DSC412

Milestone 4: Code and Final Report

Minyoung Suh

Abstract

The objective of the final project is to delegate the choice of travel destination to a machine learning model based on publicly available data. Among the global city indices available from multiple resources, 4 different criteria were adopted related to travel decisions, which were safety, cost, pollution, and hygiene. K-means clustering model was selected as an unsupervised approach since there was no definite sets of data to be imported for outputs. The selected K-means clustering model generated 5 clusters. Being the safest, reasonably affordable, and acceptably clean, one of the clusters was identified to be suitable for travel, and it included Madrid, Seoul, Tokyo, and Taipei.

Background

The tourism industry recorded the worst year in 2020 due to the covid-19 pandemic and relevant travel restrictions. Showing excellent resilience, however, global tourism has had consistent gradual recovery in the past few years and is expected to surpass pre-covid levels of prosperity by the end of 2024 [1]. The recovery has been brought by a significant increase in tourist demand worldwide together with more available flights, better international openness, and increased interest and investment in natural and cultural attractions. Despite post-pandemic growth, however, continued challenges have been addressed by the World Economic Forum that the recovery varies by region [2].

Risk and uncertainty play an important role in the choice of travel destinations. Risk is defined as the possibility to experience certain (negative) events, while uncertainty is associated with incomplete knowledge during the decision-making process. The risk and uncertainty come from the fact that it is not possible for tourists to predict or anticipate the situation at a destination before traveling. Therefore, they cannot help but rely on information from external sources, such as acquaintances, social media, or travel organizations. The choice of destinations is highly affected by the socio-demographic characteristics of tourists. High educational levels and high travel frequencies are distinct characteristics of risk-affine tourists, while higher age groups are more dominant in risk and uncertainty-averse tourist types [3].

The objective of the final project is to delegate the choice of destination to a machine learning model based on publicly available data. Global city indices are accessible from multiple resources such as OECD, UN, Gallup, Numbeo, and World Population Review. As prime determinants related to travel decisions, 4 different criteria were identified, which were safety [4], living cost [5], pollution [6], and hygiene [7]. Each index is detailed in Table 1.

Table 1. Global city indices adopted as inputs

Criteria	City Index	Description	# of Cities
Safety	Crime Index	Calculated based on the crime rate per 1,000 population for all crimes in a specific neighborhood or city	311
	Safe Cities Index	Measuring a city's safety across five areas: personal, infrastructure, health, digital, and environmental security	
Living Cost	Living Cost Plus Rent Index	Comparing the overall living cost in a particular location, including both the price of essential goods and services (like food, transportation, utilities) and the cost of rent, allowing for a comprehensive comparison of how expensive it is to live in one place compared to another.	218
	Groceries Index	Tracking the average price of a basket of common grocery items over time, as a gauge for the cost of groceries in a given location compared to others	
	Restaurant Price Index	Tracking and comparing menu prices for restaurants	
Pollution	Pollution Index	Measuring the level of air pollution and the associated health concerns in terms of 6 major air pollutants: particle pollution, ground-level ozone, carbon monoxide, sulfur dioxide, and nitrogen dioxide	249
Hygiene	Clean City Score	a numerical ranking assigned to a city based on its environmental performance, usually considering factors like air quality, waste management, public transportation options, energy efficiency policies, and overall cleanliness	50

Model Selection

K-means clustering model was selected as an unsupervised approach since there was no definite sets of data to be imported for outputs. K-means clustering method is expected to group the cities into several clusters based on their condition of safety, living cost, pollution, and hygiene.

Data Analysis and Augmentation

Since the information on every city in the world is not available from the resources, the list of highly ranked cities was imported for each index and as a result, each list includes a different subset of random cities in different numbers as indicated in **Table 1**. Compared to other indices, the hygiene index was accessible with a considerably smaller number of cities. This would restrict the length of the complied list, which could lead to incompetent outcomes at the end. Taking this into consideration, the models need to be established in 2 different ways: with and without the hygiene index considered.

After compiling the city indices, only 30 cities remained on the list when the hygiene index was included, but the compiled list ended up having 178 cities when the hygiene index was excluded. The average indices were estimated slightly differently depending on the number of cities on the list (**Table 2**). The data distribution also showed marginal distinction between 2 complied lists of cities (**Figures 1 and 2**).

Table 2. Average and standard deviation of city indices

	Crime	Safety	Living Cost	Groceries	Restaurant	Pollution	Hygiene
30 cities	44.51 (±15.4)	55.49 (±15.4)	47.71 (±18.9)	60.98 (±19.4)	54.68 (±20.5)	49.30 (±14.6)	63.76 (±9.1)
178 cities	42.31 (±15.1)	57.68 (±15.1)	38.86 (±18.1)	51.46 (±20.8)	46.32 (±21.8)	52.14 (±21.4)	N/A (N/A)

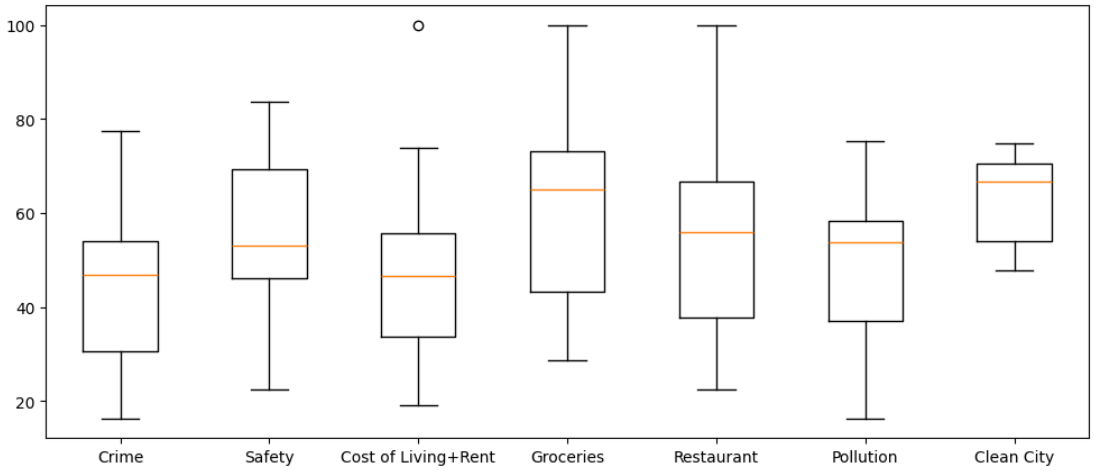

Figure 1. Summary of city indices from 30 cities

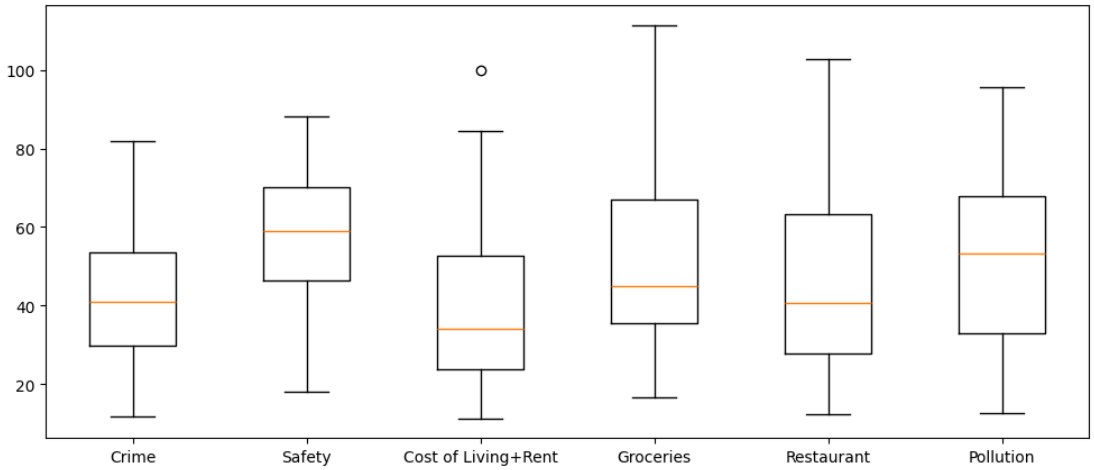

Figure 2. Summary of city indices from 178 cities

Results

The sum of square errors was estimated when the hygiene index was included and excluded, respectively, to choose the optimal number of clusters. Based on these, 5 and 7 clusters were selected accordingly (**Figure 3**).

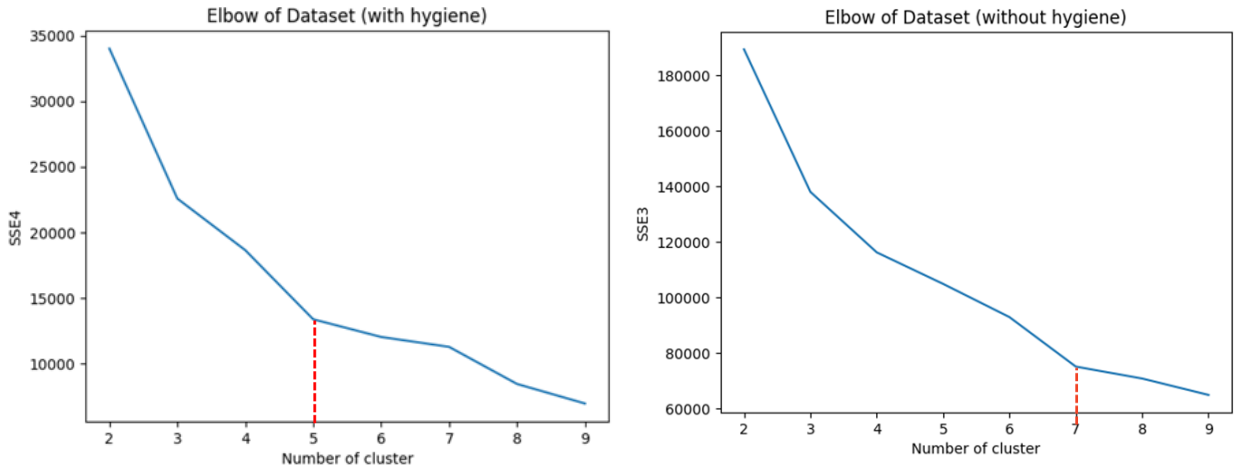

Figure 3. Sum of square error with (left) and without (right) hygiene index considered

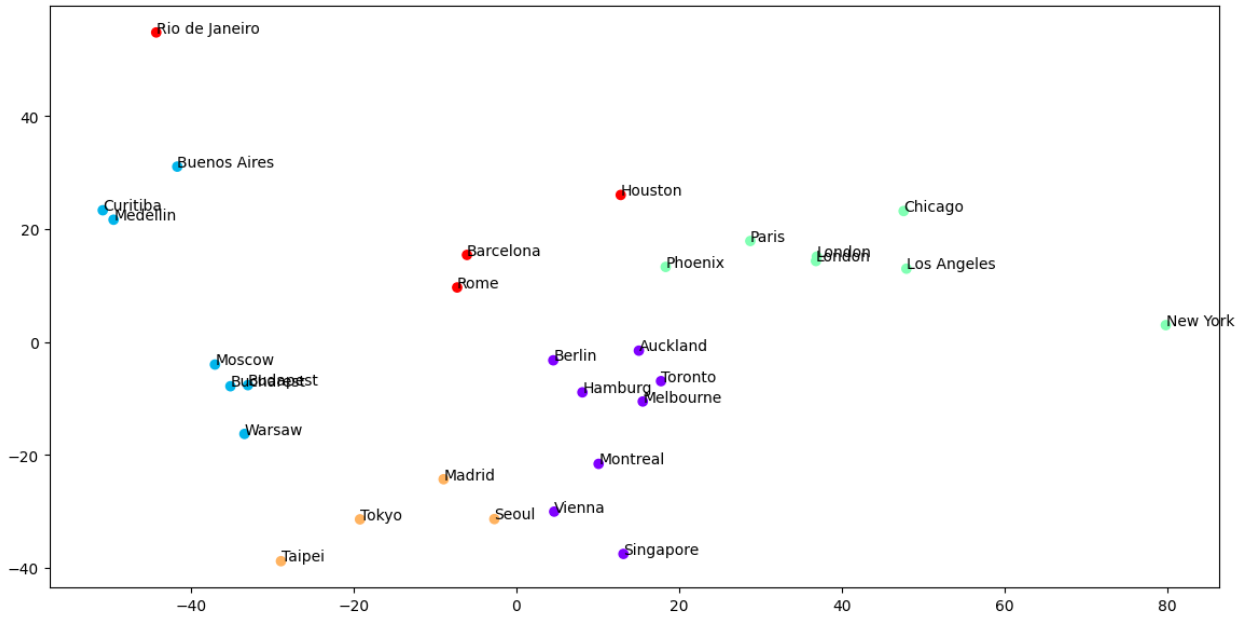

Figure 4. 30 cities sorted into 5 clusters with hygiene considered

Through principal component analysis, the clustering results are visualized in **Figure 4** and the average indices of each cluster are shown in **Table 3**. Based on the average indices, each cluster

was characterized as follows. Cluster 0 includes cities that are safe and clean, but relatively high costs are expected. Cities in Cluster 1 look most affordable but with a low level of hygiene. In contrast, cities in Cluster 2 are cleanest but most expensive. Cluster 3 showed the highest safety, while keeping relatively low costs. Cities in Cluster 4 are also affordable but safety is in question. Having 4 to 8 cities assigned to each cluster, the model outcomes were manageable and intuitive to understand.

Table 3. Average city indices of each cluster when hygiene was considered

	Crime	Safety	Living Cost	Groceries	Restaurant	Pollution	Clean City	Count
0	38.69	61.31	52.69	69.39	60.83	30.86	67.73	8
1	43.20	56.80	26.60	35.44	34.04	55.70	52.81	7
2	55.69	44.31	71.59	79.51	81.80	59.16	70.75	7
3	22.93	77.08	40.85	66.95	36.85	46.38	63.61	4
4	60.45	39.55	39.78	50.43	48.88	60.65	62.88	4

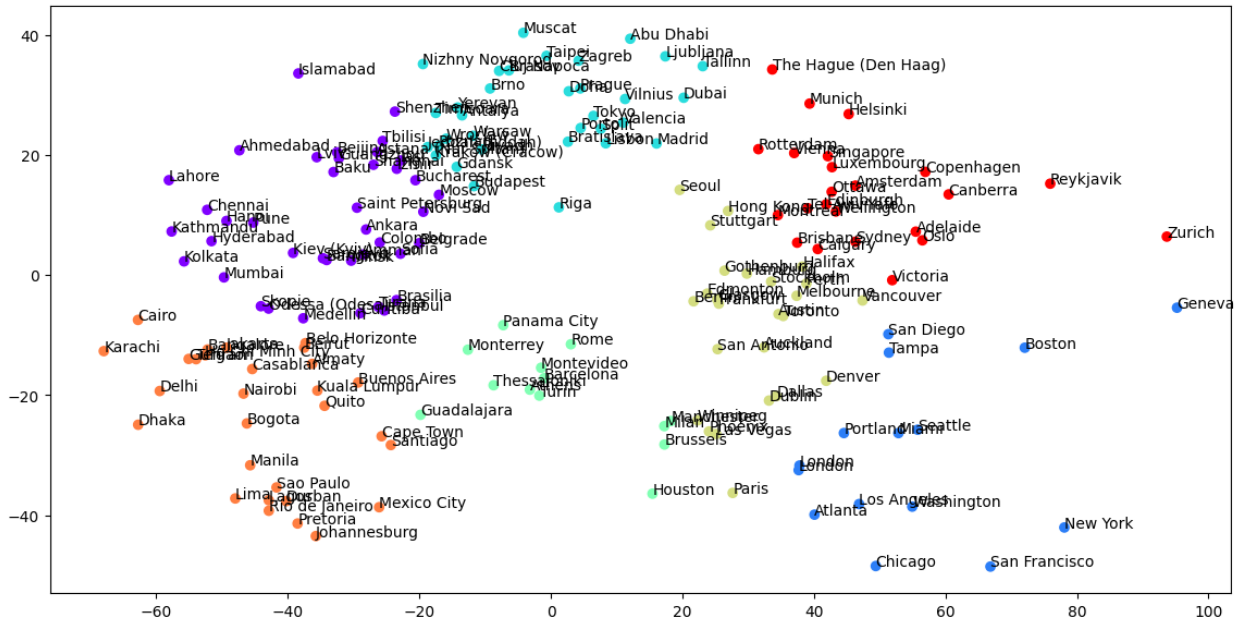

Figure 5. 178 cities sorted into 7 clusters without hygiene considered

In a similar manner, clustering results and their characteristics are summarized in **Figure 5** and **Table 4** when the hygiene index was not considered. According to the average indices, Cluster 0 is cheapest, Cluster 1 is costliest, Cluster 2 is safest, Cluster 5 is dirtiest, and Cluster 6 is cleanest.

However, this method leaves Clusters 3 and 4 in the grey area without establishing any distinctive identities to characterize those groups. As there are so many cities belonging to each cluster, as shown in Table 4, it also introduces additional dimensions of decision making within a selected cluster.

Table 4. Average city indices of each cluster when hygiene was not considered

	Crime	Safety	Living Cost	Groceries	Restaurant	Pollution	Count
0	39.39	60.61	22.36	33.20	25.75	66.69	40
1	52.27	47.73	74.11	86.40	82.61	44.03	15
2	25.53	74.47	34.05	43.21	41.51	43.52	33
3	53.99	46.01	40.88	55.68	54.08	60.55	13
4	45.12	54.88	52.85	70.32	61.92	38.51	25
5	63.01	36.99	22.44	32.47	25.18	76.08	29
6	29.26	70.74	55.85	73.35	70.63	24.36	23

The 2 different cluster models showed a certain degree of agreement with each other. For example, having the highest safety with reasonable affordability and acceptable cleanness, the most attractive cities seem to be Clusters 3 when hygiene was considered and Cluster 2 when hygiene was not considered. Cluster 3 has only 4 cities that are Madrid, Seoul, Tokyo, and Taipei. Out of these 4, 3 cities are included in Cluster 2, which are Madrid, Tokyo, and Taipei.

Conclusion

Having unsupervised models established, there are no numerical measures that can evaluate the success of models. However, K-means clustering has narrowed a long list of cities down to a manageable number of groups so that tourists can screen the cities based on the weighted values of their own. Since the final decision needs to be convinced and confirmed by humans, a good lesson was learned that having too many options might not be always helpful in decision making.

Future Work

Climatic conditions, such as mild temperature and low probability of rainfall, were thought to be a part of input data as an important determinant at the beginning. However, since the weather information significantly varies depending on the month and season of the year, the modeling seemed to go beyond the capability of K-means clustering. It has been left for future work after achieving advanced levels of machine learning techniques.

Citations

[1] F. Richter. “International Tourism to Surpass Pre-Pandemic Levels in 2024”, statista.com. Accessed: November 26, 2024. [Online.] Available: https://www.statista.com/chart/21793/international-tourist-arrivals-worldwide/

[2] World Economic Forum. “Travel & Tourism Development Index 2024”, weforum.org. Accessed: November 26, 2024. [Online.] Available: https://www.weforum.org/publications/travel-tourism-development-index-2024/

[3] M. Karl. “Risk and Uncertainty in Travel Decision-Making: Tourist and Destination Perspective”, *Journal of Travel Research*, vol. 57, no. 1, pp. 129-146, 2018, doi: https://doi.org/10.1177/0047287516678337

[4] Numbeo. “Crime Index by City 2024 Mid-Year”, numbeo.com. Accessed: October 31, 2024. [Online.] Available: https://www.numbeo.com/crime/rankings.jsp

[5] Numbeo. “Cost of Living Index by City 2024 Mid-Year”, numbeo.com. Accessed: October 31, 2024. [Online.] Available: https://www.numbeo.com/cost-of-living/rankings.jsp

[6] Numbeo. “Pollution Index by City 2024 Mid-Year”, numbeo.com. Accessed: October 31, 2024. [Online.] Available: https://www.numbeo.com/pollution/rankings.jsp

[7] World Population Review. “Cleanest Cities in the World 2024”, worldpopulationreview.com. Accessed: October 31, 2024. [Online.] Available: https://worldpopulationreview.com/world-city-rankings/cleanest-cities-in-the-world