ATLANTA YOGA STUDIO LOCATION PROSPECTING

IBM Data Science Capstone Project

Abstract

We are looking to expand our yoga studio franchise into the Atlanta area and are looking to identify the best zip codes in the city based on population, income, and existing competition.

Jason Shorb

April 3, 2020

Table of Contents

I. Introduction

This is my capstone report for completing the final course of the IBM Data Science Specialization, a 9-course series created by IBM and hosted on the Coursera platform.
The capstone project instructions were as follows:

Project Instructions: Be as creative as you want and come up with an idea to leverage the Foursquare location data to explore or compare neighborhoods or cities of your choice or to come up with a problem that you can use the Foursquare location data to solve.

For my project report:
I am looking to help a friend who has recently completed her yoga instructor certification and is now looking to open up her own studio franchise in the city of Atlanta with one of her friends. She has asked if I would run some data analysis to identify the number of yoga studios in the Atlanta city area by zip code and determine which zip code area(s) I would recommend for opening her new studio based on existing competition and neighborhood population size. If average household income or wages could be included in the analysis, that would be helpful as well.

II. Data Description

To help address this question, I first looked for data sources that could help break down the Atlanta city area either by neighborhood or zip code. I then looked for data sources that could provide population and household income based on zip code or neighborhood.

After doing some research, I was able to identify population density by zip code using 2010 US Census Data and household wages by zip code using information on Zipatlas.com. Additionally, I will use Foursquare location data to understand where existing yoga studios are located at today within the city.

Data Sources:

- **2010 US Census Zip Code data** with city, state, latitude, longitude, and total wages
  - Data Source: US Census data on www.kaggle.com
- **Atlanta Zip Code Population data**
  - Data Source: Zipatlas.com
  - http://zipatlas.com/us/ga/atlanta/zip-code-comparison/population-density.htm
  - Note: I was unable to scrape the site data so copied data to worksheet on Github
- **Foursquare location API data**

Data Importing:

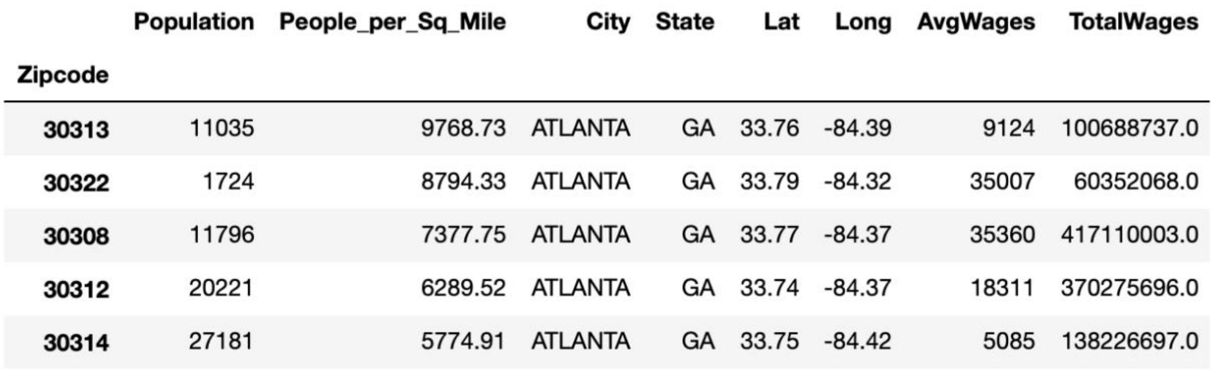
1. Upload US zip code wage data file (81,000 records)
2. Upload Atlanta zip code population data file (37 records)

Data Cleansing:

1. Update Zip Code column data types to be integer for both zip code data files
2. Fix zip code so that all codes are 5 digits and aren't missing any leading zeroes
3. Drop unnecessary columns to help simplify the analysis
4. Drop duplicate zip code records from US Census Zip table
5. Merge information from two zip code data sources into a new data source with both population and wage information
6. Identify and replace any NaN data within population and wage columns
7. Calculate Avg Wages based on Total Wages divided by Population for better comparison between zip codes
8. Change Avg Wages column data type from float to integer to remove decimals
9. Set Zipcode column as index field

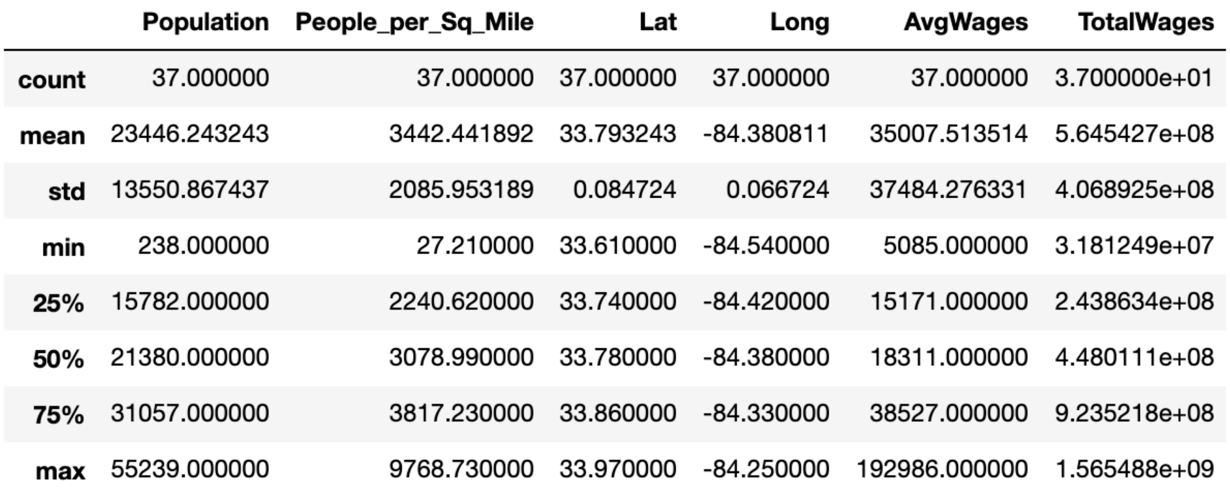
The final merged data set has just 37 records and 9 columns for the city of Atlanta. Below is a sample of the final data set along with the descriptive statistics.

Figure 1. The Final Data Set – Sample



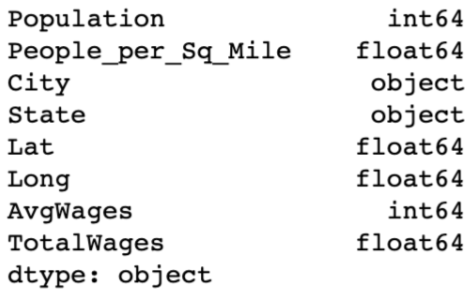
	Population	People_per_Sq_Mile	City	State	Lat	Long	AvgWages	TotalWages
Zipcode								
30313	11035	9768.73	ATLANTA	GA	33.76	-84.39	9124	100688737.0
30322	1724	8794.33	ATLANTA	GA	33.79	-84.32	35007	60352068.0
30308	11796	7377.75	ATLANTA	GA	33.77	-84.37	35360	417110003.0
30312	20221	6289.52	ATLANTA	GA	33.74	-84.37	18311	370275696.0
30314	27181	5774.91	ATLANTA	GA	33.75	-84.42	5085	138226697.0

Figure 2. Descriptive Statistical Analysis



	Population	People_per_Sq_Mile	Lat	Long	AvgWages	TotalWages
count	37.000000	37.000000	37.000000	37.000000	37.000000	3.700000e+01
mean	23446.243243	3442.441892	33.793243	-84.380811	35007.513514	5.645427e+08
std	13550.867437	2085.953189	0.084724	0.066724	37484.276331	4.068925e+08
min	238.000000	27.210000	33.610000	-84.540000	5085.000000	3.181249e+07
25%	15782.000000	2240.620000	33.740000	-84.420000	15171.000000	2.438634e+08
50%	21380.000000	3078.990000	33.780000	-84.380000	18311.000000	4.480111e+08
75%	31057.000000	3817.230000	33.860000	-84.330000	38527.000000	9.235218e+08
max	55239.000000	9768.730000	33.970000	-84.250000	192986.000000	1.565488e+09

Figure 3. Column Data Types



```
Population              int64
People_per_Sq_Mile    float64
City                   object
State                  object
Lat                   float64
Long                  float64
AvgWages                int64
TotalWages            float64
dtype: object
```

(The rest is in progress for Week 5.)

III. Methodology

The intention is to identify the zip code areas in the Atlanta city area that have the best potential for a new yoga studio based on population size, household wages, and existing yoga studio competition. Therefore, the proposed approach is to first identify and evaluate the makeup of the Atlanta area based on population and wages to determine which areas stick out as being potential opportunity areas. Based on this information, we may narrow the focus to the top five or so zip codes that seem the most promising. Then I will leverage the Foursquare location data to pull in existing yoga studio locations to see if that leads to one or two zip codes being the most promising.

Correlation of Population vs Average Wages

A little bit outside the scope of the request, I wanted to see if there were any correlation between population size by zip code and the household average wages. If so, we may be target lower population zip codes with higher average wages, or find that zip code with greater population tend to have higher average wages as well.

Data Visualization

Next, I wanted to run some visualizations against the data to see the population and average wage levels by zip code. So I ran bar charts for both and can see the results below. (any insights here?)

Next, and more importantly, I wanted to display a map of Atlanta with the Chlorpleth Map visualization to show shading by zip code neighborhood based on the population levels in each zip code. I then wanted to rerun the same visualization with average wages. (Note: I wasn't sure if it was possible to run them both at the same time and wasn't able to figure out how to do it, so ran these individually instead.). Unfortunately, either due to having an incorrect Atlanta neighborhood geomap or some other reason, the Chloropleth map wouldn't work correctly for me.

Figure 3. Chloropleth map

Any Inferential Statistical Testing performed?
If any, and what machine learnings were used and why?

IV. Results

V. Discussion

VI. Conclusion