Best City Contest: research by Alexander Kosenkov
alexander@kosenkov.com

Looking at the data

Original dataset, provided by EIU consists of 3 tables, 75 data columns in total, available for 140 cities. The goal is to find single ranking for all cities.

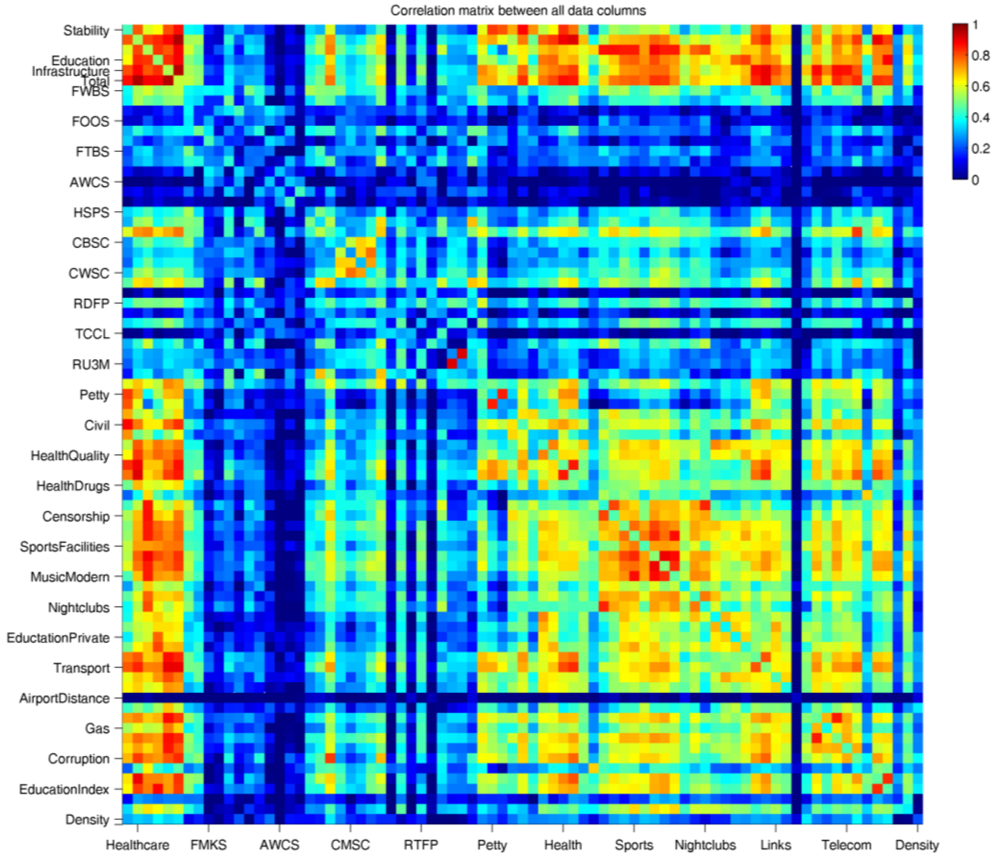
One approach, selected by the Economist is to assign some coefficients to particular columns and calculate weighted average. This method heavily depends on analytic's point of view and is inherently arbitrary. Nevertheless, such rating proved to be useful, so I will compare my results with this original rating.

I decided not to introduce any weights or arbitrary coefficients to the ranking criteria, but to calculate ranking from the database itself.

Technically speaking, the task is to reduce the complexity of data: combine as much parameters as possible to a single score, which describe cities almost as good as the complete parameter set. This problem is known as Data Dimensionality Reduction problem.

Source Data

First, let's take a look at our data. The following chart shows linear correlation between all parameters:



Parameters shown here are: Liveability scoring (5 variables + Total), data for Cost of living (30 parameters), City Liveability indicators (40) and 3 my added variables that turn out to be not so useful. The diagonal contains mean values.

Several normalization techniques were used to prepare the data. Few missing values in Costs database were estimated by average value of same parameters in 5 most similar cities. For details, please check normalizeData.m in the source code archive.

From this chart we see that the whole Cost Data block hardly correlates to anything and shall be treated separately. Additionally, Airport Distance will be removed.

Cost Factors

The sub-goal is to have a few parameters, which describe aggregated cost of living for every city. For this I used non-linear dimensionality reduction methods, available in Matlab Toolbox by Laurens van der Maaten.

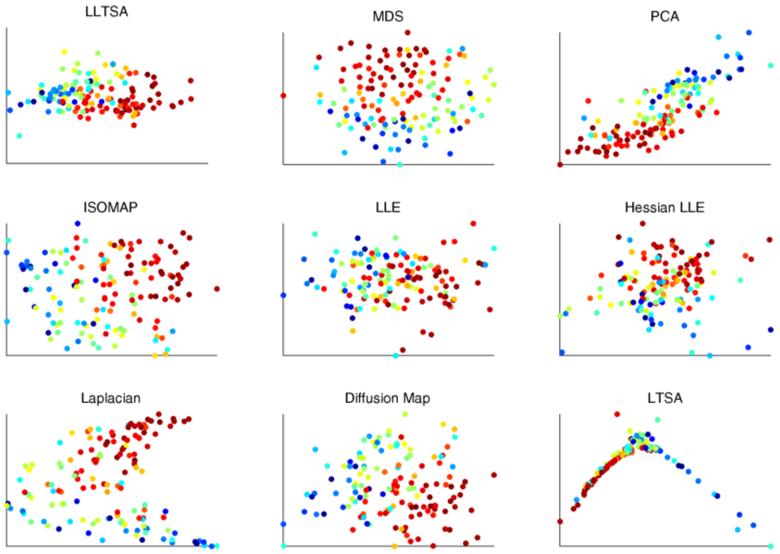
LLTSA is an unsupervised machine learning algorithm that flattens n-dimensional data into lower number of dimensions, based on how similar are parameters of different objects (cities). Similarity is defined in terms of Euclidean distance, so data normalization is important.

It looks like prices are distributed according to Log-Normal law, so I used standard normalization for every price column: newX = (log(1 + X) – mean(X)) / std(X)

I applied LLTSA method to normalized data to extract 3 new parameters ('factors').

It is amazing, that this fully autonomous algorithm created Cost Factor 1, which has 75.3% correlation with original City Liveability Total! Even though The Economist Intelligence Unit's Liveability index does not directly include any of the original Cost parameters.

The following chart shows comparison of 9 different Dimensionality Reduction algorithms:

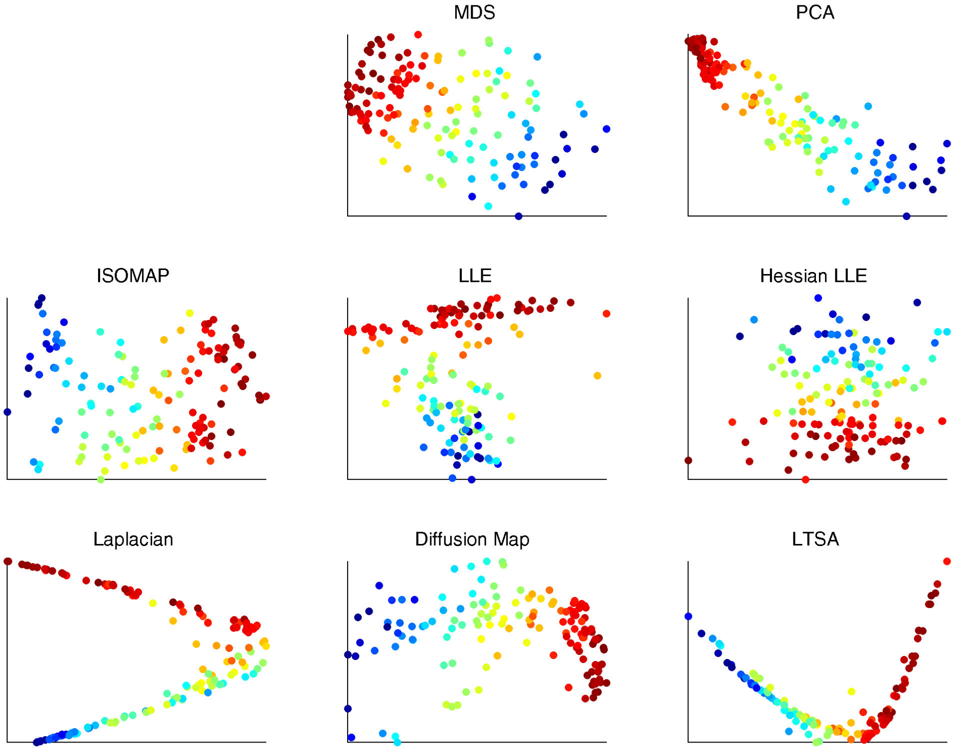


Each dot represents a city; its color – original Liveability Total score. Two axes of each sub-plot are reduced dimensions, as proposed by each method. For more details, please see the link to an MANI tool. LLTSA (linear local tangent space alignment algorithm) gives the best result here.

Comparing different Data Dimensionality Reduction methods

The goal is to reduce number of parameters even more.

Applying the same machine learning techniques to remaining parameters plus re-mixed Cost data shows that simple Principal Components Analysis gives fairly good and explainable linear results:



To my surprise, every method achieved quite good correlation with existing ranking, even though input data contains neither pre-defined coefficients nor summary Liveablity data that is immediately used in EIU index.

At this point I decided not to continue with sophisticated non-linear methods, mainly because their predicted 'top 10' of cities is not numerically stable. Small changes to source data can cause significant re-arrangements of best cities, while total correlation is still kept rather good.

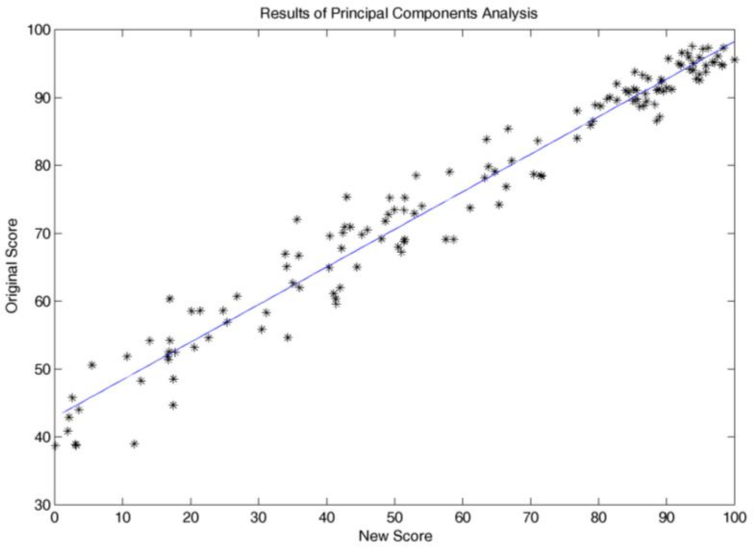
Results

Zurich is calculated to be the best city, followed by **Vienna**. Complete ranking along with parameters for final calculation is uploaded to BuzzData.

Obviously, many people might not agree with this ranking or criteria. My answer is based purely on the data and its inner relations. The next step would be to explicitly include reader's personal preferences into the analysis to produce custom rankings.

Conclusion

Resulting main factor, calculated by PCA gives the score with 98% correlation to original EIU rating. This clearly shows that domain experts of The Economist Intelligence Unit managed to produce very reasonable weighted average of source data columns. In my research, on the other hand, I used only the raw data and unsupervised algorithms to come up with similar conclusions:



In the end, my approach could be reused for many similar databases and problems, resulting in insights, comparable to expert-level analysis. The submission also includes some parts of the Matlab source code that I used along this research.

I hope I could raise your interest in modern data analysis and visualization tools with this article. If so, you are very welcomed to explore the links shown on BuzzData Articles page.

Appendix 1: Alternative Approaches

In earlier versions of the research I gathered additional data parameters for every city:

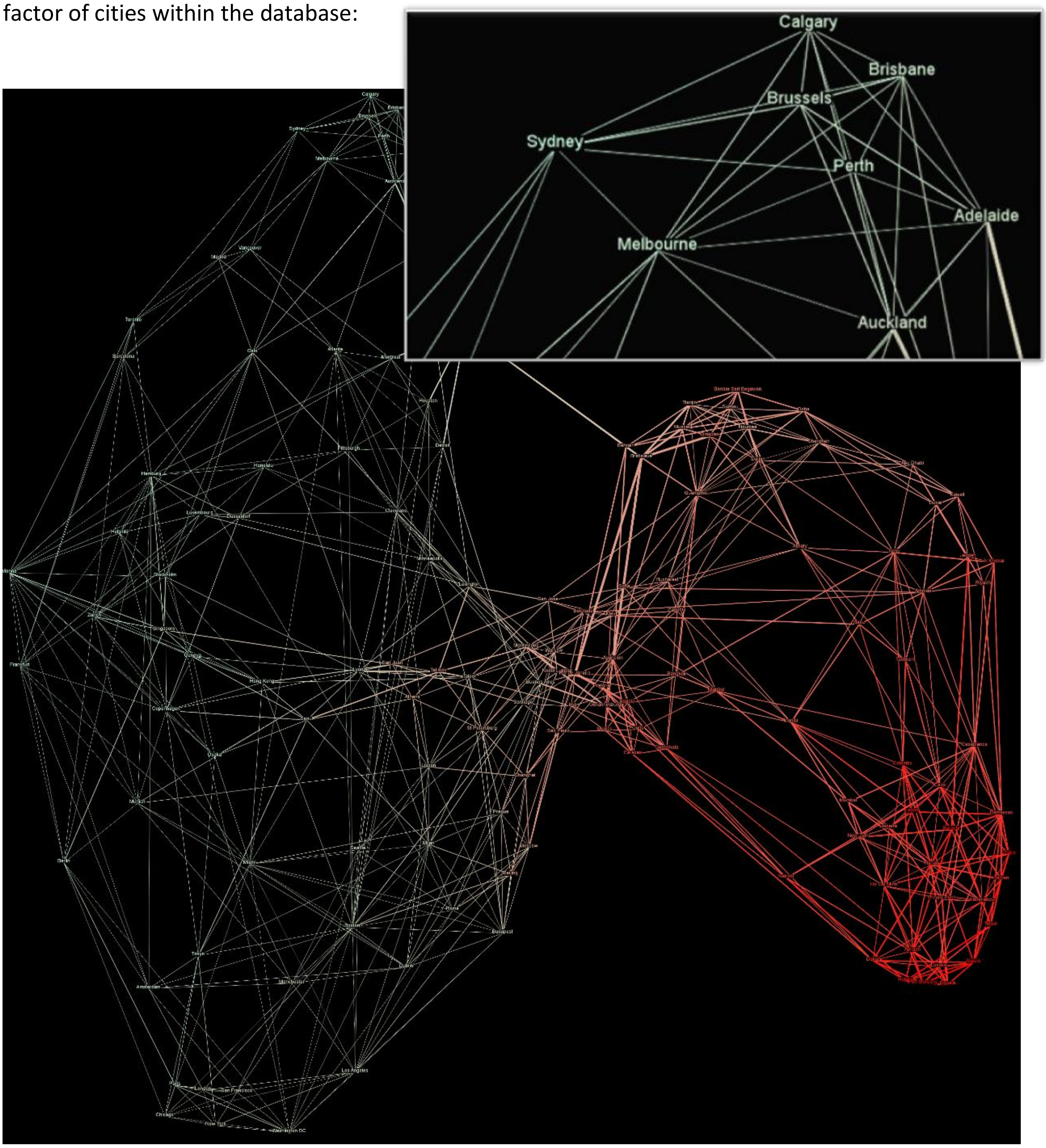
- Population Density
- Estimated number of photos found online for every city
- Estimated average ticket price to 15 major destinations, normalized by distance

Unfortunately, those weakly correlate with achieved results: Population Density seems to be uniform among European cities and vary a lot elsewhere.

Online popularity of the city is mostly built by tourists and active travelers. Nowadays there are few limits for worldwide tourism, so number of photos represents something else but quality of living.

Appendix 2: City Similarity Graph

As an illustration of non-linear data analysis algorithms, I made some visualizations of similarity factor of cities within the database:



Each city here has at least 6 most strong connections to other cities; strength is proportional to averaged difference in the scores and color corresponds to the new city Liveability Index.

Russian writer Leo Tolstoy told that “Happy families are all alike; every unhappy family is unhappy in its own way.” This similarity graph, unfortunately, shows the opposite: only wealthy cities can afford individuality.