SWP 2 - Reducing Data Complexity

In the following we will analyze the given dataset, regarding the correspondent's opinion about a variety of cities from a touristic viewpoint. Those specific cities will also be contextualized in a geographical, purely relational manner. In order to better structure our analysis and the subsequent findings and interpretations to maintain inner structure and logical consistency, we will be adhering to the following sequence of contents. Initially we will set the stage for the entirety of the analysis by giving an economic context. Going on we will detail the data cleaning process used to increase informational interpretability through a data subset. Furthermore, an overall and case specific explanation of MDS will set the stage for application in our specific case. The results will then be contextualized, and their usefulness critically analyzed. Consequently, we will perform a Factor Analysis with the subset data, whereupon we will cluster the cities by four main factors. The economic interpretation of our factor clustering will close our analysis followed by limitations of the given data interpretation and methods.

1. Economic Context

In order to better direct our analysis of the given dataset in a contextually valuable direction and form a stringent conclusion, we will use an economic context as a driving force behind our analysis. Without a distinct goal in mind, it is very hard to gain conclusive insight into any given dataset, due to the sheer amount of possible differing interpretations. Utilizing a goal in mind with a core objective of gaining insight into the respective problem area, greatly enhances the inherent logical structure, as well as the case-bound informative value of any achieved results or final assertions. Our choices and scientific procedure in this analysis will all be based upon the assumption of the economic case as follows.

We are hired by a city tour operator from Europe to gather insights into the preferences of a younger target audience regarding the most common cities for travel across Europe. While they are already operating in several cities, they conducted a questionnaire among students and young adults to gain further insights needed in order to potentially expand their company. In order to do so in a meaningful manner, we will have to adjust our dataset to the given circumstances.

2. Dataset and Data Clearing

When working on an economic case as the above, it is key to exploit the given data pool to its full potential and reduce fringe inaccuracies and non-relevant datapoints as much as possible. In order to do so, and to better be able to come to more conclusive results regarding the underlying economic problem it's been proven to be best practice to further segment the dataset. In a marketing setting this step usually involves a so-called customer segmentation, thus separating the entire customer base by certain parameters such as social status, psychological profiles, social reference groups or economic status. Members of one of those subsegments of the customer base are more likely to share certain characteristics that lead to similar preferences in products or services and similar responses to differing product offerings or price fluctuations.

In our case, when analyzing the given dataset, we noticed that most respondents seemed to belong to a customer group of "students" from a major city or adjacent/similar groups. In order to ensure greater homogeneity among our analytics dataset, we preselected by the following parameters to form a subset from the given data.

1. *Age range*
   We only included respondents in our subset who were aged between 18 and 29. Generational segmentation is widely accepted in marketing theory due to similar circumstances while growing up and their implications for prevalent characteristics

2. *Occupation*
   As for the occupation of respondents, we chose to only include those in our subset that explicitly stated they were students. This selection is mainly an indication of economic status, posing implications for disposable income and overall average buying power. For any respondents where the occupational status was unclear, we chose to not include their respective datapoints, as a smaller subset with greater homogeneity is far informationally significant than one with falsely included datapoints to the chosen economic subset.

3. *Nationality*
   In our new subset, we only included datapoints from respondents that indicated they are living in Germany currently. Because the questions are of relational nature, the implications of given answers are vastly different whether a respondent gives answers from the perspective of a local, when asked about their home city, or one of a tourist, when asked about any other city. Furthermore, by limiting the respondents to one geographical factor, cultural homogeneity further increases.

3. Multidimensional Scaling

Multidimensional Scaling (MDS) is a way of analyzing the data to examine the similarities and the dissimilarities between the objects of a dataset by calculating the distances between these objects i.e., similarity matrix, using primarily the Euclidian distance measures and then visualizing these distances in a lower dimensional space to better interpret it. Although, at first, it was only the metric MDS that was available for research purposes and it wasn't very common among the researchers. This was mainly because of non-metric MDS being found in the 1960s as well as the advancements in technology and programming languages, it gained huge popularity among different disciplines such as psychology, sociology, education and in our case most importantly, marketing.

In general, there are some advantages and disadvantages of the Multidimensional Scaling method. One of the advantages is that the results of the method are rather simple. Resulting in the fact that even the people who might not have specific knowledge about the data or the method can comprehend and comment on the distances between the objects in the map. This also applies to the similarities of the respondents' perceptions, therefore it is reasonable to say

that it is fairly easy to interpret it. Another advantage is that although in cases that the dataset has relatively small amount of pairwise similarities, scatter plot could be used too, in the cases where the dataset is bigger, for example with more than 10000 objects or features in the dataset. This is both because of the method's own nature and it is practical function of dimensionality reduction. MDS offers a more compact method, in which the data can be divided into subgroups, making it easier to compare the similarities between the datapoints.

MDS can be divided into two subtypes, metric and non-metric MDS, which differ in the way the dissimilarities are transformed into distances. The metric MDS creates a linear relationship i.e., if one doubles the dissimilarity one also doubles the distance in the cartesian space, whereas the non-metric MDS uses an ordinal scale, it is therefore a representation of the similarity of rankings instead of the actual distances.

Dimensionality reduction of MDS is perceived as one of its core advantages. However, because while the decrease in the number of dimensions increases the interpretability of the model, the error in the representation of the data increases, which adds a feature of trade-off to the model it can be interpreted both as an advantage and disadvantage depending on the situation.

In the City Trip Questionnaire, the respondents were asked to answer indirect questions and rate different attributes, which is more useful than asking direct questions of how they perceive these cities. The first reason for this is that people might not be able to answer directly how they perceive a product or a brand, in our case the cities, unless they are addressed more specific questions about it, and this would create no insights that could be used for making informed managerial decisions. The second is that asking indirect questions also helps to explain the different ways, in which the cities were perceived. This is important because it would help us to target a specific dimension that the consumers don't perceive it as we want them to do.

As a result, both because of the reasons above and the fact that the Questionnaire data is based on 20 different attribute ratings about the cities, which are non-metric, in our analysis, we have used non-metric MDS.

3.1 MDS Application

In the first part multidimensional scaling (MDS) was applied, which is a visual representation of the distances or dissimilarities between objects in a cartesian space based on the distances between each of them. Objects are more similar to each other, the closer they appear in the cartesian space.

MDS is always being applied to a matrix, in this case a triangular matrix of Euclidian distances between the cities based on their average attribute scores. Both metric and non-metric MDS were computed and displayed in a 2-dimensional cartesian space using ggplot2.

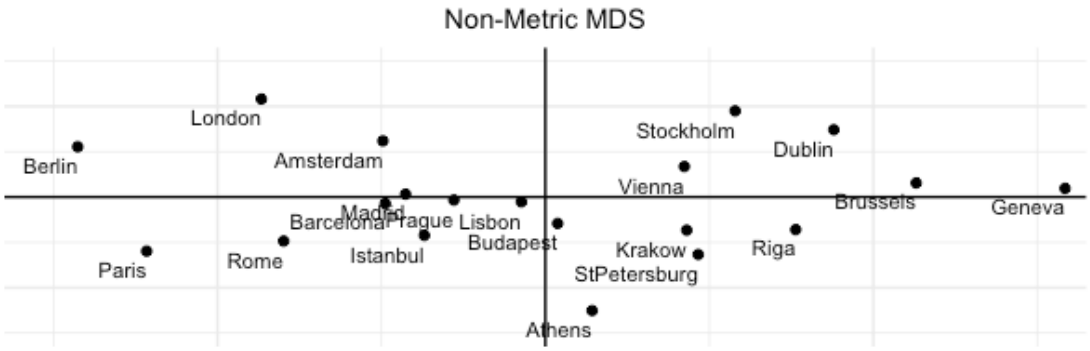

Figure 1. Non-Metric MDS

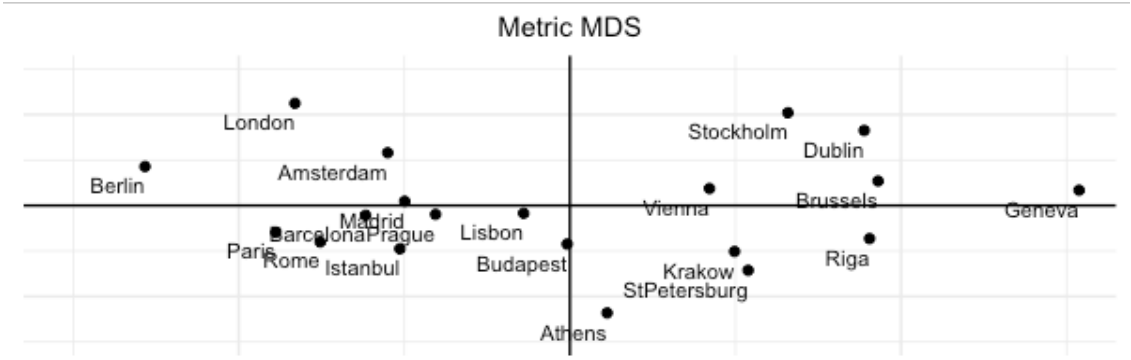

Figure 2. Metric MDS

As to be expected the results are fairly similar for both types of MDS. Geographical proximity seems to affect the placing on the cartesian space to a certain degree, possibly because of cultural similarities. One can see that Madrid, Lisbon and Barcelona are close, just like Brussels and Geneva. Furthermore, Budapest, Krakow, St. Petersburg and Riga as well as Berlin, Amsterdam and London appear in the same quadrant, which suggests a high degree of similarity. In the following section property fitting will shed some light on which attributes may be associated with the positions of the individual cities.

3.2 Property Fitting

Property fitting is a method of analyzing the proximities based on dimension to get an idea of which attributes may have affected the coordinates of each city. In our case it makes it possible to plot the attributes and preferences into the cartesian space that resulted from the MDS.
First it is necessary to relate the position in the cartesian space to the individual properties. A linear regression with the attributes as the dependent variables, and the coordinates of the cities as the independent variables is used to achieve this goal. The regression summary for the attribute ‘romantic’ is displayed below.

Response romantic:
Coefficients:

	Estimate Std.	*Error*	*t value*	*Pr(>|t|)*
x	*-0.1196*	*0.5596*	*-0.214*	*0.833*
y	*-0.1639*	*0.7239*	*-0.226*	*0.823*

Multiple R-squared: 0.005356, Adjusted R-squared: -0.1052
F-statistic: 0.04846 on 2 and 18 DF, p-value: 0.9528

In general, the impact of the cities' coordinates on the attributes is statistically insignificant. The standard error (ranging from 0.52 to 0.97) is relatively high compared to the estimates, the t-statistic values (ranging from -0.64 to 0.66) are fairly close to zero and relatively small compared to the standard error, the R-squared (ranging from 0.001 to 0.027) is very low and the p-values (ranging from 0.51 to 0.97) are very high.

However, the purpose of the regression is not to get a particularly good fit, which would be unlikely given the nature of the data, but rather to give a rough idea of the relation of the attributes to the coordinates. The graph below shows the results of the regression plotted over the results of the non-metric MDS using ggplot.

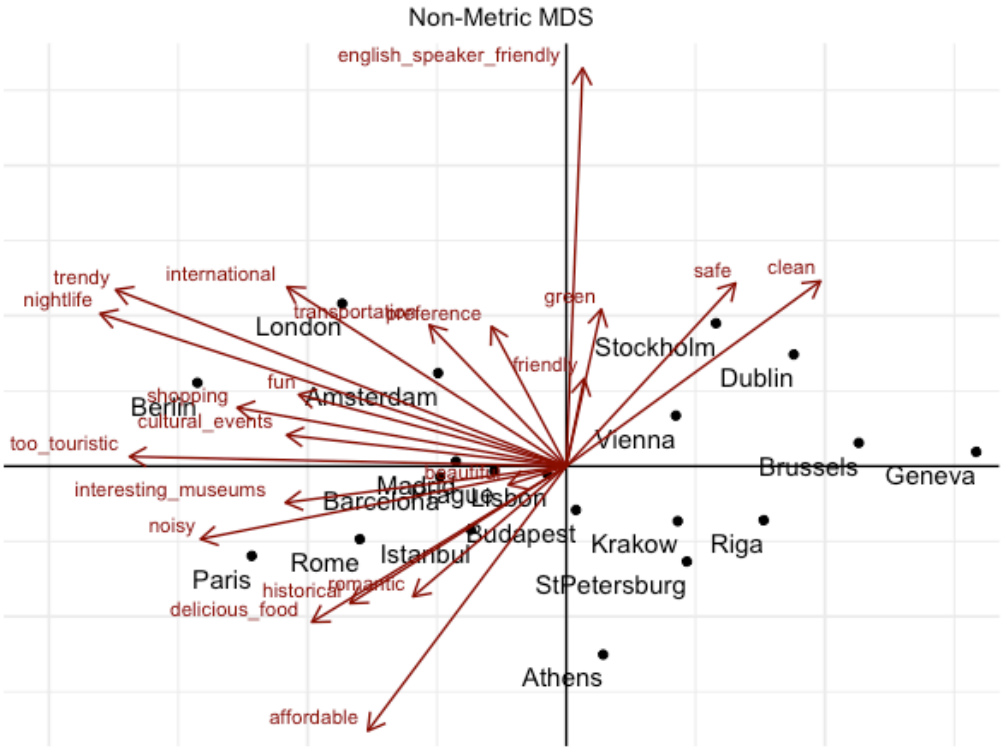

Figure 3. Results of the non-metric MDS

To interpret the visualization, one needs to project the location of the city onto the individual attribute by drawing a perpendicular line to the arrow. The distance from the city to the arrow is meaningless and irrelevant for the interpretation. The only relevant thing is the point in which the line meets the arrow.

For example, in the given graph, Brussels would be considered cleaner than Vienna, because the point of intersection is closer to the tip of the arrow, even though it is further away from the arrow itself and the direction it is pointing towards. It is also worth pointing out, that the arrows do extend past the origin, so Geneva is less trendy than Vienna.

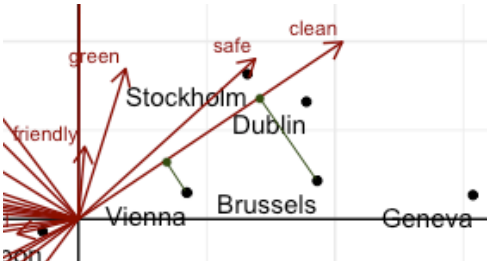

Figure 4. Projecting locations onto an attribute

The results of the property fitting are in line with what one might expect. Paris and Rome are related to the attributes ‘romantic’ and ‘historical’. London and Dublin are outstandingly ‘English speaker friendly’. Stockholm, Brussels and Geneva score high on being ‘clean’ and ‘safe’. Berlin, Amsterdam and London are both ‘fun’, ‘trendy’ and ‘international’, but also ‘too touristic’. Lastly, the cities in the second quadrant appear to be particularly correlated with the preferences for a good city trip of the queried respondents.

3.3 Key Findings and Usefulness of the Analysis

First of all, the results of the MDS visualize the similarities of the cities in the dataset. This can be used in order to build a recommender system for city trips, as it may be likely that customers will want to visit a city which is similar to the one, they have visited previously (or not). It could also prove useful to offer city trips in bundles of similar cities, as this might increase the probability of them being bought by a single customer, who’s preferences are aligned with the properties of the cities.

The property fitting allows to judge which attributes of the cities are causing their proximity in the cartesian space and also how the participants of the questionnaire view the cities. Additionally, it gives an insight into which aspects of a city are responsible for its popularity. Those insights may be used to focus on these very aspects when advertising a trip to a specific city or choose the most reasonable target audience for certain advertisements.

For a city trip provider, it can be of great value to know which cities are particularly affordable. They may want to keep their business risks as low as possible or attract younger customers with very inexpensive vacation trips.

Finally, the property fitting gives an insight into which cities are most in line with the customers preferences. This can be used to focus a marketing campaign on the less popular cities or calibrate economical decisions. A small city trip provider who is scaling his business may be interested in focusing only on the most popular of cities.

4. 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.

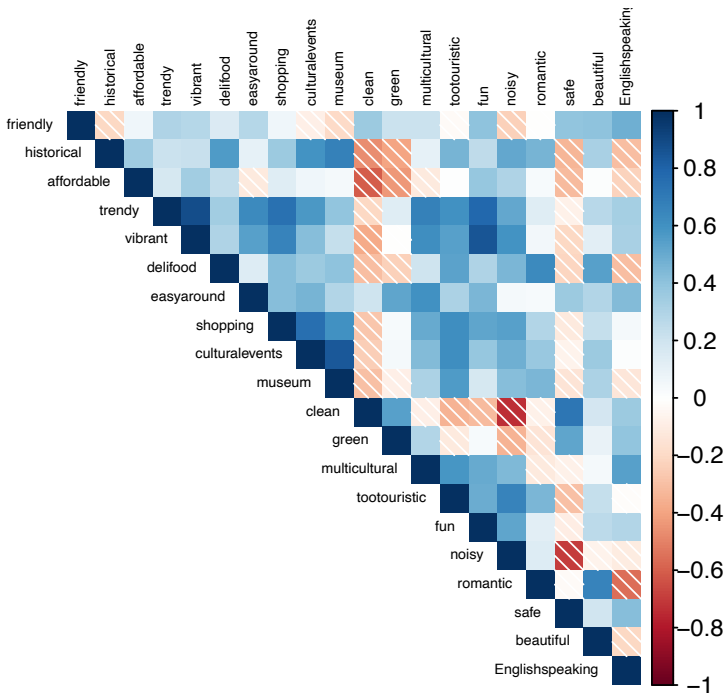

Figure 5. Correlation Plot

As shown in figure 5, 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 5 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.

4.1 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.

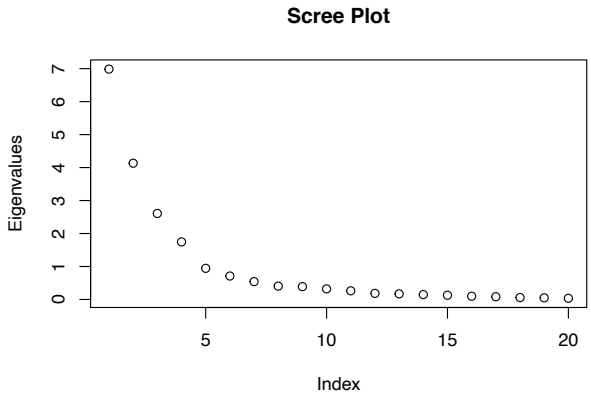

Figure 6. Scree-Test

Figure 6 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.

4.2 Factor Interpretation

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 *'urban'*. 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 *'wealthy'*. 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. We decided that **factor 3** represents the latent variable *'nostalgic'*. This leads to the hypothesis that the factor values are high here 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) which 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 these loadings, we declared **factor 4** as *'young'*. To approve the factor definitions as well as to classify the cities in the 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.

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

Figure 7. 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 which 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.

4.3. Economic Analysis

The economic implications and further usability that the clustering of the cities by Factors pose are as follows. In any given company portfolio, the marketing research team is also tasked with finding the preferences and priorities that distinct groups of customers have. Those can then be used to either improve upon a preexisting product portfolio or expand upon one, with objectively better predictive accuracy.

In this specific case, the clustering done, for the three most fitting cities regarding Factor Values for 'urban', 'wealthy', 'charming' and 'young' can be effectively used to separate the different city trips offered by a company by the demand profiles the respective customers have. In application this means, that travelers who are looking for an 'urban' city trip experience, would most likely enjoy Berlin, London and Amsterdam most as a destination. Furthermore, the product identity, in this case the offerings for trips to Berlin, London and Amsterdam should be similarly including visits to the trendy neighborhoods, Bars, Cafes and Clubs, because that is what the respective customers are most likely looking for.

So, when entering a new market, of young 'budget' travelers and students, it makes the exploration and introduction of new products and services vastly simpler, if all offered trips are structured along the recommended Factors, and the product specifications in accordance to respective customer preferences.

On a second level this analysis can also serve as a type of recommender feature, most easily used by customer representatives, when advising on a next trip. By gaining insight into previous and most well-liked travel destinations of a certain customer, the representative can give

analytically sound recommendations, based on factor clustering. A given customer that stated their favorite travel destination was Paris, due to the romantic and historical nature of the city, would be then recommended Rome as a next trip. The similarity between the two cities in primary attributes ultimately increases potential customer satisfaction.

5. Critical Comparison

In a case study like the one presented in this paper, it is important to understand the data and its underlying implications as much as possible. Therefore, performing not only MDS but also Factor Analysis leads to a more precise and thus better analysis. In the MDS method as well as in the Factor Analysis we try to identify similarities between the city regarding every attribute given in the data. The visualization derived from the property fitting is less intuitive as it demands an additional application of orthogonal lines in order to interpret the results properly. However, when performed properly it allows for deeper insights into the interdependencies. MDS itself is rather impractical in a real-life case study, because by its nature it does not group or cluster the results at all. Hence, the Factor Analysis is a helpful way to make the analysis more applicable to the case study. Through the Factor Analysis we are not only able to find similarities, but to also group the datapoints in a meaningful way. However, it might be problematic to only perform the Factor Analysis since it is a very subjective way of data interpretation, whereas MDS as a method produces results that are inherently numerically accurate and leave no room for further interpretation. As a result, the methods complement each other rather than contradict one another.

For Data Analysis of any scale the interpretability of any given results is always to be set in a context regarding the data basis. In our case we did choose to prefilter our data basis into a subset, which implicitly reduces the quantity of interpretable datapoints. The question of representativeness of any subset is always dependent on two main factors: The ratio of data reduction between pre-selection and post-selection datapoints. And the goal that is intended for that selection procedure. In our case we filtered out based on the target, to interpret only students from Germany, and reduced the data basis by a few hundred datasets. Since the overall magnitude of datapoints stayed roughly the same, the informative value is also roughly similar. This is only an issue if the reduction changes orders of magnitude (from 10^6 to 100 datasets for example).

The more general issue that overarches this entire data analysis is the question, whether initial representativeness of the given dataset existed in the first place. A survey with only 266 participants is not necessarily representative, especially in a Market research/Market segmentation analysis. While it might be, there is doubt to the informative value, given there is no information about the pre-selection of questionnaire participants and whether that selection was made in a representative manner. So, based on the assumption that the initial given dataset, the city trip questionnaire, was sufficiently representative, we can conclude, that based upon our professional assessments, the results are to be described as sufficiently representative and of economical information value, given the analysis.