YOGA STUDIO LOCATION PROSPECTING

IBM Data Science Capstone Project

JASON SHORB | APRIL 2020

AGENDA

OBJECTIVE

DATA OVERVIEW

METHODOLOGY

INSIGHTS

RECOMMENDATION

OBJECTIVE | IDENTIFY ATLANTA ZIP CODE FOR NEW YOGA STUDIO

- A friend has recently completed her yoga instructor certification and is now looking to open 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 which zip code areas might provide the best opportunity, based on existing competition and neighborhood population size.
- If household income or wages could be included in the analysis, that would be helpful as well.

DATA OVERVIEW | POPULATION & AVERAGE WAGE ZIP CODE DATA

- The data sources below were used to help determine population and household income by zip code within the Atlanta city area either by zip code.

1. **2010 US Census Zip Code data** - Includes city, state, latitude, longitude, and total wages by zip code
   - Data Source: US Census data on www.kaggle.com

2. **Atlanta Zip Code Population data**
   - Data Source: Zipatlas.com (http://zipatlas.com/us/ga/atlanta/zip-code-comparison/population-density.htm)

3. **Foursquare location API data** – Used to understand where existing yoga studios are located

DATA OVERVIEW | DATA CLEANSING STEPS

- To properly work with the data, the following data cleansing steps were completed:
  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 data from two zip code data tables into a new table frame 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 zip code column as index field

The final data set has just 37 records and 9 columns for the city of Atlanta. Here is a sample of the final data set.

Zipcode	Population	People_per_Sq_Mile	City	State	Lat	Long	AvgWages	TotalWages
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

METHODOLOGY | POPULATION & AVERAGE WAGE ZIP CODE DATA

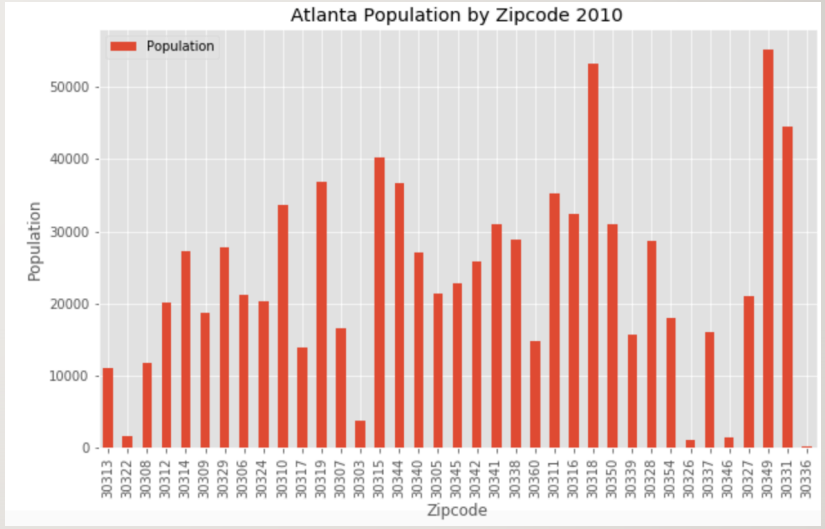
- The proposed approach is to first identify and evaluate Atlanta zip code areas based on population and wages to determine which areas stick out as being potential opportunity areas.
- Based on this information, we may narrow the focus to the top five or so zip codes that seem the most promising.
- We will leverage the Foursquare location data to identify existing yoga studios to see if one or two zip codes are most promising.

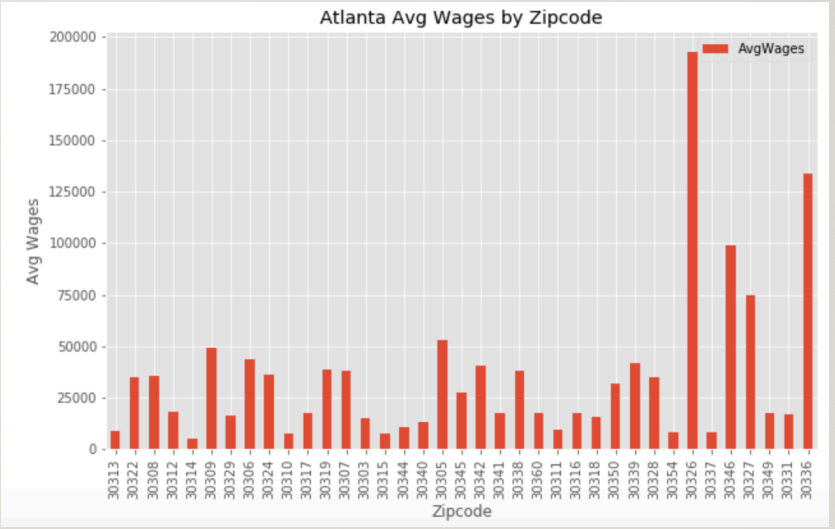
My data analysis methodology approach:

- Data Visualization – Charts and Mapping
- Inferential Statistical Analysis via Regression and Correlation
- Clustering
- Foursquare Location Data Integration

INSIGHTS | POPULATION & AVG WAGE DATA OUTLIERS

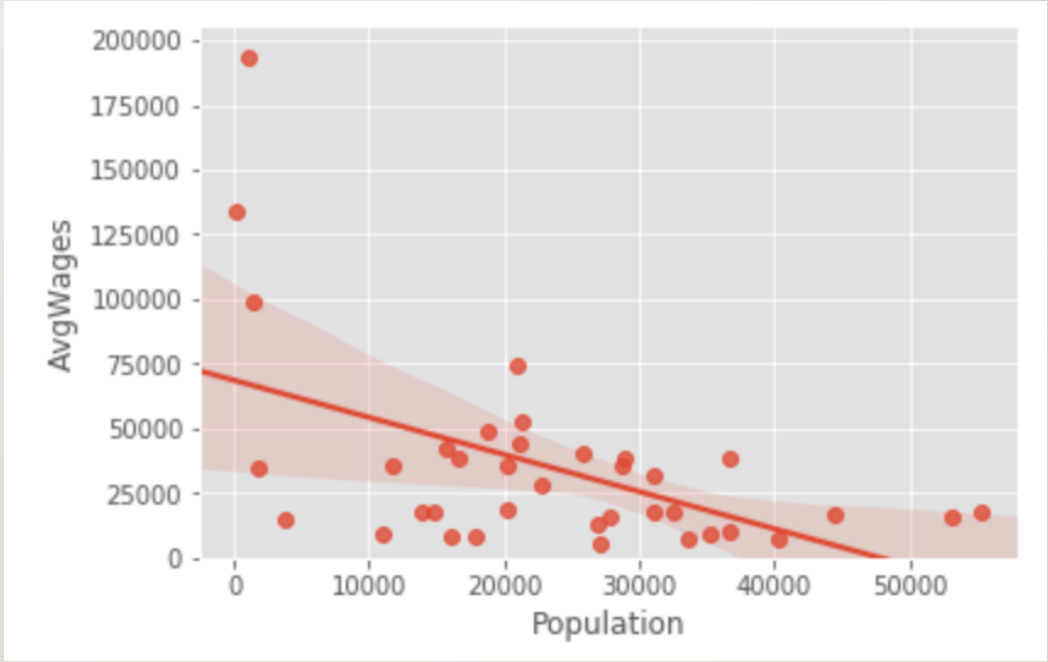
- The first thing I wanted to do was run data visualizations to see the population and average wage levels by zip code.
- Results showed noticeable zip code outliers for both population and average wages.
- For instance, zip code 30326 has a very high average wage level, but the population size is also the lowest.
- The extreme difference here makes me wonder if there might be some distortion or issues in the data from the US Census, particularly for this zip code.





INSIGHTS | CORRELATION & REGRESSION STATISTICAL ANALYSIS

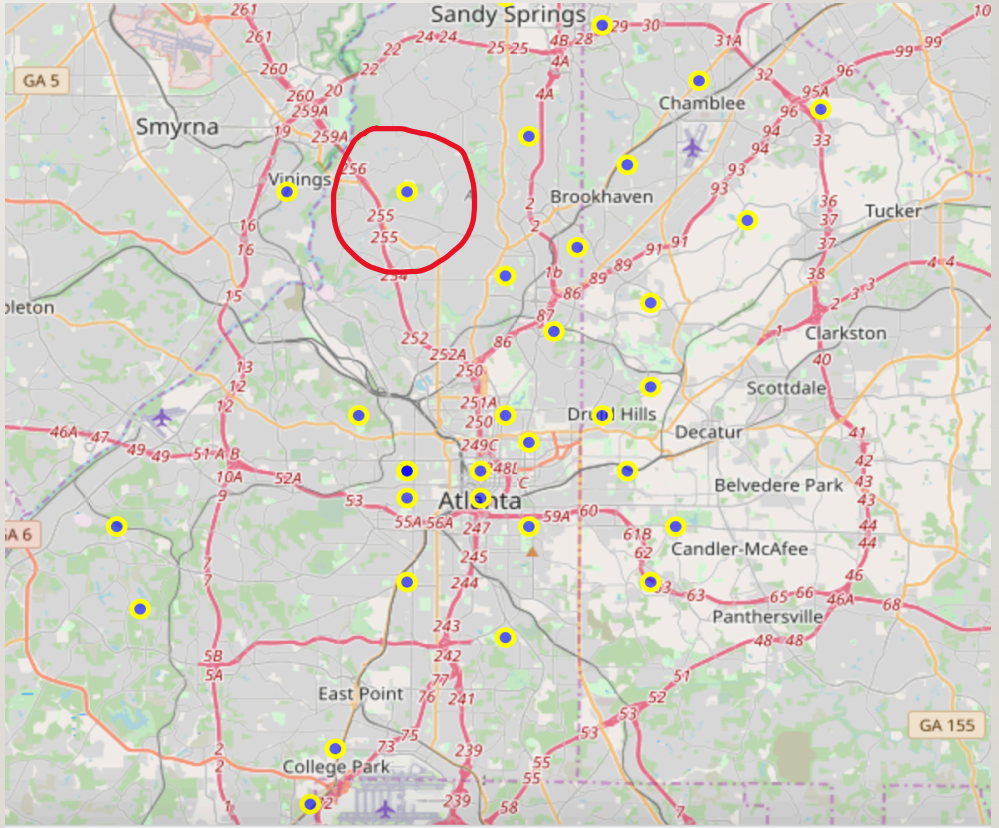
- The next thing I wanted to see if there were any correlation between population size and the household average wages.
- If so, we may be able to target lower population zip codes with higher average wages or find that zip codes with greater population tend to have higher average wages as well.
- While the Pearson Correlation Coefficient score indicates a moderate downhill (negative) relationship, the P-value is extremely small.
- Since the P-value is ≤ 0.05, it indicates strong evidence against the hypothesis that Population may have an impact on Average Wages.



Pearson Correlation Coefficient is -0.518
P-value of P = 0.001

CONCLUSION & RECOMMENDATION | ANALYSIS SUGGESTS 30327

- Based on the population and income levels by zip code, the recommended zip code area for opening a new yoga studio would be 30327 as it has a larger income level and a population over 20,000.
- Unfortunately, I was unable to successfully complete the following analysis to further validate the recommendation:
  - Chloropleth Mapping to display zip code areas based on population and/or average wages.
  - Foursquare API data to identify existing yoga studios and complete the k-means clusters.



THANK YOU!