***2. Factor Analysis***

With the aim to reduce the complexity of the extracted dataset, we decided to perform an Exploratory Factor Analysis on the dataset. The Factor Analysis allows us to diminish the high number of variables in the dataset to a small number of fundamental factors which correspond to uncorrelated latent variables, thus they allow us a straightforward interpretation of the given data. The trade-off underlying this approach is to maximize the explained variance while reducing the data complexity.

In order to perform a Factor Analysis, we first calculated the correlation matrix given the twenty variables. The matrix enables a first assessment of the interdependencies in the data, thus gives a first impression on the number of fundamental factors.



Figure X. Correlation Plot

As shown in figure X, there are some highly relevant correlations to take into account. ‘Trendy’ and ‘vibrant’ with a correlation coefficient of 0.89 as well as ‘vibrant’ and ‘fun’ (0.86) implicates an existing factor corresponding these attributes. An indication for a second fundamental factor is derived by the correlation between the attribute ‘clean’ and the attribute ‘safe’ with a positive correlation of 0.73. Furthermore, the attribute ‘green’ also seems to act on this factor with a correlation coefficient of 0.54 regarding ‘clean’ and a correlation coefficient of 0.53 regarding the attribute ‘safe’. Figure X also gives information about highly negative correlated attributes like ‘noisy’ and ‘clean’. With a correlation coefficient of -0.73, we can conclude that these two attributes together will not form a fundamental factor.

***2.1 Factor Extraction***

In this Factor Analysis we did not know how many factors we have to extract in order to reduce the data complexity while maximizing the explained variance. Therefore, we executed a Scree-Test on the data to compare the eigenvalues of the factors and to see from what number of factors the marginal gain in explained variance does stagnate. The following figure shows the Scree-Test performed on the given dataset.



Figure X. Scree-Test

Figure X determines the factor number to be 4. Implying, that from factor 5 on, there is no significant gain in the explained variance. To validate the Scree-Test results we compared the cumulative variances in the Factor Analysis across different factor numbers. Within every Factor Analysis the orthogonal ‘varimax’ rotation was used in order to sustain the non-correlation between the extracted factors and maximize the variance within a factor. The results confirmed the findings derived by the Scree-Test. With 4 factors the cumulative variance of the FA is 72%. Performing the same analysis with 5 factors leads to a cumulative variance of 76%, which corresponds to a rise in explained variance of 4%. This also holds for a factor number of 6, with a cumulative variance of 80%. Therefore, adding more than 4 factors in the FA is not efficient because the additional explained variance does not compensate for the loss in model simplicity.

***2.2 Factor interpretation*** so richtig glücklich bin ich noch nicht mit den Benennungen der Faktoren :D ich überlege da mal noch ne Runde

The identification of the 4 factors extracted depends on the loadings of the attributes on the culled factors. By the fact, that the variables are standardized, with each variable having a mean value of 0 and a standard deviation of 1, the interpretation of the factor loadings follows an intuitive structure; The higher the factor loading of a variable the higher its affiliation to the corresponding factor. The attributes *‘trendy’*, *‘vibrant’*, ‘*multicultural’* and ‘*fun’* load very high on **factor 1** with loadings of 0.90, 0.86, 0.82 and 0.74, respectively. Accordingly, we declared factor 1 to represent the latent variable *‘dynamic? Ich würde hier gerne was anderes als young nehmen, aber hip/cool klingt so sehr nach Boomer :D’*. The highest factor loadings regarding factor 2 are as follows. The attributes *‘clean’* with a factor loading of 0.88, *‘safe’* with a loading of 0.78 and *‘green’* with a loading of 0.69. Additionally, the attribute *‘affordable’* has a factor loading of -0.60. Therefore, we defined **factor 2** as *‘modern’.* Factor 3 could be formed by the high loadings of the variables *‘romantic’* (0.88) and ‘beautiful’ (0.79). However, *‘englishspeaking’* loads negative on the factor with a loading of -0.50, indicating that **factor 3** represents the latent variable *‘charming’* leading to the hypothesis that the factor values are high for the old and romantic cities. The interpretation regarding factor 4 has proven to be challenging since the loadings were moderately high and did not imply a clear factor definition. The attribute with the highest loading was *‘friendly’* (0.71) followed by *‘fun’* (0.45) what leaves little room for interpretation. Consequently, it was crucial to take a look at the negative loaded variables. The attributes *‘museums’* and *‘culturalevents’* were counter-rotating to factor 4 with the most negative loadings of -0.51 and -0.39, respectively. By virtue of this loadings, we declared **factor 4** as *‘young’*. To approve the factor definitions as well as to classify the cities in data dataset according to these definitions we aggregated the component scores across the cities. The factor values confirmed the extracted factor interpretations, as shown in the following figure.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **‘dynamic’ (F1)** | | **‘modern’ (F2)** | | **‘charming’ (F3)** | | **‘young’ (F4)** | |
| **City** | **FV** | **City** | **FV** | **City** | **FV** | **City** | **FV** |
| Berlin  London  Amsterdam  Riga  St.Petersburg  Geneva | 2.21  1.81  1.09  -1.13  -1.16  -1.45 | Stockholm  Vienna  Geneva  Budapest  Athens  Istanbul | 2.19  1.61  1.44  -0.86  -1.32  -1.63 | Rome  Paris  Vienna  Berlin  Dublin  Brussels | 1.90  1.31  1.11  -1.22  -1.44  -1.93 | Amsterdam  Prague  Lisbon  London  Brussels  Paris | 1.23  0.98  0.87  -1.08  -1.22  -1.67 |

Figure X. Factor values

The higher the Factor Values (FV) of a city the better does the city fit into the factor definition. Berlin, known for its multicultural and vibrant lifestyle has an extremely high FV regarding F1. Whereas Geneva is known as a traditional city what explains the negative FV of -1.45 in F1. We can therefore state that the extracted factor seems to be efficient and can be used for further analysis. The same holds for the other factors extracted.