Walk It Off: The Relationship Between Public Transit, Walkability and Wellbeing in 100 U.S. Cities

SI 330 Final Project Report
By Haley Johnson
April 14, 2022

Motivation

America consistently lags behind other wealthy nations when it comes to public transit. Countries in Europe and Asia report higher ridership, longer service hours and more consistent coverage than in the United States. Some have attributed this to the United States's car-centric society, which has prioritized cars over other forms of transportation when designing cities.

Meanwhile, Americans frequently complain about long commutes, heavy traffic, and noise pollution from vehicles. This poses the question: is our car-centric society making Americans unhappy? To what extent do these environmental factors impact our overall health and wellbeing?

This report attempts to answer these questions by examining correlations between happiness, walkability, and access to public transportation in 100 U.S. cities.

Data Sources

Happiness Scores

To measure happiness, I used WalletHub's 2022 ranking of the happiest cities in America. Happiness is a complex and subjective phenomenon that no metric can ever perfectly capture. Nevertheless, WalletHub's ranking was developed in consultation with experts in psychology, organizational management and human behavior, and their ranking is made up of over 30 relevant metrics. In addition to reporting overall scores, the dataset also ranks cities in three categories: emotional and physical wellbeing, income and employment, and community and environment.

To access this dataset, I saved WalletHub's web page as an HTML file. There was no missing data in the dataset. Scores are only reported for the top 100 cities in the U.S., so it only included 100 records. WalletHub revises these rankings annually and this dataset was produced using the most up-to-date information for all the metrics used to calculate overall scores. Within the individual metrics, however, it's unclear when the most recent data was published. For example, it is not stated if WalletHub used life-expectancy data from 2022 or from 2021.

Walkability & Access to Transit

Walkability and access to public transit were measured using the Environmental Protection Agency's Smart Locations dataset. The dataset reports information for every census block group used in the 2019 census and was last updated in January 2021.

I downloaded the dataset directly from the EPA's website as a comma-separated value file. Since data is reported at such a granular level, the full dataset included 203,645 rows and 122 columns. I was particularly interested in the following columns:

- **NatWalkInd** = National Walkability index scores for each census block
- **D1B** = Population density on unprotected land, in people per acre
- **D3B_Ranked** = Ranking of the density of crosswalks on a scale of 1-20
- **D4A** = The distance of the population center from the nearest public transit stop, measured in meters
- **D4A_Ranked** = Ranking of the distance of the population center from the nearest public transit stops on a scale of 1-20

I had already identified the 'CBSA_Name' (Core-Based Statistical Area) column as the one I'd be using to join the two tables together, so any rows where 'CBSA_Name' was null were dropped. There were no other rows with missing data. There were, however, cases where the dataset used placeholder values to represent an absence. If there were no transit stops in a census block group,, column D4A was reported as -99999.00. I converted -99999.00 to NULL so these rows would be ignored when I used aggregation functions (i.e. calculating the mean distance). None of the other columns I used had missing values or placeholders for missing values.

The EPA uses a fairly complex naming system for the smart locations dataset. Rather than renaming over 100 columns, I familiarize myself with their naming conventions. A comprehensive user guide is available here.

Data Manipulation Methods

Happiness Scores

Minimal manipulation was required on the happiness dataset. WalletHub's original table included a column called 'location,' which included the name of the city and the state it's located in. I used the .split() method to turn 'location' into two columns: one for the name of the city and one for the name of the state.

Similarly, the original naming scheme included spaces in many column names. Spaces can sometimes make it difficult to reference columns using their name, so I renamed them for ease of use.

```
happiness_df = happiness_df.rename(columns = {'City': 'Location', 'Total Score': 'total_score',
                                               'Emotional & Physical Well-Being':'emotional_physical_wellbeing',
                                               'Income & Employment': 'income_employment',
                                               'Community & Environment': 'community_environment'})

happiness_df['state'] = happiness_df['Location'].apply(lambda s: s.split(",")[-1])
happiness_df['city'] = happiness_df['Location'].apply(lambda s: s.split(",")[0])

happiness_df = happiness_df.drop(columns = ['Location'])
```

Code to rename and transform columns in the WalletHub dataset.

Walkability & Access to Transit

The smart locations dataset required more manipulation to prepare it for analysis. First, the four columns that contained rankings — D2A_Ranked, D2B_Ranked, D3B_Ranked and D4A_Ranked — were converted to categories. Since those columns contain numeric rankings from 1 to 20, pandas automatically casted those column integers.
The 'category' datatype is a better representation of what the underlying data represents.

```
rankings = ['D2A_Ranked', 'D2B_Ranked', 'D3B_Ranked', 'D4A_Ranked']

for rank in rankings:
    locations_df[rank] = locations_df[rank].astype('category')
```

Code to convert ranking columns to categories.

Data was reported at the census block level and the column CBSA_Name reported the name of the metropolitan area that the census block fell in. In many cases, census blocks were associated with more than one metropolitan area.

For instance, one record's CBSA_Name was 'New York-Newark-Jersey City, NY-NJ-PA.' Again, I used string manipulation to separate the CBSA_Name into two columns: one with a list of all the cities it was associated with the record and one with a list of all the states associated with the record. I then used the .explode() method so that each metropolitan area would be its own row. This significantly increased the size of the dataset After using .explode(), there were 687,736 records.

```
exploded = locations_df.explode('city')
exploded = exploded.explode('state')
exploded = exploded.drop_duplicates()
```

```
exploded['city'] = exploded['city'].str.strip().str.upper()
exploded['state'] = exploded['state'].str.strip().str.upper()
```

Code to extract and normalize city and state names.

Joining

The two datasets were joined together using city and state names. Before joining, the city and state columns were converted to be in uppercase text. I also removed whitespace from those columns.

There is some ambiguity in how city names are reported. For instance, one dataset may use 'New York City, NY,' while another could label that same location, 'New York, NY.' To check for consistency between the datasets, I examined how the following edge cases were labeled:

- New York City, NY
- St. Paul, MN
- Washington, D.C.

After examining these edge cases, I determined that there appears to be a high degree of consistency between the two datasets.

Analysis & Visualization

Motivations

Urban planners and elected officials alike aim to institute policies that will improve their constituents' wellbeing. Given the crucial role that transportation infrastructure plays in our day-to-day lives, it is worth examining its relationship to subjective measures of happiness. Subjective measures of happiness, life satisfaction, etc. are typically not measured by government agencies. Thus, it's often necessary to pull in information from outside sources. My analysis of the impact of public transportation and walkability on wellbeing is not possible by looking at just one of the data resources I used.

Manipulating Combined Dataset For Visualization

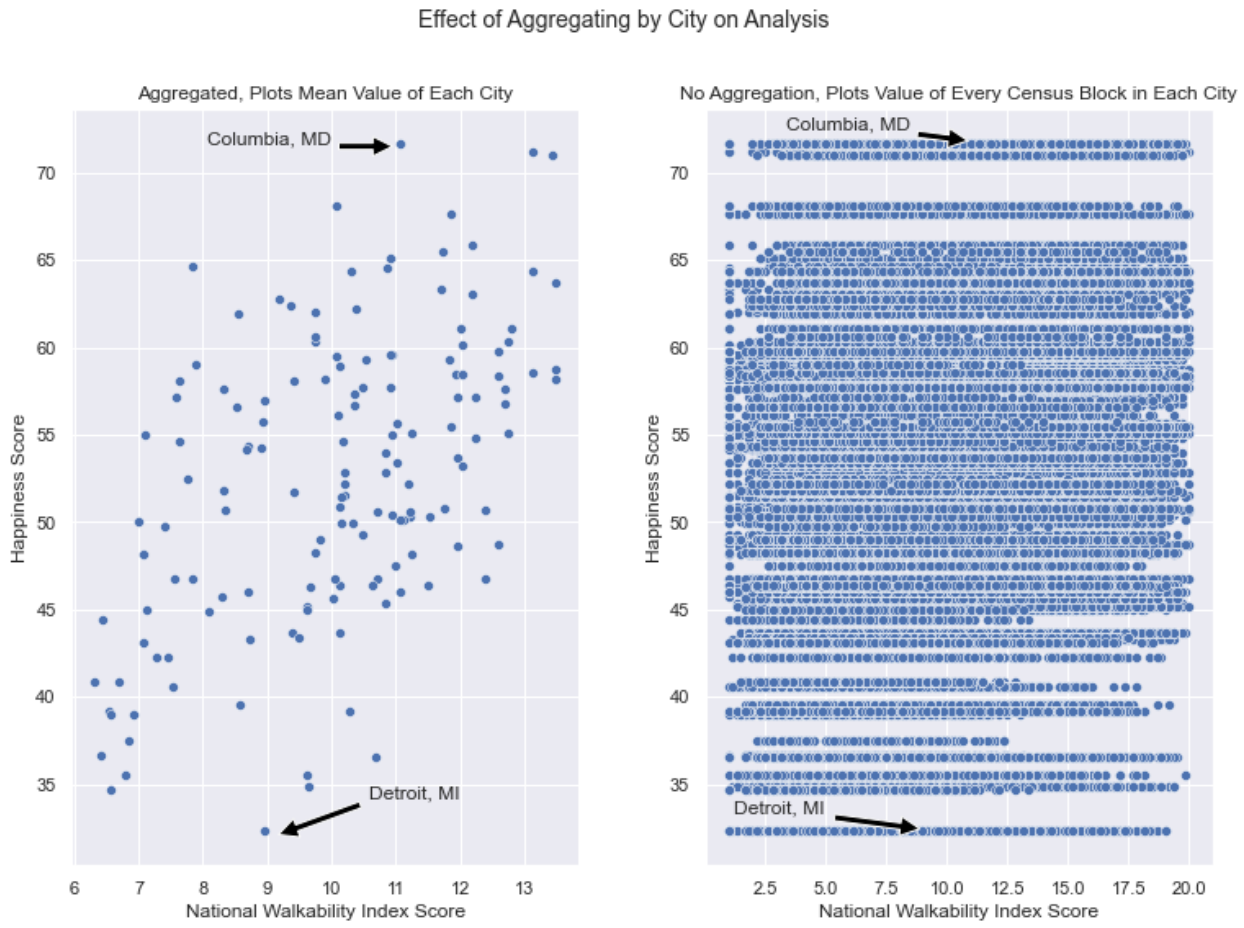
There was a many-to-one relationship between the smart locations dataset and happiness scores dataset. That is, each city contains multiple census blocks. Therefore, each row in the WalletHub dataset matched multiple rows in the EPA dataset. For some analyses, it made sense to look at the average value of a variable across all the census blocks in the city. For instance, downtown cores are likely much more walkable and central to public transit than the outskirts of a city. Aggregating across census blocks provides a much more accurate picture of the overall state of transportation in a city, because it allows these geographic variations to cancel out. To get this aggregate view, I created a pivot table. Each row contained the mean value of each column for every city in the dataset.

```
matches = exploded.merge(happiness_df, on = ['state', 'city'])
```

```
vals = list(matches.columns)[11:-8] + list(happiness_df.columns)[:-2]
df = pd.pivot_table(matches, index = ['city', "state"], values = vals, aggfunc = 'mean')
```

Code to create dataframe with aggregate values for each city.

Aggregating also made visual analysis much easier. Since data is reported at the census block level in the EPA dataset but at the city level in the happiness dataset, it made it difficult to meaningfully visualize the relationship between variables that came from different datasets.



Effect of aggregating by city on analysis. The graph on the right is much more readable and interpretable than the graph on the left, which contains a point for every census block in the dataset.

To illustrate this challenge, say that City A has 100 census tracts and City B has 200. A scatterplot of variable 'x' vs happiness scores will show 100 points where $total\ score\ =\ City\ A's\ Score$ and 200 data points where $total\ score\ =\ City\ B's\ Score$. When taking the city's mean value, however, both City A and City B will be represented by one point on the scatterplot.

Finally, to examine regional trends in access to transportation and walkability, states were assigned to one of the nine geographic regions designated by the U.S. Census Bureau. Puerto Rico appears in the dataset but is not assigned to any region, so it is labeled as a 'U.S. territory.'

National Trends

There was a positive association between happiness scores and national walkability index ratings. Similarly, there was a negative relationship between the distance to public transit stops

Scatterplots of happiness score vs. walkability index rating and happiness score vs. distance to public transit.

and happiness scores, indicating people are happier when transit is more readily available. However, this relationship was much weaker than the relationship between walkability and happiness.

Examining the linear correlation between happiness scores and walkability and happiness scores and access to transit further supported this. Walkability was over twice as strongly correlated with happiness as access to public transportation was.

```
1  corr, _ = pearsonr(df['total_score'], df['D4A'])
2  print(corr)
```

```
-0.2046791238931847
```

Correlation between happiness score and access to public transportation. The linear correlation between happiness scores and the distance between the population center and the nearest transit stop is -0.2047. As the distance to transit stops decreases, we expect that residents, on average, will be happier.

```
corr, _ = pearsonr(df['total_score'], df['NatWalkInd'])
print(corr)
```

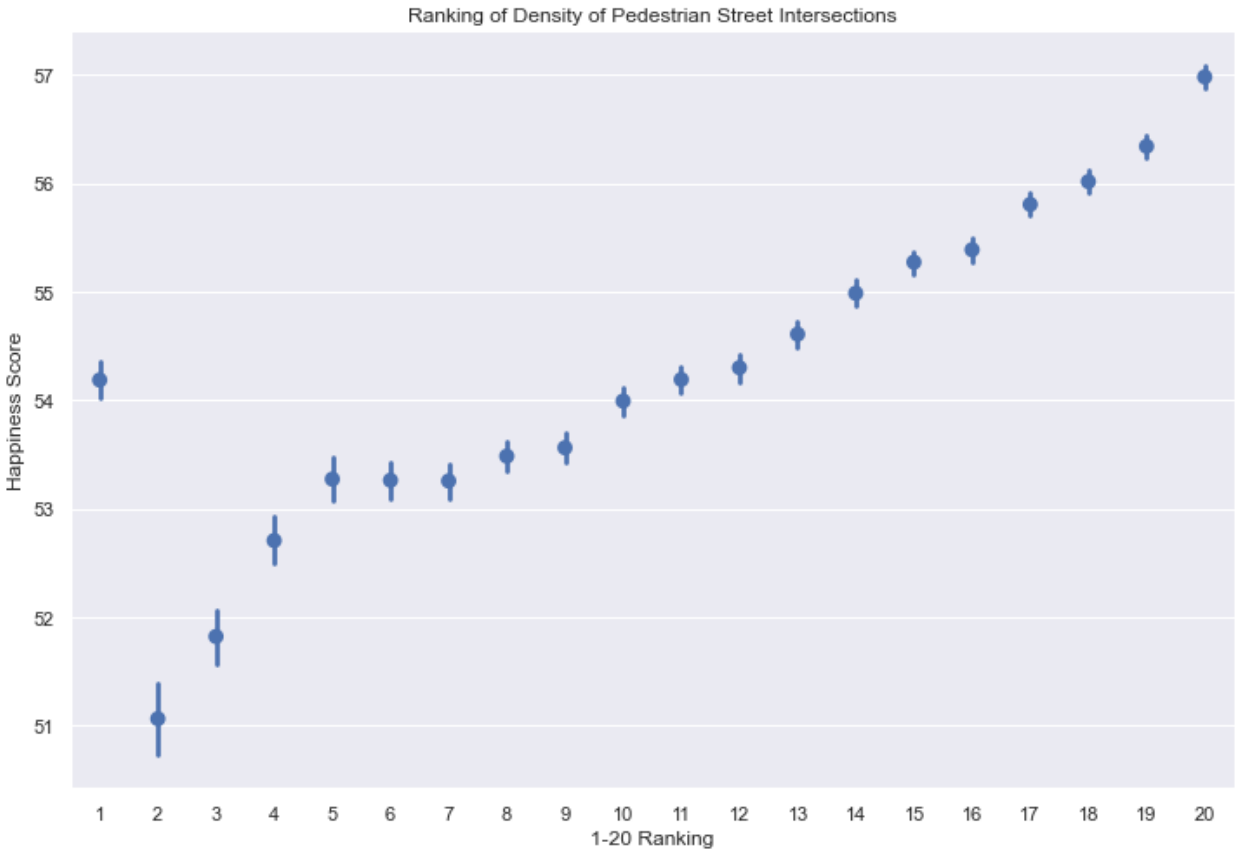
```
0.5252033893429153
```

Correlation between happiness score and walkability index rating. The linear correlation between happiness scores and walkability index ratings is 0.5252. As walkability increases, we expect that residents, on average, will be happier.

Pedestrian Intersections

In addition to the walkability index, the EPA also reports a metric that ranks the density of crosswalks in an area on a scale of 1-20, with 1 being the lowest and 20 being the highest.

The figure below is a pointplot; each point is the mean happiness score at a particular walkability ranking, and the bars sticking out of the point are the confidence interval.



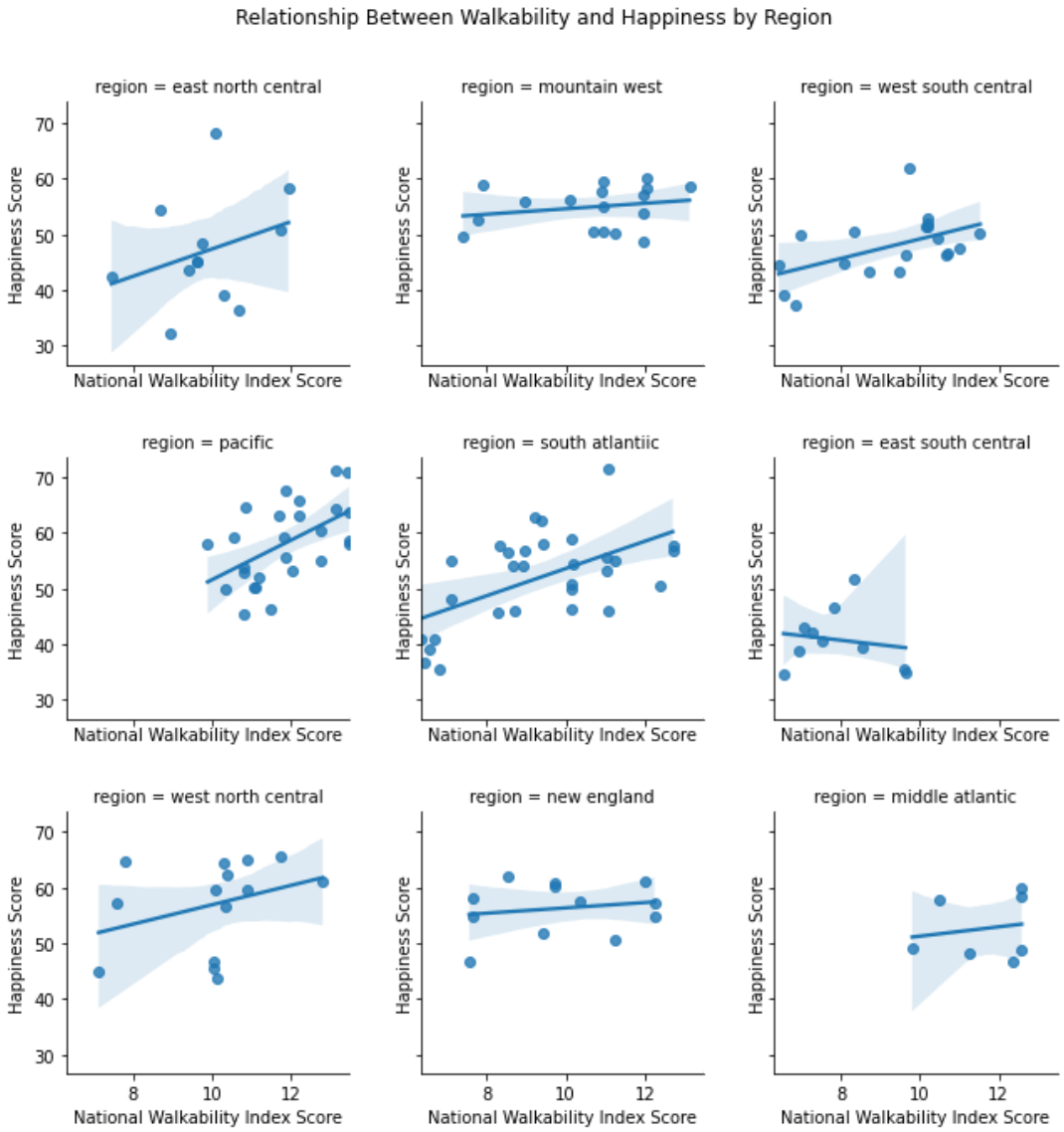
Pointplot of average happiness score by crosswalk density ranking.

Notably, cities with the lowest density ($ranking = 1$) were about as happy as cities where $ranking = 10$. One possible explanation for the sharp decline between $ranking = 1$ and $ranking = 2$ is that people living in very rural areas with virtually no crosswalks may be

happier than people living in suburban or exurban areas with limited pedestrian infrastructure. In other words, it may be that people would rather have *no* crosswalks than just have a few crosswalks.

Regional Variations in Walkability

Walkability varies significantly across the United States. But to what extent do these regional differences impact resident wellbeing? Do regional associations match the national trend? I chose to examine regional differences in walkability rather than in transit, because geographic features can limit the feasibility of building transit systems in certain parts of the country. For example, trains and buses connect the east coast corridor from Boston to Washington D.C., but mountain ranges in the American West make it difficult to connect Denver to Salt Lake City. Pedestrian infrastructure is comparatively easier to build, regardless of a city's landscape.



Scatterplots of happiness score vs walkability rating by region.

Error bars are shaded in blue around the line of best fit. With the exception of East South Central, all regions showed a positive association between walkability and happiness. The slope of the line of best fit, however, still varied amongst regions with a positive association. The impact of walkability on happiness appears to be the largest in the South Atlantic and Pacific regions. There was also significant variation in the size of error bars across regions. Error bars were considerably larger in the East North Central region than in the Pacific region, even though both exhibited positive associations.

Landscape architecture can help explain these regional differences. Research has shown that people prefer certain "kinds" of sidewalks. Pedestrians like to walk on clean, well-maintained, shaded paths away from traffic. If pedestrians only have access to cracked, unpleasant sidewalks that are too close to traffic, they're less likely to use sidewalks and enjoy the benefits of living in a walkable community.

Results from an ANOVA test confirmed observations gleaned from the graph; there is a statistically significant difference in the mean National Walkability Index Score between regions in the United States.

```
lm = ols('total_score ~ NatWalkInd', data = df).fit()
regions_lm = ols('total_score ~ Region + NatWalkInd', data = df).fit()

table1 = anova_lm(lm, regions_lm)
print(table1)
```

```
   df_resid          ssr  df_diff      ss_diff         F    Pr(>F)
0     145.0  7380.089644      0.0          NaN       NaN       NaN
1     137.0  5658.967775      8.0  1721.121869  5.208408  0.000011
```

ANOVA test for differences in the mean national walkability index rating by region. The p-value for the difference between regions is 0.000011.

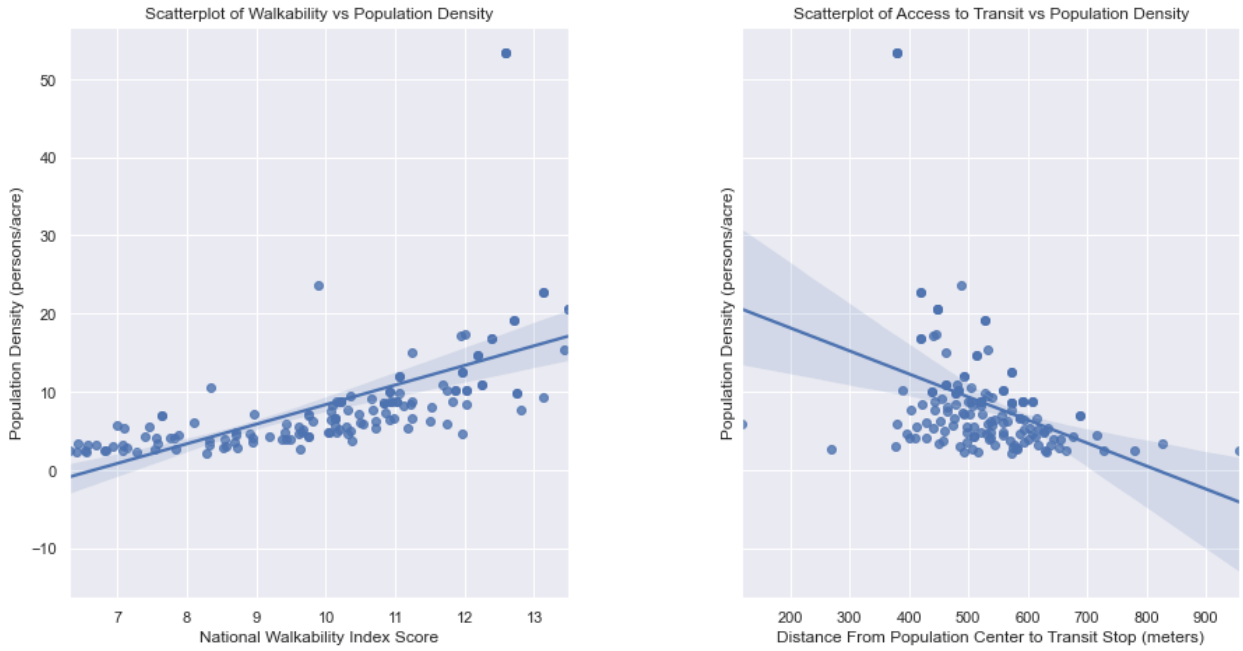
```
==============================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
Intercept                     29.7579      3.960      7.515      0.000      21.927      37.588
Region[T.East South Central]  -2.8345      2.835     -1.000      0.319      -8.440       2.771
Region[T.Middle Atlantic]      2.4603      3.124      0.788      0.432      -3.717       8.638
Region[T.Mountain West]        6.5328      2.439      2.678      0.008       1.709      11.356
Region[T.New England]          9.2368      2.624      3.520      0.001       4.048      14.425
Region[T.Pacific]              7.6860      2.353      3.266      0.001       3.033      12.339
Region[T.South Atlantiic]      5.8849      2.204      2.670      0.008       1.527      10.243
Region[T.West North Central]   9.6142      2.529      3.802      0.000       4.613      14.615
Region[T.West South Central]   1.8981      2.379      0.798      0.426      -2.806       6.603
NatWalkInd                     1.7505      0.355      4.929      0.000       1.048       2.453
```

Regression output for the model regions_lm (total score ~ Region + NatWalkInd). Not all regions had a statistically significant impact on the relationship between happiness scores and national walkability index ratings.

External Factors: Population & Population Density

Population density, of course, is an important confounding variable. It's no coincidence that 66.4% of New Yorkers walk or take transit to get to work *and* that New York is one of the nation's densest cities. Moreover, not all large cities are dense. Los Angeles is the second-largest city in the nation and sprawls over nearly 500 square miles of land. Washington D.C., on the other hand, is condensed into just 61 square miles, and its population density is nearly 40% higher than Los Angeles's. The fact that Washington D.C. has high levels of ridership and Los Angeles does not is in part due to D.C.'s population density.

To explore what effect population density has, I plotted population density against walkability index scores and the distance from the population center to the nearest transit stop.



Scatterplots of population density vs. walkability index rating and population density vs. distance to public transit

The linear relationship between population density and walkability appears to be much stronger than its relationship with distance to public transit. This can partially be explained by the fact that adding transportation infrastructure to a city requires more planning and money than adding sidewalks, crosswalks, and other pedestrian infrastructure. This is especially true for mid-sized cities, which may lack the ridership levels needed to sustain frequent, high-coverage transit service.

Conclusion

Well-being is, by definition, subjective. Factors that improve the quality of life in one city may have a limited impact on residents in another area. While access to public transportation and walkability are just two factors impacting happiness, this analysis shows that there is a clear relationship between these variables.

Generally, the relationship between walkability and happiness was stronger than the relationship between transit access and happiness. Future work should attempt to control for other variables that impact access to transit (i.e. municipal spending, location of jobs relative to residential areas, etc.) to gain a better understanding of how it impacts wellbeing.

In the meantime, cities of all sizes can improve sidewalks, crosswalks, and other pedestrian infrastructure to boost resident happiness. City planners should keep in mind, however, that there is some evidence that suggests residents would rather live in completely unwalkable communities than ones with limited or inadequate pedestrian infrastructure. Simply put, cities should do pedestrian infrastructure "the right way" or not bother with it at all.

Resources

All code and materials used in this project are available here. I have noted in my Jupyter Notebook file whenever I used resources or based my code off of snippets other have shared online. A full list of resources is available here as well:

- Stackoverflow post describing how to create subplots using seaborn regplots
- Tutorial on calculating correlations using the Python statsmodel package
- Matplotlib tutorial on adding annotations to plots
- U.S. Census Bureau map of regions