Final Report :

An Analysis of Walkability and Air Quality in the United States

Quinn Jones and Clayton Nolan

Introduction

Our final investigated two datasets, one a collection of factors contributing to the overall Walkability- a quantified measure ranging from 1 to 20 of the ease pedestrians have traversing the area- and their Walkability rankings, the second a collection of Air Quality ratings of different regions, spanning over multiple decades. The Air Quality index we specifically used was the PM2.5 which measures exposure to pollution from particles smaller than 2.5 micrometers . We chose these two to explore the relationship between Walkability and Air Quality, what correlations existed between demographic factors such as transit access and auto ownership and how those factors did or did not affect either measurements. Of particular interest to us was finding the areas with the worst or best of either measurements and researching, through further analysis and outside sources, why these regions were at either extreme.

Methods

Cleaning:

Walkability: We got both datasets from data.gov, so there were few null values. This meant that while cleaning the walkability we were less so concerned with reducing the number of points we had and more so concerned with narrowing down what we could use and consistent data collection. For example, according to the documentation all the data collection for the U.S territories were collected in different ways, skewing the data so we made the decision to drop all of those entries. Additionally we dropped 48 columns due to their lack of use for our project. Most of these were specific information on Employment which we did not need to investigate in our project. We then added a few more columns we thought could be useful such as combined State and County Fips codes and the percentages of medium and high wage workers living in an area.

Air Quality: The air quality dataset we found contained no null values and included a wide range of data collection methods, measurements, and locations across the US, covering almost every county. We decided, for simplicity, to compare it to the complicated walkability dataset we collected during the data collection method, namely the PM2.5 annual average. This is the measurement often used by organizations such as the EPA to assess air quality. We averaged this number across counties from 1999-2013, giving us the average PM2.5 for almost every county in the US.

Storage:

For storing our datasets we opted for MongoDB for two major reasons. Our first reason was the amount of columns in our Walkability dataset. Even after removing so many there were 71 total columns. Creating a schema for all of those in SQL was going to be extremely time consuming for little pay off. Additionally, analysis was going to be more straightforward and versatile in MongoDB with its seamless integration with pymongo and the aggregation pipeline methods. The one downside to not using SQL was the complexity of merging the tables however once we found a workaround through the pandas merge methods it was smooth sailing.

Analysis and Findings

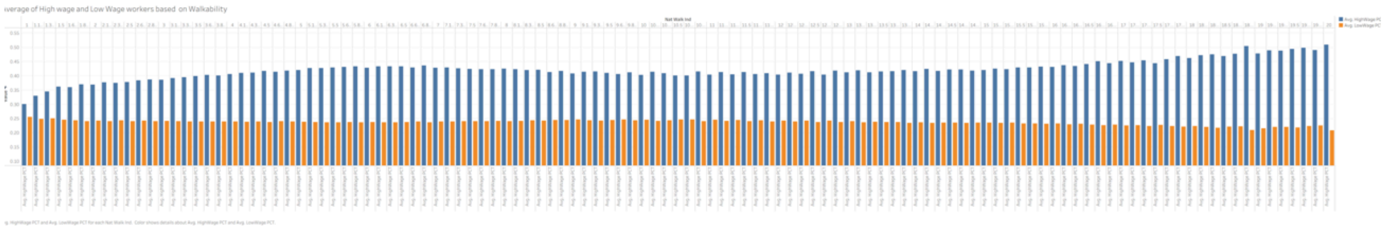
Walkability:

We explored four questions with exclusively the Walkability data set:

- *Q1: Which cities or counties rank as the most walkable in the country, and what factors contribute to their scores?*
- *Q2: How does public transport access affect walkability?*
- *Q3: How does public transport access affect the amount of cars owned per household?*
- *Q4: What is the average Walkability of each County*

Here is how we investigated each and what we found.

Q1: To answer this question I first found where the highest areas of walkability are. Within the top 10, 6 are in California. This coincides with our graph of the Average High Wage and Low Wage workers depending on Walkability. The visualization shows that the more walkable a place is, the higher the percentage of High wage workers live there and California, with the largest economy in the United States, follows this trend. These both suggest a correlation between the wages of the people who live there and the Walkability. I go on to explore this idea but first I check and see if the total protected land of an area changes the Walkability, which it does not seem to do. I investigated this by checking the Total protected land of areas with high walkability and upon reviewing the values there seemed to be no correlation. I then went on to investigate the relationship between the wages and Walkability. I did this by finding the average walkability of the higher wage communities and Lower wage communities and determined which were which by using a match stage. For high wage communities the match selected entries that had a greater percentage of high wage workers than low wage workers and for lower wage it was the reverse. Upon finding the average of these two groups you can see that they are quite similar. There is slightly higher average Walkability within low wage communities but this is contradicted and given more context by our Visualization.



g. HighWage PCT and Avg. LowWage PCT for each Nat Walk Ind. Color shows details about Avg. HighWage PCT and Avg. LowWage PCT.

This Visualization shows us that no matter the Walkability there isn't much fluctuation between the amount of Low wage workers in a region however there is a positive correlation between the number of high wage workers in a region and higher Walkability. This may suggest two things, one that people who have a higher income and can afford to move, choose to move to areas with greater walkability, or that economically prosperous areas that have high wage jobs and need more High wage workers, like cites, are in turn more walkable.

Q2: To answer this question, I calculated the average walkability of the 20 areas with the best transit access and 20 areas with the worst transit access. This revealed a drastic change, showing that areas with poor transit access have much worse Walkability, coming in with an average of 8.4 and areas with good public transit access have considerably better Walkability, coming in at 14.8.

Q3: For this question I once again found the best and worst 20 areas for transit and then found the average amount of people who owned no cars, one car or two or more cars. We can see through the results that the average amount of people without cars increases in areas with good transit access, and in areas with poor transit access the average amount of people with 2 or more cars also greatly increases.

Air Quality:

- *Q1: Which county has the worst average air quality?*

-Q2: Which state has the worst average air quality

-Q3: What is the average air quality for every region in the US

-Q4: Which areas have the worst air quality, and what environmental, industrial, or demographic factors might explain these conditions?

Q1: This was investigated by sorting by PM2.5, then taking the data in reverse order. This told us that Riverside, California, has the worst air quality, with a score of 21. This can lead to severe health side effects over a long period of time.

Q2: We used a MongoDB aggregation pipeline to group the data by state, calculate the average PM2.5 for each state, sort the results from highest to lowest average PM2.5, and limit the output to the state with the highest average PM2.5. This analysis showed that Ohio has the worst average air quality in the dataset, with an average PM2.5 level of approximately 13.91, meaning it experiences higher fine particulate pollution than the other states on average.

Q3: I calculated the average PM2.5 level for each U.S. region by grouping states into the Northeast, Midwest, South, and West, then averaging their long-term PM2.5 values. The results show that the South has the highest average PM2.5 levels, followed by the Northeast and Midwest, while the West has notably lower levels. The West stands out as an interesting outlier, likely because its larger geographic size results in more dispersed populations and industries, which reduces sustained average pollution levels. However, this lower average can conceal short-term spikes from events like wildfires, which don’t fully appear in long-term regional averages.

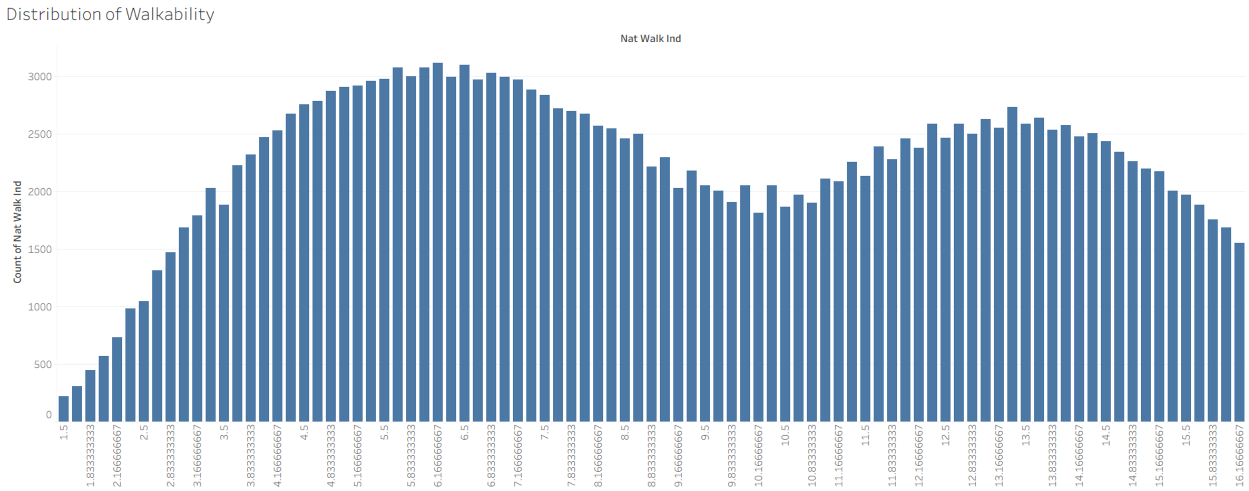
Q4: We used a MongoDB aggregation pipeline to group the data by state and county, calculate the average PM2.5 level for each county, sort counties by pollution from highest to lowest, and select the top 20 worst cases. This analysis shows that the worst air quality is concentrated in specific counties rather than entire regions, with many of the highest PM2.5 levels found in California counties such as Riverside, Kern, Los Angeles, and those in the Central Valley. These areas are likely affected by factors like wildfire smoke, heavy traffic, industrial and agricultural emissions, and geographic features that trap pollution. Other high-pollution counties in states like Pennsylvania, Ohio, Illinois, and Michigan suggest that dense populations and industrial activity also play significant roles, showing that regional averages can mask serious local air-quality problems.

Combined:

We explored four questions with both data sets:

- *Q1:Do areas with poorer air quality tend to have lower walkability scores compared to areas with cleaner air?*
- *Q2: How does population affect air quality?*
- *Q3:How does income affect air quality?*
- *Q4:How does access to public transport affect air quality?*
- *Q5:How does car ownership per household affect air quality?*

Q1: To answer this question I sorted the Air Quality in ascending order and calculated the average Walkability of the top 50 areas with the highest Air Quality. Then I did the reverse, sorting it in descending order and calculating the average Walkability of the worst 50 areas. The two resulting numbers 6.5 and 7.3 are very close to each other, less than a degree off, suggesting that the Walkability of an area is not a large factor in the Air Quality. Interestingly enough if we take a look at our Walkability Index distribution visualization, we can see both of these averages are right within the most common Walkabilities, further implicating that Walkability is not a factor in the Air Quality.



Q2: To answer this question I calculated the total population of different counties and average Air Quality by using the group stage of the aggregation pipeline and then sorted it by population in ascending order. As we can see in the results, there is a correlation between worsening Air Quality and higher population.

Q3: We examined how income relates to air quality by using the percentage of low-wage workers as a proxy for income and splitting areas into two equal groups: those with a lower share of low-wage workers and those with a higher share. We then compared the average 15-year PM2.5 levels between these two groups. The results show that PM2.5 levels are very similar across both groups, suggesting that at this aggregated level, income differences do not strongly correspond to differences in air quality, and that broader regional averages may be masking more localized income-based disparities.

Q4: We analyzed how access to public transportation relates to air quality by measuring transit access based on distance to the nearest transit stop and grouping areas into good, medium, and poor transit access categories. We then calculated the average 15-year PM2.5 levels for each group. The results show that PM2.5 levels are highest in areas with good transit access and lowest in areas with poor transit access, suggesting that higher pollution is associated with better transit access. However, this pattern likely reflects higher urban density, traffic, and economic activity in these areas rather than transit access itself being a direct cause of increased air pollution.

Q5: We examined how household car ownership relates to air quality by grouping census areas into four U.S. regions and calculating regional averages for PM2.5 levels, alongside the percentages of households with 0, 1, or 2+ cars. The results show that the West has the lowest average PM2.5 levels despite having the highest share of multi-car households. In contrast, the Midwest and Northeast have higher pollution levels with lower multi-car ownership. This suggests that car ownership alone does not explain differences in air quality, and that broader factors such as urban density, industrial activity, transportation patterns, and geography play a larger role, highlighting the limitations of using regional averages to infer direct causal relationships.