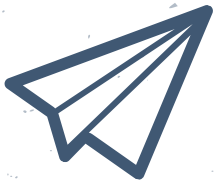


CITY RECOMMENDER

WHERE SHOULD I LIVE IN THE NEW WORK-FROM-HOME ERA?

Evanya Wilson

Data Science Capstone, 2021

PROBLEM STATEMENT

How can we compare cities by certain attributes, and group them by their similarities?

BACKGROUND

Living through the 2020 pandemic, it has become quite evident that we can easily work remotely, and no longer need to go into the office. Many companies are now opting for completely remote work, or a hybrid model of some sort. By now, most of us have probably considered the idea of working from a new place, some even long term.

With this project, I would like to encourage people to take advantage of moving to a new city, allowing them to consider a new place to live feasibly based on their current living expenditures and environment. Ultimately, I will be recommending another city that has similar attributes to their current city.

DATA

The data consists of three tables pulled from numbeo.com. Numbeo.com provides cost of living and quality of life information on different cities across the world. Most of the features are indexed based on New York City (NYC) standards. NYC being 100. So a value of 120 would be 20% more than NYC cost, and a value of 80 would be 20% less than NYC cost. The data also accounts for the currency exchange rate.

The Cost of Living table has 8 variables and 573 rows. The Quality of Life table has 11 variables and 243 rows. The Property Prices table has 9 variables and 486 rows. The data dictionary in the supplemental notebook provides information on the tables in more detail.

PRE-PROCESSING AND EXPLORATORY DATA ANALYSES

When joining the three tables on the 'City' column, I ended up with a total of 230 rows, as not all the tables had the same cities within. Although a considerable number of rows were discarded when merging, I felt this was the best option as I would not have been able to impute information accurately. After merging, there were no missing values in the data. After dropping the duplicate Quality of Life column, the final dataset had 22 columns. I ran a summary to view some statistical information on the data in order to get a better understanding and compare the scale of each feature. Many of the features ranged from single digits to triple digits.

A correlation heatmap allowed me to easily see which features correlated to others. It was found that purchasing power, restaurant and grocery costs are highly correlated with cost of living . This make sense as each would contribute to cost of living expenses. Healthcare quality is positively correlated with quality of life, and pollution is highly negatively correlated with quality of life.

Viewing distributions of each numerical variable also allowed me gain better understanding. The distributions varied, being right skewed, left skewed and also normally distributed. It was determined appropriate to scale the data before moving onto modeling.

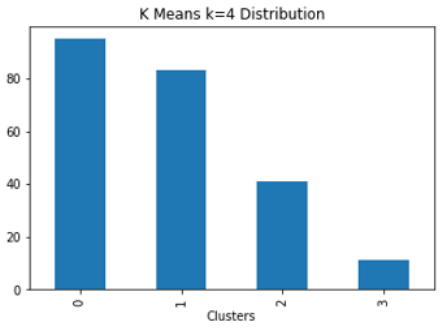
MODELING

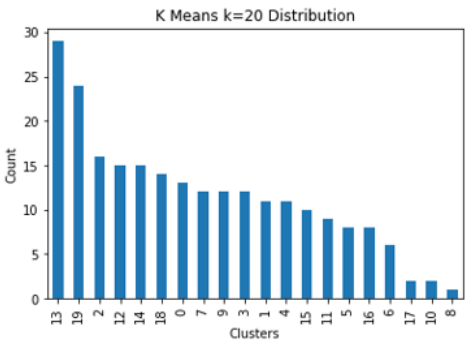
When considering a recommender system, there are typically three main types that come to mind; content based, user-item, and item-item. The nature of my data doesn't allow for any of the typical recommender system algorithms, and so, I explored a number of unsupervised learning models for this project. I ran clustering algorithms to determine similarities between cities, and group them in order to recommend similar options to users. Some of the clustering models I explored were the K-Means Model, DBSCAN, Gaussian Mixture Model, and Hierarchical Clustering. I applied dimensionality reduction using a Principal Component Analysis (PCA) and t-SNE, which allowed me to visualize the models. I also computed other distance models like Cosine Similarity and Euclidean Distances. I will discuss each in more detail.

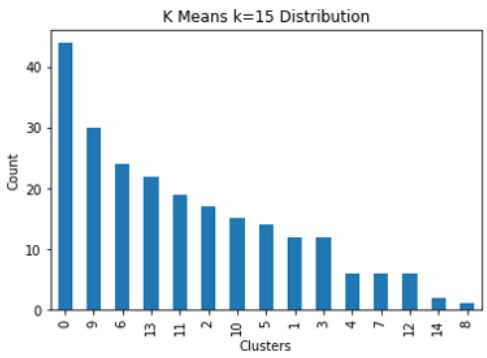
K-Means Model

Looking at the silhouette score and the inertia, it was determined that 4 was the ideal number to group the cities into. I decided to run the K-Means model with this recommendation. When looking at the 4 separate clusters, it was interesting to see how they were grouped. The majority of Cluster 0 consisted of cities in Eastern Europe, while Cluster 1 was mostly cities in the USA. Cluster 2 was the smallest and had cities in East Asia. The final cluster was quite mixed.

Now, considering the recommendation system, I felt it more appropriate to increase the clusters as not to overwhelm users with too many options when recommending similar cities. I tried k=20, but moved forward with 15 groups for the recommendation system as I noticed that most clusters with k=20 had cities from the same region. While this makes sense, I reduced the number of clusters a little in hopes to capture cities from different regions within the same group. This could make the suggestions a little more exciting for the user. Interestingly, the smaller groups were similar for both k=4 and k=15, so it seemed those cities were distinctly different from the rest.

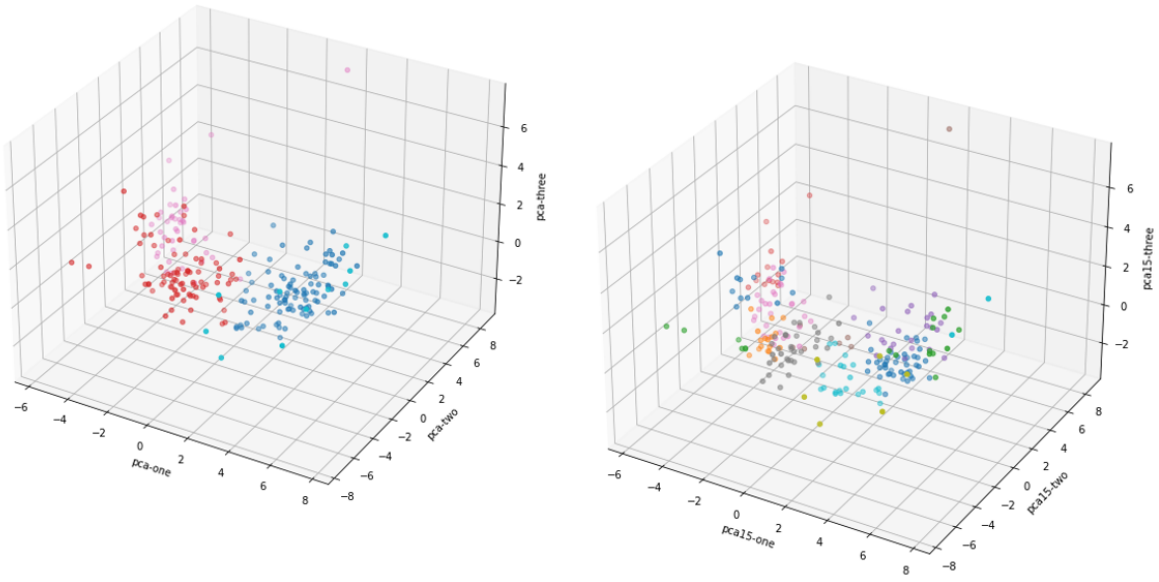






Visualization with Dimensionality Reduction
PCA & t-SNE

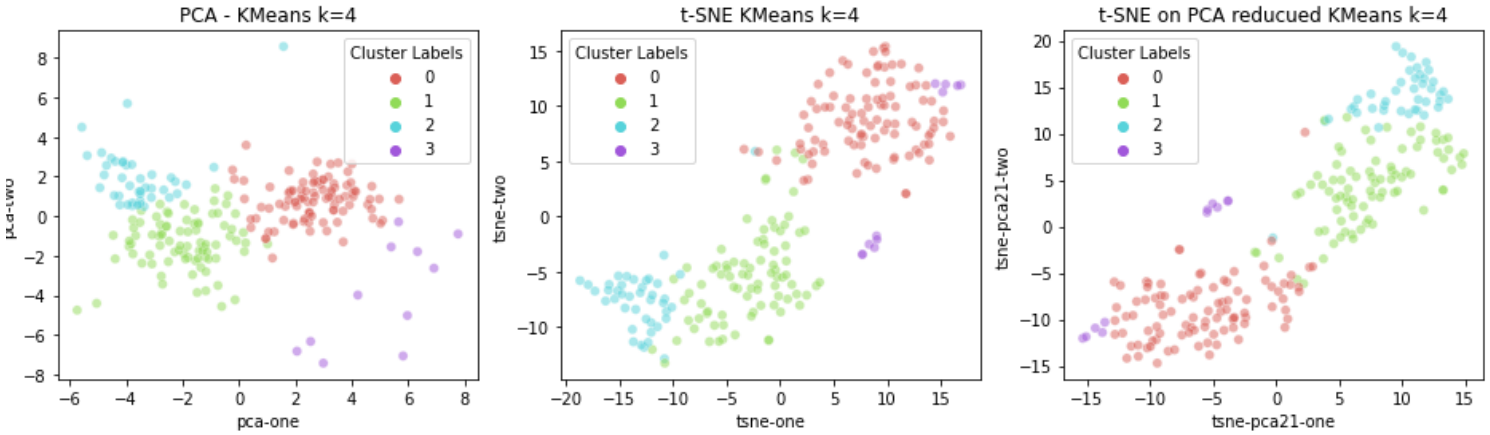
In order to visualize the K-Means clusters, I applied a PCA to the already scaled data. The first two principal components account for 65% of the total variance in the data. The first three components account for 75% of the variance in the data. I was able to visualize the clusters on both 2D and 3D graphs.



I also applied another dimensionality reduction method, t-SNE, which is known to have better visualization results.

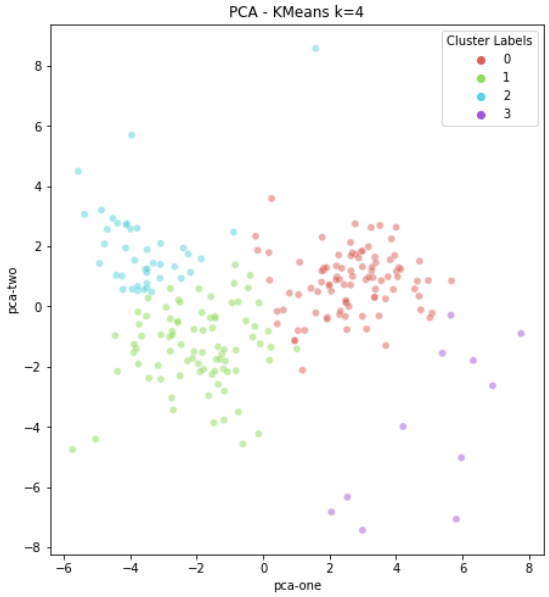
t-SNE on PCA Reduced Data

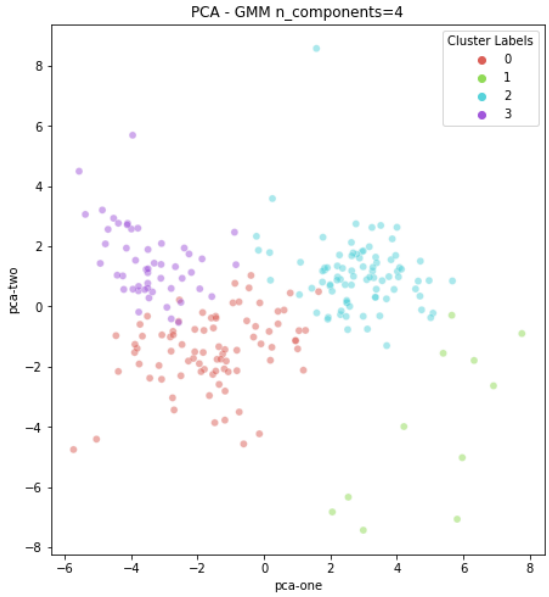
To go one step further, I applied t-SNE to PCA reduced data to see what I could find. I used all features as components in the PCA, then reduced once more with t-SNE. The visuals were a quite different.



Other Unsupervised Clustering Models - DBSCAN, Gaussian Mixture Model, Hierarchical Clustering

When attempting DBSCAN and Hierarchical Clustering, I found that they were not as successful in separating the data into groups. When specifying the number of clusters, the models would put upwards of 95% of the data into one cluster. I deemed these inappropriate models and continued with the Gaussian Mixture Model. With 4 clusters, the Gaussian model was almost visually identical to the K-Means model. It also identically labelled the cities.





Cosine Similarity & Euclidean Distance

For checks and balances purposes, I calculated Cosine Similarity and Euclidean Distance on my data. Entering specific cities allowed me to see the top 20 most similar cities. Many of these were found within their K-Means and Gaussian clusters. I felt Euclidean Distance was of best use out of the two, as it calculates a closer distance between points.

FINDINGS & CONCLUSION

After analyzing this data in depth, it has been made quite clear which cities are very similar, to be considered an alternative living destination based on one's current salary and expenses. I was able to compare each of the clustering models to see which cities they deemed similar, and so grouped. In comparing my results to my initial goals, I am satisfied with the outcomes. Through these clustering models, I was able to provide recommendations of similar cities as I had envisioned.

Clustering performance evaluation requires knowledge of the ground truth classes, which is almost never available in practice, or requires manual assignment by human annotators. When applying unsupervised methods, rather than optimize against a particular metric like accuracy, the goal is to instead look for consistency and agreement across the algorithms, suggesting the underlying patterns being found in the data are real. I found this consistency in the K-Means and Gaussian models, as well as the Euclidean Distance calculation. Another evaluation metric to consider is maximizing the silhouette score. I made sure to optimize the model appropriately based on this.

Limitations

Limitations to my project include lack of world cities in the dataset. My data only has 230 cities in the world. I would love to include more cities if information on them becomes available. Also, the data used in this project is ever-changing. The data used is from the current year, 2021. Many events can and will contribute to significant changes in costs and quality of life in each city over time. In order to keep similarities as true as possible, the data would need to be continuously updated.

NEXT STEPS

Moving forward, I would like to determine which features contribute most to the grouping of each cluster. In trying to decide the number of clusters for the recommender system, I would consider consensus clustering which will provide results based on multiple k's.