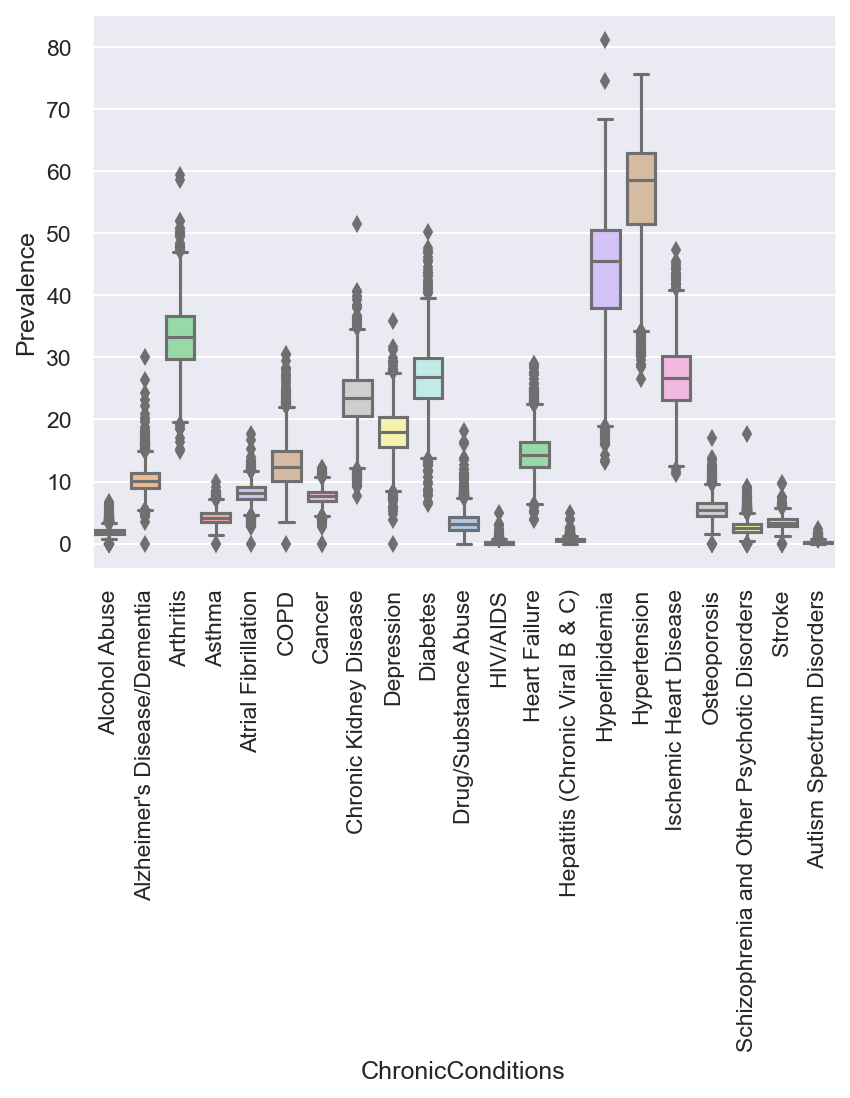
To see if there is a correlation between any of the socioeconomic factors found in the ‘walkability’ data set and the prevalence of different concurrent chronic condition amounts, I split the merged data set into four samples based on their value for the ‘ChronicConditionAmount’ column. Then I created a heatmap of the correlation matrix between each of the 4 samples’ variables. There was a very weak correlation between the ‘Prevalence’ column and any of the variables from the ‘walkability’ data set.

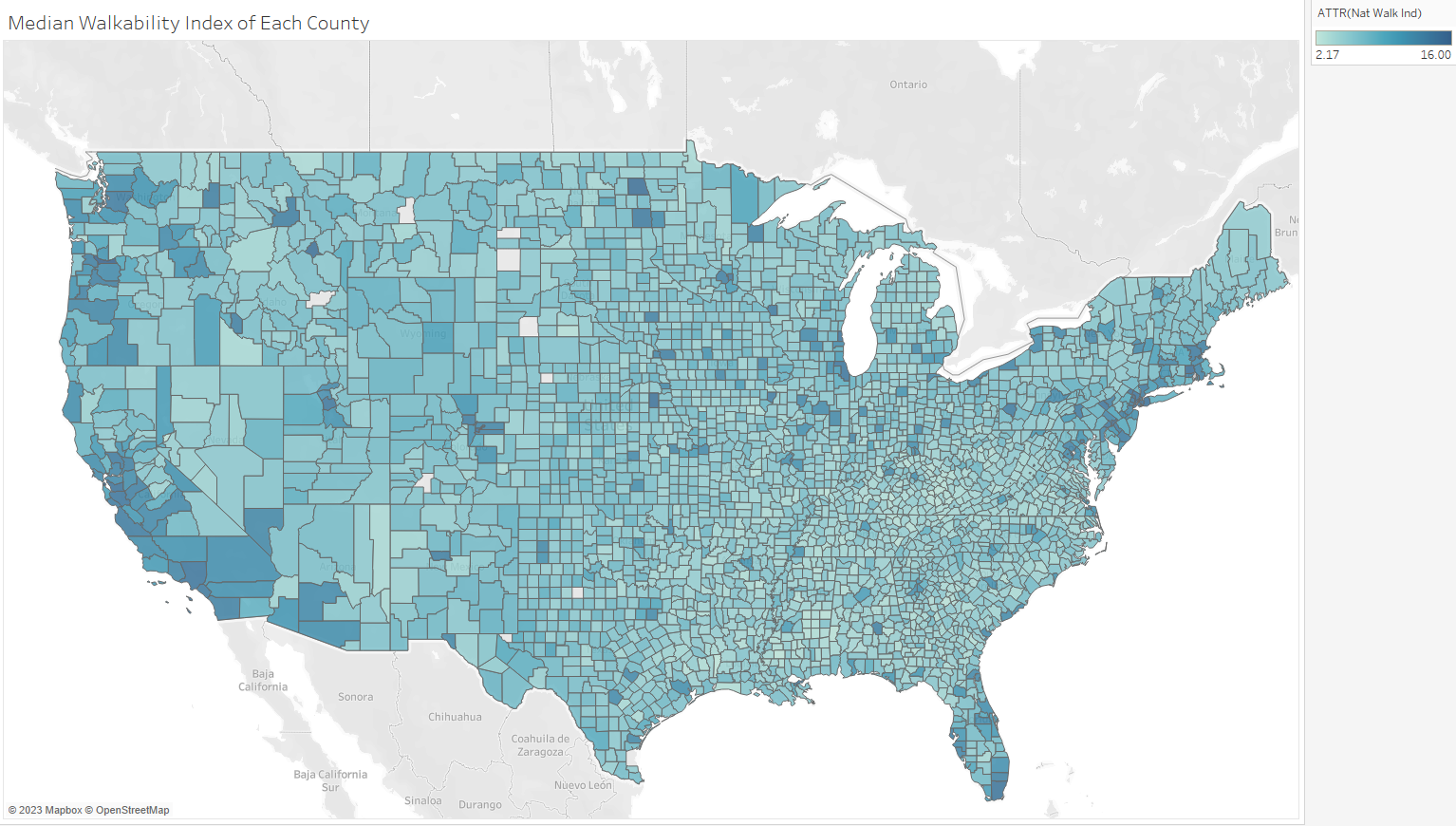




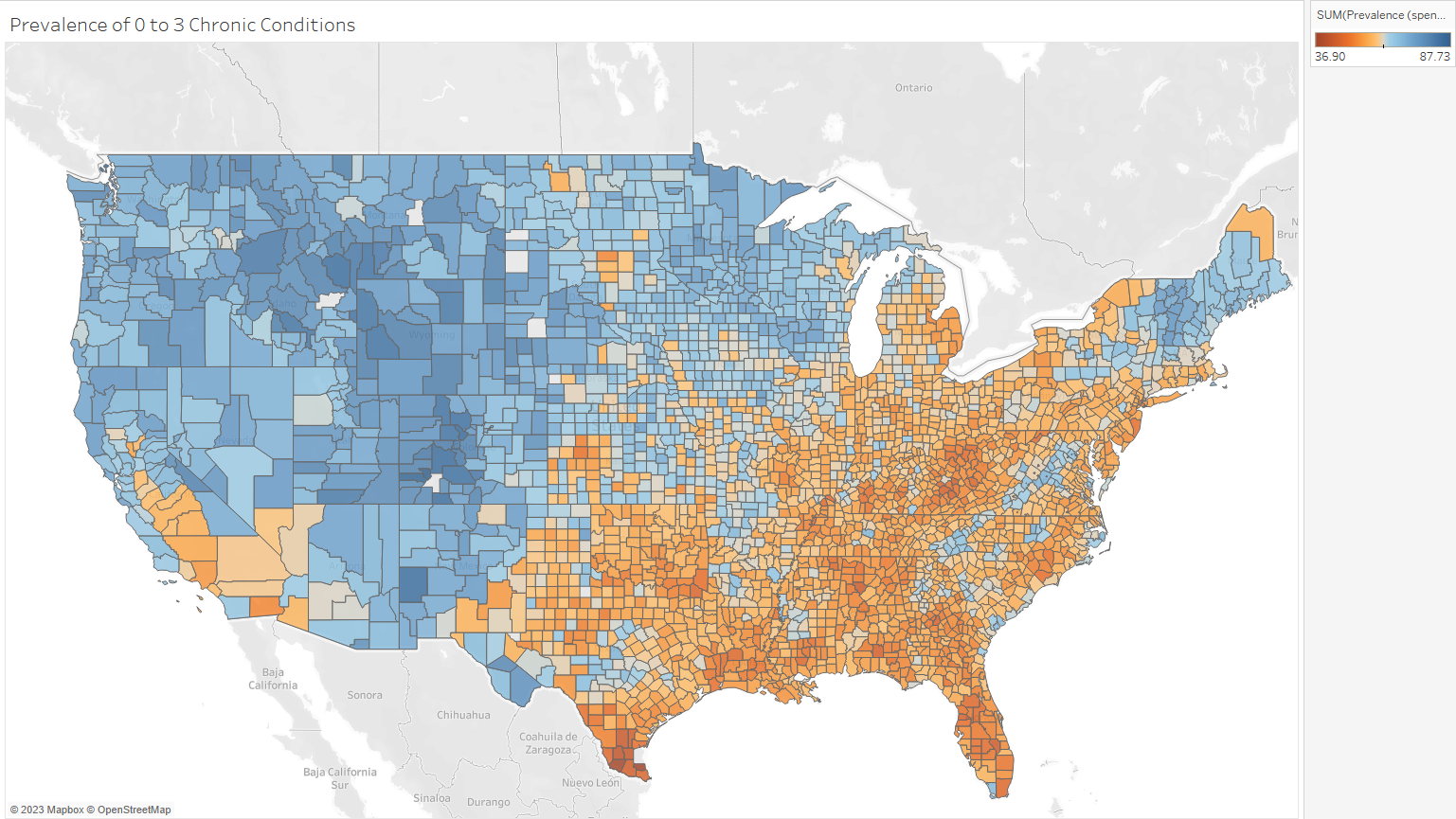




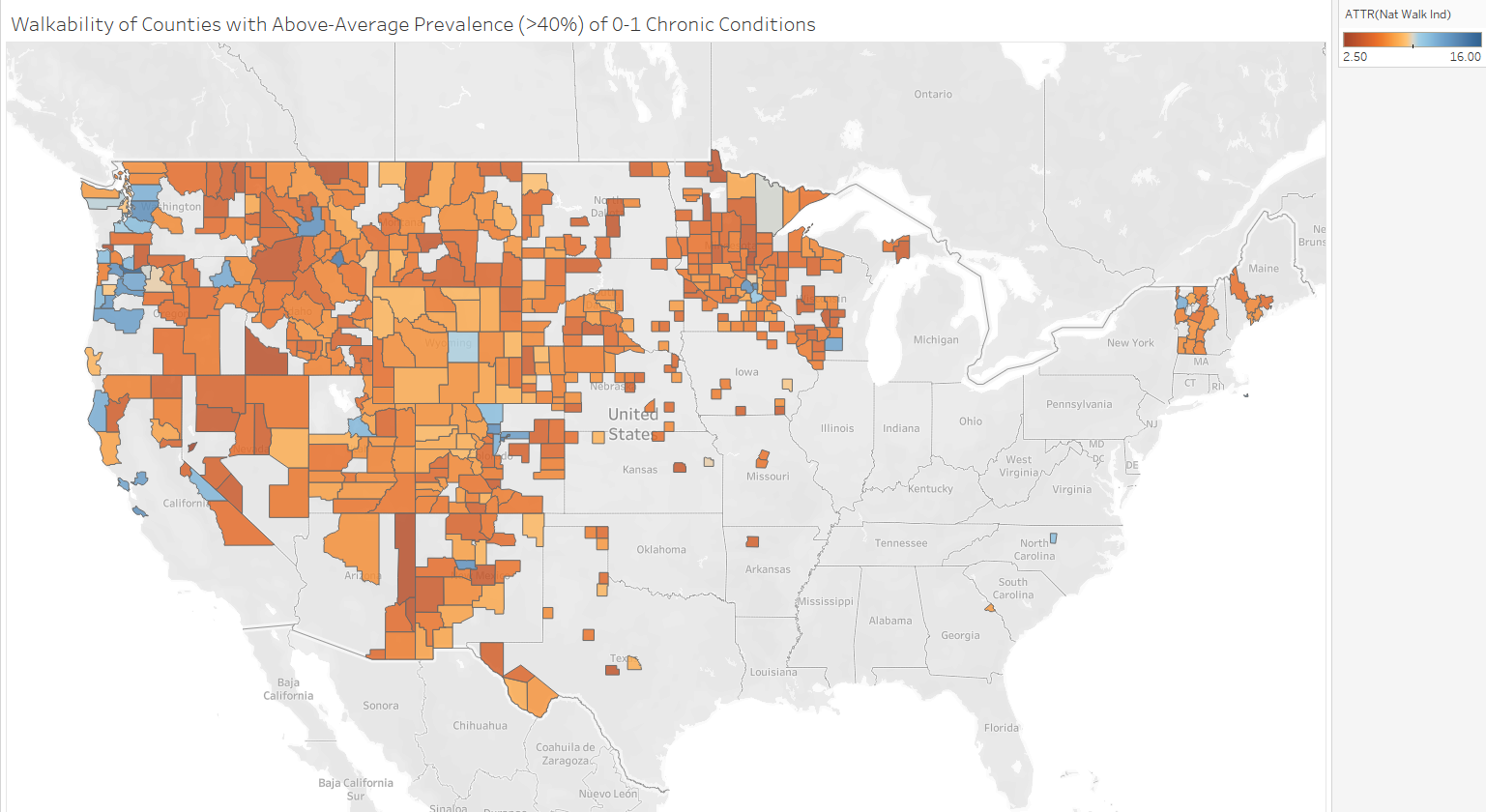
 I also wanted to see which specific chronic conditions are the most prevalent across the country. I created a boxplot showing the prevalence distribution for each chronic condition. On average, the top three most prevalent were hypertension, hyperlipidemia, and arthritis.

 I imported the data sets into Tableau so I could explore the data further and in a geographic context. I plotted the walkability of each county to find any geographic patterns behind walkable urban planning. Unsurprisingly, the most walkable counties in the U.S. where located along the coasts and in or around metropolitan areas.

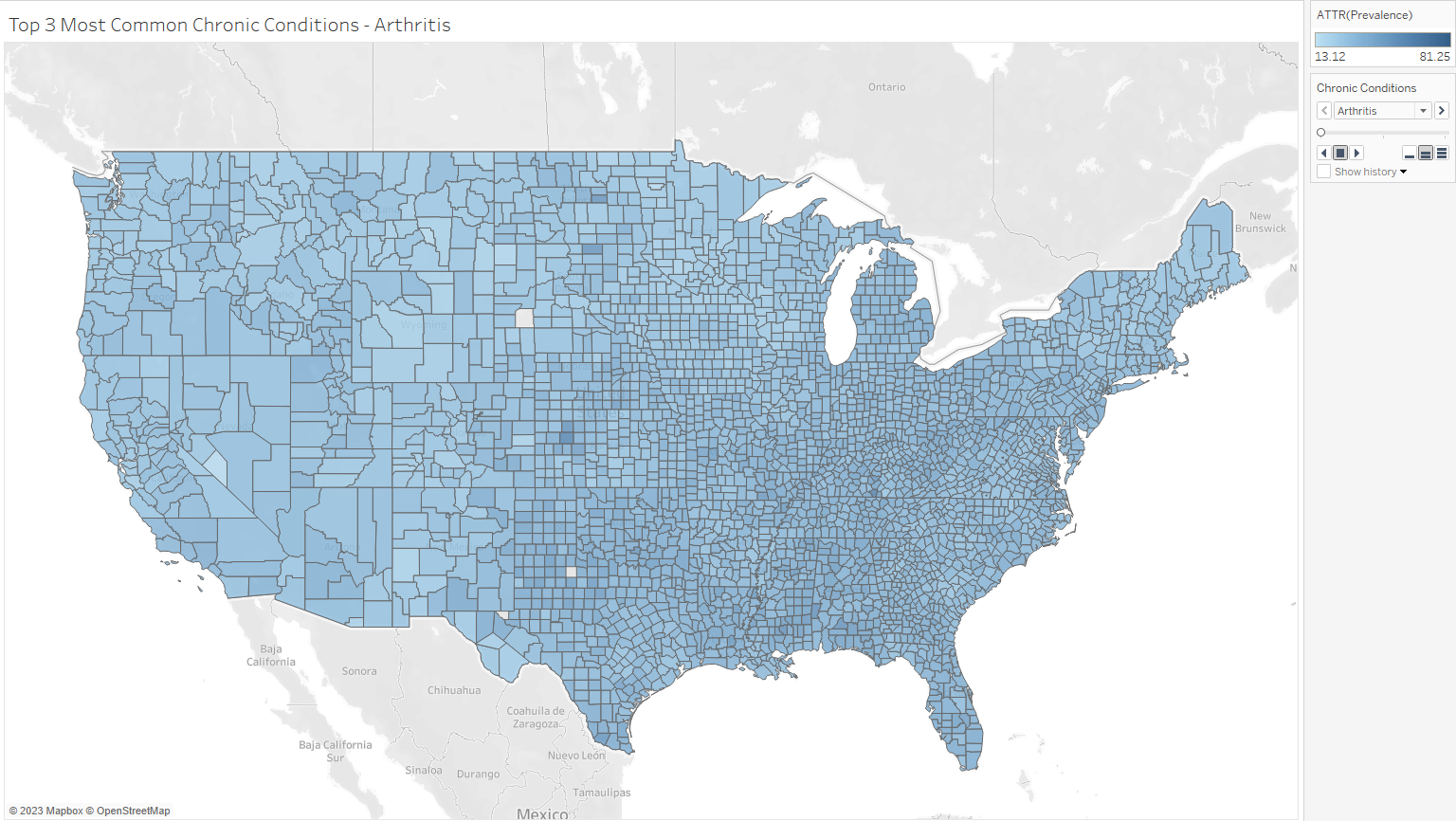
I then plotted the combined prevalence of 0 to 3 chronic conditions to dichotomize the chronic conditions variable and make it easier to understand. An interesting geographic pattern emerged where the West generally has a higher prevalence of 0 to 3 chronic conditions than the East. High concentrations of 4+ chronic conditions are seen in the South and Southeast while the opposite is found in the Midwest area.

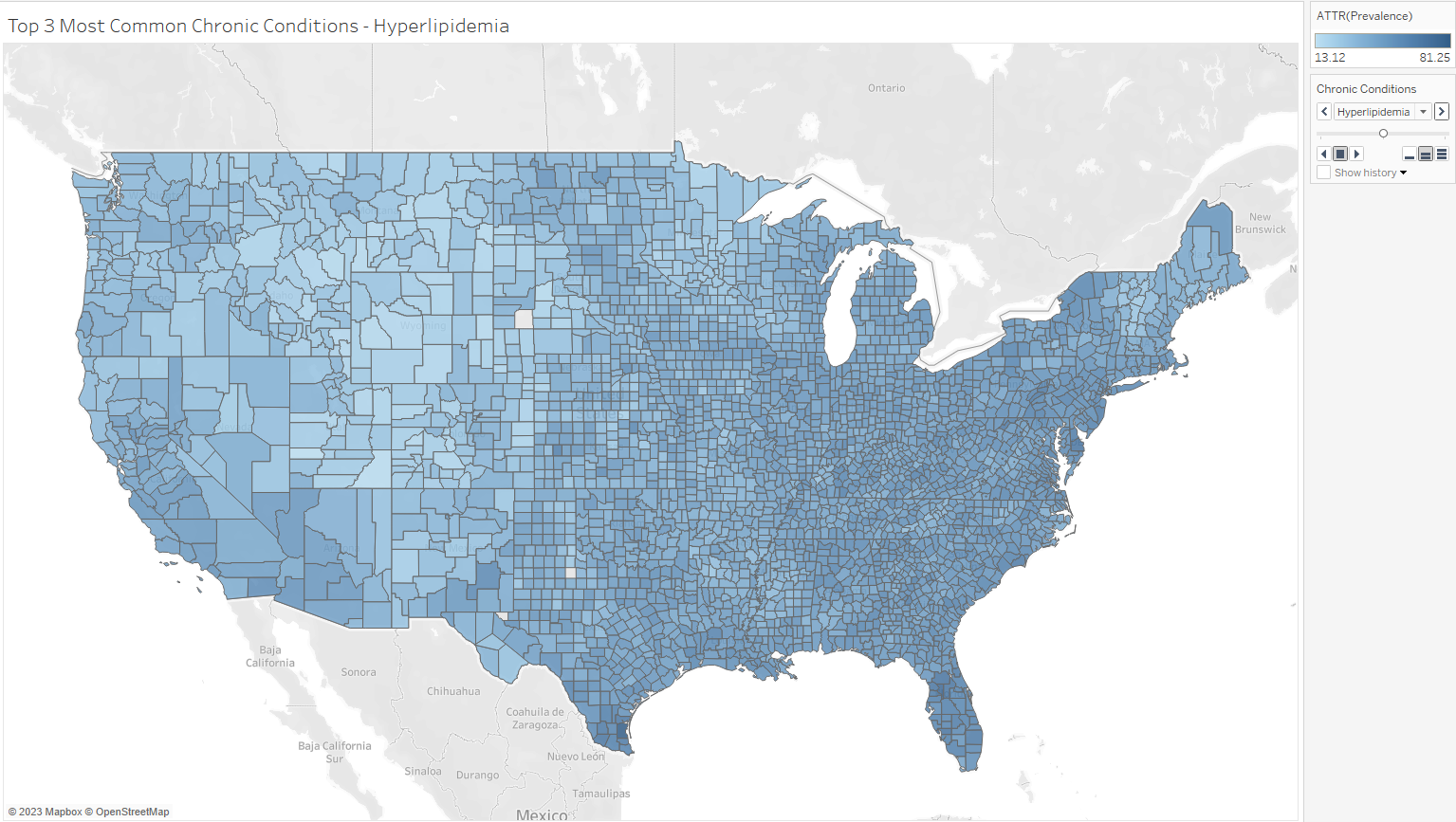
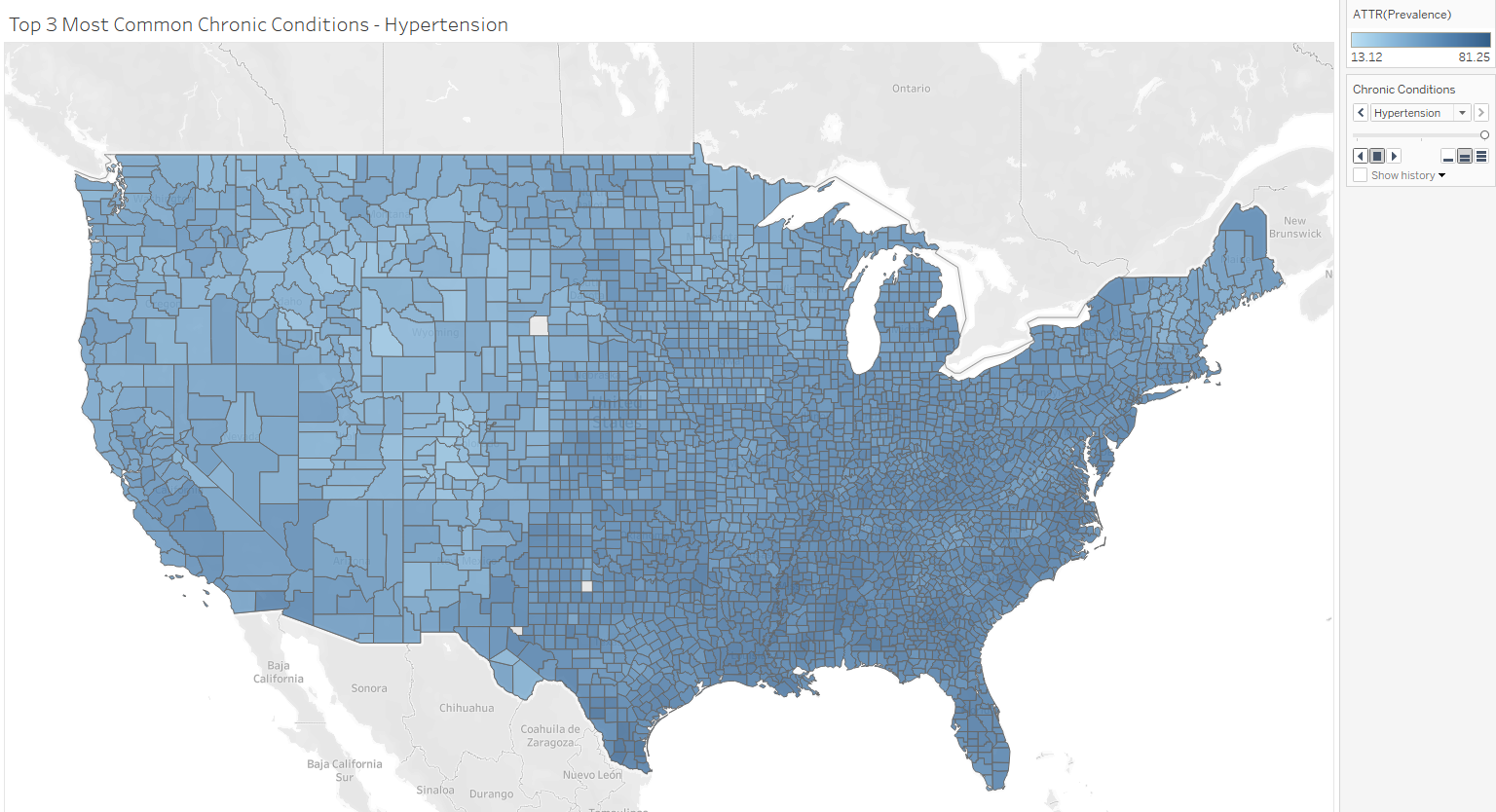


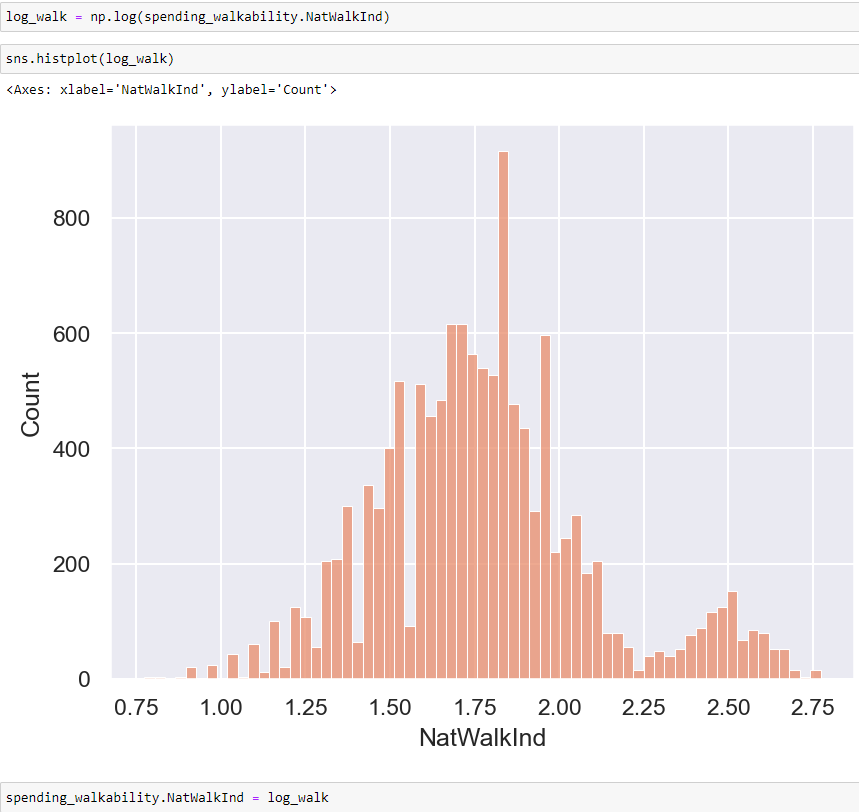
To see if walkability had any noticeable relationship with prevalence, I plotted the walkability of counties with an above-average prevalence of 0 to 1 chronic conditions. I also plotted the walkability of counties with an above-average prevalence 6+ chronic conditions. In both of the plots, the majority of counties were on the low-end of walkability. This is not to say that walkability causes one or the other, but just that the vast majority of counties in the United States are not very walkable.

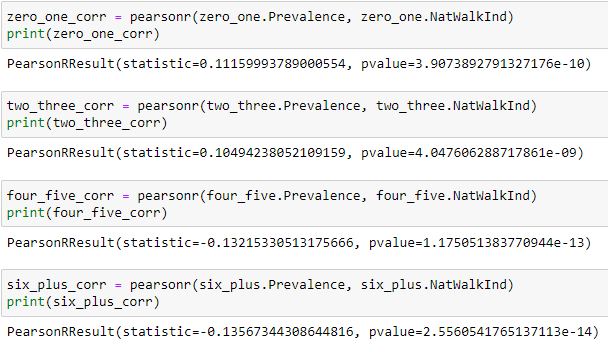


I also plotted the prevalence of the top three most common chronic conditions on a map. As expected from the previous graphs, the Southeast generally had the highest concentration of each chronic condition.

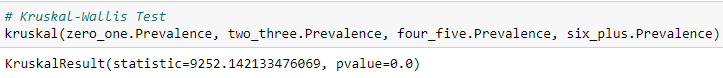




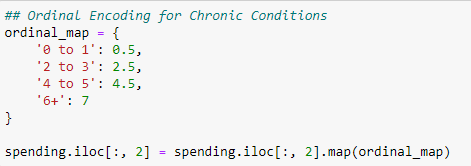
 Now the last thing to do before I begin performing tests, is addressing the issue of the main predictor variable that I am centering my analysis on, ‘NatWalkInd’, being heavily right-skewed. I applied a log transformation to the entire column and plotted the new distribution. This resulted in a much more even data distribution so I decided to replace the walkability index variable with a log-transformed version of itself.

 After sufficiently exploring the data, I set out to test my hypothesis. I performed the Pearson correlation test using the ‘pearsonr’ module from the SciPy library. For each of the 4 samples that I split up based on their amount of chronic conditions, the absolute value of the derived correlation coefficient was around 10.5% to 13.5%, suggesting that there is a weak relationship between the prevalence of chronic conditions and the walkability index of a county.

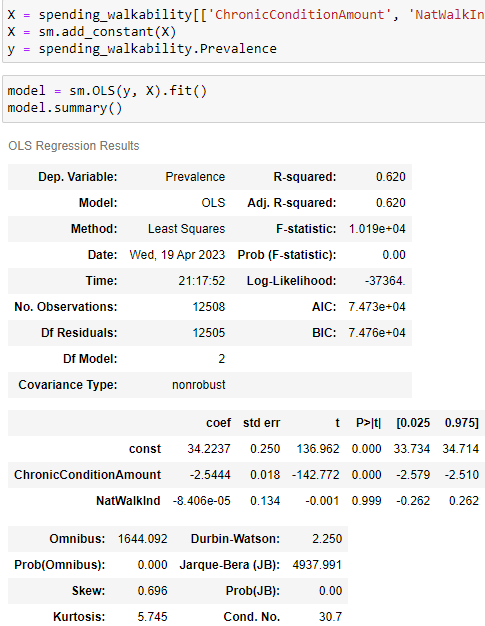
Using the ‘kruskal’ method of the SciPy library, I performed the Kruskal-Wallis test with the four samples. The advantage of the Kruskal-Wallis test is that it is roughly equivalent to one-way ANOVA and acts as a nonparametric alternative for when the normality assumption of ANOVA is not met. According to Webb (2023), the disadvantage of nonparametric statistical tests is that they are generally less efficient and require larger sample sizes than their parametric counterparts. The response variable is the ‘Prevalence’ column, I am testing to see if the median prevalence between the four groups is different. The test results show a test-statistic of 9252 and a p-value of 0.0. The high test statistic and a p-value of 0.0 confirms that there are statistically significant differences in the prevalence of different amounts of concurrent chronic conditions



Finally, I look to perform multiple linear regression to see if the log-transformed ‘NatWalkIndex’ variable can reliably predict the prevalence of different amounts of chronic conditions. First, I performed ordinal encoding of the ‘ChronicConditionAmount’ variable, allowing me to use it as an independent variable in the linear regression to represent the 4 groups of concurrent chronic conditions. I set the new values to represent the average amount of chronic conditions in numeric form, so the string ‘0 to 1’ would be represented by 0.5.



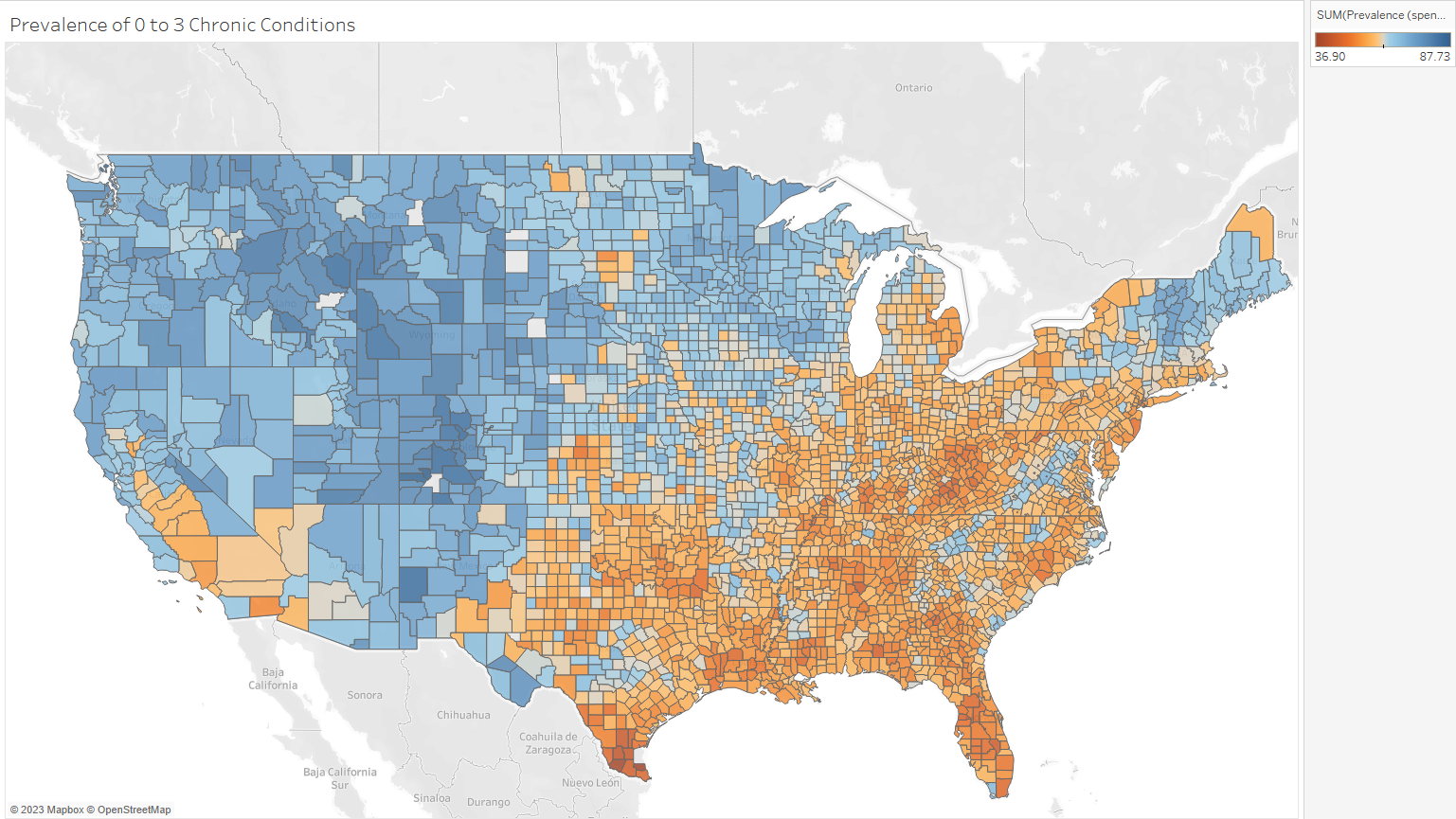
Multiple linear regression assumes that there exists a linear relationship between the independent variables and the dependent variable. In this case, the ordinal encoded ‘ChronicConditionAmount’ has a strong negative correlation with the response variable ‘Prevalence’ but the walkability index predictor variable has a weak correlation as shown from the previous correlation tests. Nevertheless, I performed ordinary least squares regression using the ‘OLS’ module from the Statsmodels library. The summary is shown below.



**E: Data Summary and Implications**

The resulting model has a fairly high R-squared value of 62% and the p-value of the F-statistic suggests that the model is statistically significant, but the coefficient of the predictor variable ‘NatWalkInd’ is very small and the p-value of 0.999 strongly suggests this predictor is not statistically significant. This confirms the suspicion from earlier tests and plots that the walkability index cannot reliably predict the prevalence of different amounts of concurrent chronic conditions in a county.

One big limitation of my analysis is that out of around 100 possible predictor variables from the Environmental Protection Agency’s data set, I only chose 5 of them to look into and ultimately settled on using just the walkability index variable for the tests. I did this to simplify the data cleaning process and to keep my analysis centered on walkability. The index developed by the EPA considered only four factors, the density of intersections in an area, the proximity to transit stops, and the diversity of land use. I believe a direction of future study of the data sets is to include more environmental factors into the analysis, of which the EPA’s dataset contains many. Another promising direction that should be considered is looking into possible causes of the difference of chronic condition prevalence between areas in the East and West of the United States.



My recommended course of action for the Centers for Medicare and Medicaid Services would be to provide data that is grouped on a city-level or even using the Census Block Group system that the EPA uses to allow greater precision and easier joining of the two datasets. Walkability can greatly vary between the neighborhoods of a city, let alone an entire county. With a higher degree of granularity, more accurate analysis can be done of the urban design of the neighborhoods that Medicare beneficiaries live in. Based on my analysis, walkability does not have a significant relationship with the prevalence of chronic conditions in a county.

**F: Sources**

Buttorff, C., Ruder, T., & Bauman, M. (2017, May 26). *Multiple chronic conditions in America*. RAND Corporation. Retrieved April 16, 2023, from https://www.rand.org/pubs/tools/TL221.html

Cassella, C. (2022, February 27). *Here's the impact that walkable neighborhoods can have on your physical health*. ScienceAlert. Retrieved April 16, 2023, from <https://www.sciencealert.com/here-s-the-impact-that-walkable-neighborhoods-can-have-on-your-physical-health>

Centers for Medicare &amp; Medicaid Services. (2022, January 10). Dashboard. CMS. Retrieved April 17, 2023, from https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Chronic-Conditions/CCDashboard

Cortright, J. (2020, January 5). *Walkable places are growing in value almost everywhere*. Strong Towns. Retrieved April 16, 2023, from https://www.strongtowns.org/journal/2020/1/22/walkable-places-are-growing-in-value

Healy, K. (2018, January 1). FIPS-codes/state\_and\_county\_fips\_master.CSV at master · KJHEALY/FIPS-codes. GitHub. Retrieved April 17, 2023, from https://github.com/kjhealy/fips-codes/blob/master/state\_and\_county\_fips\_master.csv

U.S. Environmental Protection Agency. (2021, July 3). Walkability Index. Data.Gov. Retrieved April 17, 2023, from <https://catalog.data.gov/dataset/walkability-index>

Webb, R. (2023, March 12). 13.1: Advantages and disadvantages of nonparametric methods. Statistics LibreTexts. Retrieved April 20, 2023, from <https://stats.libretexts.org/Bookshelves/Introductory_Statistics/Mostly_Harmless_Statistics_(Webb)/13%3A_Nonparametric_Tests/13.01%3A_Advantages_and_Disadvantages_of_Nonparametric_Methods>