COST OF LIVING

Sukhmani Mangat
Keir Martin
Dina Mattei
Nisha Mistry

INTRO AND PROBLEM STATEMENT

With success in business goals and initiatives, organizations may choose to expand their locations to accommodate new branches and more employees. The expansion of a company may require thorough analysis of the proposed new location and how it will affect the potential new employees. Specifically, the company may be interested in analyzing the cost of living to scope out where the optimal location for a new office may be. The following Kaggle dataset entitled “Cost of Living by City 2022” includes data from various cities throughout the world. The dataset includes attributes related to the cost of living, groceries, restaurants, and local purchasing power. The company can utilize this information to see which city is the most suitable for their new location. The problem statement focuses on investigating which U.S city is ideal for a company to open a new location for its employees. It encompasses the analysis of the dataset’s attributes to target the top cities for the company’s expansion. Criteria for optimal cities will be determined by a low cost of living, grocery, and restaurant index. Employees are more likely to live in a city that is affordable in those attributes. Similarly, employers would more likely want to establish a new location that is more profitable for their organization.

DATA SOURCES AND DATA PREPARATION

Cost.of.Living.2022 × | Untitled1* ×

Filter

	Rank	City	Cost.of.Living.Index	Rent.Index	Cost.of.Living.Plus.Rent.Index	Groceries.Index	Restaurant.Price.Index	Local.Purchasing.Power.Index
95	NA	El Paso, TX, United States	55.92	23.17	40.56	54.45	48.18	118.77
94	NA	Wichita, KS, United States	58.92	24.26	42.67	53.08	57.42	119.24
93	NA	Little Rock, AR, United States	59.26	25.60	43.48	57.28	64.63	131.07
92	NA	Akron, OH, United States	62.20	22.90	43.78	63.55	55.56	102.89
80	NA	Toledo, OH, United States	64.99	22.91	45.26	64.78	63.45	90.45
90	NA	Tulsa, OK, United States	62.35	28.92	46.68	58.89	60.48	132.05
88	NA	Lexington, KY, United States	62.92	29.27	47.15	69.21	55.65	102.40
84	NA	Tucson, AZ, United States	64.34	27.66	47.15	61.91	65.88	83.69
70	NA	Dayton, OH, United States	68.03	27.92	49.23	66.09	59.01	79.21
86	NA	Mesa, AZ, United States	63.66	32.99	49.28	63.20	72.07	103.42
82	NA	Oklahoma City, OK, United States	64.94	31.83	49.42	69.97	57.78	127.95
91	NA	Memphis, TN, United States	62.29	34.91	49.45	58.08	69.54	109.95
87	NA	Albuquerque, NM, United States	63.44	33.91	49.60	64.60	64.07	122.44
76	NA	Saint Louis, MO, United States	66.83	32.58	50.78	67.03	66.47	123.20
78	NA	Kansas City, MO, United States	66.07	33.50	50.80	60.66	75.16	127.20
62	NA	Des Moines, IA, United States	69.33	30.03	50.91	72.40	65.94	108.35
73	NA	Louisville, KY, United States	67.71	32.24	51.09	70.16	69.26	109.65
89	NA	San Antonio, TX, United States	62.59	38.27	51.19	57.25	69.45	137.67
47	NA	Rochester, NY, United States	71.57	29.13	51.68	71.98	69.13	100.12
79	NA	Chattanooga, TN, United States	65.69	37.25	52.36	69.00	60.43	91.70
63	NA	Fresno, CA, United States	68.97	34.02	52.59	64.72	72.91	116.40
75	NA	Indianapolis, IN, United States	67.14	36.12	52.60	67.62	67.64	121.53
68	NA	Spokane, WA, United States	68.11	35.09	52.63	64.88	70.34	112.36
83	NA	Cincinnati, OH, United States	64.39	39.74	52.83	64.25	63.76	125.17
45	NA	Syracuse, NY, United States	72.53	31.30	53.20	79.21	65.02	78.44
34	NA	Buffalo, NY, United States	74.29	29.56	53.32	72.40	73.62	124.36

The dataset shows the cost-of-living index by city from 2022. It includes 578 records of United States and internal cities. The variables presented include rank, city, cost of living index, rent index, cost of living index plus rent index, groceries index, restaurant price index, and local purchasing power.

Link to Kaggle dataset:

https://www.kaggle.com/datasets/kkhandekar/cost-of-living-index-by-city-2022

DATA EXPLORATION, VISUALIZATION, CLEANSING, AND TRANSFORMATION

Initial data preparation entailed filtering for only U.S cities. International cities were removed from our evaluation as we wanted to focus on U.S cities only. The variable Rank was removed from our dataset as the values were labeled as null.

```r
gproj <- Cost.of.Living.2022
head (gproj, 5)
```

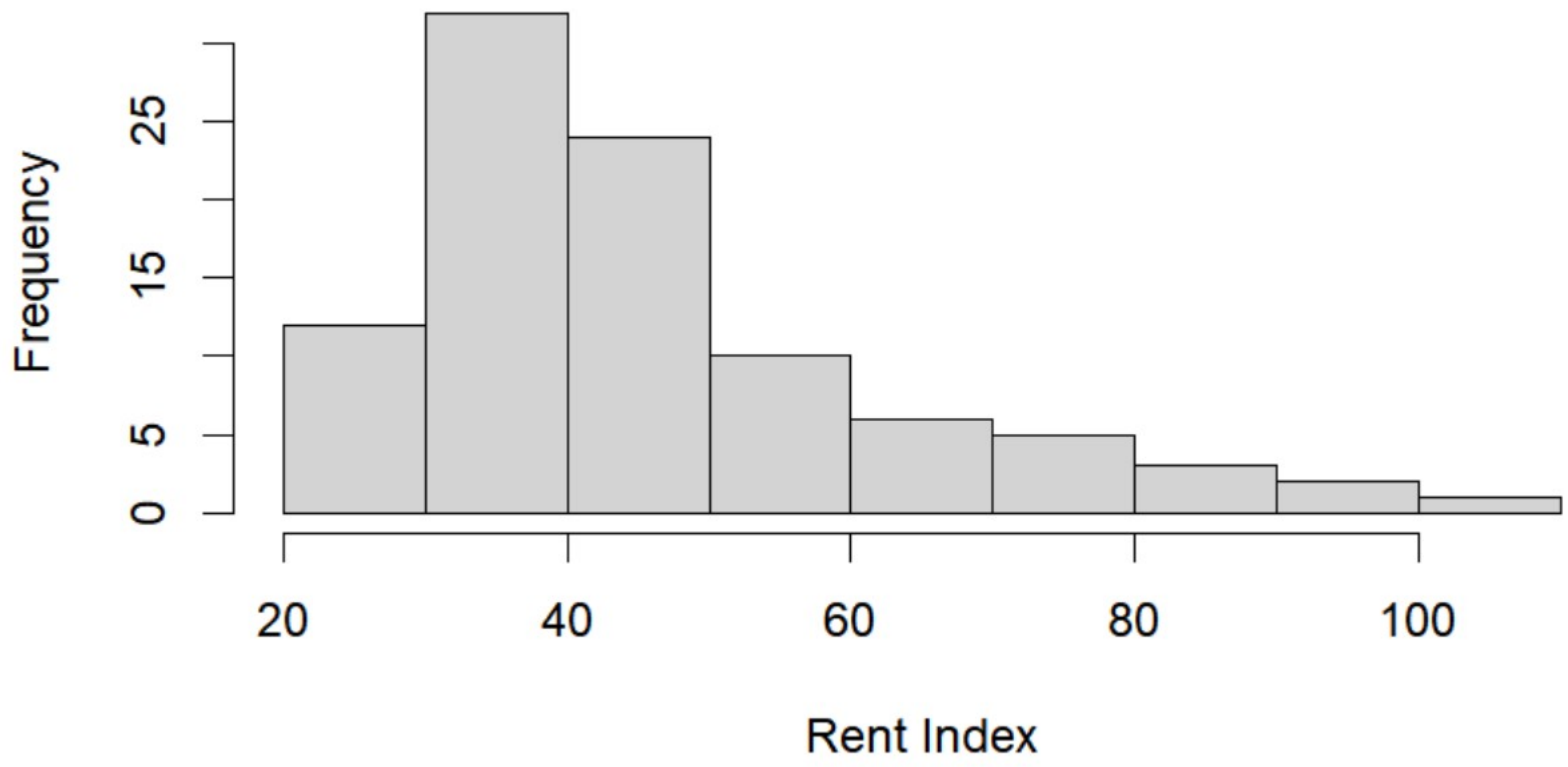
	City <chr>	Cost.of.Living.Index <dbl>	Rent.Index <dbl>	Cost.of.Living.Plus.Rent.Index <dbl>
1	Honolulu, HI, United States	103.65	65.07	85.56
2	New York, NY, United States	100.00	100.00	100.00
3	Santa Barbara, CA, United States	95.01	78.42	87.23
4	Berkeley, CA, United States	94.36	88.22	91.48
5	San Francisco, CA, United States	93.91	108.42	100.72

5 rows | 1-5 of 7 columns

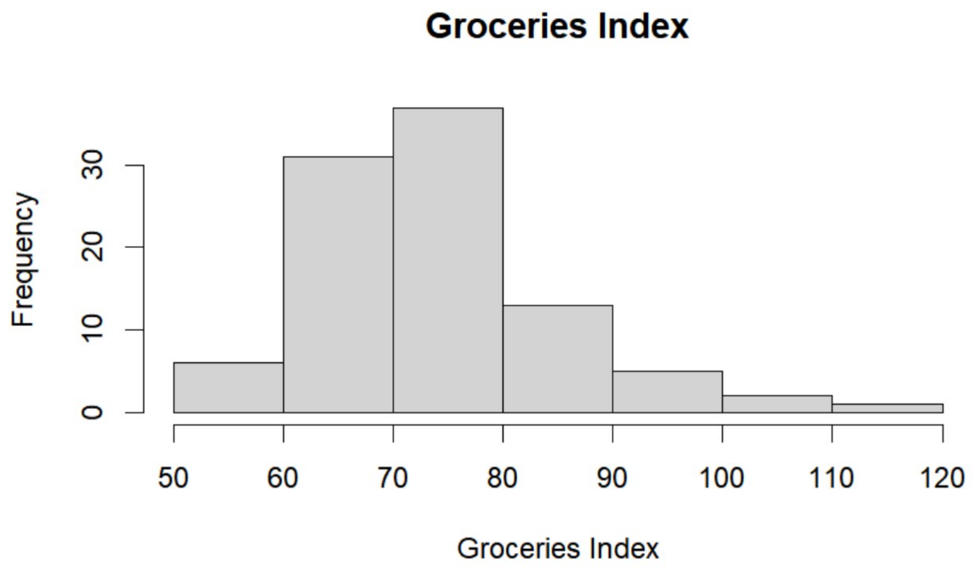
METHODOLOGY AND MODELING

This project consisted of four different analysis techniques. First, we did a histogram of the rent, groceries, and restaurant indexes to show the frequency distribution of each attribute that would be analyzed in the future models.

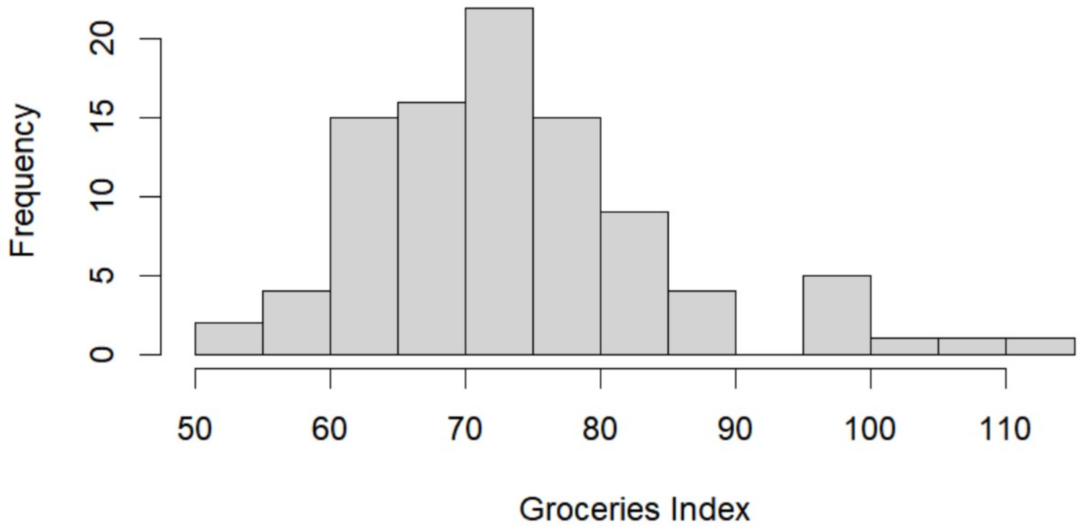
RENT INDEX W/ BREAKS AT 10



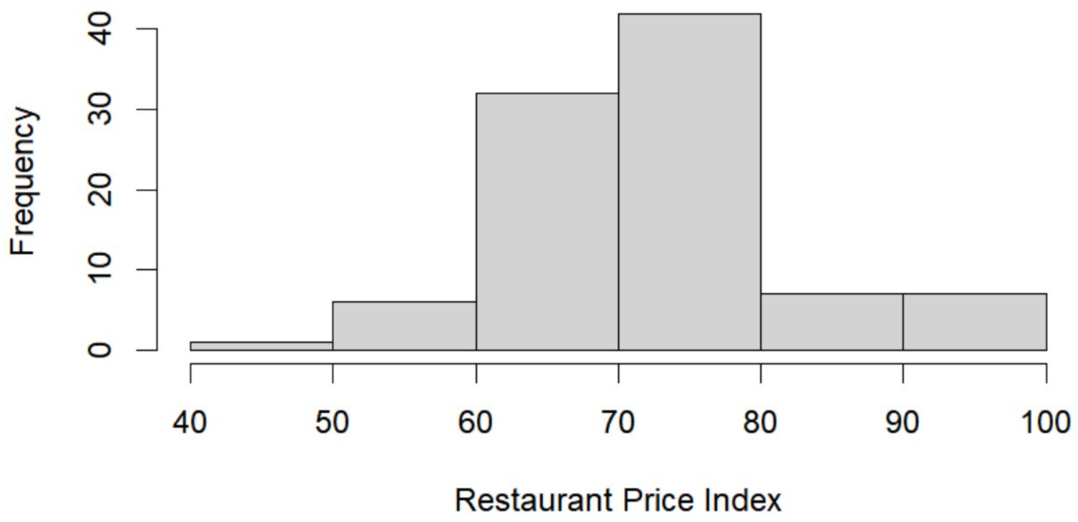
GROCERIES INDEX W/ BREAKS AT 5



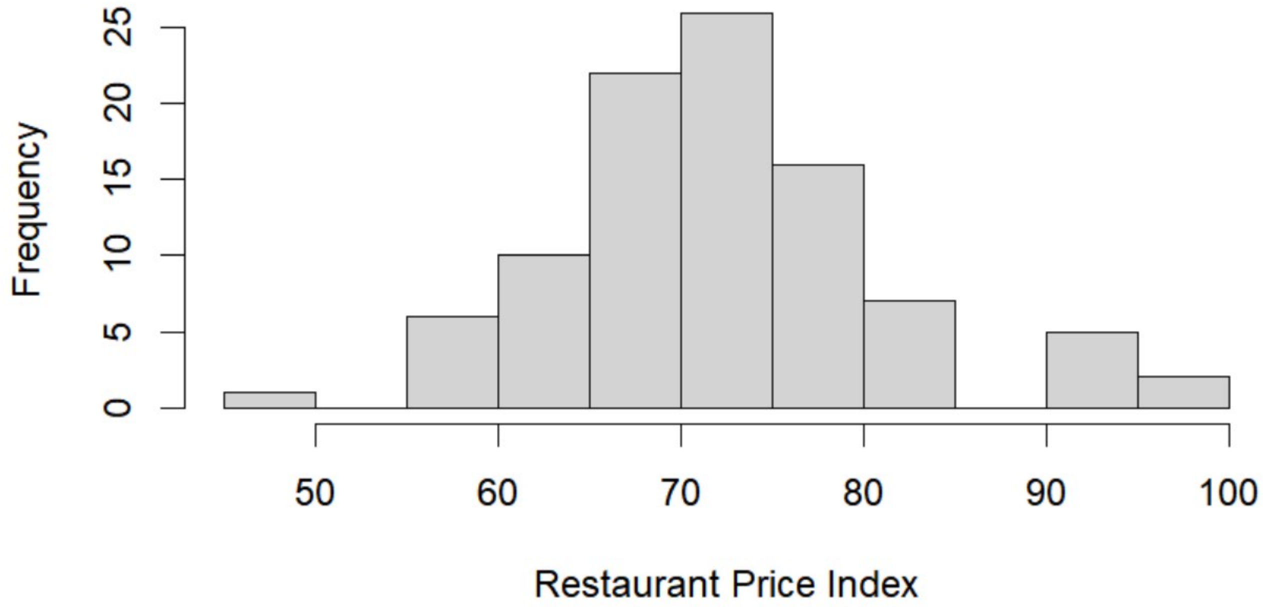
GROCERIES INDEX W/ BREAKS AT 10



RESTAURANT PRICE INDEX W/ BREAKS AT 5

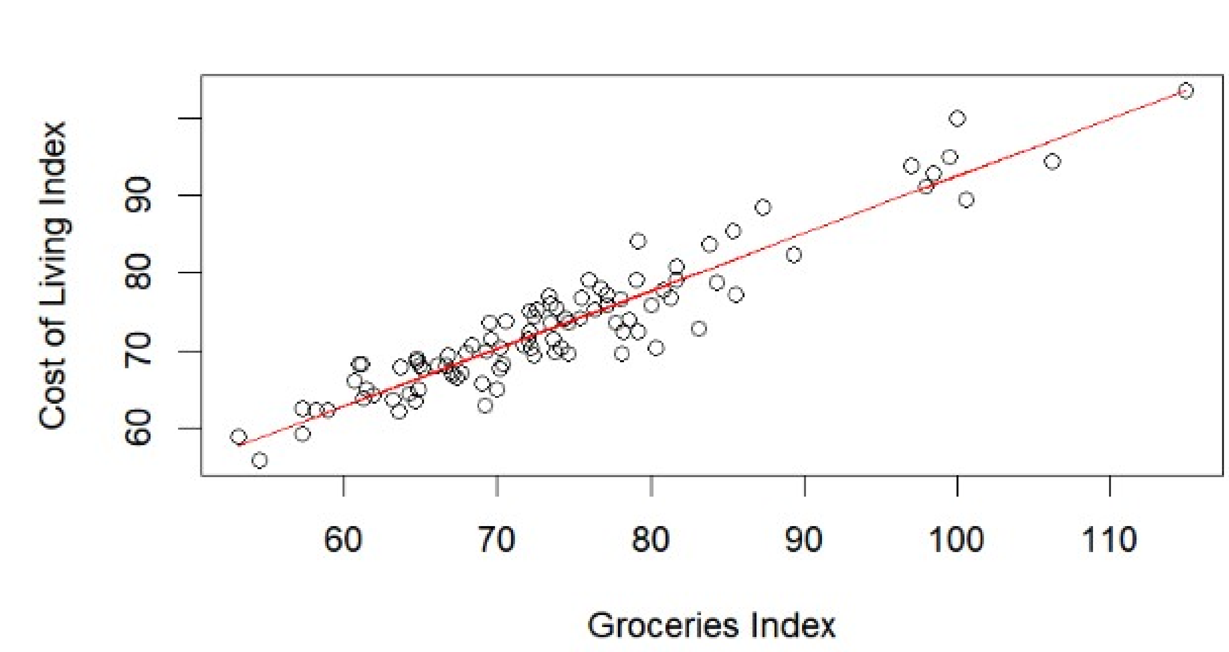


RESTAURANT PRICE INDEX W/ BREAKS AT 10



METHODOLOGY AND MODELING

LINEAR REGRESSION PLOT AND CORRELATION OF GROCERIES



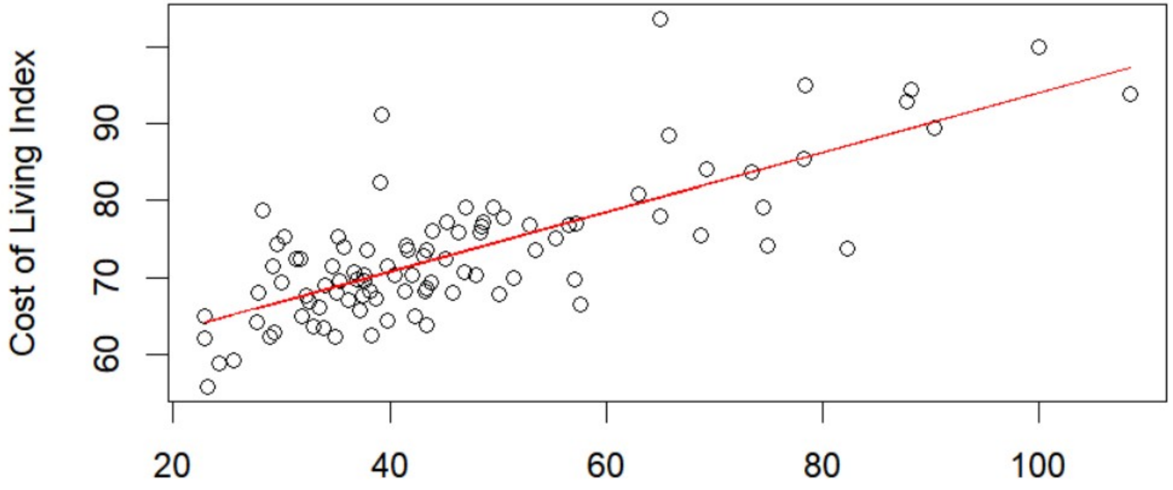

```
grocx <- gproj$Groceries.Index
y <- gproj$Cost.of.Living.Index
cor(grocx,y)
```

```
[1] 0.9423183
```

For the second analysis, we did a scatter plot with a linear regression line to show the correlation of the rent, groceries, and restaurant price index to the cost-of-living index. Below each index scatterplot, we used the correlation function in R to calculate the relationship strength between the two variables.

9

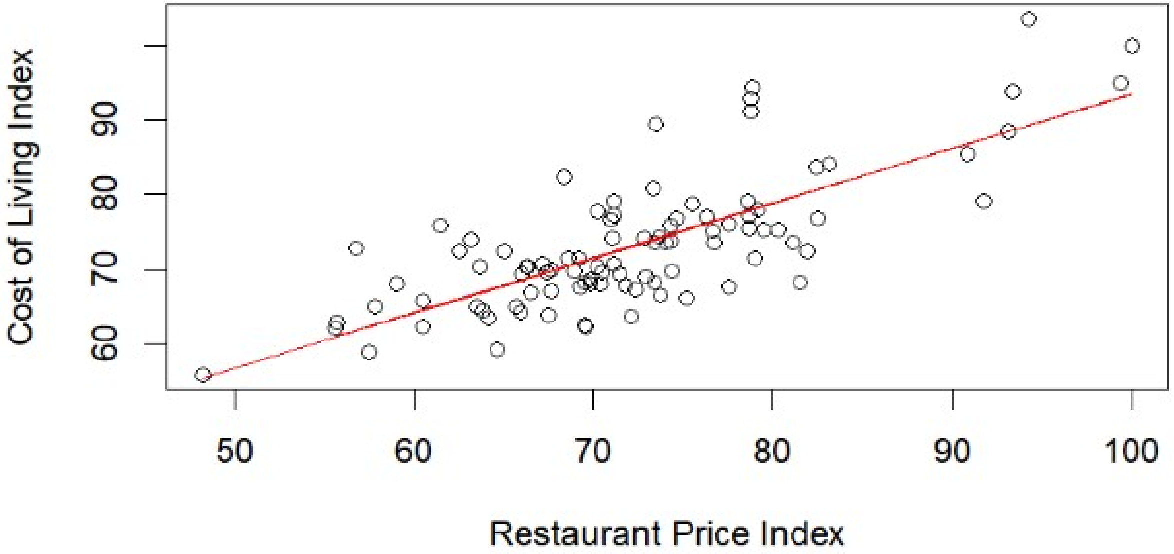
LINEAR REGRESSION PLOT AND CORRELATION OF RENT




```
rentx <- gproj$Rent.Index
y <- gproj$Cost.of.Living.Index
cor(rentx,y)
```

```
[1] 0.7641735
```

LINEAR REGRESSION PLOT AND CORRELATION OF RESTAURANT PRICE




```
restx <- gproj$Restaurant.Price.Index
y <- gproj$Cost.of.Living.Index
cor(restx,y)
```

```
[1] 0.7602267
```

METHODOLOGY AND MODELING

	City	Cost.of.Living.Index	Range
1	Honolulu, HI, United States	103.65	High
2	New York, NY, United States	100.00	High
3	Santa Barbara, CA, United States	95.01	High
4	Berkeley, CA, United States	94.36	High
5	San Francisco, CA, United States	93.91	High
6	Oakland, CA, United States	92.93	High
7	Anchorage, AK, United States	91.23	High
8	Santa Clara, CA, United States	89.41	High
9	Seattle, WA, United States	88.52	High
10	Boston, MA, United States	85.47	Medium
11	Queens, NY, United States	84.02	Medium
12	Washington, DC, United States	83.74	Medium
13	Pittsburgh, PA, United States	82.36	Medium
14	Jersey City, NJ, United States	80.79	Medium
15	Philadelphia, PA, United States	79.19	Medium
16	Los Angeles, CA, United States	79.19	Medium
17	Minneapolis, MN, United States	79.08	Medium
18	Birmingham, AL, United States	78.82	Medium
19	Miami, FL, United States	78.00	Medium
20	Sacramento, CA, United States	77.88	Medium
21	Charleston, SC, United States	77.26	Medium
22	Asheville, NC, United States	77.25	Medium
23	Chicago, IL, United States	77.06	Medium

For the third analysis, we grouped the cost-of-living indexes into ranges. This was done using the cut function in R which divided the cost-of-living index into 3 levels. These levels were labeled high, medium, and low.

Assigning Low, Medium and High labels to levels

Hide

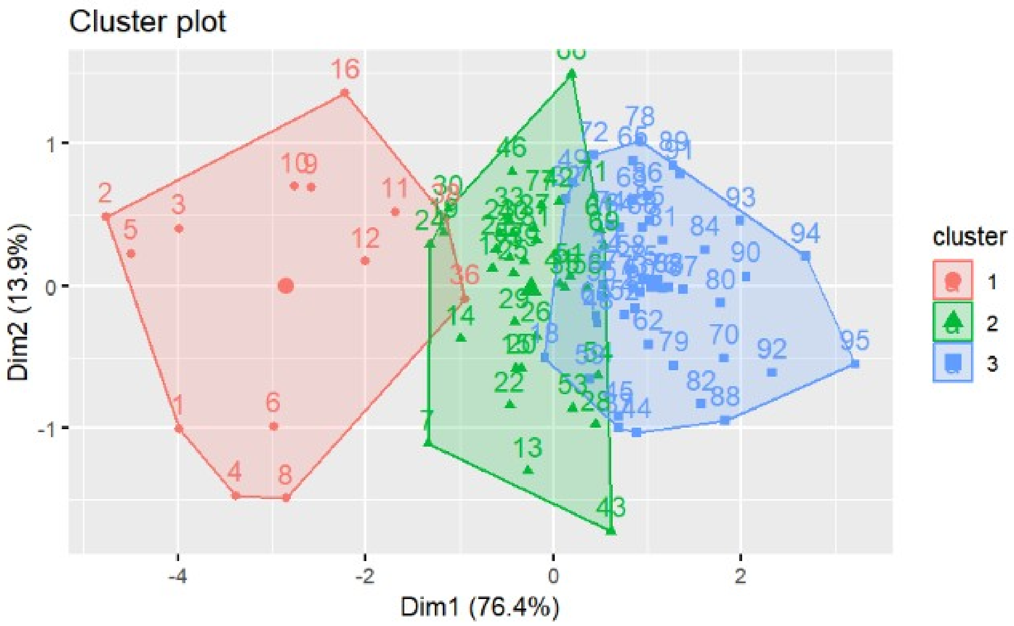
```
cut(costindex,3, labels = c("Low", "Medium", "High"))
```

```
 [1] High   High   High   High   High   High   High   High   High
[10] Medium Medium Medium Medium Medium Medium Medium Medium Medium
[19] Medium Medium Medium Medium Medium Medium Medium Medium Medium
[28] Medium Medium Medium Medium Medium Medium Medium Medium Medium
[37] Medium Medium Medium Medium Medium Medium Medium Medium Medium
[46] Medium Low    Low    Low    Low    Low    Low    Low    Low
[55] Low    Low    Low    Low    Low    Low    Low    Low    Low
[64] Low    Low    Low    Low    Low    Low    Low    Low    Low
[73] Low    Low    Low    Low    Low    Low    Low    Low    Low
[82] Low    Low    Low    Low    Low    Low    Low    Low    Low
[91] Low    Low    Low    Low    Low
Levels: Low Medium High
```

METHODOLOGY AND MODELING

	City	Cost.of.Living.Index	Range	km.cluster
1	Honolulu, HI, United States	103.65	High	3
2	New York, NY, United States	100.00	High	3
3	Santa Barbara, CA, United States	95.01	High	3
4	Berkeley, CA, United States	94.36	High	3
5	San Francisco, CA, United States	93.91	High	3
6	Oakland, CA, United States	92.93	High	3
7	Anchorage, AK, United States	91.23	High	2
8	Santa Clara, CA, United States	89.41	High	3
9	Seattle, WA, United States	88.52	High	3
10	Boston, MA, United States	85.47	Medium	3
11	Queens, NY, United States	84.02	Medium	3
12	Washington, DC, United States	83.74	Medium	3
13	Pittsburgh, PA, United States	82.36	Medium	2
14	Jersey City, NJ, United States	80.79	Medium	2
15	Philadelphia, PA, United States	79.19	Medium	2
16	Los Angeles, CA, United States	79.19	Medium	3
17	Minneapolis, MN, United States	79.08	Medium	2
18	Birmingham, AL, United States	78.82	Medium	1
19	Miami, FL, United States	78.00	Medium	2
20	Sacramento, CA, United States	77.88	Medium	2
21	Charleston, SC, United States	77.26	Medium	2

For the fourth analysis, we ran the rent, grocery, and restaurant price indexes through the kmeans clustering model. Three clusters were created and a kmcluster column was added next to the ranges column. These two columns were compared to each other to see if the cluster model aligns with generated cost of living index ranges.



RESULTS

The scatterplot visualizations and correlation function show a strong correlation of each attribute to the cost-of-living index. As the independent variable (attribute) on the x-axis increases, then the dependent variable (cost of index) increases. The correlations for the rent and restaurant price index were 0.76 which indicates a strong correlation. However, the highest correlation was the groceries index at 0.94. The ranges for the cost-of-living ranges are as follows: 55.9 – 71.8 for the low range, 71.8 – 87.7 for the medium range and 87.7 – 104 for the high range. There were no cost-of-living values that matched the borderline values of 71.8 and 87.7. As a result, all the cost-of-living index values were grouped in the correct range. After the attributes were processed through the kmeans clustering model, there was significant overlap between clusters 1 and 2 and clusters 2 and 3. There was no overlap between clusters 1 and 3. Due to this overlap between the clusters, the kmeans model did not align with the cost of index ranges. This was seen after adding the clustering column to the dataset. Cluster 1 consists of high and medium range values. Cluster 2 consists of high, medium, and low range values. Cluster 3 consists of medium and low range values.

EVALUATION

The Rent Index histogram with break = 10 is positively skewed which means that there are more data with higher values than there are with lower values. In the case of a rent index histogram, this means that there are more expensive accommodations than there the affordable one. Another thing that can be interpreted that most of the people pay rent between 30 to 40. The groceries index histogram with break =10 is almost a bell-shaped barring some anomalies at right side of the histogram. If we understand it as a bell shape histogram, we can say that the data is evenly distributed about the means and it is relatively affordable. But if we reduce the bin side i.e. decrease the break to 5 then we see the histogram becomes skewed on the right side, this means that affordable groceries are rarer. The restaurant price index histogram can be analyzed as a distribution where restaurants with very low prices are also present and restaurants with higher prices are also present. The affordable price restaurants are in abundance. The regression line for rent index shows upward trend, i.e. as the rent price increases the cost of living is likely to increase. Another thing we see here is that majority of data is concentrated at the lower part of the regression line, this indicates that more people are paying a affordable rent and less people are looking for higher price apartments. Another thing we can get here is that majority of people drawing an average salary. The regression line for groceries index is having a steeper slope with co-relation 94% than rent and restaurant index where c-relation is about 75%, this indicates that cost of living is affected more by groceries rather than rent or restaurant index

AFTERTHOUGHTS

Major challenges and solutions

The major challenge is to relate local purchase power to other index indicators. The k-mean cluster does not magically recommend any city to live in where the cost of living is less. It has to interpreted by some other method.

Conclusion and future work

After analyzing our data with different methods, we learned that data analytics is not an exact science. Considering the vast amount of available modeling tools, it is hard for us beginners to determine which method is the best. I feel as though we did an acceptable job with comparing different methods and interpreting the data. In the future, familiarizing oneself with more models may