Assessing Ideal U.S. Northeast States for Living

Irith Chaturvedi

Irith2004@gmail.com

Introduction

***Motivation*:**

The Northeastern region of the USA comprises states including Pennsylvania, Maine, New York, New Jersey, Massachusetts, Connecticut, Rhode Island, New Hampshire and Vermont. They are historically and culturally important as a place of confluence of different populations. Nevertheless, people in such cities as New York continue facing these problems while trying to afford their income. Daniel shows this struggle, noting “65% of Latino households having trouble with cost of living compared to 58% of black households and 51% of Asians or native Hawaiian/ pacific islanders” (Parra, 2023).

Alternative Places in the U.S. Northeast:

Why Stay in the Northeast:

This project highlights the Northeast's historical prosperity, providing reassurance to those considering relocating, while also exploring more affordable alternatives

Focus on Affordability:

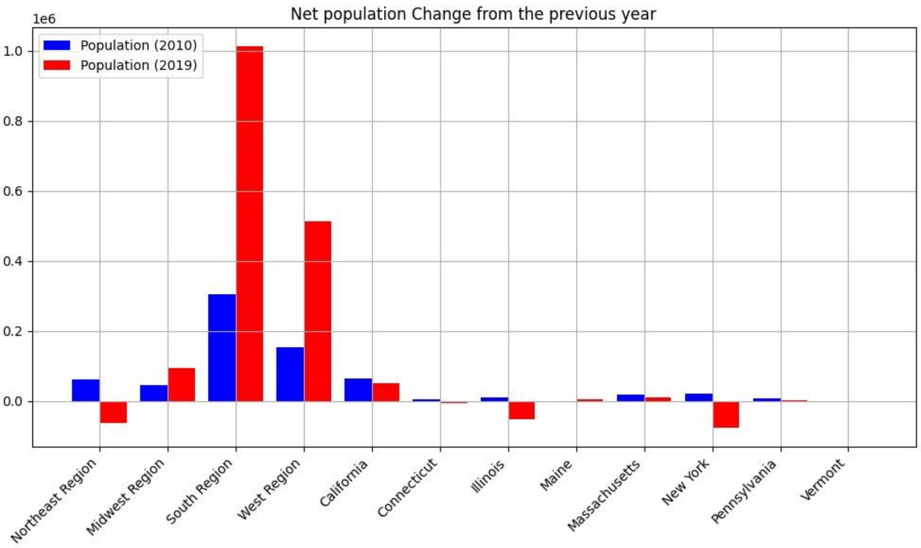
I wanted to spotlight cities that won't strain the budget, aiming to provide viable options for career development without incurring excessive costs.

Avoiding Newly Saturated Cities:

I would suggest caution against moving to popular Southern cities like Austin andDallas, suggesting consideration of sister cities for better career prospects.

Strategic Move from New York:

Encouraging individuals to make strategic decisions when leaving New York.Underscoring the importance of exploring nearby cities with growth potential to ensure a smooth transition.



In the table above, the Y-axis describes the population change in numbers, while the X-axis is the regions being compared. As we can see, the South has seen a drastic change in population even from one year to the other. In comparison, the American Northeast has been relatively unchanged in total population. So I would suggest moving to cities in the Northeast from New York because there has been less of a drastic change in migration to those locations, this also shows stability with regards to socio-economic factors.

Methods:

Data Retrieval:

I made the use of four datasets for my analysis. Two datasets, on the Per capita income per US state and the second dataset on the updated house prices of Northeastern US states were retrieved as a CSV file from Kaggle. The dataset on the Population statistics of US states, territories and regions, was downloaded from Data.gov, a government subsidiary containing millions of regularly updated datasets. The dataset used to create a pipeline machine learning model on Pittsburgh home prices was retrieved from census.data.gov.

Data Cleaning:

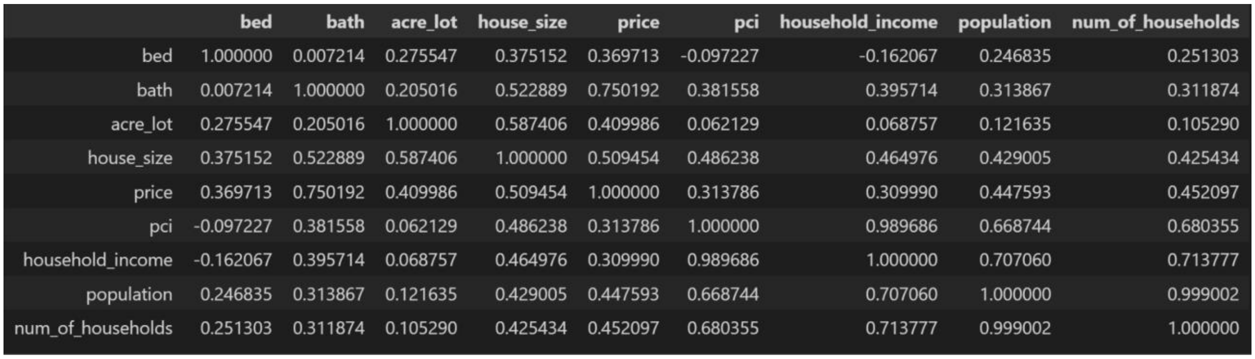
Data cleaning involved the preparation of the datasets for analysis, visualizations and machine learning. For the machine learning datasets, Only the columns of price and neighborhood were used from the Pittsburgh dataset to train and test the Logistic regression model. Mainly, all datasets had null and empty records removed, unwanted columns dropped.

Data Sorting:

Datasets on Per capita income, housing information and Population changes had values grouped together based on each US state. The numerical values were aggregated based on the mean values. The population change dataset was parsed only to contain columns having information for the years 2010 and 2019. This was specifically done to establish a significant change in data over a considerable period of time, such as a decade.

Analysis and Visualizations:

The analysis was carried out using the Python PANDAS library to manipulate dataframes created from the datasets. The Folium framework, was used to create the spatial visualizations on US states by using the coordinates of the state borders as the outline feature for the maps. The gradient scale of the map was given a red base. Matplotlib was also used to create the double bar plots. Double bar plots offer an amazing way to portray changes in the same variable based on an external variable, such as time in this case.
To choose the adequate variables for analysis, a correlation coefficient table was created, this table provides insight as to how one variable directly affects the other.

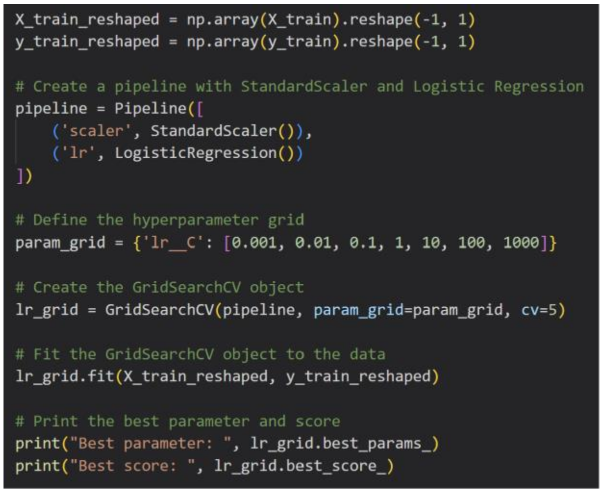


	bed	bath	acre_lot	house_size	price	pci	household_income	population	num_of_households
bed	1.000000	0.007214	0.275547	0.375152	0.369713	-0.097227	-0.162067	0.246835	0.251303
bath	0.007214	1.000000	0.205016	0.522889	0.750192	0.381558	0.395714	0.313867	0.311874
acre_lot	0.275547	0.205016	1.000000	0.587406	0.409986	0.062129	0.068757	0.121635	0.105290
house_size	0.375152	0.522889	0.587406	1.000000	0.509454	0.486238	0.464976	0.429005	0.425434
price	0.369713	0.750192	0.409986	0.509454	1.000000	0.313786	0.309990	0.447593	0.452097
pci	-0.097227	0.381558	0.062129	0.486238	0.313786	1.000000	0.989686	0.668744	0.680355
household_income	-0.162067	0.395714	0.068757	0.464976	0.309990	0.989686	1.000000	0.707060	0.713777
population	0.246835	0.313867	0.121635	0.429005	0.447593	0.668744	0.707060	1.000000	0.999002
num_of_households	0.251303	0.311874	0.105290	0.425434	0.452097	0.680355	0.713777	0.999002	1.000000

The correlation coefficient value is between -1 and 1, the higher and positive the coefficient vale, the positive and highly related the established relationship between variables.

Machine Learning Model:

A machine learning model to predict the best Pittsburgh neighborhood to live in, based on the supplied budget was created using Logistic Regression. Contrary to its name, logistic regression is a parametric model that performs a classification task based on categorical variables. To train the model, Data of the best neighborhoods and the current listings of each property were provided to the model, based on the training data, an input budget value would classify the best neighborhood to live in the city of Pittsburgh, Pennsylvania. It may be noted that my results showed why moving to a state neighboring New York or relatively close to it in the Northeast helped us settle on the state of Pennsylvania. As a precaution, a metropolitan city such as Pittsburgh is suggested as an alternative so that there is not a huge change in lifestyle while moving from New York City.



```python
X_train_reshaped = np.array(X_train).reshape(-1, 1)
y_train_reshaped = np.array(y_train).reshape(-1, 1)

# Create a pipeline with StandardScaler and Logistic Regression
pipeline = Pipeline([
    ('scaler', StandardScaler()),
    ('lr', LogisticRegression())
])

# Define the hyperparameter grid
param_grid = {'lr__C': [0.001, 0.01, 0.1, 1, 10, 100, 1000]}

# Create the GridSearchCV object
lr_grid = GridSearchCV(pipeline, param_grid=param_grid, cv=5)

# Fit the GridSearchCV object to the data
lr_grid.fit(X_train_reshaped, y_train_reshaped)

# Print the best parameter and score
print("Best parameter: ", lr_grid.best_params_)
print("Best score: ", lr_grid.best_score_)
```

Data Reshaping:

- X_train_reshaped and y_train_reshaped are created by reshaping the training data (X_train and y_train) using np.array and reshape.
- This converts the data into NumPy arrays and reshapes them to one dimension (-1) with a single column (1). This is done to prepare the data for the pipeline.

Pipeline Creation:

- StandardScaler(): This step standardizes the features in the data by removing the mean and scaling to unit variance.
- LogisticRegression(): This step trains a logistic regression model on the data.

Hyperparameter Tuning:

- A dictionary called param_grid is defined.
- This dictionary contains a single key, 'lr_C', with a list of values to try for the C parameter of the LogisticRegression model. These values range from 0.001 to 1000.

GridSearchCV:

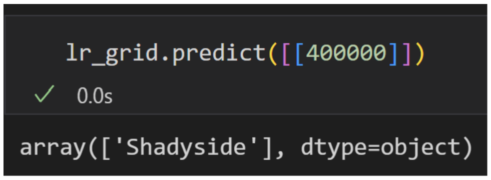
- This object takes the pipeline and the parameter grid as arguments.
- It will then train the logistic regression model with each combination of parameter values in the grid and choose the one that performs the best on a held-out set of data using 5-fold cross-validation in this case.

Model Fitting and Evaluation:

- The GridSearchCV object is fitted to the reshaped training data.
- This trains the logistic regression model with all possible combinations of hyperparameter values from the grid.

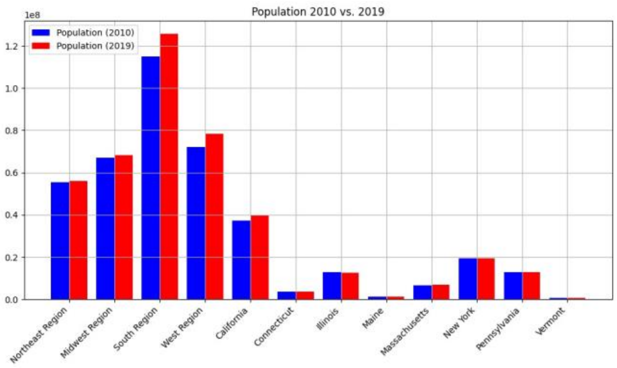
Results:

When a budget value in USD is supplied to the model for a prediction, a Pittsburgh neighborhood that best suits the monetary value is returned.

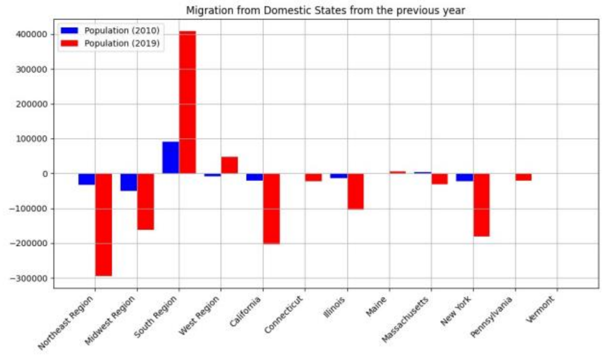


```
  lr_grid.predict([[400000]])
✓ 0.0s
array(['Shadyside'], dtype=object)
```

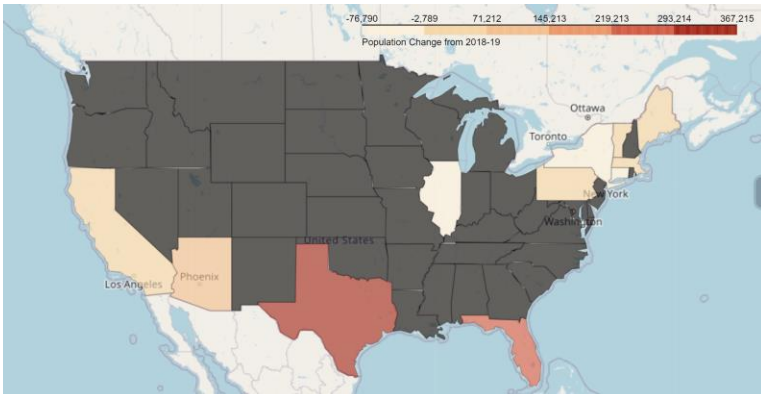
Results



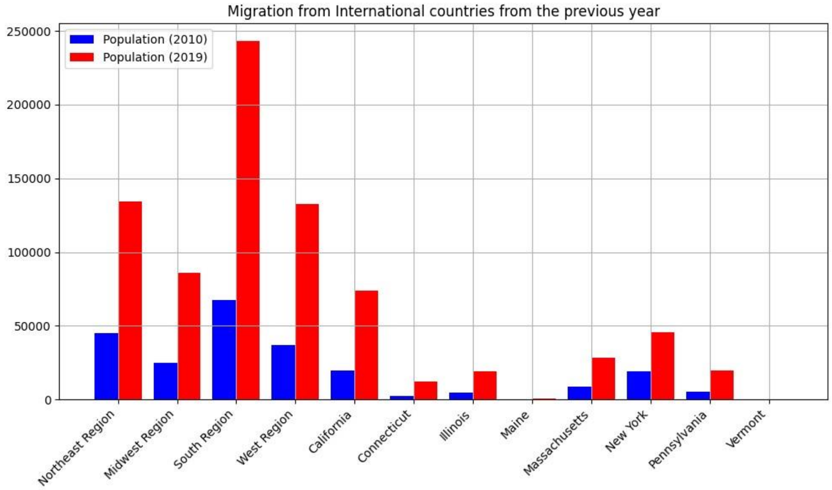
In terms of my results, I have found there to be an increase in population in regions where the cost of living is reasonable. For example, the southern region of the USA has experienced a greater increase in population between 2010 and 2019 as compared to New York, as the cost of living in the south is more reasonable than New York. Such trends serve as evidence that a more reasonable cost of living can be tied to an increase in population.



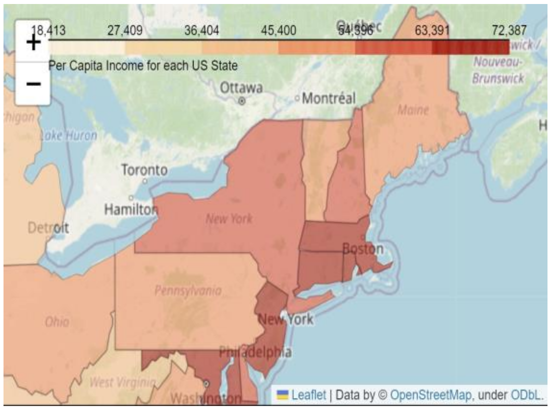
As the cost of living increases in New York, we can see that less people from other states are inclined to move to New York. However, the cost of living stays relatively reasonable in areas such as the south, which leads to higher populations of people migrating to the southern region from other states. This further solidifies my statement regarding the cost of living being tied to changes in population, with migration considered.

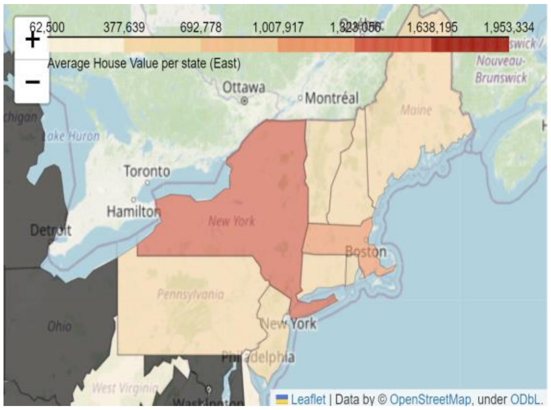


To further demonstrate our claim, we can see that the highlighted states in the visualization above are all known to have a decent job market, yet it's the states with a reasonable cost of living that have an increase in population, whereas the states with the higher cost of living have experienced a decrease in population from 2010 to 2019.



In terms of international migration, areas with a steady job market are more favorable, as well as areas with a reasonable cost of living. As a result, Northeastern regions (excluding New York due to the higher cost of living) and southern regions are more favorable for people migrating to the US from international countries.





The reason why Northeastern states aside from New York are favorable to move to can also be attributed to the fact that the average income is sufficient when compared to the average cost of living which is fairly reasonable.

Conclusion

Given that I mainly focused on "open availability" related to employment opportunities and housing during my data analysis, it is important to explain this before I present my two main recommendations. I identify the US Northeast as a dominant economy and target areas with steadily expanding job markets but stagnant populations. They are based on the recognition that in these places, the economic conditions for their business are favorable.

In addition, my suggestions focus on relatively cheaper housing regions. The decision has been geared in offering an appropriate solution to people planning to leave New York city. In directing people towards affordable housing areas, I want to focus on cities that will still be economically efficient for an individual despite being cheaper than NYC (New York City).

Pittsburgh, Pennsylvania, and Hartford, Connecticut, serve as the target cities in this study. This includes Pittsburgh, where a house goes for an average price of $160k and thus becomes a great place to move to from NYC as one can still make savings. The city also has a respectable culturally diverse population with the biggest minority category is black or African American, making up approximately 22% while Asians are at 5.5%. On average house value, the mean salary for white-collar/corporate jobs coincides while amounting to $64k.

As the analysis progresses, I arrived at Hartford, Connecticut, where the average house price is $190K—significantly more affordable than New York City. Hartford boasts a diverse demographic, with 34% of its population identifying as Black or African American and around 20% as Hispanic. Additionally, salaries for white-collar jobs remain competitive. Both Pittsburgh and Hartford offer vibrant cultural scenes, rich diversity, and a favorable balance between housing costs and salaries, making them excellent alternatives for those moving from New York City.

References

City Limits. (2023, April 26). More Than Half of Immigrant New Yorkers Can't Afford Basic Needs, Report Finds. Retrieved from https://citylimits.org/2023/04/26/more-than-half-of-immigrant-new-yorkers-cant-afford-basic-need s-report-finds/

Kaggle. (n.d.). USA Real Estate Dataset. Retrieved from https://www.kaggle.com/datasets/ahmedshahriarsakib/usa-real-estate-dataset

Kaggle. (n.d.). United States Counties by Per Capita Income. Retrieved from https://www.kaggle.com/datasets/kabhishm/united-states-counties-by-per-capita-income

U.S. Census Bureau. (n.d.). Population Estimates Program: 2010s State Total. Retrieved from https://www.census.gov/data/datasets/time-series/demo/popest/2010s-state-total.html

iStockphoto. (n.d.). [Link to the main page of iStockphoto]. Retrieved from https://www.istockphoto.com/photos/

Data USA. (n.d.). Pittsburgh, PA. Retrieved from https://datausa.io/profile/geo/pittsburgh-pa/#:~:text=The%205%20largest%20ethnic%20groups,(Hispanic)%20(1.61%25).

Data USA. (n.d.). Hartford, CT. Retrieved from https://datausa.io/profile/geo/hartford-ct#:~:text=The%205%20largest%20ethnic%20groups,(His panic)%20(9.41%25).

Redfin. (n.d.). Shadyside Housing Market, Pittsburgh, PA. Retrieved from https://www.redfin.com/neighborhood/156434/PA/Pittsburgh/Shadyside/housing-market

Data.gov. (n.d.). Property Data with Geographic Identifiers. Retrieved from https://catalog.data.gov/dataset/property-data-with-geographic-identifiers/resource/ed31b5da-6c4c-48ba-bcf8-718777778ba9