Environment Systems and Decisions (2020) 40:526–543
https://doi.org/10.1007/s10669-019-09753-z

# Exploring good cycling cities using multivariate statistics

**Andrew J. Collins[1] · Craig A. Jordan[1] · R. Michael Robinson[1] · Caitlin Cornelius[2] · Ross Gore[1]**

Published online: 3 December 2019


**Abstract**
Some U.S. cities are excellent for cycling, like Portland, and some cities are not so good. This observation raises the question: what are the characteristics of a city that make it good for cycling? This study investigates the characteristics of 119 cities to explore what factors help make a city good for cycling. What "good" means in terms of cycling cities is subjective and we use the popular Bicycling Magazine ranking of cities for this purpose. We collected a variety of data sources about our cities including geographic, meteorology, and socioeconomic data. These data were used to conduct cluster analyses and create multivariate generalized linear regression models. We hypothesized that geographic and meteorology factors were important in determining good cycling cities. However, our hypothesis was proved wrong because socio-economic factors, like house pricing and obesity rates, play a more important role. For example, hilly cities, like San Francisco, can have excellent cycling infrastructure. The analysis shows what cities are like each other, regarding our considered characteristics; thus, city planners might wish to look at similar cities to help determine forecasts of expected use and public benefit of cycling. We use a case study of the Hampton Roads region of Virginia to show the application of our regression models.

**Keywords** Bicycling · Cycling · City planning · Cluster analysis · Multivariate regression

## 1 Introduction

When a city planner is determining how to improve their cycling infrastructure, they must draw from a variety of information about the practicalities, both technical and social, of any proposed plan. Technical practicalities include building cost and disruption to existing traffic flow. Social practicalities include deciding on where, in the city, to build the cycle path network, likely usage rates, and public support. This paper hopes to help city planners by determining what factors make up a good cycling city and provide information on which cities with successful cycling programs are similar to their own. There are over 3000 cities in the United States, and it is not immediately obvious what factors should be considered to determine the similarity of cities. What role does population play? How influential are median economic and education levels? Do topography, daily temperatures, and average precipitation rates impact cycling rates? The analysis presented in this paper provides insight into these questions using cluster analysis, and these factors are ranked using regression analysis. We collected a variety of data, both geographic, meteorology, and socio-economic, for 119 U.S. cities, of which 50 are considered good bicycling cities (Bicycling Magazine 2017). Using these data, we have conducted a variety of statistical analyses to determine which of our variables make a good cycling city. Figure 1 shows the geographic coordinates of the cities considered in this study. The solid red circles are those cities that are considered the best for cycling (Bicycling Magazine 2017) and the size of city's circle indicates its population size. What makes a cycling city "good" is subjective and we have chosen to use Bicycling Magazine's ranking for this analysis; the subjective nature of this measure does reduce the validity of the results which should be interrupted as indicators as opposed to definite findings.

Our initial hypothesis was that geographic and meteorology factors would play an important part in determining what makes a "good" cycling city. For example, if a city is too hot or hilly, then it would not be good for cycling. As our analysis shows, this is not the case, i.e., hilly cities, like San Francesco, can be considered a good cycling city in terms of Bicycling Magazine's ranking. Thus, city planners might what to look at other factors, like socio-economic, when

✉ Andrew J. Collins
ajcollin@odu.edu

[1] Old Dominion University, Norfolk, VA, USA

[2] Norfolk, USA

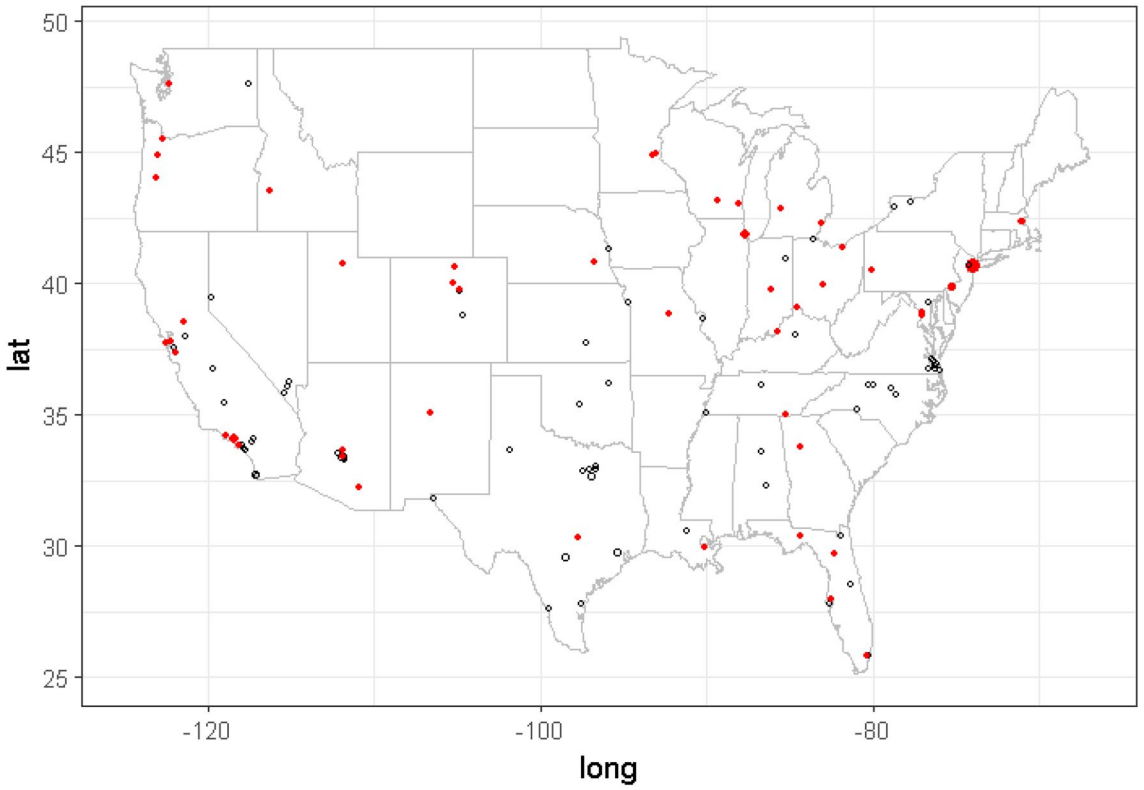


**Fig. 1** Geographic distribution of Cities considered in the study. Size indicates population, and the solid circle indicates the top U.S. bicycling city, according to Bicycling magazine

trying to determine whether their city has the potential to become a great cycling city.

The next section gives some background in bicycling research. That section is followed by a detailed description of our data collection sources and process. The results section outlines the statistical analysis conducted in the study, including correlation, cluster analysis, and multivariate regression. Finally, conclusions are given.

## 2 Background

In the United States, several cities are making efforts to become more bicycle-friendly to reduce pollution, decrease costs, and increase health benefits. Some studies have attempted to understand the concerns of residents, both bicyclists and non-bicyclist alike. In San Diego, for example, it was found that 58% of households had bikes, of those, 68% were adult riders. However, having a bicycle is only a resource, and a bicycle is only useful if owners can use them (Clark et al. 2018).

The majority of respondents who cycled said that cycle paths separated from cars were their preferred bicycle route, 75% said they do not ride at night, and only 15% cycled for transportation purposes (Jackson and Ruehr 1998). Other studies were geared towards helping city planners develop bicycle-friendly programs (Harkey et al. 1998). One key component of bike-friendly cities then, as revealed by the *bicycle stress level index* (Geelong Planning Committee 1978; Sorton et al. 1998) as well as the Federal Highway Agency index model, would be the total miles of bike lanes available, plans to create more bike-friendly roads, and the perceived safety and low stress level of the cyclist as they traverse these bikeways. Thus, the health of a city's cycling network has an impact on whether or not a city is great for cycling (Pucher et al. 2010; Schoner and Levinson 2014; Marqués et al. 2015; Pucher and Buehler 2016).

Other items besides bicyclist perceived safety/stress levels and city bicycle path planning/construction contribute to what constitutes a bicycle-friendly city. Cyclist demographics, household demographics, residential location, season, bicycle amenities at work, and bicycle-friendly facilities and business all contribute to the likelihood that an individual will choose to bike (Sener et al. 2009). Individuals aged 25–45 are the most likely group to cycle, as are white males who spend fewer hours at work (Moudon et al. 2005). Further, high-income households are more likely to cycle when compared to lower-income households (Parkin et al. 2008). In terms of the decision to bike or not to bike, a one-degree increase in morning temperature is linked with a 3% increase

in the chance of an individual deciding to bike to work and a one miles per hour (mph) increase in wind speed results in a 5% decrease in the chance that an individual decides to bike to work in northern states (Sears et al. 2012). Lastly, bike-friendly businesses tend to emerge in response to the increased presence of cyclists, which is the case for 15% of tourists visiting the Outer Banks, NC, accounting for 680,000 people annually and is responsible for $60 million in business and supports 1407 jobs (Meletiou et al. 2005).

These items, taken in conjunction with each other, would suggest that our measurements of bike-friendly businesses, weather, hilliness, property values, and health measurements are reasonable determinants of bicycle-friendly cities, and provide a comprehensive index for determining similarities between cities.

A final note is that bicycling is not always a positive thing for the environment; Ferguson (2008) points out that mountain biking can be destructive to forested landscapes. Hence, we have avoided discussing the environmental impact of bicycling in this paper.

### 2.1 Data collection

According to Statista, there are 19,505 cities, towns, and villages (incorporated places) within the United States in 2015 (Statistica 2017). Of the municipalities, 754 has a population greater than 100 K, with ten cities over one million. Our study collected data on 119 cities. These cities were chosen because they belonged to one or more of the following groups:

- Top 100 most populous cities in the U.S. (Pierce and Kolden 2015)
- Bicycling magazines top 50 cycling U.S. cities of 2016 (Bicycling Magazine 2017)
- Seven cities of Hampton Roads

No cities were in all three groups, that is, none of the Hampton Road's cities are in top 50 cycling cities. Only three of the cities had a population less than 100 K: Suffolk, VA (85 K); Portsmouth, VA (96 K); and Boulder, CO (97 K). A full list of the cities can be seen in Table 1, and their geographic spread can be seen in Fig. 1. Due to our selection criteria, not all states were included in this study; for example, many northern Midwest states and non-contiguous states were not included. Future work would include data from all fifty states.

Which cities that were included in our analysis could vary based on which data source was used. For example, the ranking of cities, by population, changes over time. We used the most populous cities as outlined by Pierce and Holden (2015) so that we were able to have a complete set of hilliness data for these cities, as provided by the paper. Bicycling Magazine was chosen for determining the "good" bicycling city's ranking.

Bicycling magazine, the U.S. most popular cycling periodical, produces an annual ranking of U.S. cities and publicly announces the top fifty cities. Factors that they used to determine this ranking included miles of bicycle lanes per square mile, bicycling-friendly business including cyclist-friendly bars, number of female cycle commuters, and people per bike share. Their website did not offer a complete dataset or was the ranking calculation method. However, we decided that this would be a good measure for determining what makes a good cycling location, due to the popularity of the magazine, and included the 2016 results within our list of cities. Of the fifty cities, 35 wherein the top 100 most populous cities.

Bicycling magazine is not the only organization to provide a ranking of the cities, in terms, of bicycling. The League of American Bicyclists, a bicycling society, and BikesForPeople, a bicycling advocacy group, also construct annual rankings of locations. BikesForPeople produces a ranking of cities (BikesForPeople 2019). There are significant differences between how these rankings are created, which is discussed and analyzed in the discussion section of this paper. We could have chosen to create a mixture of these different city rankings for our own city ranking but decided to use only a single source due to the subjective nature of combining dataset, e.g., determining weightings. Bicycling Magazine was chosen as our single source because of its longevity (started in 1961) and popularity (it is the most popular bicycling magazine). Though other data sources use more quantitative datasets in their ranking measures, it should be pointed out the selection of those measures is subjective. For example, BikesForPeople uses a measure called Acceleration, which represents "how quickly a community is improving its biking infrastructure and getting people riding;" obviously, this measure is biased against already established cycling cities that have previously invested a lot in cycling infrastructure and now, understandably decreases this investment.

Weather data were collected from U.S. Climate Data provided by your weather service (Your Weather Service 2017). Four indicators were used in this analysis: Average highest/lowest temperature in the hottest/coldest month, and average rainfall in the driest/wettest month. Eleven cities did not have weather information on the site; however, all eleven cities were effectively suburbs of other cities, so the major cities' weather data were used, for example, Garland and Plano assume to have the same weather as Dallas.

Topological (hilliness) data were extracted from Pierce and Kolden (2015). In their paper, they ranked the hilliness of the 100 most populous cities. Factors used included relief elevation range within city boundaries,

**Table 1** List of all U.S. cities considered in the study including bicycling magazines 2016 ranking

| City | State | Rank | City | State | Rank | City | State | Rank |
|---|---|---|---|---|---|---|---|---|
| Albuquerque | NM | 35 | Garland | TX | NA | Orlando | FL | NA |
| Alexandria | VA | 34 | Gilbert | AZ | NA | Philadelphia | PA | 15 |
| Anaheim | CA | NA | Glendale | AZ | NA | Phoenix | AZ | NA |
| Arlington | VA | 25 | Grand Rapids | MI | 33 | Pittsburgh | PA | 20 |
| Atlanta | GA | 43 | Greensboro | NC | NA | Plano | TX | NA |
| Aurora | CO | NA | Hampton | VA | NA | Portland | OR | 3 |
| Austin | TX | 7 | Henderson | NV | NA | Portsmouth | VA | NA |
| Bakersfield | CA | NA | Hialeah | FL | NA | Raleigh | NC | NA |
| Baltimore | MD | NA | Houston | TX | NA | Reno | NV | NA |
| Baton Rouge | LA | NA | Indianapolis | IN | 13 | Riverside | CA | NA |
| Birmingham | AL | NA | Irvine | CA | NA | Rochester | NY | NA |
| Boise City | ID | 27 | Irving | TX | NA | Sacramento | CA | 37 |
| Boston | MA | 17 | Jacksonville | FL | NA | Salem | OR | 47 |
| Boulder | CO | 10 | Jersey City | NJ | NA | Salt Lake City | UT | 14 |
| Buffalo | NY | NA | Kansas City | MO | NA | San Antonio | TX | NA |
| Cambridge | MA | 8 | Laredo | TX | NA | San Bernardino | CA | NA |
| Chandler | AZ | NA | Las Vegas | NV | NA | San Diego | CA | NA |
| Charlotte | NC | NA | Lexington | KY | NA | San Francisco | CA | 2 |
| Chattanooga | TN | 30 | Lincoln | NE | 44 | San Jose | CA | 26 |
| Chesapeake | VA | NA | Long Beach | CA | 28 | Santa Ana | CA | NA |
| Chicago | IL | 1 | Los Angeles | CA | 24 | Scottsdale | AZ | 48 |
| Chula Vista | CA | NA | Louisville | KY | 31 | Seattle | WA | 5 |
| Cincinnati | OH | 36 | Lubbock | TX | NA | Spokane | WA | NA |
| Cleveland | OH | 41 | Madison | WI | 16 | St. Louis | MO | NA |
| Colorado Springs | CO | NA | Memphis | TN | NA | St. Paul | MN | 32 |
| Columbia | MO | 42 | Mesa | AZ | NA | St. Petersburg | FL | NA |
| Columbus | OH | 39 | Miami | FL | 40 | Stockton | CA | NA |
| Corpus Christi | TX | NA | Milwaukee | WI | 46 | Suffolk | VA | NA |
| Dallas | TX | NA | Minneapolis | MN | 6 | Tallahassee | FL | 38 |
| Denver | CO | 11 | Montgomery | AL | NA | Tampa | FL | 45 |
| Detroit | MI | 50 | Nashville | TN | NA | Tempe | AZ | 22 |
| Durham | NC | NA | New Orleans | LA | 19 | Thousand Oaks | CA | 49 |
| El Paso | TX | NA | New York | NY | 4 | Toledo | OH | NA |
| Eugene | OR | 18 | Newark | NJ | NA | Tucson | AZ | 23 |
| Fort Collins | CO | 12 | Newport News | VA | NA | Tulsa | OK | NA |
| Fort Wayne | IN | NA | Norfolk | VA | NA | Virginia Beach | VA | NA |
| Fort Worth | TX | NA | North Las Vegas | NV | NA | Washington | DC | 9 |
| Fremont | CA | NA | Oakland | CA | 21 | Wichita | KS | NA |
| Fresno | CA | NA | Oklahoma City | OK | NA | Winston-Salem | NC | NA |
| Gainesville | FL | 29 | Omaha | NE | NA | | | |

Melton Ruggedness Number, and the standard deviations of elevation within various radius of the city center. They used a weighted formula method for determining the ranking from these factors, which can be found in the paper. Though theoretically possible to collect the data needed for the remaining 19 cities (including five Hampton Roads cities), we found difficulty finding it and decided to use a different method to fill in these data gaps.

The method employed was to use another dataset that contains elevation data relating to cycling to determine which cities, with a hilliness rank, are like the cities that do not have a rank. Each unassigned city was assigned a rank based on this comparison. The elevation difference data, from "map my ride" (www.mapmyride.com), a popular ride mapping service offer by Under Armor, Inc., were used. Map my ride was ranked one of the best cycling apps

by cyclist weekly (Elton-Walters and Wynn 2017) so was considered a credible source for the data. Using this comparison approach, the remaining 19 cities' hilliness data were filled out.

A new dataset for hilliness could have collected that covered all 119 cities in this study. However, as Pierce and Kolden (2015) point out, determining which metric approximates hilliness is subject; hence, the need for their study in the first place. As such, we did not think that any other source of topological data was appropriate for determining hilliness.

For geographic and climate information, we considered social factors in our model. There are quite literally thousands of socio-economic factors that could have been included in this study. As such, we only considered a sampling of factors in this analysis, and we accept that this could have to lead us to miss interesting relationships. The factors we covered were mainly related to health, as we believed that this was likely to have the most significant impact on an individual decision to cycle or not. Other factors included the number of bicycle-friendly businesses in a city and average house prices. As mentioned in Sener et al. (2009), bicycle amenities available have an impact on a person's willingness to cycle hence the inclusion of the number of cycling-friendly businesses within a city in our dataset. Local economics and household demographics also have an impact on bicycling rates, hence the inclusion of house prices in our analysis. We have purposely avoided any inclusion on ethnicity information due to the potential complexity that this could introduce to interrupting the results; for example, the fear, within different communities, of facing potential hostility from other road users while cycling can affect an individual's decision to cycle (Community Cycling Center 2012); hence there was a desire to keep the data as macro as possible. We leave it to future work to incorporate ethnicity factors into our model.

Median property values were collected for each city, from Zillow (www.zillow.com), a nation-wide real estate service. These data points were all collected in early June 2017, due to the constantly changing market values, to put the property price dataset as close to the other datasets collection times as possible.

Several social factors were collected for the cities, including smoking rates, obesity rates, lack of sleep, no leisure exercise, blood pressure, and physical and mental health. Details about all these variables can be found in Table 2. These data were collected from the Centers for Disease Control and Prevention (CDC)'s 500 cities project (https://www.cdc.gov/500cities/). These variables were selected due to their believed connection to active transportation, i.e., cycling.

The final variable considered was the number of cyclist-friendly businesses in the city. This variable was used in the bicycling magazine's ranking. The variable was determined by a weighted sum of bicycle-friendly business in the city with a platinum business being worth four, a gold worth three, a sliver worth two, and a bronze worth one. The business awards are given by the League of American Bicyclist (www.bikeleague.com), which is a non-profit organization dedicated to promoting bicycling in America.

All data inputting were completed for all variables except for 'Miles' and 'Shared.' These were not completed due to the difficulty in getting the data, and only the data for the fifty best cyclist cities were used.

## 3 Method

We conducted a variety of statistical analyses on the dataset, starting with some descriptive statistics. This basic descriptive analysis was first applied to the fifty best cycling cities subset's data and then to all city data. Cluster analysis and statistical regression modeling were also conducted on the dataset. The cluster analyses focused on clustering the cities into groups. The purpose of this clustering was to observe any characteristics of the groups. Hierarchical clustering and k-mean clustering were used in the analysis (Everitt and Dunn 2010). The number of clusters used in k-mean clustering was determined using the Elbow test (*15*). Logistic regression was used to determine which variables affected membership to Bicycling magazine's top fifty cycling cities. Since cluster analysis is less common than regression analysis, we provide a brief introduction to both k-mean cluster analysis and hierarchical clustering.

The focus of this research is on exploratory data analysis as opposed to inferential statistics because it was not clear what hypothesis should be made about what makes a city a great cycling city. Tukey (1980) and others have long argued the importance of exploratory data analysis as an approach to better understand the data and its underlying system/phenomenon. The findings of this research show several interesting phenomena which we hope are of interest to the reader. Given the vast array of possible dependent variables that could have been used in this analysis, the authors accept that this analysis is a starting point for a deeper understanding what makes a great cycling city.

### 3.1 K-mean cluster analysis

K-mean cluster analysis organizes the cities, based on the variables specified in Table 2, into groups (or clusters). K-mean clustering places the dataset into 'k' partitions such that the distance (Euclidean) between the characteristics of the cities and their partitions mean is minimized (Everitt and Dunn 2010). The number of partitions ('k') is determined based on the elbow test.

**Table 2** Table describing all the variable considered in the analysis

| Variable name | Description | Source |
|---|---|---|
| City | 100 most populous cities, 50 ranked cycling cities, and seven Hampton roads cities | 1, 2 |
| State | State City belongs to | 1, 2 |
| Population | Population in 2016 | 3 |
| Rank | Best cycling city 2016 rank | 1 |
| House Price | Zillow Home value index; median value for the city in June | 6 |
| Miles | Miles of bike lanes per square mile | 1 |
| Shared | Number of shared bike available per 1000 residents | 1 |
| Annual low temp | The average temperature of the coldest month | 4 |
| Annual high temp | The average temperature of the warmest month | 4 |
| Driest | Lowest average precipitation in inch | 4 |
| Wettest | Highest average precipitation in inch | 4 |
| Lat | Approximate latitude of the city | 5 |
| Long | Approximate longitude of city | 5 |
| Hilliness | If the city was in top 100 hilliness cities then hilliness was determined according to (2). Missing data were filled in using analysis approach discussed above | 2 |
| Smoking | Current smoking among adults aged ≥ 18 years | 3 |
| Obesity | Obesity among adults aged ≥ 18 years | 3 |
| Sleeping | Sleeping less than 7 h among adults aged ≥ 18 years | 3 |
| No exercise | No leisure-time physical activity among adults aged ≥ 18 years | 3 |
| Illness | Physical health not good for ≥ 14 days among adults aged ≥ 18 years | 3 |
| Mental illness | Mental health not good for ≥ 14 days among adults aged ≥ 18 years | 3 |
| High Blood Pressure | Taking medicine for high blood pressure control among adults aged ≥ 18 years with high blood pressure | 3 |
| Cycling-friendly businesses | Number of weighted platinum (4), gold (3), silver (2), and bronze (1) cyclist-friendly business in the city | 1, 7 |

1. https://www.bicycling.com/culture/news/the-50-best-bike-cities-of-2016
2. Pierce and Kolden (2015)
3. https://chronicdata.cdc.gov/500-Cities/500-Cities-Local-Data-for-Better-Health/6vp6-wxuq
4. www.usclimatedata.com
5. www.google.com
6. www.zillow.com
7. www.bikeleague.com

#### 3.1.1 Elbow test

The elbow test, first developed by Thorndike (1953), determines the optimal number of groups (or clusters) of the cities based on the variables specified in Table 2. This method involves performing k-means clustering on a range of partitions and calculating the within-cluster sum of squares (WSS). A plot of the WSS for each cluster is evaluated for determining the point of a bend that looks like an elbow. The elbow represents the suggested number of clusters that should be used in the analysis.

### 3.2 Hierarchical cluster analysis

Hierarchical clustering is an iterative approach that builds up groups of points that are close together or clustered. At each iteration, the two closest objects, not already in the same group, are joined. This step is repeated until all objects are in a single group. Our analysis used the Euclidean distance with complete-linkage clustering.

## 4 Results

A variety of statistical analysis was conducted on the dataset. Two versions of the dataset were considered: one containing all cities (without the 'Miles' and 'Shared' variables as they were not available for all cities); and the other dataset containing only Bicycling magazine's fifty best cycling cities, known as the best50 group. The fifty best cities were analyzed first to gain an understanding of what characteristics make up a great cycling city (as determined by Bicycling magazine). Three types of analysis were conducted: descriptive statistics, cluster analysis, and regression modeling. The descriptive statistical analysis focused on correlation (Everitt and Dunn 2010); the cluster analysis was conducted on the

fifty best cycling cities first then all the cities; similarly, the regression analysis was conducted on best50 group first.

## 4.1 Descriptive statistics

Table 3 shows the statistics for all the common variables for the best group and all other cities considered. Noticeable findings from these statistics are that standard deviation for the best city population is larger than the mean population (due to outliners of Los Angeles and New York), all health indicators are better for the best cycling city (which was tested using Student's *t* test at 95% level), and the best cycling cities tend to be in the north of the country. In terms of weather information, the best cycling cities are observing slightly lower temperatures but have a wetter rainy session. Since rain is not associated with good cycling conditions, this was a slightly surprising result. These descriptive statistics give a flavor of data but require further investigation before any conclusions can be made.

### 4.1.1 Spearman's rank-order correlation coefficients

Correlation analysis was conducted on both sets of data. The dependent variable for the best cycling city data was its reverse ranking (i.e., Chicago's reverse ranking is 50 as it is the best cycling city). Reverse ranking was used to make correlations easy to read (a positive correlation implies a better ranking). The dependent variable for all city data was whether the city was ranked (1) or not (0) in Bicycling magazine's list, which is a binary variable. Since ranked data are involved, Spearman's Rank-Order Correlation was used though similar results were found using Pearson's correlation. The correlation values are shown in Table 4.

There are only a few statistically significant correlations that are present in both sets of data, namely: Cycling-friendly business, high blood pressure, obesity, and illness. The worse the health measures, the lower a city is ranked (or the chance of being a "good for cycling" city); this result is as expected. Hilliness and most of the weather data were not in correlation to ranking which surprised the researchers as we felt that topology and meteorology would play an important part in determining the best cycling cities. The more north a city is, the more chance it will be a good cycling city; this might indicate a bias within Bicycling magazine's ranking. Regarding the best cycling cities, the larger the population and the richer the population (home value), the higher is ranked; however, Bicycling magazine has an incentive to make the larger and richer cities more highly ranked as, at some level, ranking a cities highly is a form of advertising to that cities' populous (to understand this point, consider the reverse, an individual might be less likely to purchase Bicycling magazine if it does not mention their home city).

Correlation analysis was also conducted between the variables. The obvious variables were correlated, for example, obesity and smoking rates. The only significant, slightly unusual correlation was that house prices negatively correlated with smoking and obesity, which are also correlated. This correlation could be due to obesity and smoking being more associated with poor people though we only speculate this relationship here and do not conclude it.

### 4.1.2 Interruption of the correlation analysis

Correlation does not mean causation; as such, there are several possibilities to explain the correlation between two variables, namely: A causes B, B causes A, third-party effects, self-reinforcing cycle, and by chance. We will discuss these possibilities for the correlation relation with regard to both the health factors and the cycle-friendly businesses significant correlations observed in the above results.

**Table 3** Descriptive statistics of the fifty best cities and the other 69 cities

| City group | Statistic | Population | Hilliness | Home value | Annual low temp | Annual high temp | Driest | Wettest | Lat | Long |
|---|---|---|---|---|---|---|---|---|---|---|
| Best50 | Mean | 682,600 | 48.34 | $339,938 | 29.92 | 85.86 | 1.43 | 4.54 | 38.42 | − 96.78 |
| Others | Mean | 409,436 | 54.87 | $229,179 | 33.81 | 90.86 | 1.59 | 4.15 | 35.38 | − 96.33 |
| Best50 | SD | 1,253,161 | 26.94 | $247,775 | 12.34 | 6.85 | 1.13 | 1.91 | 5.00 | 17.23 |
| Others | SD | 349,268 | 30.30 | $153,246 | 9.69 | 7.66 | 1.28 | 1.96 | 4.02 | 15.80 |
| | | Smoking | Obesity | Sleeping | No exercise | Illness | Mental illness | High blood pressure | Cycling business friendly | |
| Best50 | Mean | 18.52 | 23.46 | 34.93 | 21.91% | 11.84% | 12.32% | 28.68 | 16.48 | |
| Others | Mean | 19.31 | 24.83 | 36.77 | 25.53% | 12.84% | 12.58% | 31.02 | 3.12 | |
| Best50 | SD | 4.29 | 4.63 | 5.11 | 5.61 | 2.65 | 2.10 | 5.71 | 19.57 | |
| Others | SD | 3.36 | 4.36 | 3.72 | 4.65 | 2.30 | 1.96 | 4.49 | 5.49 | |

The correlation analysis indicates a negative relationship between negative health factors and being a good cycling city (which implies a positive relationship between positive health factors and being a good cycling city). Since cycling is supposed to improve health, it might be reasonable to assume that the good cycling cities have improved health and lower negative health factors, e.g., lower obesity rates, hence implying B causes A. However, the relationship, between being a good cycling city and negative health factors, cannot purely be explained by this explanation due to the spread of the data. Simply put, there just are not enough cyclists to explain the difference in obesity rates. From our city data, the percentage of obesity ranges from Fermont, CA (12.9%) to Detroit, MI (36.4%). Since only about 0.6% of the US population commutes to work (Pucher et al. 2011) and only about 5% biking (14 million) of the US cycles at least twice a week (Breakaway Now Research Group 2015), there simply is not the number of cyclists to explain this 23.5% spread in obesity rates. Hence, something else must be occurring for this relationship to be happening. We would argue that healthier populations in the US are more likely to have individuals take up cycling; hence there is an 'A causes B' relationship as well. We would argue that the correlation between best cycling cities and negative health factors is a reinforcing one: better cycling facilities mean healthier populations (through more cycling), and healthier populations tend to cycle more, resulting in public demand for better cycling facilities. It is important to note that Bicycling magazine did not include any health indicators in their derivation of the top 50 cycling cities; as such, we can conclude that this correlation is not a consequence of how the subjective ranking of cities was formed in the first place.

The relationship between the number of cycling-friendly business, in a city, and whether the city is a good cycling one can be explained by the way that Bicycling magazine generates its rankings. Bicycling magazine uses the number of cycling-friendly businesses in its calculations to determine the city rankings; as mentioned previously, we do not the exact method for doing this calculation. This inclusion, in Bicycling magazines calculations, could explain the correlation found above, and the reason that the cycling-friendly business factor appears in the regression models, which we will discuss later. To determine if the numbers of cycling-friendly businesses really matter in determining if a city is a great cycling city, we could use another cycling city ranking dataset, for example, BikesForPeople has its own city ranking, which does not include cycling-friendly business (BikesForPeople 2019). We leave this further check of this relationship for future research.

## 4.2 Cluster analysis on the fifty best cycling cities

Both Hierarchical and K-mean cluster analyses were conducted on the fifty best cycling cities data. The purpose of the analysis to determine if there were any obvious groupings of the cycling cities that might give us insight into what makes a great cycling city.

### 4.2.1 K-mean cluster analysis

K-mean cluster analysis was conducted to determine if there were any geographical groups that emerge. The latitude and longitude coordinate information were removed before performing the analysis. K-mean clustering places the dataset into 'k' partitions such that the distance (Euclidean) between the points (cities) and their partitions mean is minimized (Everitt and Dunn 2010). The variables determine the dimensions of the space which meant our points are placed in the space of 17 dimensions (excluding latitude and longitude). Each partition, shown in Fig. 2, is identified by a different color for readability, which we, ironically, graph using latitude and longitude.

Figure 2a shows a split between the east and west. This split implies that there are distinct differences between the two sides of U.S.A. beyond their geographic location. To determine if the weather data introduced some geographical bias on the results, we repeated the analysis removing the four meteorological variables. As seen in Fig. 2b, the split is still apparent, but with some crossover, e.g., Boston is now in the western group. Thus, the socioeconomic conditions of the two coasts create distinct conditions for the cities. Note that hilly cities are spread across the U.S.A., so topological information was not excluded from the analysis. This finding implies that eastern cities would be advised to look at other eastern cities when looking for success criteria, for developing their bicycling infrastructure plans, because of their similarities between the cities and similarly for western cities.

The number of groups, for both k-means analyses, was determined by the "elbow" method, which was first developed by Thorndike (1953). The method looks for the most rapid change in gradient when the total error for each number of partitions, Fig. 3, it can be seen that a single clear "elbow" occurs with two partitions. This was the criterion used for the selection of two partitions ($K=2$) in our cluster analysis.

### 4.2.2 Hierarchical clustering

Hierarchical cluster analysis was also conducted for the best cycling cities, and the dendrogram can be seen in Fig. 4. Latitude and Longitude data were removed from the analysis like in the k-mean clustering. Like the k-mean

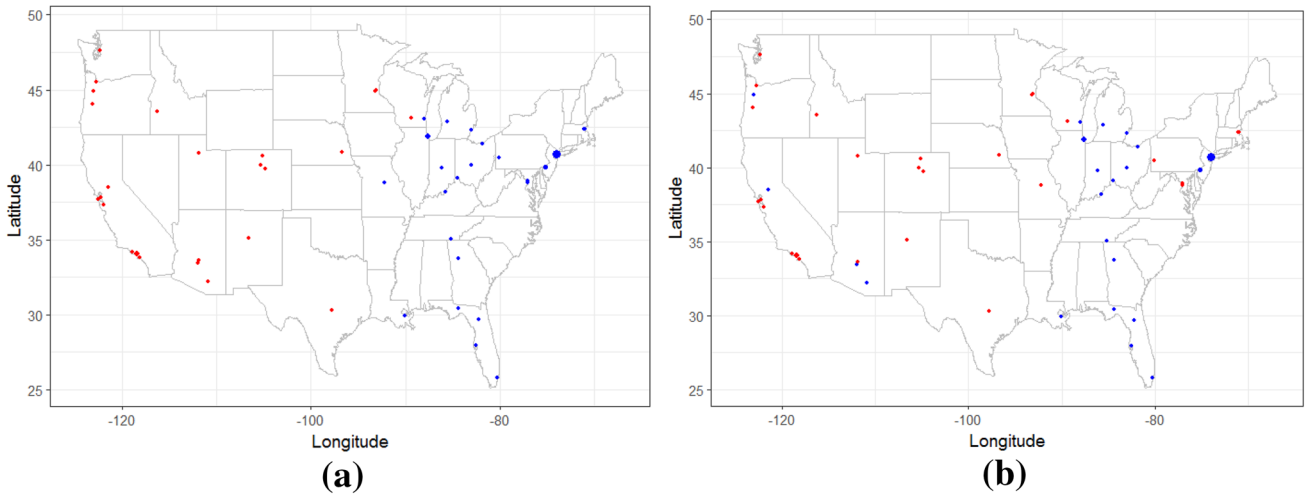


**Fig. 2** K-mean cluster analysis results from the best bicycling cities data **a** excluding latitude/longitude and **b** excluding latitude/longitude and weather data. Size indicates population and solid circles indicate the best U.S. bicycling city, according to Bicycling Magazine

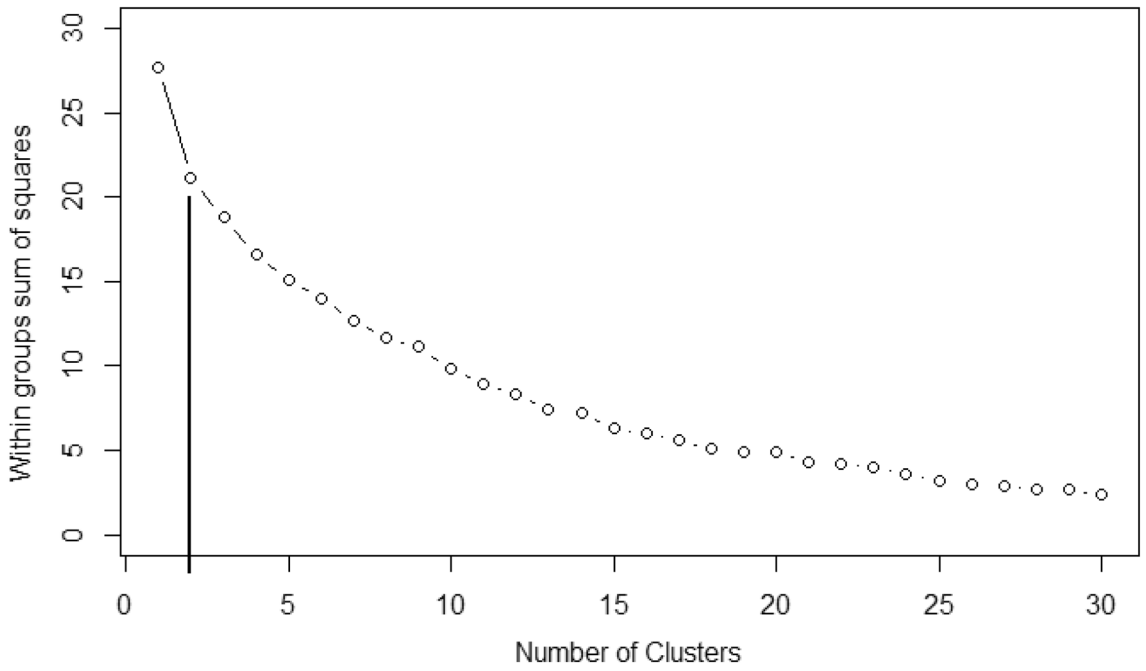


**Fig. 3** Sum of squared errors used in "Elbow" method for determining the number of groupings for the best cycling cities

cluster, there is a split between eastern and western cities in the groupings, with the eastern cities being in the first major grouping and the western cities in the second major grouping.

The dendrogram also revives some other interesting groupings. For example, the coastal Californian cities are all grouped along with Albuquerque, New Mexico. Also, Boston and Washington are more similar to each other than their suburbs, Cambridge, and Alexandria. We believe this dendrogram supports our claim that there is a difference between the eastern and western cities.

### 4.3 Cluster analysis including all cities

The second set of cluster analysis was conducted on all the city data. We were unable to obtain 'cycle path miles per square mile' and 'number of people per shared bicycle' information, for all cities, so these variables were removed

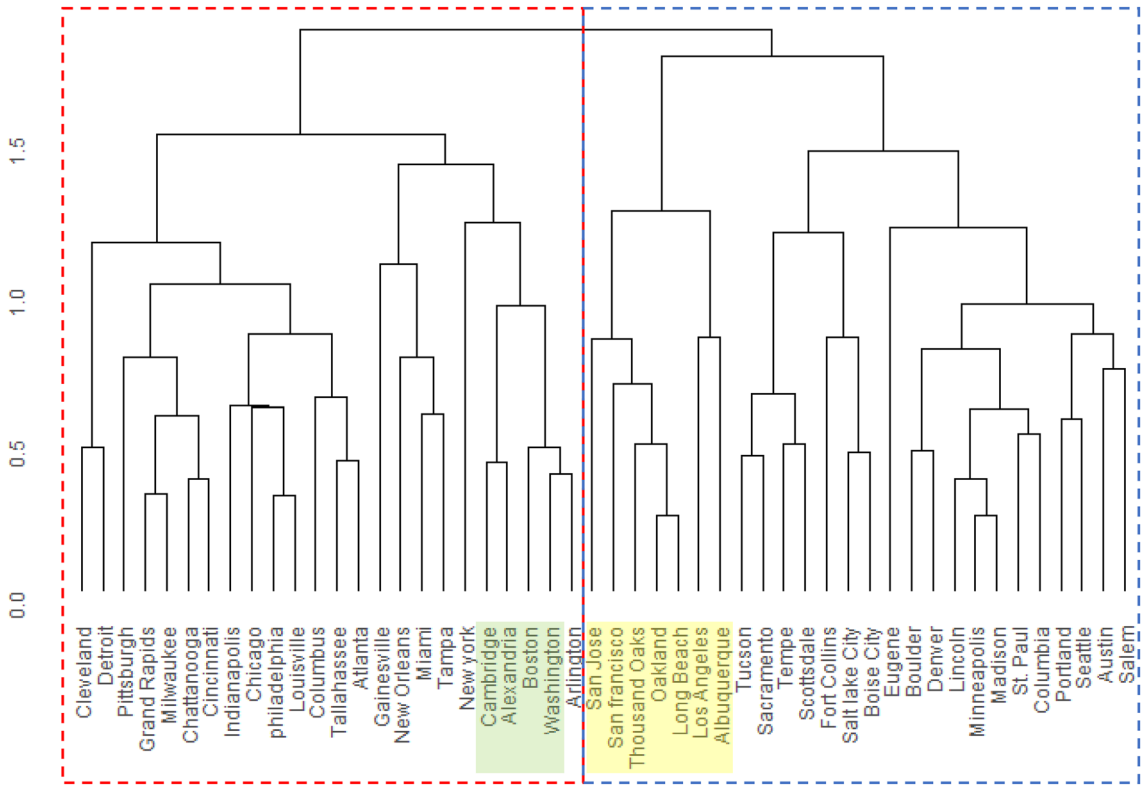


**Fig. 4** Dendrogram of the fifty best cycling cities, excluding latitude/longitude data

along with longitude and latitude. As with the 50 best cities dataset, we conducted cluster analysis using k-mean clustering and hierarchical clustering.

#### 4.3.1 K-mean cluster analysis

Using the elbow method, we determined that there should be either two or seven clusters, as shown in Fig. 5.

The two cluster case splits along the east and west divide again, shown in Fig. 6a, similar to the previous result.

The seven clusters are shown in Fig. 6b with the 50 best bicycling cities shown as filled out circles and the others as circle outlines. It should be noted that every group contains at least one of the best cycling cities. There is a striking geographic split of the cities, especially given that all coordinate data were removed from the analysis. These groups can be approximated to a California coast group (red); a southern coastal group (including Hampton Roads) (blue); a non-coastal eastern state group (orange); a big eastern cities group (gray); a northern Midwest group (purple); a southern Midwest group (green); and a Arizona/inland California group (black). These descriptions are not perfect and are only approximate, for example, North Carolina is in big cities group, but Los Angeles is not.

#### 4.3.2 Hierarchical clustering

To gain a better understanding of the cluster formation, we conducted a hierarchical clustering analysis, including all the cities' data. However, due to the total number of cities, we excluded the resultant dendrogram because it was difficult to read. A reduced version is provided in Fig. 7. Performing hierarchical cluster analysis on a subset of the data does not affect the results (it would change for k-mean clustering).

The dendrogram shows that the Hampton Roads cities can be split into two groups, namely Chesapeake and Virginia Beach in one group and the rest were in another group. Chesapeake and Virginia Beach seem to be very similar to the northern Florida cities of Gainesville and Tallahassee, whereas the rest seem to be similar to Louisiana cities and Louisville, KY.

From the two sets of cluster analysis, it would be suggested that the Hampton Roads city planners look at the cities of the southern coast when looking for comparable cities, especially Louisiana and Northern Florida. There is a multitude of other statistical analysis methods that could have been applied to the dataset including factor analysis, principal component analysis (PCA), and Bayesian model. We leave this analysis to future research. The research also acknowledges that other measures could also be included,

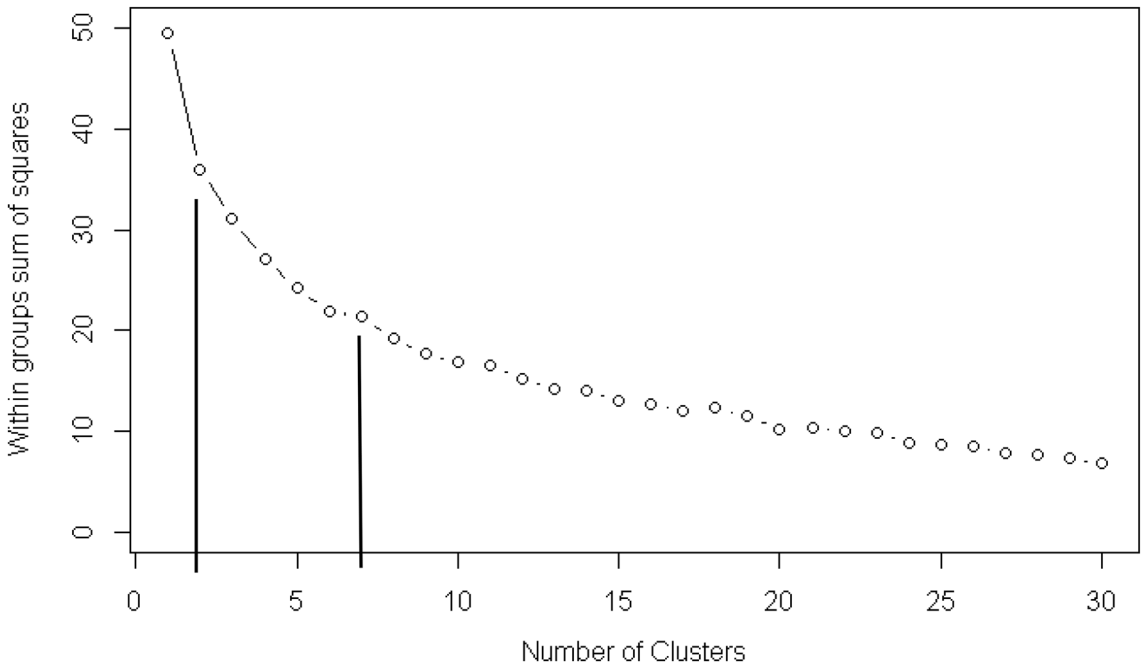


**Fig. 5** Sum of squared errors used in "Elbow" method for determining the number of groupings for all cities

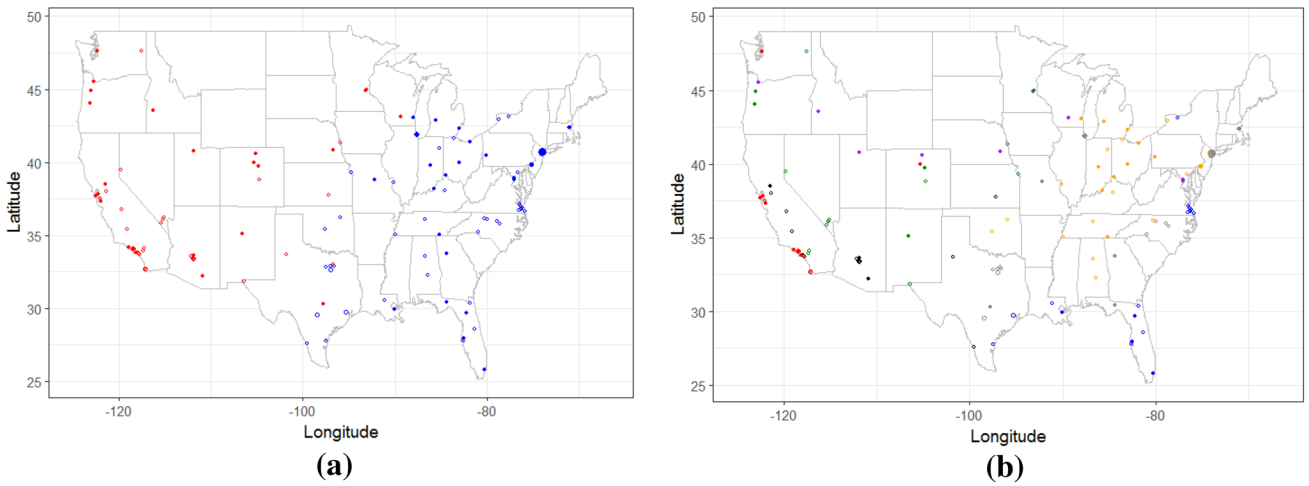


**Fig. 6** K-mean cluster analysis results for all cities data using all variables excluding latitude/longitude with **a** two clusters and **b** seven clusters

including student population, political will, etc.; these extra variables could have changed our results.

## 4.4 Evaluation of cluster models

Since cluster analysis is a descriptive statistical approach, there are no associated hypothesis tests. To evaluate the clusters, we performed an analysis on the Average Silhouette Width. Average Silhouette Width (ASW) is a measure of the coherence of a k-means clustering solution (Kaufman and Rousseeuw 2009). It ranges from − 1 to 1. A high ASW value (> 0.5) means that the clusters are homogeneous (all cities are close to their cluster center) and that all clusters are well separated. Lower values represent less structure in clusters. The ASW values for the fifty best cycling cities model and the all data model were 0.23 and 0.21, respectively. These results imply some salient patterns in the characteristics of the cities are captured within the clusters (Rousseeuw 1987). The low scores are probably due to the bias of ASW method towards to n-spheres, a shape our clusters do not follow.

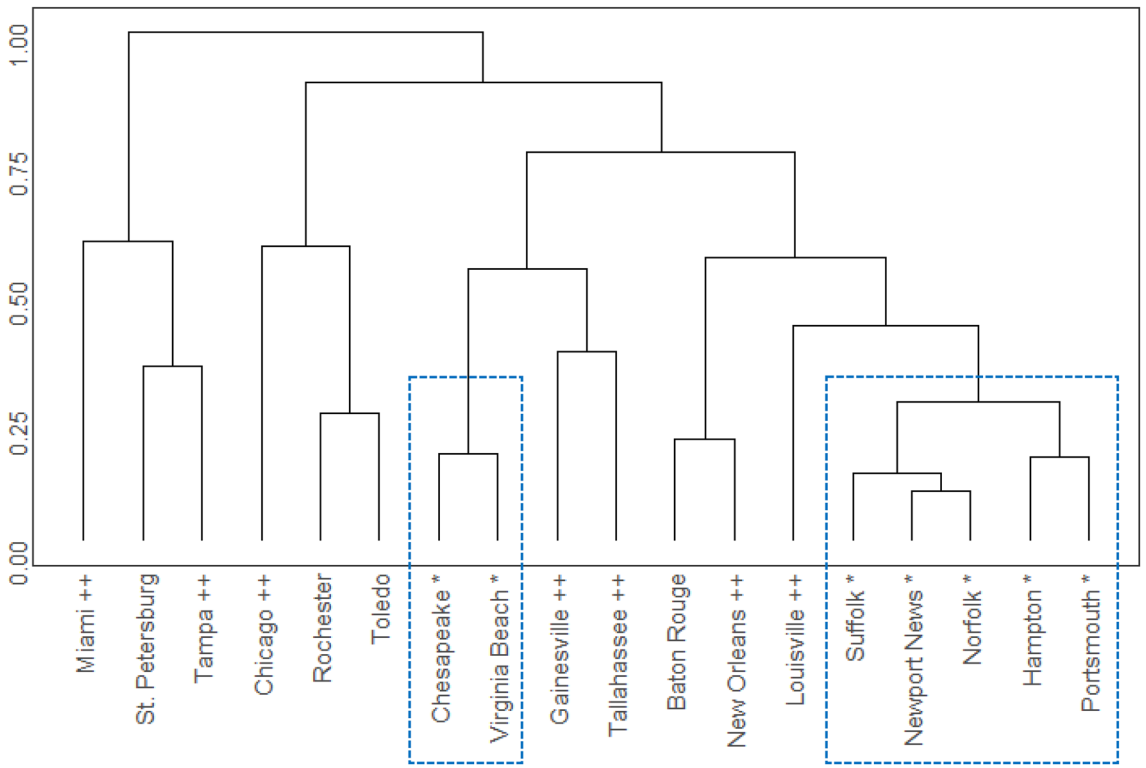


**Fig. 7** Dendrogram of all cities data with all variables excluding latitude/longitude. The distance measure is shown on the y-axis. Hampton Road cities' names are followed by a '*' and the best cycling cities are followed by a '++'

Variations of our cluster models were constructed using different variable combinations. When only the health and weather variables were considered in the models, ASW values of 0.25 and 0.34 were found. However, we decided to present the results that showed all the variables in this paper to give the reader a more generalized view. There were some statistically significant correlations between the variables, especially the health-related variables. These correlations could bias the cluster analysis towards the health variables.

Hierarchical clustering does not use the partition mean and, thus, will produce slightly different results than the k-mean cluster analysis. As such, we argue that the two approaches help triangulate on the results. The clustering approach could be used by a city planner to help them find similar cities for inputs into their planning processes. If the city planner were not from a city listed, this would require the planner to collect data specified in Table 2. These data would be added to our existing dataset, and the cluster analyses repeated (i.e., normalization of the data, the elbow method for determining the number of clusters, etc.). Once these similar cities are found, the planner could consult literature related to these cities to help them make a more informed decision about future additions and modifications to his/her bicycling infrastructure plans. Future work on our project could include developing an automated tool that would conduct the cluster analysis and present which cities are similar; thus, requiring the city planner to only need to collect and enter the data (Table 4).

### 4.5 Regression model of the fifty best bicycling cities

To get a better understanding of what makes a high ranking cycling city, several multivariate linear regression models were constructed using the best cycling city data. Initially, a subset of the variables was created to remove duplicate of effects due to multicollinearity. This subset of variables was used to construct a series of regression models with the non-statistically significant variables iteratively removed until we reached the following model:

$$y_{\text{best}} = -1.021x_{\text{blood}} + 0.233x_{\text{business}} \quad (1)$$

There '$y_{\text{best}}$' is the reverse ranking of the fifty best cites (according to Bicycling magazine); thus, a higher number implies a better city for cycling. The independent variables were '$x_{\text{blood}}$,' the percentage of the city taking medicine for high blood pressure control, and '$x_{\text{business}}$,' the total scoring of cycling-friendly business in the area. The relationship with cycling-friendly business is unsurprising due to

**Table 4** Spearman's rank-order correlation coefficients compared to reverse ranking of 50 cities

| Variable | Best50 | All | Variable | Best50 | All |
|---|---|---|---|---|---|
| Population | 0.29* | 0.075 | Hilliness | −0.021 | −0.116 |
| Home value | 0.448** | 0.209 | Smoking | −0.334* | −0.126 |
| Miles | 0.21 | | Obesity | −0.401** | −0.18* |
| Shared | 0.22 | | Sleeping | −0.232 | −0.182* |
| Annual low temp | −0.093 | −0.179 | No exercise | −0.268 | −0.341** |
| Annual high temp | −0.241 | −0.366** | Illness | −0.326* | −0.206* |
| Driest | 0.02 | −0.038 | Mental illness | −0.327* | −0.064 |
| Wettest | −0.056 | 0.037 | High blood pressure | −0.459** | −0.257** |
| Lat | 0.256 | 0.35** | Cycling-friendly businesses | 0.457** | 0.488** |
| Long | −0.075 | −0.033 | | | |

*, ** Imply statistical significances at 95% and 99% levels respectively

**Table 5** Linear regression outputs for the best cycling cities

| Variable | Coefficient | SE | *P* value |
|---|---|---|---|
| High blood pressure | 1.021 | 0.074 | 0.000*** |
| Cycling-friendly Businesses | −0.233 | 0.085 | 0.009** |

**Significant at a 99% confidence level, and ***Significant at 99.9% level

**Table 6** Logistic regression outputs for the best cycling cities

| Variable | Coefficient | SE | *P* value |
|---|---|---|---|
| (Intercept) | −6.057 | 2.589 | 0.019* |
| House price | 6.88E−06 | 2.20E-06 | 0.002** |
| Obesity | 0.331 | 0.121 | 0.006** |
| No exercise | −0.21 | 0.071 | 0.006** |
| Cycling-friendly businesses | 1.27 | 0.039 | 0.001** |

*Significant at 95% confidence level, and **Significant at 99% level

Bicycling magazine's using this variable in its determination of the best cycling cities. The statistical significance of these results can be found in Table 5. The adjusted R-squared score was 82.54%; implying that most of the variability amongst the cities was captured in our model. The correlation between the '$x_{\text{business}}$' and '$x_{\text{blood}}$' variables was −0.23.

However, from our dataset, it is not clear whether the relationship was actually post hoc, that is because a city is great for cycling results in more cycling-friendly businesses and a decrease of high blood pressure in its residents. The authors suspect that the relationship is self-reinforcing, e.g., a fitter (low blood pressure) population might be more amiable to having cycle paths, which, in turn, makes it easier for the population to exercise (lowering blood pressure). This relationship is discussed above in the interruption of the correlation analysis section.

## 4.6 Logistic regression for all cities

A regression analysis was conducted on the dataset of all the cities. In this case, whether a city is in the top fifty cycling cities ("1") or not ("0") was the dependent variable in this analysis. Since this variable was binary, a normal linear regression model would not be appropriate as it assumes the dependent variable is continuous. As such, a generalized linear regression, specifically a logistic regression, model was used. A logistic regression model produces an s-curve from the dependent variables, which can be approximated to either 1 (is a good city) or 0 (not a good city).

Logistic regression models were iteratively created from all the dependent variables by removing all the dependent variables that were not statistically significant (at the 95% level) at each step. Unfortunately, our final model resulted in multicollinearity effects, which are discussed in detail below. The final logistic model was

$$I_{\text{best}} = \frac{1}{1 + e^{-(0.00000688x_h + 0.331x_o - 0.221x_n + 1.27x_b - 6.057)}} \quad (2)$$

The dependent variable '$I_{\text{best}}$' is the model's prediction of whether a city has the characteristics of being a good city for cycling. The independent variables are '$x_{\text{h}}$' for house prices, '$x_{\text{o}}$' for obesity rates, '$x_{\text{n}}$' for no exercise rates, and '$x_{\text{b}}$' for bicycle-friendly business. Again, bicycling-friendly business numbers are expected to relate to the binary output variable due to the way that Bicycling magazine determined its calculation of the best cycling cities. The statistical significance of these results can be found in Table 6. The results incident that more wealthy cities and cities with bicycling-friendly businesses are likely to be good cycling cities; as discussed previously, this might be due to bias in Bicycling magazine's methodology. It also shows that cities, where a lot of the population does not exercise, are not likely to be good cycling cities. Controversially, it shows that more

obese cities are better for cycling, which we will be discussed in more detail.

Determining how good this logistic model helps us determine its usefulness. Since logistic regression makes different distribution assumptions than normal regression, normal model goodness measures cannot be used, i.e., adjusted *R*-squared. We, therefore, had to use a different measure of goodness of fit. The McFadden's Log Likelihood statistic (or McFadden's pseudo-*R*-squared statistic) is a standard measure for logistic regression (McFadden 1973). Unlike the *R*-squared measures, it compares the log likelihood of the model to the null model (that is, the model where the dependent variable is considered to be constant. This is done as follows:

$$1 - \ln \mathcal{L}\,(\text{Model}|\boldsymbol{x})/ln\,\mathcal{L}\,(\text{Null Model}|\boldsymbol{x}) \tag{3}$$

Since likelihood is a probability, it takes values between zero and one, inclusive. A perfect model would have a likelihood of one, resulting in McFadden statistic of one. Assuming the null model has the worst likelihood, this would have a McFadden statistic of zero. Thus, the higher the McFadden, the better the data fit the model.

Our model scored a McFadden statistics of 0.313, which implies it is a weak fit to the data and that other important factors are missing. However, this result is not surprising as we did not include a lot of factors used by the Bicycling Magazine to determine the best cycling cities, for example, the number of female cycle commuters, and people per bike share (which was removed from the all city dataset). Our analysis is about using geographic, meteorology, and socio-economic data to determine what makes good cycling cities beyond the obvious cycling-related measures. As such, we are satisfied with the McFadden statistics as a first try at this analysis and would want to improve its value in future work by incorporating more variables into our analysis.

A reader might interrupt the positive coefficient for the obesity variable to imply that more obese a city, the better it is for cycling. However, as our model is multivariate, the situation is more complex due to multicollinearity effects between obesity rates and the 'lack of exercise' variable. It should be noted that just because a model has multicollinearity effects does not invalidate it and we hope to provide a justification for the relationship between 'no exercise' and obesity rates in our model.

If only univariate regression models were considered, then the coefficient signs would be the same as those observed in the correlation analysis. We constructed a univariate logistic regression model using only obesity as the independent variable, and its coefficient was negative, as expected, in this model. Thus, the positive coefficient for obesity is due to the interplay with the other variables. Scatter graphs of the dependent variables are shown in Fig. 8.

The scatter graphs show a positive linear relationship between obesity rates and lack of exercise, for the city data. However, though there appears to be a negative relationship between the home value and obesity/no exercise, this relationship is exponential, not linear. A logistic regression model is, at its heart, still a generalized linear model and thus is looking at linear relationships between the dependent and independent variables. If the relationship between home value and obesity/no exercise had been linear, there would have been no need for one of the obesity or no exercise variables (as their effect on the dependent variable would have already been taken into account with the remaining variable). However, this is not the case, and, as such, minimizing mean squared error algorithm, used in determining the coefficient variables, compensates by introducing a damping variable, which, in our case, is the obesity variable. In an ideal world, we would want all our independent variables to be uncorrelated; unfortunately, we do not live in an ideal world and must accept some correlation between our variables.

Other models were constructed that had little correlation between variables, but the models were deemed not as useful because either the variables were not statistically significant or McFadden's Log Likelihood statistic was very low. For example, if the obesity variable was removed, the McFadden's Log Likelihood statistic would be 0.20 and only the 'bicycle-friendly business' variable remained significant. This indicates that the complex interplay between obesity rates and the 'lack of exercise' variable is important in understanding what makes a great cycling city. A simplistic level, a slimmer city population, that does not exercise, is less likely to be a great cycling city compared to a fatter city population. This phenomenon may be due to the accessibility of bicycling to obese people over other forms of exercise, e.g., jogging or calisthenics exercise.

Why is the obesity variable the damping variable and not the 'lack of exercise' variable? We believe it happens because the obesity variable is not significantly different between the two populations (ranked and not ranked). This can be seen in Fig. 9. A Welch's *t* test for unequal variances was conducted on the obesity values for the two populations, and it was shown to not be statistically significate ($P$ value = 0.05198). Thus, the obesity variable has a little impact on determining if a city is ranked (correlation of 0.18), which is required to dampen the effects of 'no exercise' within the model without unduly affect the model's output.

#### 4.6.1 Regression model predictions for the Hampton Road region

The logistic regression model was applied to cities of Hampton Roads in Virginia to give an example application of our model. The results of this application are shown in Table 7.

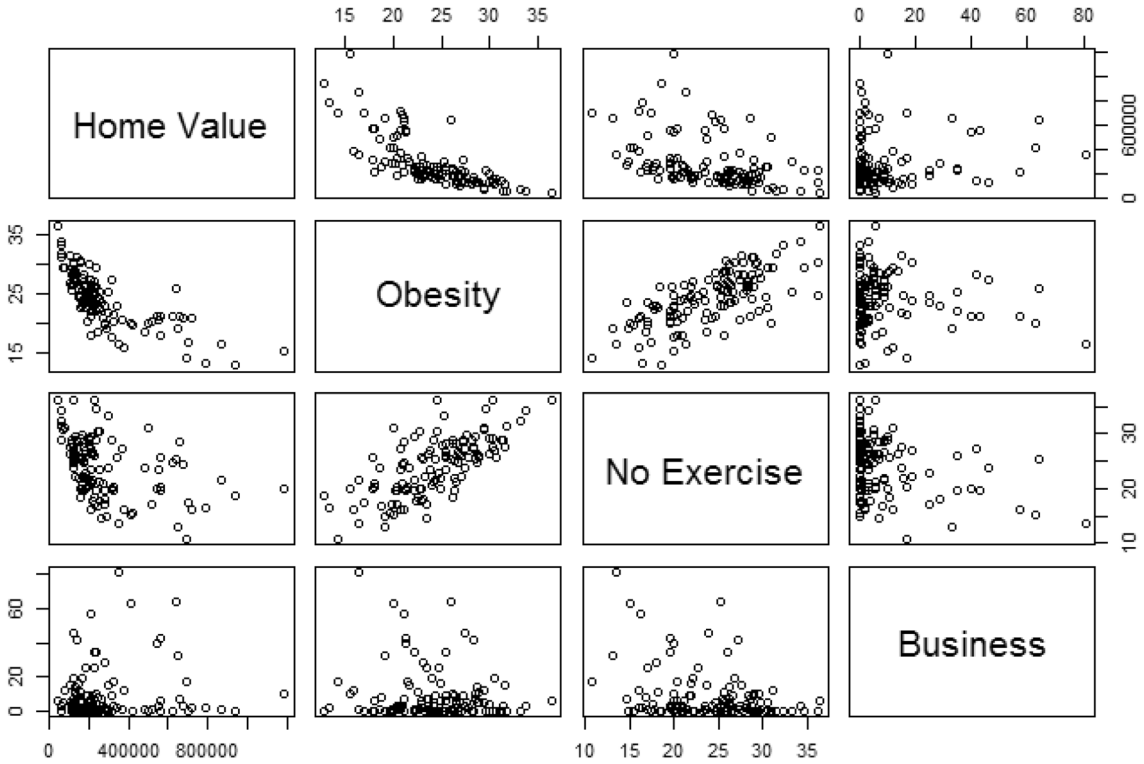


**Fig. 8** Scatter graphs of the dependent variables from the logistic regression analysis

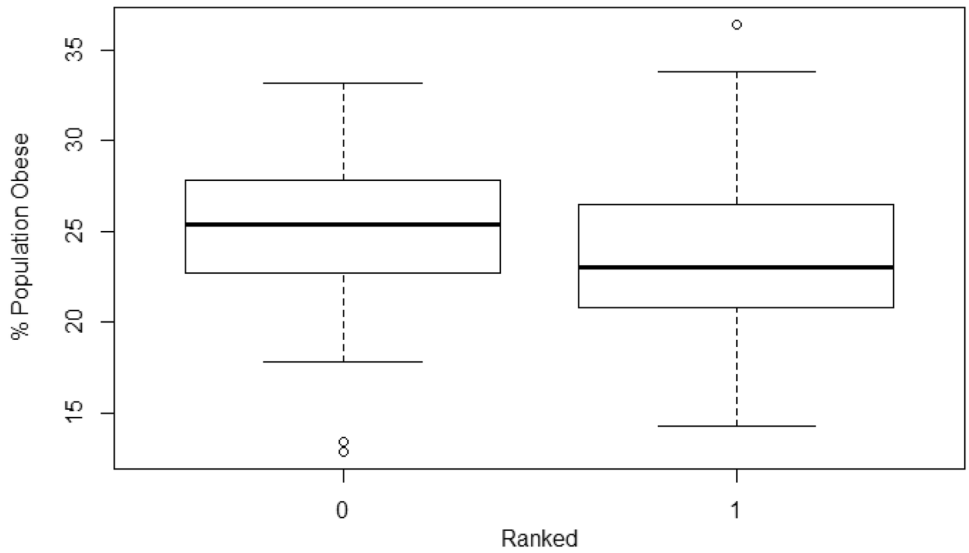


**Fig. 9** Box plots of the percentage of the population that is obese for the best cycling cities ("1") and the others ("0")

The scoring indicates the percentage chance we would expect the city to be "ranked," that is, the chance it is a good city for cycling.

The city of Norfolk is a significantly higher chance of becoming considered a good cycling city. This is mainly due to Norfolk being the only city in the Hampton Roads region with recognized bike-friendly businesses (by the League of American Bicyclist). This result validates our model somewhat because, in recent years, Norfolk has begun an extensive effort to expand its cycling infrastructure with the introduction of many new cycle paths with its cycle plan (City of Norfolk 2014). Figure 10 shows some new cycle paths laid out in Norfolk in the last few years.

**Table 7** Application of the logistic regression model to the Hampton Roads region

| City | Score (%) |
|---|---|
| Norfolk | 79 |
| Hampton | 45 |
| Virginia Beach | 32 |
| Chesapeake | 31 |
| Portsmouth | 30 |
| Suffolk | 27 |
| Newport News | 20 |

## 5 Discussion

The paper has looked at several different modeling approaches, including cluster analysis, to investigate what makes a good bicycling city and, more importantly, to understand the relationship between cities, in terms of cycling. What is a good bicycling city is subjective, and we use a single data source, Bicycling magazine, to make this determination for us. Our clustering approach could be used by a city planner to help them find similar cities whose existing plans they could use as a guide for their planning process. Currently, this would require the planner to collect data specified in Table 2 on their city. These data would be added to the existing dataset, and the cluster analysis process repeated (i.e., normalization of the data, elbow method for determining the number of clusters, etc.). Once these similar cities are found, the planner could consult literature related to these cities to help them make more informed decisions about future additions and modifications to his/her bicycling infrastructure plans. Future work on our project could include developing an automated tool that would conduct the cluster analysis and present which cities are similar; thus, requiring the city planner to only need to collect and enter the data.

The dataset used, from Bicycling magazine, is far from inclusive of all possible effects that determine a good bicycling city; for example, it has been shown that cycle network structure has an important impact (Pucher et al. 2010; Schoner and Levinson 2014; Marqués et al. 2015; Pucher and Buehler 2016). We did contact Bicycling Magazine about how they exactly calculated their rankings but did not receive a response; as such, their rankings were considered a "blackbox" in terms of the output variable. However, Bicycling magazine is the most widely circulated periodical on bicycling and, thus, their opinions carry weight within the community, and they have established themselves as a legitimate source of cycling information within the community for the last 50 years.

As mentioned throughout the analysis, Bicycling magazine may have bias within it is ranking, for example, giving a higher ranking to the more populous wealthy cities or cities with lots of bicycling-related businesses (that could sell the magazine). However, this might be the case, i.e., larger wealthy city might just have more people that are in to bicycling. As such, it should be remembered that what is a "good" cycling city is subjective, and any dataset will have its own bias. Even if the data and methodology used in determining what makes a "good" cycling city are publicly available, there still exists the bias in determining what measures are included in the first place. However, to add robustness to this work, it is suggested, for future work, that the analysis be complete on a ranking of cycling cities that is independent of Bicycling magazine, if that is possible.

As previously mentioned, there does exist another dataset that rank the US cities in terms of cycling, for example, BikesForPeople (2019) The BikeForPeople dataset uses different variables in determining what makes a good cycling


**Fig. 10** Cycling lanes introduced into Norfolk, VA that overcome the geographic limitations of **a** bridges, and **b** complex historic road layouts

city, i.e., ridership numbers, fatality rates, network coverage, ridership covers all demographics, and the cities' proposed future improvements (the Bicycling Magazine uses miles of bicycle lanes per square mile, bicycling-friendly business including cyclist-friendly bars, number of female cycle commuters, and people per bike share). Note that BikesForPeople dataset introduces a bias towards "up and coming" bicycling cities with its "future improvements" measure. This measure is biased against cities that have already spent, previously, development dollars, in the past, to improve the cycling infrastructure. As such, we reiterate any dataset on the ranking of a subjective measure will contain some bias.

We did conduct a quick analysis between the Bicycling magazine and the BikesForPeople data. The two datasets were different in ways beyond the variables used to calculate the score. The correlation between the two datasets score, for a given city, was only 60%. Several of the cities included in the Bicycling Magazine data were not included in the BikesForPeople data, i.e., Grand Rapids, Thousand Oaks, and San Jose. The top two bicycling cities, Chicago and San Francisco, were considered mediocre in the BikesForPeople dataset, and the top two BikesForPeople US cities, Boulder and Fort Collins, were ranked 10th and 12th place, respectively, in the Bicycling Magazine data. As such, there is merit to conduct our analysis with the BikesForPeople dataset to see if our findings still hold and we leave this analysis for future work.

Given the issued discussed with the available rankings, we could construct our own ranking. Other data sources could have used, like the League of American Bicyclists, to construct an independent ranking of the US cities based on measures that we determine. However, as already mentioned, the selection for these measures is subjective and, as such, we believe that this ranking should be left to subject matter experts within the industry to determine this value. For validation purposes, we could repeat the analysis given above using the BikesForPeople data to see how the models compare and repeat the analysis only using measures that align with both datasets; we leave this to future work.

Other validation approaches could have been employed in this research; for example, the cluster analysis models could have been validated using cross-validation. We considered this type of validation approach but decided against it due to the relatively small numbers used (120 cities), i.e., removal of 30 cities for validation be actually be a quarter of data sample. We were concerned that any removal of data for validation would either warp the initial analysis if too many were removed or produce meaningless validation results if too few were removed. Since the majority our results are from exploratory data analysis, it would have also been difficult to pre-determine what we were validating.

As the correlation analysis indicated, there would seem to be a cyclic relationship between what makes a good cycling city and the health factors. To explore this relationship further, a longitudinal analysis would seem the most appropriate way forward as future work for this research.

## 6 Conclusions

In this paper, a variety of multivariate statistical models were developed to investigate the relationship between a set of US cities, in terms of cycling, and to investigate what factors make a good cycling city, as determined by Bicycling Magazine. This included correlation analysis, cluster analysis, and regression modeling. Since determining what is a "good" cycling city is subjective, it is possible that Bicycling magazine's dataset has biases within it and these biases are reflected in our results. As such, the following results should be viewed with this consideration in mind. However, we would like to point out that any ranking of cycling cities will be subjective and susceptible to its own biases. As such, we recommend that this analysis be completed using a different independent ranking of "good" cycling cites to see any of our results persist which are independent of these biases.

The correlation analysis showed that the best cycling cities were shown to be more healthy, in general, than the other 69 cities considered in the study; this result was expected. However, both the cluster analysis and the regression analysis produced surprising results. Our cluster analysis implied that there was a difference, beyond geographic and atmospheric, between the east of U.S.A and the west, when regarding cycling. The regression analysis showed that only socio-economic factors were important when determining a good cycling city and not geographic factors, like hilliness, and weather. Thus, our hypothesis that these physical factors are important in determining good cycling cities was incorrect. For example, hilly cities, like San Francesco, are highly ranked cycling cities. Since our analysis showed that it is a population's social characteristics that determine whether a city will be a good cycling city, we suggest that city planners should look at their population, not geography, for determining whether their cycling infrastructure plans will be successful when comparing to other cities.

However, though our analysis shows that socioeconomic factors play an important part in determining which cities are good for cycling, it is not clear what are the dependent variables and what are the independent variables. For example, does having a more active population lead to a better cycling city, or does being a good cycling city lead to a more active population? We argue that the case of Norfolk, which has the correct socioeconomic factors but is not yet a ranked cycling city, implies there is some validity to the way we have ordered variables, i.e., a more active population leads to more cycling. The takeaway, for city planners, of this paper is that it does not matter how good the geographic

and atmospheric conditions are in your city if people do not go out and cycle they will not start cycling without significant incentive. Thus, city planners should be careful not to assume that just because they have ideal conditions for cycling does not mean that the population will cycle.

In terms of the actionability of the results, we would suggest that city planners use the cluster analysis results to determine which grouping their city most likely belongs to. This can be done by comparing their cities factors, as defined in Table 2, to our list of cities. Once the grouping has been determined, a city planner may wish to look at the historical plans, and any outcomes, from those cities in their grouping to determine what is likely to work for their city.

The next stage of the project is to repeat the analysis for a larger number of cities and variables to see our results persist and to repeat the analysis using a different dataset. Since clustering is a descriptive statistical method, these results are subjective, and future work will include adding more appropriate variables to the dataset, and including more cities, to counter this.

## References


Bicycling Magazine (2017) The 50 best bike cities of 2016. https://www.bicycling.com/culture/news/the-50-best-bike-cities-of-2016. Accessed 25 May 2017

BikesForPeople (2019) City ratings. https://cityratings.peopleforbikes.org/. Accessed 16 Aug 2019

Breakaway Now Research Group (2015) U.S. Bicycling participation benchmarking study report, pp 1–64

City of Norfolk (2014) City of Norfolk bicycle and pedestrian strategic plan, pp 1–158

Clark SS, Seager TP, Chester MV (2018) A capabilities approach to the prioritization of critical infrastructure. Environ Syst Decis 38(3):339–352

Community Cycling Center (2012) Understanding barriers to bicycling project. Community Cycling Center, Portland

Elton-Walters J, Wynn N (2017) 17 best cycling apps: iPhone and Android tools for cyclists. https://www.cyclingweekly.com/news/product-news/best-cycling-apps-143222. Accessed 9 June 9, 2017

Everitt BS, Dunn G (2010) Applied multivariate data analysis. Wiley, New York

Ferguson K (2008) The destructive impact of mountain biking on forested landscapes. Environmentalist 28(2):67–68

Geelong Planning Committee (1978) Geelong Bikeplan. Geelong Planning Committee, Geelong

Harkey D, Reinfurt D, Knuiman M (1998) Development of the bicycle compatibility index. Transp Res Rec 1636:13–20

Jackson M, Ruehr E (1998) Let the people be heard: San Diego County Bicycle use and attitude survey. Transp Res Rec 1636:8–12

Kaufman L, Rousseeuw PJ (2009) Finding groups in data: an introduction to cluster analysis. Wiley, New York

Marqués R, Hernández-Herrador V, Calvo-Salazar M, García-Cebrián J (2015) How infrastructure can promote cycling in cities: lessons from Seville. Res Transp Econ 53:31–44

McFadden D (1973) Conditional logit analysis of qualitative choice behavior. Wiley, New York

Meletiou M, Lawrie J, Cook T, Obrien S, Guenther J (2005) Economic impact of investments in bicycle facilities: case study of North Carolina's Northern Outer Banks. Transp Res Rec 1939:15–21

Moudon AV, Lee C, Cheadle AD, Collier CW, Johnson D, Schmid TL, Weather RD (2005) Cycling and the built environment, a US perspective. Transp Res Part D 10(3):245–261

Parkin J, Wardman M, Page M (2008) Estimation of the determinants of bicycle mode share for the journey to work using census data. Transportation 35(1):93–109

Pierce J, Kolden CA (2015) The Hilliness of US Cities. Geogr Rev 105(4):581–600

Pucher J, Buehler R (2016) Safer cycling through improved infrastructure. American Public Health Association, Washington, DC

Pucher J, Dill J, Handy S (2010) Infrastructure, programs, and policies to increase bicycling: an international review. Prev Med 50:S106–S125

Pucher J, Buehler R, Merom D, Bauman A (2011) Walking and cycling in the United States, 2001–2009: evidence from the National Household Travel Surveys. Am J Public Health 101(S1):S310–S317

Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20:53–65

Schoner JE, Levinson DM (2014) The missing link: Bicycle infrastructure networks and ridership in 74 US cities. Transportation 41(6):1187–1204

Sears J, Flynn B, Aultman-Hall L, Dana G (2012) To bike or not to bike. Transp Res Rec 2314:105–111

Sener I, Eluru N, Bhat C (2009) Who are bicyclists? Why and how much are they bicycling?". Transp Res Rec 2134:63–72

Sorton AA, Walsh T (1998) Bicycle stress level as a tool to evaluate urban and suburban bicycle compatibility. Transp Res Rec 1438:17–24

Statistica (2017) Number of cities, towns and villages (incorporated places) in the United States in 2015, by population size. https://www.statista.com/statistics/241695/number-of-us-cities-towns-villages-by-population-size/. Accessed 25 May 2017

Thorndike RL (1953) Who belongs in the family? Psychometrika 18(4):267–276

Tukey JW (1980) We need both exploratory and confirmatory. Am Stat 34(1):23–25

Your Weather Service (2017) U.S. Climate data. www.usclimatedata.com. Accessed 9 June 2017