Project 1 Data Analysis - Group 2

|  |  |
| --- | --- |
| **Team Name:** | Data\_Slack{ers}=False |
| **Project Title:** | Where are we going to live next?: A 10 city analysis to find the most desirable city to live in |
| **Team Members:** | Ben Schatz | Sean Daily | Vraj Patel | Ashley Cooper |
| **Git Repo Link:** | https://github.com/benschatz97/group\_project\_1.git |
| **Git set up by:** | Ben Schatz **Collaborators:** Sean Daily | Vraj Patel | Ashley Cooper |
| **Datasets used:** | Cost of Living in the US - total cost of living for housing cost, food, transit, childcare taxes and other necessities Crime Rate for Each City - major crime rate for each city  Urban Connectivity - includes the city walk score, transit score and bike score Population Growth - population growth of the city from 2020-2024 |
| **Dataset links:** | [Cost of Living in the US (kaggle.com)](https://www.kaggle.com/datasets/wayneitprohatter/cost-of-living-in-the-us) [United States Crime Rates By City Population (kaggle.com)](https://www.kaggle.com/datasets/kabhishm/united-states-crime-rates-by-city-population?select=crime_250_plus.csv) [US Cities Urban Connectivity (kaggle.com)](https://www.kaggle.com/datasets/vellis1/us-cities-urban-connectivity) [Population of all US Cities 2024 (kaggle.com)](https://www.kaggle.com/datasets/dataanalyst001/population-of-all-us-cities-2024) |

|  |
| --- |
| **Hypothesis:** Nashville is the most desired city to live in based on its ranking for cost of living, crime rate and urban connectivity.  **10 Cities Analyzed**: Atlanta, Austin, Boston, Chicago, Denver, Los Angeles, Nashville, New York, San Diego and Tampa |

|  |
| --- |
| **Questions:**   1. Do we want to live in a large (population) city or small? 2. **​​**As the population increases in the city how does it affect the total crime for the city? 3. As the population increases in the city how does it affect the walk score? 4. As the population increases in the city how does it affect the transit score? 5. As the population increases in the city how does it affect the bike score? 6. As the population increases in the city how does it affect the total cost of living? |

|  |
| --- |
| **Data cleaning -** Ben, Sean, Vraj, Ashley Cost of Living in the US - Ashley  Crime Rate for Each City - Sean  Urban Connectivity - Ben  Population Growth - Vraj  **Analysis Summary (bar charts, regression analysis) -** Ben, Sean, Vraj, Ashley  **Analysis Summary -** Ben, Sean, Vraj  **Presentation -** Ben, Sean, Vraj, Ashley |

**Project Overview:**

For this project we decided to pick 10 US cities, with varying populations, to move and settle down. Our goal is to determine which of the 10 cities we selected is the best option for us. The cities that we picked out were Atlanta, Austin, Boston, Chicago, Denver, Los Angeles, Nashville, New York, San Diego, and Tampa. We are going to use data sets we found from Kaggle.com to create several different bar and scatter plots accompanied by regression analysis to visualize our analysis. To analyze each city, we decided to compare three different lifestyle categories against each city’s population. The categories we chose were crime rate, cost of living, and urban connectivity. Urban connectivity was broken into three different categories including bike score, walk score and transit score. We ultimately combined these three scores to come up with a single urban connectivity rank. We chose to use these categories because we thought they are areas pertinent to most people’s everyday living.

Our regression statistics will be calculated based on 2024 population compared to all our other variables for each city. We decided to break down the project into the four categories because we have four team members. Each team member would be responsible for completing EDA, charts, regression analysis and conclusions for each section of the project. This was done not only for simplicity, but to give each team member a complete understanding of the project and to be able to discuss any portion of the project when it comes to questions. Specific roles for the project included: Vraj Patel: population and introduction to the project, Ashley Cooper: cost of living, Sean Daily: crime rates, Ben Schatz: Urban Connectivity and conclusion.

**Analysis:**

Our group project goal was to see which city in our top 10 that we chose as a group is the most livable city based on our data sets that we have chosen. The results are based on the data sets of population, urban connectivity, crime per city, and cost of living.

This report compares several key metrics across 10 selected cities: New York, Chicago, Boston, Nashville, Atlanta, Tampa, Austin, and others. The focus areas include population, urban connectivity (walk, transit, and bike scores), crime rates, and cost of living. These metrics provide insights into the livability and desirability of each city in terms of accessibility, safety, and affordability.

The first data set focuses on population. Among the 10 cities, New York ranks as the most populated city with an estimated population of 8.09 million. New York’s massive population underscores its role as a global hub for business, culture, and tourism. Conversely, Tampa is the least populated city, with approximately 408,510 residents, highlighting its comparatively smaller size and lower density. These population figures set the stage for understanding how urban infrastructure, safety, and cost of living interact with the size of the population.

A graph with blue and orange squares

Description automatically generated

The population comparison bar graph shows the population of the cities in 2020 compared to 2024. We can see that New York is the most populous city and yet it has the greatest change (decrease) in population from 2020 to 2024. Another interesting note from this graph is that Austin, Tampa, and Atlanta were the only cities with substantial increases in population over the last four years, with Atlanta and Tampa being the least populated of our ten cities.

Urban connectivity is a key indicator of how well people connect and move throughout a particular city. While there are dozens, if not hundreds of different factors that go into calculating a city’s urban connectivity, we decided to focus on walkability, bike-ability, and public transit. In this analysis, we broke down the urban connectivity into three categories: walk score, transit score, and bike score. We then compiled a total urban score using the sum of these three categories.

The walk score shows that New York scored an 88 (the highest of the ten cities), indicating a highly walkable city where most errands can be completed on foot. This makes sense for a densely populated and well-planned city like New York, with a concentration of amenities and workspaces. On the opposite end, Nashville ranked a mere 1/10, suggesting heavy reliance on cars due to poor walkability and a lack of pedestrian-friendly infrastructure. When running our regression analysis for 2024 population against the walk score, we are given a Pearson R-value of .64. This is a moderately strong and positive correlation between the city’s population and walk score. This means, when the city’s population increases it is reasonable to assume the walk score will increase.

A graph of a city

Description automatically generated

A white background with black text

Description automatically generated

Transit Score pertains to the public transportation, Bus, trains and other rail systems. New York again ranked the highest at 10/10. This reflects its comprehensive public transportation network, including subways, buses, and regional trains. Nashville, once again, ranked the lowest with a 1/10 ranking. This is likely due to its limited public transportation options, making it heavily car dependent. When running our regression analysis for 2024 population against the transit score, we are given a Pearson R-value of .74. This is a strong and positive correlation between the city’s population and transit score. This means, when the city’s population increases it is very reasonable to assume the transit score will increase.

A graph of different cities

Description automatically generated

A white background with black text

Description automatically generated

Chicago took the top spot for biking, ranking a 10/10, thanks to its extensive network of bike lanes and bike-sharing systems. Conversely, Nashville remained at the bottom with a 1/10 rank, indicating poor biking infrastructure and limited accessibility for cyclists. However, the correlation value for population and bike score is weak. When running our regression analysis for 2024 population against the bike score, we are given a Pearson R-value of .36. This is a positive but weak correlation between the city’s population and bike score. This means, when the city’s population increases the bike score may increase but will have much more variability than our transit and walk scores.

A graph of a city bike score

Description automatically generated

A white background with black text

Description automatically generated

When looking at the overall urban connectivity, New York emerges as the top performer across the board, making it one of the most accessible cities, particularly for those who prefer walking or using public transit. Chicago and Boston also score highly across walkability, transit, and biking, positioning them as strong contenders in terms of urban mobility. On the other hand, Nashville consistently ranks at the bottom across all three categories, indicating limited urban connectivity and a car-dependent lifestyle.

A graph of different cities

Description automatically generated

**Urban Connectivity Summary**

Our urban connectivity scores were used against the 2024 city population. We wanted to determine if cities with larger populations were easier or more difficult to get around in and interact with. Pearsons R correlation was the main statistic we used to determine these factors. Our results for our three different categories came in mixed. Our transit score came in with the highest R value of .74. This is a very good indication that more populated cities have better public transportation systems. Conversely our bike score r vale came in quite low at only .36. This is a relatively weak correlation meaning that population size is not a great indicator of how well a city ranks for biking. One reason for this may be the more populated a city the less room there is for things that may be considered luxury or non-essential perks such as bike lanes and bike paths. The walk score correlation was closest to our total urban score calculation. The walk score came in at .65 and the total urban score r value came in at .61. These values are both moderately strong indications that more populated cities are easier to get around on foot and overall are better for moving around and interacting with a particular city. From an urban connectivity standpoint, the more populated cities are more desirable.

The next category in the analysis is crime rates. Like urban connectivity, our crime rate data set was broken down into several subcategories of crime. We chose to focus on the total number of major crimes rather than in individual crime rates such as robbery, assault or murder. Atlanta ranked as the city with the highest number of major crimes among the 10 cities, making it the most dangerous in terms of overall crime. This could be due to various factors, including socioeconomic conditions, urban planning, and policing strategies. On the contrary, New York is ranked as the safest city in terms of the total number of major crimes. This may come as a surprise considering its large population and reputation for being a bustling metropolis. However, over the past decades, New York has significantly reduced its crime rate, especially in comparison to other major cities.

When running a regression analysis for total major crimes in each city against their 2024 population we are given some interesting results. The Pearson R-value for this relationship is -.42. There are two factors when analyzing an R-value of a particular relationship. The strength of the absolute value of the number (0-1). The closer the number is to zero the weaker the correlation and the closer the number is to 1, the stronger the correlation. The second factor is whether the number is positive or negative. We see that our R-value for crime versus population is a negative value. While this is not a strong correlation, the fact that it is negative is an indication that as population increases, overall crime numbers decrease. Public influence could be explained as a large factor in why when the population increases the overall crime numbers decrease.

These findings highlight that while Atlanta may offer certain economic opportunities and cultural attractions, its high crime rates may make it a less appealing option for residents concerned with safety. New York, despite being the most populated city, provides a relatively safer environment, which adds to its overall livability score.

A screenshot of a data

Description automatically generated

Looking at the data frame table above we can see that:

* Atlanta is the worst city in terms of total crimes at 5,727.52 total crimes.
* New York ranked at the top as the lowest crime rate city at 1,987.49 total crimes

A better representation of this data frame is the bar graph below:

A graph of blue bars with white text

Description automatically generated

Key takeaways from the crime rate visualizations and statistics include:

* New York is the safest city out of the 10 cities we have chosen for this project while also having the largest population.
* Atlanta is the most dangerous of the 10 cities we have chosen while being the second smallest city in terms of total population.
* Correlation between population and overall crime is moderate but negative

Finally, we examine the cost of living across the cities. In this category, cities were ranked based on a total cost index that includes housing, groceries, transportation, and other living expenses. Just as in the crime and urban connectivity categories, cost of living had several subcategories. We decided to focus on the overall cost of living. Austin emerged as the city with the most affordable cost of living among the 10 cities, making it a top choice for individuals or families looking for a more cost-effective place to live. Nashville followed closely behind, indicating that despite its poor scores in walkability, transit, and biking, it remains an attractive option for its affordability. New York is ranked as the most expensive city in terms of cost of living. This finding aligns with its status as a global financial hub and a high-demand market for real estate. While New York offers unparalleled access to job opportunities, culture, and urban connectivity, its high cost of living may be a significant deterrent for many people considering relocation.

A graph of a number of living states

Description automatically generated with medium confidence



One city that stood out was Boston. Boston is one of the least populated cities and yet it has one of the highest living costs. This is even more surprising when we factor in the overall correlation between population and total cost of living is strong at r = .69. Overall population may not be the best factor in this case. Population per square mile would likely be a better option. The Pearson R-Value of .69 is a strong and positive correlation. This tells us that as population increases, the cost of living will also increase.

When combining the data across these four major categories; population, urban connectivity, crime, and cost of living—it becomes clear that each city offers distinct advantages and disadvantages.

* New York stands out for its high population, excellent urban connectivity, and safety, but it comes with a significantly high cost of living.
* Chicago and Boston offer strong alternatives in terms of urban connectivity, with slightly lower living costs than New York. This makes them appealing choices for people who prioritize walking, biking, or public transit.
* Austin and Nashville may be more attractive to people looking for affordability, though both cities face challenges in terms of urban mobility and, in the case of Nashville, very low walkability and transit scores.
* Atlanta’s high crime rate may overshadow some of its other qualities, making it a less attractive option for those prioritizing safety
* The following data set (next page) summarizes all the compiled rankings and gives a final overall city rank for each of our pre-chosen cities.
* A screenshot of a graph

  Description automatically generated
* NOTE: Remember our ranking system ranks the most favorable city in each category as 10 and the least favorable as 1.
* We decided not to put any extra importance on any one category.

To find the best city to live in we added all the scores from each category and sorted the cities. One interesting note from the table above is that not one of our cities ranked in the top half for all three categories.

**Conclusion:**

New York and Tampa came out as the top ranked cities, each sharing a total city score of 21. This provides another interesting outcome. If we look back at our population numbers, we see that Tampa is the least populated city and New York is the most populated city. This answers our second main question of whether we want to live in a big or small city. This clearly shows we can be happy living in the most populated or least populated city.

**Learning experiences and Limitations:**

Throughout the duration of the first project our group came across multiple learning experiences in which we needed to troubleshoot, along with some limitations for creating, developing and implementing the project. These experiences include:

* Finding relevant data sets that match our project goal
* Cleaning data from a certain data set to allow for merging of datasets
* Merging code between partners – In the beginning as our team was working on code, some variables and data frames were given very generic call names. This made merging the code very messy and altered the data. Renaming the call variables and data frames to more category specific names solved this problem.

**Things to added or altered:**

* Subcategories could have been used for all three of our major categories. More major categories including education, weather and tax rates could have been chosen
* A larger number of initial cities could have and should have been used to give a better representation of the true correlations.