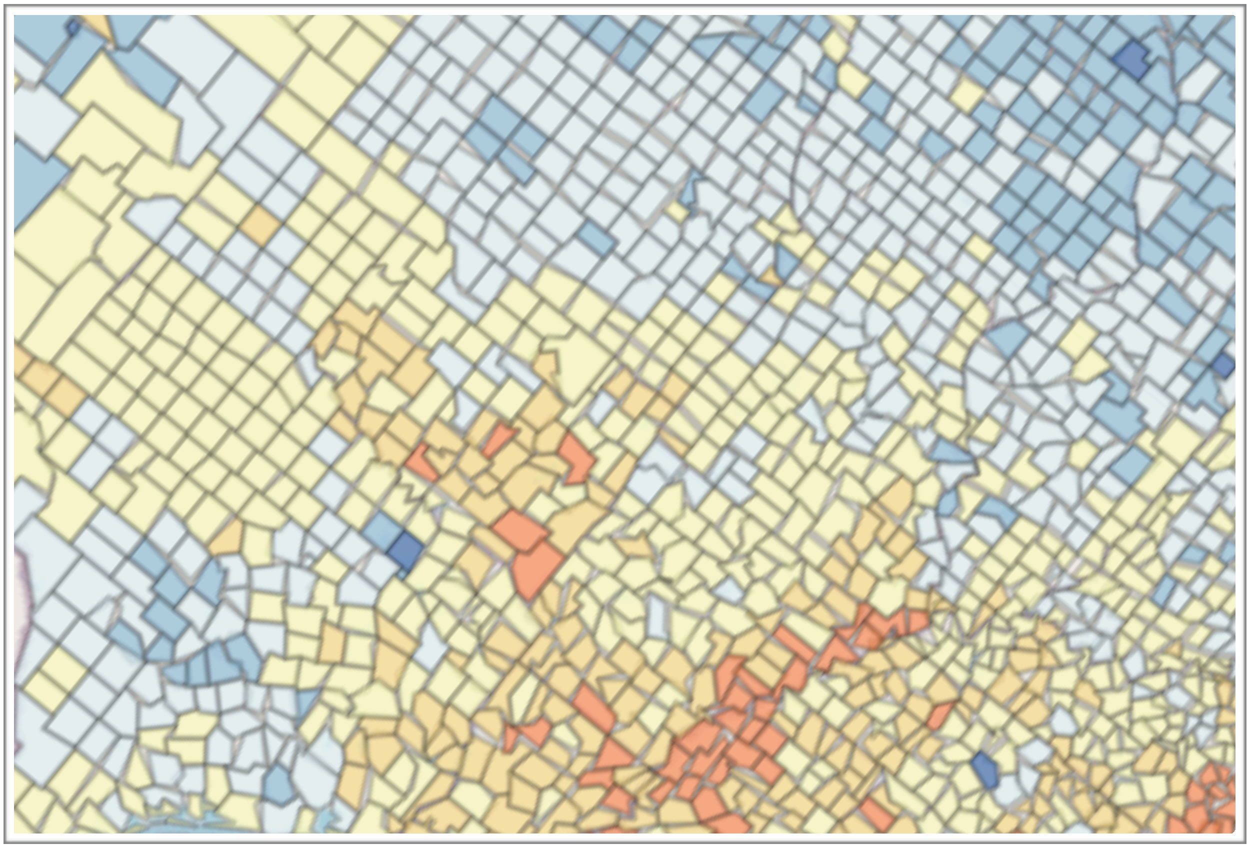
# Your Neighborhood Is Killing You

*Can Foursquare data predict life expectancies?*

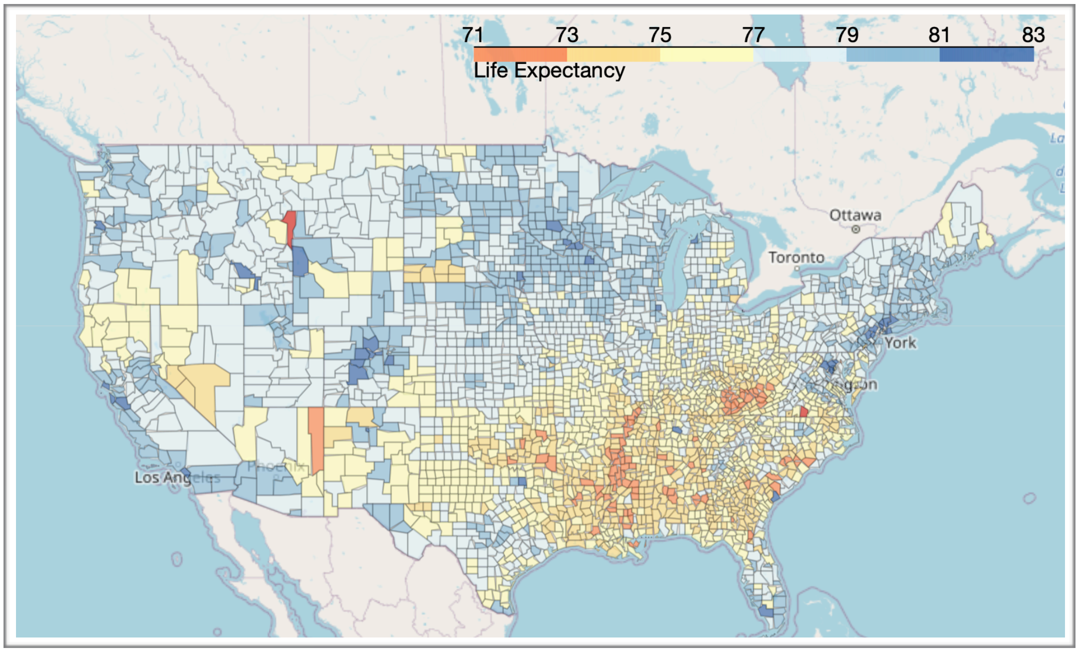


Aaron Chaiclin

May 2020

# Introduction

No one lives forever. Throughout human history we have worked to increase our life expectancy and have succeeded in increasing the average around the globe.



Despite our progress, the life expectancy varies by gender, ethnicity and country. In fact, there are different life expectancies regionally within the United States.

Research by the University of Washington's Institute for Health Metrics and Evaluation (IHME) has pinpointed life expectancies at the county level. These life expectancies vary from 68.4 to 83.3 years old. This ~20% difference is a significant amount when talking about human life. In a first world country, it seems counterintuitive that there would be such disparity depending upon where you live.

For example, the life expectancy in Shelby county, Alabama is 79.15 while neighboring Talladega county is 73.3. What could cause a 7% drop in life expectancy in an area on the other side of the river?

There are many factors that play a role in how long an individual lives. Extrapolated across an entire county the factors influencing life expectancy is countless. In order to improve life expectancies, we need to understand why there are regional differences. We can look at traditional metrics (e.g., income, ethnicity, health insurance access), but can we determine that there are other regional differences that account for higher or lower life expectancies? Do regions that have more fast food restaurants have lower life expectancies? Do more vegetarian restaurants or salad places attribute to a higher life expectancy? Do more gyms lead to better fitness and longer lives? What about more hospitals or doctor offices per capita?

Using Machine Learning we hope to create a model that can predict the life expectancy for a county primarily using Foursquare data.

If we can determine that restaurant types and the availability of specific services are correlated with life expectancies it could shed a new light on the way we think of planning for cities and counties. Certainly many government officials as well as many NGOs would have in interest in this type of study.

# The Data

This project focuses on two main data sets. Life expectancy data from IHME and location data from Foursquare.

IHME has provided life expectancies at the county level, available on their website. This data set also includes fitness and obesity statistics. Although their data is a longitudinal study we are only interested in the most recent data. IHME used small area estimation methods to produce annual life tables and calculate age-specific mortality risk at the county level for the United States. De-identified death records from the National Center for Health Statistics (NCHS) and population counts from the census bureau, NCHS, and the Human Mortality Database were used in the analysis. Results of the study were published in JAMA in May 2017. This data was available for download in Excel format. In addition to the life expectancy information each county has a fitness and obesity measure. This data would be used as a comparison metric to the foursquare data.

Foursquare provides location data for various businesses, and other landmarks, through an API (API doc link). The API provides real-time access to over 105MM places available across 190 countries and 50 territories. Foursquare refers to this as "venue" data. Foursquare categorizes the venues and provides geo coordinates for each location. The API will provide up to 50 results (per call) within a specified radius of a geo location. Since counties are not circular shaped, we need to calculate a big enough radius to capture all the venues and then provide a secondary process to verify that the geolocation is within the bounds of the county.

The Foursquare categories will be used to classify the venues, which will be aggregated at the county level. Since this project is limited in the amount of data we can gather from Foursquare, due to API and licensing limits, we are limit the data pulls to specific "seed" categories. The categories are listed below.

Seed Categories

| | | | |
|---|---|---|---|
| fast food | bike trail | juice bar | ski area |
| fried chicken | park | gluten free | trail |
| pizza | pedestrian plaza | salad place | factory |
| hot dog | playground | vegetarian | industrial estate |
| fish & chips | rec center | gym | medical center |
| bbq | rock climbing | smoke shop | bar |
| er | spiritial center | hospital | gun shop |

The seed categories were selected based on venue types that provide access to fitness (e.g., gyms & trails) or medical care, imply a certain diet (e.g., fast food or vegetarian), or be a potential health hazard (e.g., factory). Although these are the seed categories, we do not limit the actual categories returned from the API, as the data only shows the "primary" category for the venue.. For example, we may request the category "pizza" and receive a restaurant who's primary category is "italian food". In this example, we would then use the category of "italian restaurant" for that venue, not pizza. No presumptions have been inserted into the model, for example we might expect a correlation between medical access and longer life expectancies, but if the model finds a negative correlation, then we accept the truth of the data as is.

In addition to our two main datasets, we will import GeoJSON shape files (available from from a fellow data wrangler, Eric Celeste) and population counts for the counties, from the US Census. The GeoJSON data is for location validation and the population data is to provide per capita statistics of the venues, in addition to the raw counts.

# Data Cleaning

There are a number of challenges in working with these data sources. Whenever working with multiple sets of data there is typically a challenges of getting the data to fit together. In this section, we'll cover how we joined these sets together and then cover each dataset's specific challenge.

When working with US county data, there is a standard unique ID, called a FIPS code, that can be used to reliably join the data when it is available from the dataset. Of course, our life expectancy source did not provide a FIPS code. FIPS codes were scraped from the wikipedia. A unique key was created by concatenating the state abbreviation and the county name. After capitalizing the code and removing spaces and periods, the data joined fairly well, connecting 3,127 of 3,142 counties. Six counties required that we manually match the codes due to name variations (e.g., "DC-DISTRICTOFCOLUMBIA" vs "DC-WASHINGTON). Nine counties needed to have FIPS codes manually added because they were not in our source data. We were able to get FIPS codes for all 3,142 counties.

GeoJSON data was easily join via a FIPS code, once a prefix was added to match the GeoJSON file format.

The list of counties was used in our API calls to get the venue data and we imputed the county FIPS code into the JSON result. This allowed us to query the county shape file when validating the assigned county before aggregating the data.
It was difficult to find that county shape files that would work for generating choropleth maps as well as validating the venue locations. Once a reliable source had been found, the maps worked easily and the shape files were added to a GeoPandas DataFrame

Population data, which was retrieved from the US Census website, included 11 years of population broken down into different age groups. We simply selected the total population for most recent year for this study. Only 3,138 counties merged cleanly with the population data. We manually adjusted some FIPS codes to better align the data and 1 county was dropped because it's population was merged into another county, leaving us with 3,142.

As the data moved through the various processes we increased our number of counties by one but Foursquare only had data for 3,046 counties. In the end, we had enough data for 3,046 counties out of 3,142**.** Which accounts for over 90% percent of the total US population.

# Feature Selection

By combining the county data with the Foursquare data, we created 922 different features. This accounts for 461 that were simply raw, aggregated venue counts by category. 461 features were the per capita calculation of the aggregated venue counts.

In order make sure were were generating accurate results, we cut many different subsets of data and then ran them through our models to pick the best performing. We used two different methods for creating the subsets. SciKit Learn's RFE (recursive feature elimination) process was used to create datasets with 10, 20, 50, 100, 200, & 300 features. Pearson correlation was used to determine our top correlating features. We ranked the features by absolute value and created subsets with 10, 20, 50, 100, 300 features.

In analyzing the features, we discovered that no one particular feature has a high correlation with the life expectancy. The max correlation was only 0.35 as shown below in the top 10 features by correlation.

| Feature | Pearson Correlation | Absolute Value |
|---|---:|---:|
| Bar | 0.346542 | 0.346542 |
| Park | 0.313285 | 0.313285 |
| Trail | 0.312461 | 0.312461 |
| fastfoodPerCap | -0.262772 | 0.262772 |
| Pub | 0.232082 | 0.232082 |
| Playground | 0.228892 | 0.228892 |
| totalPop | 0.205338 | 0.205338 |
| Sports Bar | 0.203747 | 0.203747 |
| American | 0.199809 | 0.199809 |
| barPerCap | 0.198616 | 0.198616 |

The information provided for a sparse matrix of data, as only 122,271 elements of 2,707,941 contained non-zero values. The sparsity is calculated as 0.9568.

It will be difficult to create an accurate prediction with a sparse matrix, but we'll put our regression models to the test and see how close we can get.

In the end, we found that the entire set of features produced more accurate results than the culled data, regardless of the work done to determine the optimal number of features. This is likely because of the sparsity of the data along with the low correlations that make it hard to whittle down the features from the complete set.

Later we'll get into more details in the Methodology and Results sections to see how well we could predict life expectancies.