A Work Project, presented as part of the requirements for the Award of a Master's degree in Business Analytics from the Nova School of Business and Economics.

TOWARDS THE EXPANSION TO THE UPCOMING CITIES:

A CLUSTERING APPROACH FOR LUGGit

RITA DE ALMEIDA FRECHES

Work project carried out under the supervision of:

Qiwei Han

Ricardo Figueiredo

17-12-2021

Acknowledgments: To my advisor Qiwei Han for the incessant support provided throughout this thesis and to Ricardo Figueiredo, Diogo Correia, Miguel Santos, and all the members of LUGGit for presenting me this opportunity and guiding me during the whole process.

Abstract: Acknowledging the success of clustering techniques as decision support tools, this paper proposes the development of an enhanced K-means algorithm to resolve LUGGit's problem of expansion. With the intent of identifying the cities that most accurately meet the company's expectations, an extensive process of data collection, reflecting a wide-ranging market-study, was on the basis of the creation of the "Weighted K-means", a clustering method capable of weighting the various attributes based on their relative significance to each member of the team, being adjustable to the present and the future needs of the company.

Keywords: Data Science, Business Analytics, Cluster Analysis, Data-Driven Business Decisions, Sensitivity Analysis.

This work used infrastructure and resources funded by Fundação para a Ciência e a Tecnologia (UID/ECO/00124/2013, UID/ECO/00124/2019 and Social Sciences DataLab, Project 22209), POR Lisboa (LISBOA-01-0145-FEDER-007722 and Social Sciences DataLab, Project 22209) and POR Norte (Social Sciences DataLab, Project 22209).

Table of Contents

Introduction

In a time of unprecedented tourism growth worldwide, the standard patterns of traveling are progressively changing, opening a door for businesses to meet the needs of these new-era travelers with innovative, profitable solutions. The period of stay for traditional touristic destinations is shrinking, with voyagers seeking shorter getaways to multiple destinations, in an increased urge to travel the world (Almeida et al. 2021). Naturally, this type of travelers continually searches for ways to make the most of every minute of their stay, reducing to the maximum extent the time lost dealing with logistics. There is where LUGGit plays its part by saving tourists a considerable amount of time.

LUGGit is a start-up company that offers real-time luggage collection and delivery services, with the foremost objective of resolving the time lapse between a tourist's arrival at the airport and registration in the place of accommodation. Through a mobile platform, LUGGit connects travelers with independent drivers – "Keepers" – who collect and deliver luggage at the desired place and time, further offering storage solutions in the process. Founded in 2019, the Portuguese company first established its services in Lisbon and Porto, and, earlier this year, began its international expansion by integrating the city of Vienna, Austria in its range of operations. Moreover, it is important to highlight that LUGGit's target customers are, predominantly, international plane travelers who visit these cities for short-term tourism. Thus, benefitting from the huge touristic intensity in Europe, which accounts for half of the world's tourist arrivals (World Tourism Organization 2021), this start-up has been experiencing an outstanding adherence to its services. However, the team is currently facing a complex decision, which intends to be resolved through the development of a decision support model.

To validate and scientifically support this thesis, henceforth it will be adopted a methodology based on the Design Science Research Process, presented by Ken Peffers et al. in 2007. The proposed framework aims to provide a road map for design science processes,

through the creation of artifacts designed to hasten problem-solving at the intersection of IT and businesses (Peffers et al. 2007). The outlined structure comprises six steps: Problem Identification and Motivation, Objectives of a Solution, Design and Development, Demonstration, Evaluation, and Communication. Peffers et al. explicitly stated the adaptability of this methodology to specific types of research, enabling the framework to be adjusted accordingly. Therefore, within the scope of this thesis, the events that regard demonstration and evaluation are combined in the section "Interpretation of the Results". Moreover, to provide theoretical context to the solution implemented, a review of literature is included in the "Design and Development" phase, where the method to resolve LUGGit's problem of expansion is effectively created. Finally, the step reserved for communicating the conclusions to the intended audience is accomplished by means of this thesis and was previously addressed through direct consultation with LUGGit, whose feedback is included in the abovementioned section.

Problem Identification and Motivation

In the wake of LUGGit's rapid growth, the team aims to pursue international expansion, ideally amplifying its services to include three additional cities in the next two years. Nevertheless, the selection of the upcoming cities represents a puzzling problem.

As a starting point, the company established a couple of directives to guide this decision. First of all, the cities to consider should be European cities. Secondly, to capacitate a fast and exponential growth, these cities should represent larger markets than those of Lisbon and Porto[1]. Broadly speaking, the company's potential market in each city is represented by the respective tourism intensity and, consequently, is subjected to the intrinsic seasonality. Accordingly, the curation of these cities should be based on the number of arrivals from international territories to the respective airports – a portrait of the potential market size each

[1] The decision of expanding to Vienna, although fulfilling the requirements, was also influenced by external factors that met the company's needs at the time.

city represents. Notwithstanding, the main obstacle behind the complexity of this problem is the impossibility of using LUGGit's existing data to substantiate the choice of the next cities. As it solely provides information about the services carried out in the cities of Lisbon, Porto, and Vienna, no comparison embracing all the possibilities could be undertaken and, for this matter, additional data is essential. Under these circumstances, the question this thesis intends to answer is fairly straightforward: How should LUGGit's business grow?

Objectives of a Solution

Foremost, the aim of a solution is to provide a data-driven answer to the problem raised. More specifically, the approach proposed has two major objectives: to gather information to sustain the decision and, secondly, to retrieve insights from that data, using a decision support model that enables to compare the possible cities for the expansion of LUGGit's businesses and infer a selection of the three most advantageous. Furthermore, the collection of data to assess this challenge can represent a useful font of information, not only for the resolution of this business problem, but also to pilot the company's future operations. Besides, the advanced solution should be as flexible as possible considering the phase of growth LUGGit is crossing.

Despite the demanding setting, a clustering technique is a tool with the capability to identify the most fitting cities for expansion and potentiate LUGGit's growth and profits – the ultimate goal of every firm. Through the recognition of the different groups those cities intrinsically belong to, it is possible to acknowledge the most fitting cluster, keeping in mind that, preferably, it will not include the cities where the company is already established.

Design and Development

Literature Review

Clustering is probably one of the most basic abilities of humankind. (Everitt, Landau et al. 2011). Once a new object is identified, the human mind intuitively applies knowledge about

similar objects encountered in the past, with the inherent intention of recognizing similarities and differences that could allow to classify it (Kaufman and Rousseeuw 2005). Over the latest decades, the concept of cluster analysis has been broadly discussed under the scope of various fields. Although there is not a universally accepted definition, clustering can be described, in its widest sense, as the task of organizing data into groups based on similarity, with the foremost objective of creating meaningful clusters that capture the natural structure of the data. Wherefore, the degree of association is maximal between patterns within a cluster and minimal among patterns belonging to a different cluster (Jain, Murty et al. 1999).

The far-reaching applications of clustering techniques to practical problems are predominantly twofold: for understanding and for utility. Steinbach and Kumar stated that i*n the context of understanding data, clusters are potential classes and cluster analysis is the study of techniques for automatically finding classes* (Steinbach, Kumar et al. 2006, p. 487-488). Specifically, clustering methods have played a crucial role in the Business area by easing the understanding and analysis of the large amounts of information gathered to sustain decision making – this topic will be approached in detail further in this section. Employed independently or in a combined manner, clustering for utility enables the abstraction from individual data, centering the analysis on the clusters in which the objects reside (Steinbach, Kumar et al. 2006).

The major advances in technology, combined with the rising need of classifying cases in more than three dimensions, led to the emergence of a wealth of clustering algorithms, the so-called automatic classification procedures (Kaufman and Rousseeuw 2005).

K-means is one of the oldest and most widely used algorithms for cluster analysis (Steinbach, Kumar et al. 2006). This algorithm partitions the data into a pre-defined number of non-overlapping clusters – symbolized by k –, with the premise that each observation can only be allocated to one cluster. Being a form of unsupervised learning, the assignment classes – also designated by clusters' labels – are not known *a priori* and, hence, are inferred by the

algorithm with the absence of category information (Jain 2009). To be exact, firstly *k* points are set as the initial clusters' centers – commonly labeled centroids – chosen at random or according to some heuristic procedure. Subsequently, each of the remaining observations is assigned to a cluster in a way that the distance between the data point and the centroid of that cluster is minimal. Once all points are grouped into *k* clusters, the centroids are re-calculated as the mean of all the instances belonging to that cluster, inducing an iterative process that only ceases when reassignments of clusters are no longer possible or the within-cluster variation reaches a pre-determined value and, thus, is minimized (Rokach and Maimon 2005).

Algorithm 8.1 Basic K-means algorithm.

1: Select K points as initial centroids.
2: **repeat**
3: Form K clusters by assigning each point to its closest centroid.
4: Recompute the centroid of each cluster.
5: **until** Centroids do not change.

Figure 1 - Basic K-means Algorithm (source: Steinbach and Kumar 2006)

The contributions of the K-means algorithm to the most diverse fields are undoubtedly remarkable. One of its major applications in the Business area is market segmentation, which can be defined as the process of breaking a companies' potential or effective market into segments. To this extent, clustering enables a clear understanding of prospects without the need of analyzing each case individually and the gain of additional insights through the grouping of items, leading to an effective market segmentation (Kuo et al. 2002).

Despite its recognized efficiency in resolving the clustering problem, the K-means algorithm holds certain limitations that can compromise the accuracy of its results. First of all, the assignment of the initial centroids highly impacts clusters' membership: when random initialization is used, different runs of K-means typically generate distinct outcomes (Steinbach, Kumar et al. 2006). Therefore, under the scope of this thesis, a solution advanced in the K-means++ algorithm will be employed to mitigate this problem. The K-means++ algorithm

proposes an alternative approach to the original K-means method, which allows setting the initial cluster centers in an attempt to force the centroids to be as distant as possible from one another, covering the occupied data space to the furthest extent from initialization (Arthur and Vassilvitskii 2007). Further, the algorithm is not robust to outliers, whilst the presence of these data points can substantially influence the mean value and, ultimately, the value of the centroids. Lastly, K-means does not perform well with qualitative data and can be affected by the alleged course of dimensionality, the undesirable consequence of keeping a disproportionate number of dimensions relative to that of existing observations (Han, Kamber et al. 2011).

As formerly referred, empowering a flexible process of decision-making is crucial for LUGGit. Therefore, a "Weighted K-means algorithm" was assembled to improve clustering analysis across multiple data sources and factors that might have different subjective impacts to the diverse members of the company, producing dissimilar results accordingly. Thus, the development of this algorithm was further strengthened with the application of sensitivity analysis to consolidate the distinct outcomes, a broadly undertaken method to enhance decision support tools. Within this frame of reference, sensitive analysis is defined as the practice of tracing the variation of a model's outcomes as a set of model-related assumptions change (Borgonovo and Plischke 2016). Above all, sensitivity analysis provides consistency to the conclusions inferred by the model: an outcome is considered reliable if it remains coherent throughout adjustments. Thereby, the uncertainty and subjectivity inherent to both the process of decision-making and the analysis of clustered data are diminished (Abe and Gee 2014).

Taking into consideration the theoretical context provided, clustering is undeniably an adequate tool to accurately substantiate the choice of the future cities for LUGGit's expansion.

Data Collection and Understanding

As previously stated, to develop the proposed method and fulfill the objectives abovementioned, information covering all the cities under consideration had to be gathered.

The potential growth each city represents is manifold. Apart from intuitive factors, as the extension of a city's market size, existing competition, or aspects that might directly or indirectly influence costs of operations, numerous other factors had to be considered. For instance, aspects that might impact the propensity of tourists to adhere to LUGGit's services or the probability of partnerships with accommodation businesses, which senses the complexity behind a problem that, at a first sight, might seem simple. To accomplish this exhaustive analysis, information was retrieved from the most diverse sources, covering official aviation databases, articles and publications, official webpages, statistical and financial databases, and even less conventional sources such as navigation tools, from which data was manually generated. In this way, the collection and understating of data were conducted through an iterative process, as the understanding of such information repeatedly led to the necessity of collecting additional data to deepen the analysis until the final dataset was accomplished.

As prior established, the selection of the feasible cities was founded on the number of international arrivals to the respective airports and, wherefore, the 20 cities that revealed the highest number were preferred[2]. Compounded with Lisbon, Porto, and Vienna, the 3 cities where LUGGit already operates, 23 cities were under the scope of this analysis.

With respect to the attributes, they were fashioned into 7 general categories. The unique identifier of each instance is given by the combination of the columns "City" and "Country", comprised in the group "Identifiers". Each of the remaining 6 categories intends to represent a genre of features that, directly or indirectly, impact the company's business.

The attributes that regard a city's airports – LUGGit's preferable place to target its audience – were incorporated in the category "Airports". This class includes the indispensable variable "No. International Arrivals" used to determine the possible destinations towards expansion, along with the columns "Distance Airport-CC" and "Time Airport-CC", which

[2] An illustration of this criterion is provided in Appendix I, Figure 9.

characterize the route between the airports and the city center. Hence, these variables provide crucial insights by reflecting the distance and time "Keepers" need to transverse. The following category – "Airbnb" – covers the attributes that specify the way and the circumstances under which Airbnb operates in that city, containing the column "No. Airbnb Listings", which improves the perception of a city's tourism intensity by assessing the volume and occupancy of Airbnb's. In addition, the data within the category "Competition" reveals insights about the existence of companies that perform similar services, focusing on LUGGit's two main competitors. Furthermore, to address this problem in the most substantiated way possible, additional information is included in the remaining categories: "Operations", "Cost of Living", and "Additional Characteristics". In order to provide a full understanding, the entire set of variables used, along with a detailed description, is provided in Appendix I, Table 5.

Ultimately, it is important to note that this wide range of data sources granted these variables different relative importance in the process of decision-making, demanding a model capable of weighting the different team members' perceptions of the prime factors.

Data Curation

The quality of the results outputted is highly dependent on the quality of the inputted data, thus, the next step was to ensure that the dataset was properly treated. In this case, the curation of the data was not extensive, as it was gathered bearing in mind its future application.

The first attempted step was treating the existing missing values, which were replaced by zero since it mirrored its true meaning. Further, the feature "Airbnb Legal Barriers" was manually transformed into a numerical variable, allowing its posterior use in the clustering algorithm. The diverse legal restrictions were ranked by their level of strictness – attained through research –, which allowed to measure the rigidity imposed in each city and, for the cities that presented more than one restriction, their values were summed – Appendix I, figure 10. Finally, the assessment of the data utility was a continuous process, following the constant

adjustment of the business' needs. The attributes proven to be no longer relevant were dropped, namely the columns "Entrepreneurial Ecosystem", "Ease of Expansion", and "Topography". Similarly, the variable "Indirect Competition" was removed, as it held the same estimate for all cities and, thusly, added no value to the reasoning. A cleaned and treated dataset made possible the visualization of the data, which will be object of analysis in the next section.

Descriptive Statistics and Exploratory Data Analysis

Having a first look at the dataset, one can notice the heterogeneity of values across the different cities. Consequently, before diving deeper into the solution developed, it is important to conduct a detailed analysis of the dataset, the variables that compose it, and possible relations between them, to discover patterns hidden in the data and test hypothesis that can help identify the advantages and disadvantages behind that heterogeneity.

The first analysis concerns the legal barriers imposed towards Airbnb, along with the number of active listings on the platform. This first variable approximates potential variations in the market size in the long run, as it represents the rigidity of the limitations to the growth, and even maintenance, of the number of existing Airbnb's, while the second approximates a city's current market size. Figure 2 illustrates the values assigned to each city, sorted by their level of strictness, as well as the number of active listings.

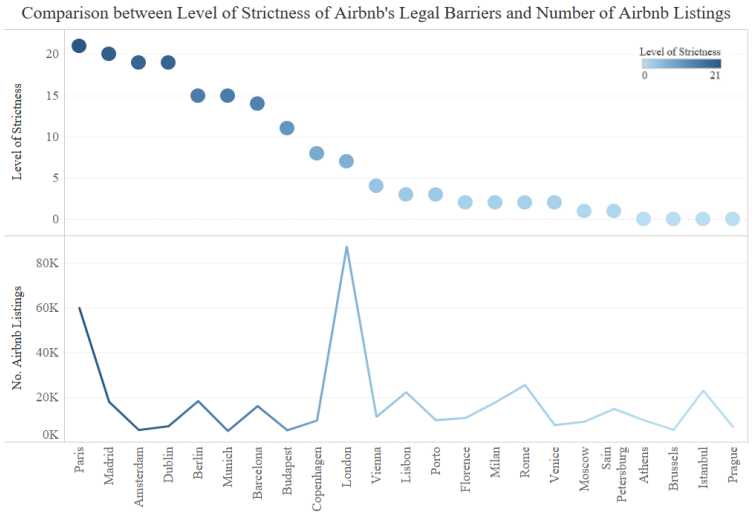

Figure 2 - Resemblance between Airbnb Legal Barriers and Number of Active Listings in the Platform

In this fashion, there is a high probability of growth stagnation in Paris. However, this high level of strictness represents a more serious drawback in the cities with minor tourism intensity, in particular, Madrid, Amsterdam, and Dublin. Moreover, the absence of Airbnb restrictions, or its low level of rigidity, alongside a significant market extent is noteworthy, especially in Rome and Istanbul, large cities with margin to expand its tourism.

The subsequent analysis aims to understand if there is a trade-off between a city's market size and the minimum time taken to complete a service, demonstrated in Figure 3.

Figure 3 - Number of International Arrivals and Time taken from the Airport to the City Center by car

In fact, one can state that the minimum duration of each service is, from a general point of view, higher when the number of international arrivals is increased. Even though a longer service also presupposes a greater price, it does not necessarily mean that it is more profitable for LUGGit due to the associated costs and, thereby, a shorter time span for each service is preferable. In this way, the cities of Paris, Rome, and Prague hold a considerable advantage, as they present a low value regarding the time that comprises the distance from the airport to the city center and a substantial market extent in comparison. Additionally, it is evident which cities account for the greatest number of international arrivals: London and Paris are evidently the

"biggest" cities, followed by Istanbul and Rome, data points that clearly represent outliers.

Finally, the average Airbnb's occupancy was examined with the intention of identifying the cities with better performance and to comprehend if the presence of seasonality in the cities with greater tourism concentration can be perceived through this variable – Figure 4.

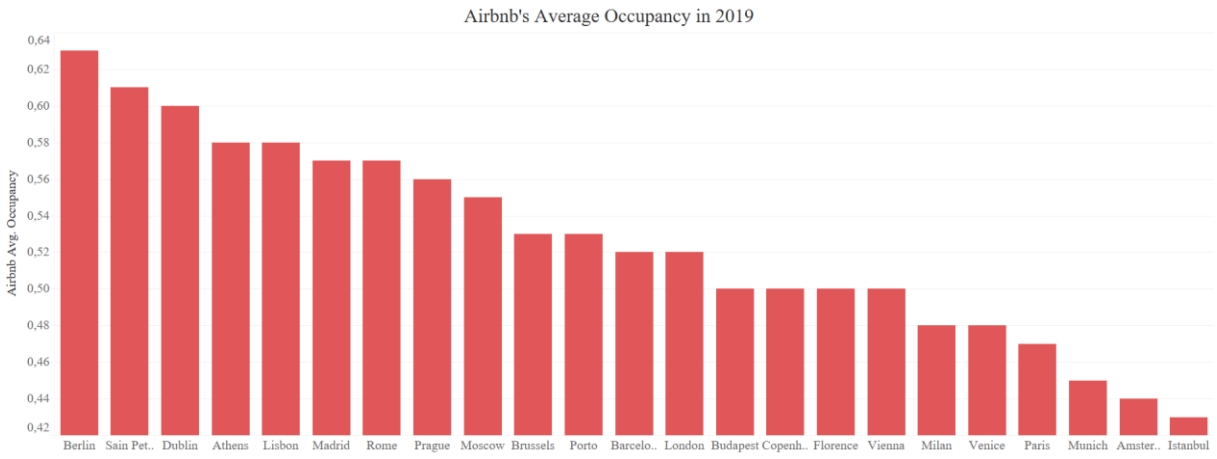

Figure 4 - Average Occupancy of Airbnb's in 2019, per city

Regarding the year of 2019, pre-pandemic, Berlin exhibits the highest average, with 63% of occupancy of its Airbnb listings, followed by Saint Petersburg, with a value of 61%. In contrast, Istanbul manifests the lowest percentage – although not insufficient, as it still presents 43% of occupancy of its vast number of listings. Thereby, one could assume a high probability of Istanbul suffering from a more intense variation of the number of tourists throughout the year. This assumption was confirmed through research and, indeed, Istanbul is a city with an acknowledged division between the so-called high season and low season, which substantiates the utility of this attribute as a perception of a large city's exitance of seasonality.

The insights retrieved from these analyses are crucial to the future interpretation of results. However, by exploring the dataset in more detail, one can also recognize the endless interconnections between variables: the time taken to complete each service is subject to the propensity of a city to register traffic congestion; likewise, the prior mentioned service cost in each city is dependent on the respective costs of performing and maintaining these operations, namely, the average monthly salary, the gas price, and the average cost per click; even the

market size each city represents, to be accurately measured, has to take into account the variables that indirectly influence it. This interdependence reaffirms the importance of a model-based solution, where all relations are considered and properly weighted to the fullest extent.

Data Modeling

As reasoned before, a clustering technique, specifically the K-means algorithm, fulfills the requirements to resolve this problem. Nevertheless, the methodology here proposed intends to go beyond this method of cluster analysis, employing the reasoning formerly described in the "Literature Review" section, although slightly altering it to make it perfectly adjustable to meet LUGGit's needs and interests. Undoubtedly, for the selection of the preferred cities for expansion a wealth number of factors need to be considered and, naturally, the team sought to include all the relevant attributes into the analysis, nonetheless, there is a perception of the most relevant factors to the business. Thence, the "Weighted K-means algorithm" empowers the addition of the relative importance of each attribute into the K-means model, through the assignment of a relative weight to each feature. To provide a full understanding of the method implemented, primarily the logic behind it will be described.

Weighted K-means Logic

Mathematically, *n* objects, represented as vectors of *p* attributes, are grouped into *k* clusters by assigning these objects to the closest centroid, based on a measure of similarity. In this regard, the Euclidean distance will be used to prescribe the proximities between the data points and the centers of the clusters and, within this distance formula, the relative weight of each feature – defined as *w* –, is applied, as represented in the equation below.

$$d(x, c) = \sqrt{\sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 * w}$$

Equation 1 - Weighted K-means Algorithm Distance Formula

Specifically, each of the 23 cities under analysis (*n*), represented by a vector containing the values of the 24 attributes (*x*), is allocated to the cluster with the nearest centroid (*c*). This proximity is measured taking into account the relative importance of the attributes, mapped into a vector with 24 weights (*w*), regarding each of the variables. In Appendix II, Figure 11 the code developed in Python is provided.

Relative Importance of the Attributes

As abovementioned, this modified approach of the K-means algorithm was developed with the purpose of completely aligning the clustering algorithm with the needs of the company, therefore, the assignment of the relative importance of each column in the dataset was settled by diverse members of LUGGit – a table with the variables' relative impact, perceived by each member, is given in Appendix II, Table 6. The weights were assigned considering an interval ranging from 0 to 2, where a value below 1 presupposes a less significant variable compared to the remaining and a value above 1 infers a greater relative importance. Hereby, the attributes with a weight of zero are not considered by the model, operating as a form of regularization.

Moreover, acknowledging LUGGit's early stage of growth and the recent expansion to a larger city, alongside the subjective nature of the weights, developing a method flexible enough to adjust itself to the perception of the most substantial variables was crucial to guarantee its utility. In this fashion, an interactive property was attached to the model to allow the weights to vary accordingly, enabling the visualization of the results being promptly recalculated as the relative significance of each column changes. The practicality of this property is twofold: on the one hand, the diverse weights perceived by each member of the team can be tested, on the other hand, the model is adjustable enough to meet the expectations of the company as its needs evolve. Besides, it shapes this data-science-based model into an intelligible tool for any member of the company.

Optimal Number of Clusters and Initialization of the Centroids

Before the deployment of the model, it is necessary to pre-determine the number of clusters to divide the data into (k). Hence, the optimal number was ascertained by applying the Elbow Method, considering 2 to 5 clusters, and, from its analysis, it was possible to conclude that dividing the data into 3 or 4 clusters would grant the most reliable conclusions – Figure 12 in Appendix II illustrates these outcomes. In addition, as previously stated, the determination of the initial clusters' centers intended to position these centroids as distant as possible from one another, targeting a group of dissimilar cities. For this matter, two considerably large cities, alongside two comparatively smaller ones, were assigned as the initial clusters' centers.

Model Deployment

Once all the processes preliminary to the model were completed, one last step was required to ensure the data was suitable to fit the model. Considering that K-means and, as a consequence, the "Weighted K-means", are not algorithms robust to outliers and do not perform well with qualitative data, to diminish the impact of the outliers present in certain attributes, the Robust Scaler from scikit-learn library was applied to the numerical features, along with One Hot Encoding to transform the categorical attributes not priorly treated into numerical ones.

After passing through this pre-processing pipeline, the data was finally inputted into the "Weighted K-means algorithm", which was initially deployed considering each variable as equally significant and, afterward, attempting the relative weights provided by LUGGit's CEO, COO, and Head of Operations, creating 3 and 4 clusters for each case. For the demonstration of these clusters, the Principal Component Analysis method was applied, projecting data into a two-dimensional space and, thus, enabling its visualization.

In the forthcoming chapter, the resulting clusters and respective analyses are presented, ultimately disclosing the most fitting cities for LUGGit's expansion.

Interpretation of the Results

First and foremost, as ambitioned, the algorithm produced distinct results depending on the inputted relative weights, especially when all attributes were defined as equally important.

The nomenclature of the resulting clusters mirrors the value of its centers in a broad manner and is common to all weighting strategies: in the "Growth Maintenance" cluster are included the cities that would allow LUGGit to maintain its growth rhythm; the cities within the "Smaller and Expensive" cluster are, typically, cities with a smaller market size and comparatively higher costs of living and operations; finally, to the "Best Cities" cluster are allocated those cities that, based on this data and the company's present interests, would potentiate LUGGit's growth to its maximum extent. Further, to identify the preferable cities within this cluster its distance to the centroids was calculated, in order to determine those closest to the center, that is to say, the optimal cities. The outcomes obtained implementing the strategy of even relative weights will not be analyzed, as they proved to be less adequate to the interests of the company – they are illustrated in Appendix II, Figure 13 and Figure 14 for comparison purposes and validation of the model.

Relative Weights provided by LUGGit's CEO

The results represented in the below figure depict LUGGit's CEO perception of the most notable attributes. The variable that regards a city's number of international arrivals was considered by the three team members the most important feature, in this case accounting a relative weight of 1.9, and, subsequently, to the restrictions towards Airbnb was assigned a relative weight of 1.7. The average cost-per-click and average monthly net salary succeeded as two of the most relevant factors for the CEO of the company, in the respective order, followed by the number of active Airbnb listings, reaffirming the importance of a city's market size. Moreover, to the minimum time required to complete a service and to the number of direct competitors were attributed significant relative weights, of 1.4 and 1.2, respectively.

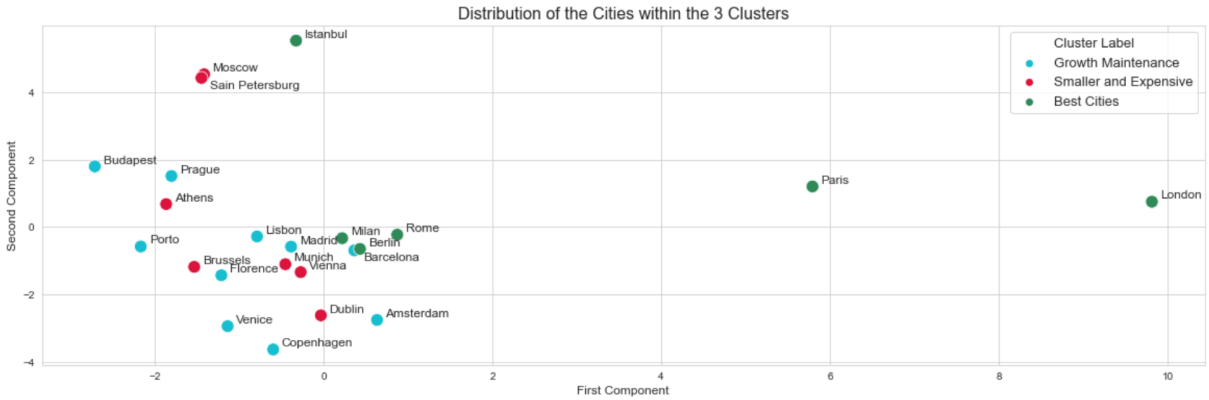

Figure 5 - Weighted K-Means Outcome considering 3 Clusters and the Relative Weights provided by LUGGit's CEO

Within this context, the cluster containing the recommended destinations for expansion includes six cities, namely, Istanbul, Rome, Milan, Paris, Berlin, and, London, ordered by their proximity to the centroid. Assigning the almost maximum relative weight to one of the attributes representative of the market size, Istanbul, one of the largest cities with low costs of living and operations, is pointed as the most profitable city. Additionally, it is important to note that, in comparison to the outcomes of the equal weights' strategy, Saint Petersburg shifts from the "Best Cities" cluster to the "Smaller and Expensive" one, which could be explained by the considerable amount of time necessary to travel from the airport to the center of Saint Petersburg, not possible to counterbalance by its smaller market extent.

Attribute Name	***Best Cities***	***Growth Maintenance***	***Smaller and Expensive***
No. International Arrivals	12 746.82 K	5 465.95 K	5 329.40 K
Time Airport-CC (minutes)	55.00	24.00	37.14
Airbnb Legal Barriers	7.833	8.20	5.71
Average CPC (US $)	0.85	0.79	0.90
Avg. M. Salary (€)	2 186.31	1 703.22	1 908.26
Direct Competition	1.83	1.4	0.43

Table 1 - Values of the Centroids considering 3 Clusters and the Weights provided by LUGGit's CEO

A portion of the centroids of each cluster is put forward in Table 1. Examining the values of the favored cluster, one can notice that, although presenting a significantly greater number of international arrivals, this leverage has associated disadvantages, in particular, the high costs

of operations – both the salary and the average cost-per-click, variables the CEO considered to be of extreme significance –, the minimal services' duration, and the average number of direct competitors. Thereby, these results do not represent a feasible solution within this context.

Figure 6 - Weighted K-Means Outcome considering 4 Clusters and the Relative Weights provided by LUGGit's CEO

Considering the partition of the data into 4 clusters, Paris and London are removed from the optimal group and reassigned to a fourth cluster labeled "Giant Cities". The dissimilarities between these two destinations and the remaining are evident in Figure 6 and, as referenced before, they justify several of the outliers in the dataset, concerning not merely the market size, but also the costs of living and maintaining the services, especially for London.

Attribute Name	***Best Cities***	***Growth Maintenance***	***Smaller and Expensive***	***Giant Cities***
No. International Arrivals	9 458 K	5 465.95 K	5 329.40 K	19 323.90 K
Time Airport-CC (minutes)	50.00	24.00	37.14	42.5
Airbnb Legal Barriers	4.75	8.2	5.71	14.00
Average CPC (US $)	0.67	0.79	0.90	1.23
Avg. M. Salary (€)	1 640.45	1 703.22	1 908.26	3 278.03
Direct Competition	0.50	1.4	0.43	4.5

Table 2 - Values of the Centroids considering 4 Clusters and the Weights provided by LUGGit's CEO

The withdrawal of Paris and London empowered the decrease of those undesirable high values: the estimated cost of hiring the "Keepers" notably declined by more than 500€, likewise, the average level of strictness applied to Airbnb services and average number of existing direct

competitors were drastically reduced. Although still elevated, even the average value of "Time Airport-CC" for that cluster went from 55 minutes to 50 minutes. Hence, within this weighting strategy, it is clear that the growth these two cities might proportion would lead LUGGit to entail elevated costs. Naturally, the extraordinary number of international arrivals previously displayed also decreased, however, the present average value is still considerably above those of Lisbon, Porto, and Vienna. Besides, as the centroid is no longer influenced by London's and Paris' huge potential markets, considering this methodology Rome becomes the most profitable destination towards expansion, followed by Milan, Berlin, and, at last, Istanbul.

Acknowledging the improvement achieved by dividing the data into 4 clusters and the fact that the "Growth Maintenance" and the "Smaller and Expensive" clusters remained unaltered, this will be the methodology henceforth implemented. The succeeding analysis concerns the outputted clusters based on the relative weights assigned by LUGGit's COO.

Relative Weights provided by LUGGit's COO

The number of international arrivals persisted the most significant factor, with a relative importance measured in 1.7. Afterward, to the average occupancy of Airbnb's was assigned a relative weight of 1.6. Lastly, it is worth mentioning the increased relative significance of the variable "Traffic Percentage", with a relative weight of 1.4 perceived by the company's COO.

Figure 7 - Weighted K-Means Outcome considering 4 Clusters and the Relative Weights provided by LUGGit's COO

In this framework, Rome prevails the most fitting city, nevertheless, Berlin is appointed as the second city closest to the center of the cluster, followed by Saint Petersburg and Istanbul, respectively. One possible reasoning for this variation within the "Best Cities" cluster in relation to the former outcome, can be deduced from, firstly, the lower relative weight assigned to the number of tourist arrivals, making Saint Petersburg – a comparatively smaller city – suitable for expansion. Secondly, the enhanced value assigned to the average Airbnb's occupancy rate, passing from 0.8 to 1.6, might sense the allocation of Berlin in second place. Ultimately, the increased relative importance of a city's propensity to register traffic congestion could explain why Istanbul, although remaining in the favored cluster, is no longer one of the three cities preferable for expansion, as it accounts for the highest value within this variable.

With regard to the key cluster and comparatively to the prior outcomes, the present weighting strategy enabled the reduction of both the minimum time required for each service and the cost of labor, at the expense of a smaller – yet high – number of international arrivals. This deviation is product of the shift between Milan and Saint Petersburg within the cluster.

Attribute Name	***Best Cities***	***Growth Maintenance***	***Smaller and Expensive***	***Giant Cities***
No. International Arrivals	8 921.48 K	5 658.22 K	5 242.78 K	19 323.90 K
Time Airport-CC (minutes)	47.50	27.92	35.00	42.5
Avg. M. Salary (€)	1362.75	1746.28	2109.15	3 278.03

Table 3 - Values of the Centroids considering 4 Clusters and the Weights provided by LUGGit's COO

Finally, one last comparative analysis, portraying the most valuable features for LUGGit's Head of Operations, will be carried out before accurately identifying the best cities for the company's expansion, based on the methodology developed.

Relative Weights provided by LUGGit's HOO

In this respect, the number of tourist arrivals holds the utmost relative weight. Further, the distance and time that characterize the route from the airport to the centers of the cities were

extremely valued, presenting relative weights equal to 1.7., succeeded by the variable "Airbnb Legal Barriers", whose relative importance was measured in 1.3. Contrarily, the start-up's HOO considered the columns "Traffic Percentage" and "Direct Competition" less substantial, attributing relative weights of 0.7 and 0.3, respectively. At last, both the price index and average monthly net salary were excluded from the analysis.

Figure 8 - Weighted K-Means Outcome considering 4 Clusters and the Relative Weights provided by LUGGit's HOO

On this basis, five cities compose the core cluster under analysis and, reiteratively, Rome is the first recommended city for expansion. The second city nearest to the centroid is Istanbul, succeeded by Milan, Berlin, and Paris, here ordered by their proximity to the cluster center. Oppositely to the former strategies, in the present approach the factors that regard the cost of living and labor were considered of small significance, which could perfectly explain Paris allocation to the preferred cluster. Accordingly, this weighting strategy provides precious insights, proving that, if the company considers incurring greater expenses, Paris could be one of the cities to consider for expansion purposes based on this methodology, however, still not one of the most fitting cities, most probably due to, in this case, its high level of strictness towards Airbnb. Moreover, Istanbul's position could be justified by the assignment of a minor relative weight to the attribute "Traffic Percentage". Ultimately, as expected, the centroid being analyzed presents a considerably greater value concerning cost-related variables and, at the same time, with regard to the features that approximate a city's market size, when contrasted with the centers resulting from the two previous approaches.

Recommendation of the Upcoming Cities for LUGGit's Expansion

After a careful analysis of the model's outcomes, considering the diverse weighting strategies individually, one can notice the subjectivity inherent to this process. Hence, to objectively identify the right course of action for LUGGit's expansion, it is crucial to establish the most fitting cities taking into account all the results simultaneously, assessing the prevalence of those cities included, at least once, in the foremost cluster.

Ranking	***Cities***
1.	Rome
2.	Berlin
3.	Istanbul
4.	Milan
5.	Saint Petersburg
6.	Paris

Table 4 - Ranking of Recommended Cities

Under the assumptions of this methodology and the company's present interests, three cities prevailed in the preferred cluster as the perceptual significance of the attributes oscillated. Unquestionably, Rome proved to meet LUGGit's needs, whether implementing a more conservative approach, heavily weighting the factors that forecast possible threats and excessive costs, or a riskier one, valorizing a rapid growth. Being the city with the fourth highest number of tourist arrivals and the third with more active Airbnb listings, Rome undoubtedly represents a large market for the company in the present day and, currently holding almost no restrictions towards the proliferation of Airbnb, for the future as well. Moreover, with a relatively low average regarding the monthly net salary and the cost-per-click, alongside a not-so-high gas price and a reduced minimum duration for each service, this city is not an expensive one to maintain operations. Therefore, Rome is the recommended first city for expansion.

As the amplification to the three cities will not be carried out simultaneously, the choice of the remaining two cities is influenced by the first destination. However, within the present

circumstances, Berlin is the recommended second city for the company to expand its operations to. This city represents a relatively smaller market with regard to the number of international arrivals and active Airbnb listings, albeit significantly larger than those of Lisbon, Porto, and Vienna. Conversely, accounts for the highest percentage of Airbnb's occupancy, which reasserts its grand tourism intensity. Another of Berlin's greatest advantages is having one of the smallest distances between the airport and the city center, which substantially reduces service's costs. Further, the advised third city for expansion is Istanbul, one of the largest cities with no current restrictions imposed on the growth or maintenance of Airbnb. The main advantages of this city are the extremely low costs of maintaining operations and of living, giving a great margin for tourists to adhere to LUGGit's services. On the other hand, the principal drawbacks of Istanbul are the long distance and duration of the route between the airport the city center, along with an increased percentage of traffic congestion. Finally, Milan, Saint Petersburg, and Paris also proved to be cities to consider in the future.

Challenges and Limitations

As aforementioned, the major challenge encountered throughout the development of this thesis was due to the lack of existing data capable of supporting the resolution of the problem proposed. As a consequence, the collection of such data represented a complex and long-lasting process, comprising information retrieved from a wide variety of sources, aligned with a wide-ranging research, with the intent of covering all the factors potentially decisive.

Furthermore, the uncertainty regarding the impact of each attribute constituted an arduous limitation to overcome, demanding the creation of an adjustable solution, flexible enough to integrate the diverse perceptions of the team members and meet future needs.

Lastly, the identification of the cluster that accurately met LUGGit's expectations and interests likewise represented a challenge by virtue of the subjectivity inherent to the resolution of selection business problems founded on clustering techniques.

Recommendations for Future Steps

With the cities that fulfill the expectations of the company identified through the model, an immediate step succeeds. Making use of existing data respecting the past services carried out until this date, already including the operations conducted in the newly city of Vienna, an estimation of the potential profitability of the recommended cities should be addressed to guarantee the accuracy of the solution proposed, making use of information the present methodology could not benefit from. Further, a more ambitious recommendation is to infer the relative importance of each variable exploiting this same data, in an attempt to understand which factors concerning the cities where LUGGit already operates are effectively impactful.

Ultimately, with the intention of delineating the most meticulous course of action and fully benefit from the practicability of the method created, in a subsequent phase to that of the expansion towards the first city, it is advised a new deployment of the model, this time taking into account the knowledge acquired with the first expansion and the consequent fresh perception of the relative significance of each attribute.

Conclusion

In essence, an extensive process of data collection, comprising a wide variety of data sources, delineated the methodology implemented to resolve LUGGit's problem of expansion. Empowering a substantiated analysis of the most fitting cities, simultaneously conferred subjective importance to the attributes, as team members perceived different prime factors for the expansion. To this extent, acknowledging the success of clustering techniques as decision support tools, an enhanced clustering algorithm was developed in an attempt to weigh the different perceptions. Through this methodology, strengthened by a sensitivity analysis approach, it was possible to overcome the foremost challenges and limitations and identify the cities that most accurately fulfill LUGGit's present expectations, as well as provide a decision-making tool capable of meeting the future needs of the company.

References

Literature References

Abe, Yasuyo and Kevin A. Gee. 2014. “Sensitivity analyses for clustered data: An illustration from a large-scale clustered randomized controlled trial in education”. *Evaluation and Program Planning* 47: 26-34.

Almeida, António et al. 2021. “Factors explaining length of stay: Lessons to be learnt from Madeira Island”. *Annals of Tourism Research Empirical Insights* 2, no. 1.

Arthur, David and Sergei Vassilvitskii. 2007. “k-means++: the advantages of careful seeding”. SODA '07.

Borgonovo, Emanuele and Elmar Plischke. 2016. “Sensitivity analysis: A review of recent advances”. *European Journal of Operational Research* 248, no. 3 (February): 869-887.

Everitt, Brian, Sabine Landau, et al. 2011. *Cluster Analysis, 5th Edition.* Wiley Series in Probability and Statistics. London: John Wiley & Sons, Inc.

Han, Jiawei, Micheline Kamber et al. 2011. *Data Mining Concepts and Techniques, Third Edition*. United States of America: Morgan Kaufmann Publishers.

Hevner, A. R. et al. 2004. “Design Science in Information Systems Research”. *MIS Quarterly* 28, no. 1: 75-105.

Jain, A. K., M. N. Murty, et al. 1999. "Data clustering: A review" *ACM Computing Surveys* 31(3): 264-323.

Jain, Anil K. 2009. “Data clustering: 50 years beyond K-means”. *Pattern Recognition Letters* 31, no. 8 (June): 651-666.

Kaufman, Leonard and Peter J. Rousseeuw. 2005. *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Kuo, R. J. et al. 2002. "Integration of self-organizing feature map and K-means algorithm for market segmentation". *Computers & Operations Research* 29, no. 11 (September): 1475-1493.

Peffers, Ken et al. 2007. "A design science research methodology for information systems research". *Journal of management information systems* 24, no. 3: 45–77.

Provost, Foster and Tom Fawcett. 2013. "Data Science and its Relationship to Big Data and Data-Driven Decision Making". *Big Data* 1, no. 1.

Rokach, Lion and Oded Maimon. 2005. "Data Mining and Knowledge Discovery Handbook", 321-352. New York: Oxford University Press, Inc.

Steinbach, Michael, V. Kumar, et al. 2006. *Introduction to Data Mining*. United States of America: Pearson.

World Tourism Organization. 2021. *European Union Tourism Trends*. Madrid: World Tourism Organization.

Data Collection References

Airbnb. 2021. "Amsterdam". Airbnb. https://www.airbnb.pt/help/article/860/amesterd%C3%A3o (accessed October 7, 2021).

Airbnb. 2021. "Hospedagem responsável na Rússia". Airbnb. https://www.airbnb.pt/help/article/2150/hospedagem-respons%C3%A1vel-na-r%C3%BAssia?_set_bev_on_new_domain=1610018182_Y2UwY2I2ZmRkYjY5 (accessed October 7, 2021).

Airbnb. 2021. “Munich”. Airbnb. https://www.airbnb.pt/help/article/1239/munique?_set_bev_on_new_domain=1610018182_Y2UwY2I2ZmRkYjY5 (accessed October 7, 2021).

AirMundo. 2020. “Top 32 most visited cities in Europe 2019”. AirMundo. https://airmundo.com/en/blog/most-visited-cities-in-europe/ (accessed September 23, 2021).

Airportr Technologies Limited. 2021. “Airline and government approved”. Airportr. https://airportr.com/en/ (accessed September 30, 2021).

B., Elena. 2015-2021. “Airbnb Management Software: Property management software for Airbnb Business”. Hosty. https://www.hostyapp.com/ (accessed September 23, 2021).

BoB. 2021. “Airports”. BoB. https://bob.io/airports/?lang=en (accessed September 30, 2021).

Bodkin, Peter. 2019. “Explainer: The new rules on Airbnb hosting come into effect today - here's what you need to know”. TheJournal.ie. https://www.thejournal.ie/explainer-short-term-lets-4704793-Jul2019/ (accessed October 7, 2021).

Cox, Murray. 2019. “Inside Airbnb. Adding data to the debate”. Inside Airbnb. http://insideairbnb.com/index.html (accessed September 23, 2021).

Emerging Europe. 2020. “Budapest becomes latest city to clamp down on Airbnb”. Emerging Europe. https://emerging-europe.com/business/budapest-becomes-latest-city-to-clamp-down-on-airbnb/ (accessed October 7, 2021).

Gesley, Jenny. 2016. “Germany: Law Restricting Airbnb and Other Vacation Rentals Takes Effect in Berlin”. The Library of Congress. https://www.loc.gov/item/global-legal-monitor/2016-05-24/germany-law-restricting-airbnb-and-other-vacation-rentals-takes-effect-in-berlin/ (accessed October 7, 2021).

Google. 2017. Google Maps. https://www.google.com/maps (accessed September 25, 2021).

Kaszás, Fanni. 2020. "Goodbye Airbnb? Budapest Plans to Limit Short-Term Renting". Hungary today. https://hungarytoday.hu/airbnb-budapest-hungary-restrictions-plan/ (accessed October 7, 2021).

Kelly, Olivia. 2020. "Airbnb-style short-term rentals to be refused applications in Dublin". The Irish Times. https://www.irishtimes.com/news/environment/airbnb-style-short-term-rentals-to-be-refused-applications-in-dublin-1.4178857 (accessed October 7, 2021).

Közgazdász. 2021. "New Regulation of Short-Term Rental in Budapest from 2020". Whispering Tree. https://www.whisperingtree.hu/blog/new-regulation-of-short-term-rental-in-budapest-2020/ (accessed October 7, 2021).

Lock., S.. 2021. "Airbnb listings in Europe by city 2020". Statista. https://www.statista.com/statistics/815145/airbnb-listings-in-europe-by-city/ (accessed September 23, 2021).

McClanahan, Paige. 2021. "Barcelona Takes on Airbnb". The New York Times. https://www.nytimes.com/2021/09/22/travel/barcelona-airbnb.html#:~:text=The%20ban%2C%20which%20took%20effect,property%20holds%20the%20appropriate%20license (accessed October 7, 2021).

MCNICOLL, Tracy. 2020. "Paris mayor's race takes a run at Airbnb-style rentals". France 24. https://www.france24.com/en/20200313-paris-mayor-s-race-takes-a-run-at-airbnb-style-rentals (accessed October 7, 2021).

Mladen Adamovic. 2021. "Average Monthly Net Salary (After Tax) (Salaries And Financing) by City". Numbeo. https://www.numbeo.com/cost-of-living/ (accessed October 3, 2021).

Obensa, Victor Rodriguez. 2013. “Preços na Europa 2021”. Preciosmundi. https://pt.preciosmundi.com/europa/ (accessed October 3, 2021).

Oltermann, Philip. 2016. “Berlin ban on Airbnb short-term rentals upheld by city court”. The Guardian. https://www.theguardian.com/technology/2016/jun/08/berlin-ban-airbnb-short-term-rentals-upheld-city-court (accessed October 7, 2021).

O'Sullivan, Feargus. 2018. “Barcelona Finds a Way to Control Its Airbnb Market”. Bloomberg CityLab. https://www.bloomberg.com/news/articles/2018-06-06/how-barcelona-is-limiting-airbnb-rentals (accessed October 7, 2021).

Robledillo, Gerardo. 2009-2021. “Expatistan's Cost of Living Map of Europe”. Expatistan. https://www.expatistan.com/pt/custo-de-vida/indice/europe (accessed October 2, 2021).

Shatford, Scott. 2021. “Short-Term Rental Analytics: Vrbo & Airbnb Data”. AirDNA. https://www.airdna.co/blog/airbnb-analytics-drive-rental-revenue (accessed September 23, 2021).

Sorin. 2018. “Denmark's New Airbnb Regulations”. Host Minded Copenhagen. https://hostminded.com/denmark-new-airbnb-regulations/ (accessed October 7, 2021).

Stevens, Paul. 2019. “Short-term rental properties under scrutiny in Athens”. Short Term Rentals. https://shorttermrentalz.com/news/rental-properties-athens-scrutiny/ (accessed October 7, 2021).

Takeaways, Key. 2021. “Airbnb statement in support of the Shine a Light Campaign”. Airbnb Newsroom. https://news.airbnb.com/ (accessed October 7, 2021).

TomTom International BV. 2021. “Traffic congestion ranking: TomTom Traffic Index”. TomTom. https://www.tomtom.com/en_gb/traffic-index/ranking/ (accessed October 9, 2021).

Toor, Amar. 2016. “It's now illegal to Airbnb your entire apartment in Berlin”. The Verge. https://www.theverge.com/2016/5/2/11564370/airbnb-berlin-illegal-apartment-housing-price (accessed October 7, 2021).

Tun, Zaw Thiha. 2020. “Top Cities Where Airbnb Is Legal or Illegal”. Investopedia. https://www.investopedia.com/articles/investing/083115/top-cities-where-airbnb-legal-or-illegal.asp (accessed October 7, 2021).

Appendix

I. Complementary Information regarding the Dataset

Variable Group	Variable Name	Description
Identifiers	City	Identification name of the city
	Country	Identification country of the city
Airports	No. International Arrivals	Total number of international arrivals to the city's airports[3][4]
	Distance Airport-CC	Average of the distance from the airport to the city center in kilometres[3]
	Time Airport-CC	Average of the time taken by car from the airport to the city center in minutes[3]
	No. Airports	Number of airports in the city
Airbnb	No. Airbnb Listings	Number of active listings in the Airbnb platform[5]
	Airbnb Avg. Occupancy	Average occupancy of the Airbnb's listed, in percentage[4]
	Airbnb Price/Night (€)	Average price per night of a room in the Airbnb platform, in euros
	Airbnb M. Revenue (€)	Hosts' average monthly revenue from listings in the Airbnb platform, in euros
	Airbnb Legal Barriers	Official legal barriers to the growth and maintenance of the number of rooms and

[3] Value computed considering the total number of airports in each city, declared in the variable "No. Airports".
[4] Data regarding the year 2019, pre-pandemic
[5] Data regarding the year 2021, affected by the pandemic

		houses listed in the Airbnb platform
Competition	Direct Competition	Number of direct competitors
	Direct Competitor Bob	Presence of the direct competitor BoB
	Direct Competitor Airportr	Presence of the direct competitor Airportr
	Indirect Competition	Number of indirect competitors
Cost of Operations	Average CPC (US $)	Average cost-per-click (CPC) in Google Ads search advertising, in US dollars[5]
	Gas Price (€/L)	Gas price per liter, in euros[5]
	Avg. M. Salary (€)	Average Monthly Net Salary, in euros[5]
Cost of Living	Price Index	Cost of Living Index[6]
	Public Transp. Ticket (€/Unit)	Price of a public transportation single ticket, in euros[5]
	Taxi Tariff (€/Km)	Cost of a taxi ride per kilometer, in euros[5]
	Taxi Min. Price (€)	Base price of a taxi ride, in euros[5]
Additional Characteristics	Avg. Days of Stay	Tourists' average number of days of stay
	Topography	Categorization of the city's topography
	Ease of Expansion	Number of cities of easy expansion to
	Entrepreneurial Ecosystem	Distance from the airport to the center of the city's entrepreneurial ecosystem
	Traffic Percentage	Percentage of time traffic congestion was registered throughout 1 year[4]
	No. Mobility Platforms	Number of existing mobility platforms

Table 5 - Variables Used and Respective Description

[6] Prague is considered as the city of reference (Prince Index = 100)

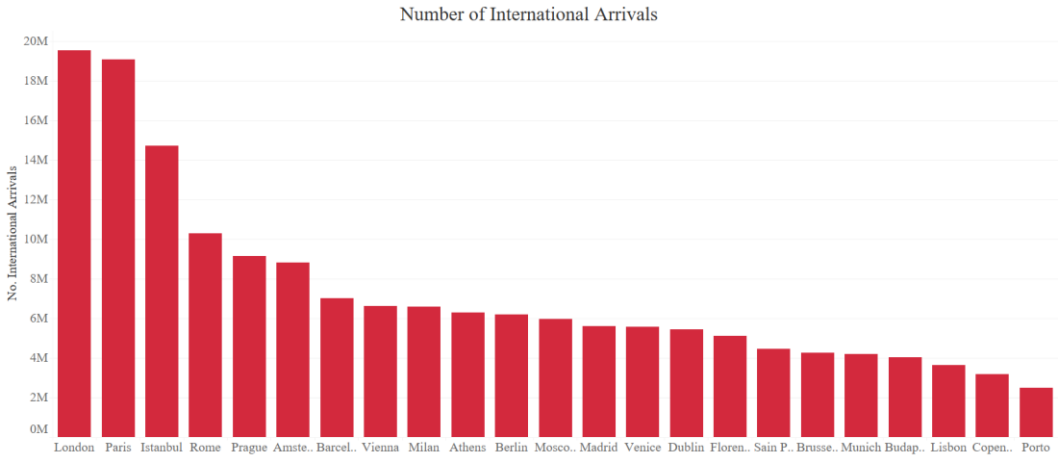

Figure 9 - Number of International Arrivals as criteria to the selection of the possible cities

City	Legal Barriers	Level of Strictness
Paris	strict rules for renting secondary homes; max 120 nights/year for primary homes	21
Madrid	prohibit rent of entire block apartments; max 90 nights/year for entire homes	20
Dublin	monitoring by local authorities; max 90 nights/year for entire homes	19
Amsterdam	max 30 nights/year for entire homes; city permit	19
Berlin	city permit; max 90 nights/year for secondary homes	15
Munich	city permit; max 90 nights/year for secondary homes	15
Barcelona	forbidden short-term private room rentals	14
Budapest	government permit	11
Copenhagen	max 70 nights/year for entire homes	8
London	max 90 nights/year for entire homes	7
Vienna	prohibition in residential zones for entire apartments	4
Lisbon	temporarily stop issuing new licenses	3
Porto	temporarily stop issuing new licenses	3
Milan	tourist tax	2
Venice	tourist tax	2
Florence	tourist tax	2
Rome	tourist tax	2
Moscow	guest registration for foreign nationals	1
Sain Petersburg	guest registration for foreign nationals	1
Prague	0	0
Brussels	0	0
Istanbul	0	0
Athens	0	0

Figure 10 - Variable "Airbnb Legal Barriers"

II. Complementary Information regarding Data Modeling

```
class KMeans_weighted:
        def __init__(self, k, tol = 0.001, max_iter = 100):
            self.k = k
            self.tol = tol  # tolarance
            self.max_iter = max_iter


        def fit(self, data, features_weight):

            self.centroids = {}
            initial = np.array([6, 17, 19, 3])

            for i in range(self.k):
                self.centroids[i] = data[initial[i]]

            for i in range(self.max_iter):
                self.classifications = {}   # stores the cities allocated to each cluster

                for i in range(self.k):
                    self.classifications[i] = []

                for city in data:
                    distances = [np.sqrt(np.abs(np.sum(((city - self.centroids[centroid])**2)*features_weight))) for centroid in self.centroids]
                    classification = distances.index(min(distances))
                    self.classifications[classification].append(city)

                prev_centroids = dict(self.centroids)

                for classification in self.classifications:
                    self.centroids[classification] = np.average(self.classifications[classification], axis = 0)

                optimized = True

                for c in self.centroids:
                    original_centroid = prev_centroids[c]
                    current_centroid = self.centroids[c]
                    if np.sum((current_centroid - original_centroid)/original_centroid * 100.0) > self.tol:
                        optimized = False

                if optimized:
                    break
```

Figure 11 - Weighted K-means Development Code

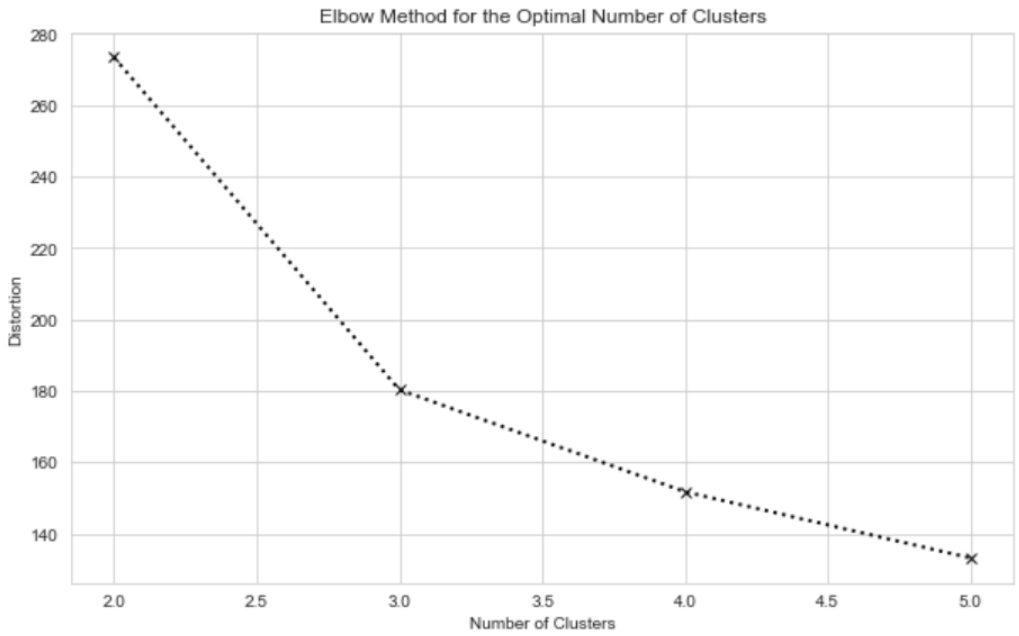

Figure 12 - Elbow Method to find the Optimal Number of Clusters

Variable Name	*CEO Weights*	*COO Weights*	*HOO Weights*
No. International Arrivals	1.9	1.7	2
Distance Airport-CC	1.1	1.2	1.7
Time Airport-CC	1.4	1.4	1.7
No. Airports	1	1.3	1.7
No. Airbnb Listings	1.5	1.5	1.3
Airbnb Avg. Occupancy	0.8	1.6	1.3
Airbnb Price/Night (€)	0.5	1.1	1.3
Airbnb M. Revenue (€)	1.2	1.1	1.3
Airbnb Legal Barriers	1.7	1.4	1.3
Direct Competition	1.2	1	0.3
Direct Competitor BoB	0.4	1.1	0.3
Direct Competitor Airportr	0.4	1.1	0.3
Average CPC (US $)	1.6	1	1
Gas Price (€/L)	1.2	1	0.7
Avg. M. Salary (€)	1.6	1.1	0
Price Index	1.5	1	0
Public Transp. Ticket (€/Unit)	0.9	0.5	0.7
Taxi Tariff (€/Km)	0.7	0.9	0.7
Taxi Min. Price (€)	0.8	1	0.7
Avg. Days of Stay	0.5	1.2	0.5
Traffic Percentage	1	1.4	0.7
No. Mobility Platforms	1.5	0.1	0.7

Table 6 - Relative Weights provided by LUGGit's Team Members

Figure 13 – Weighted K-Means Outcome considering 3 Clusters and Equal Relative Weights

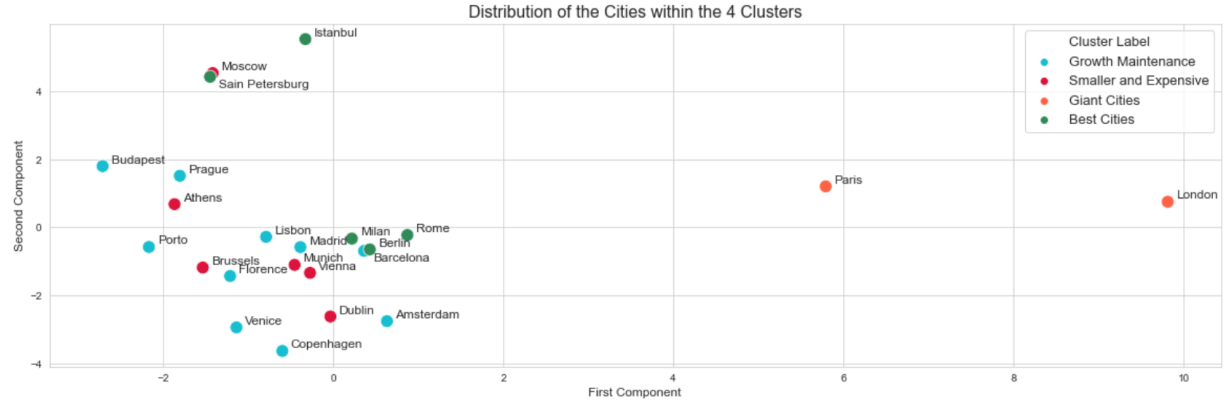

Figure 14 - Weighted K-Means Outcome considering 4 Clusters and Equal Relative Weights

Figure 15 - Weighted K-Means Outcome considering 3 Clusters and the Relative Weights provided by LUGGit's COO

Figure 16 - Weighted K-Means Outcome considering 3 Clusters and the Relative Weights provided by LUGGit's HOO

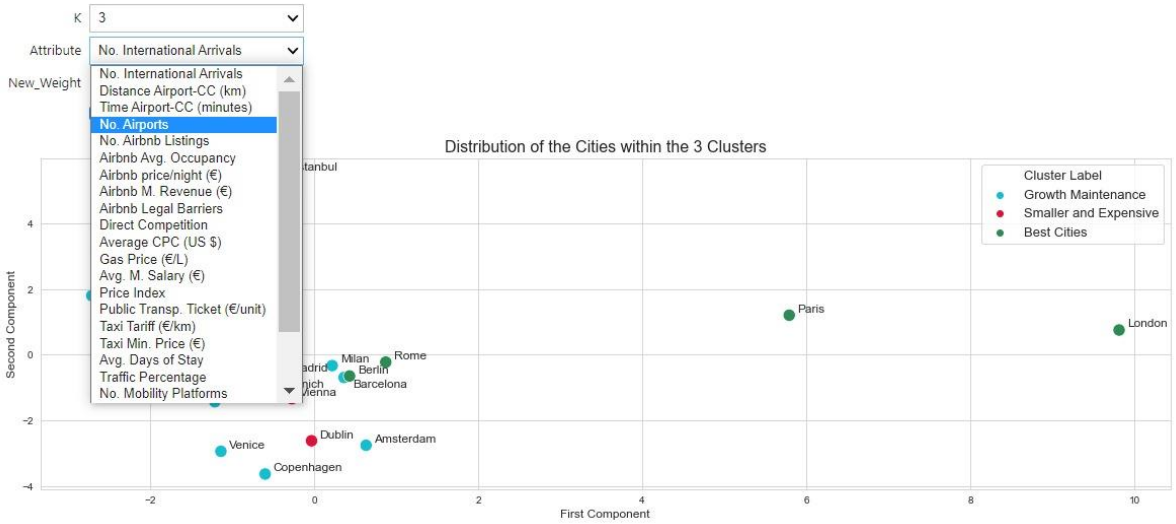

Figure 17 – Interactive Property in the Weighted K-Means

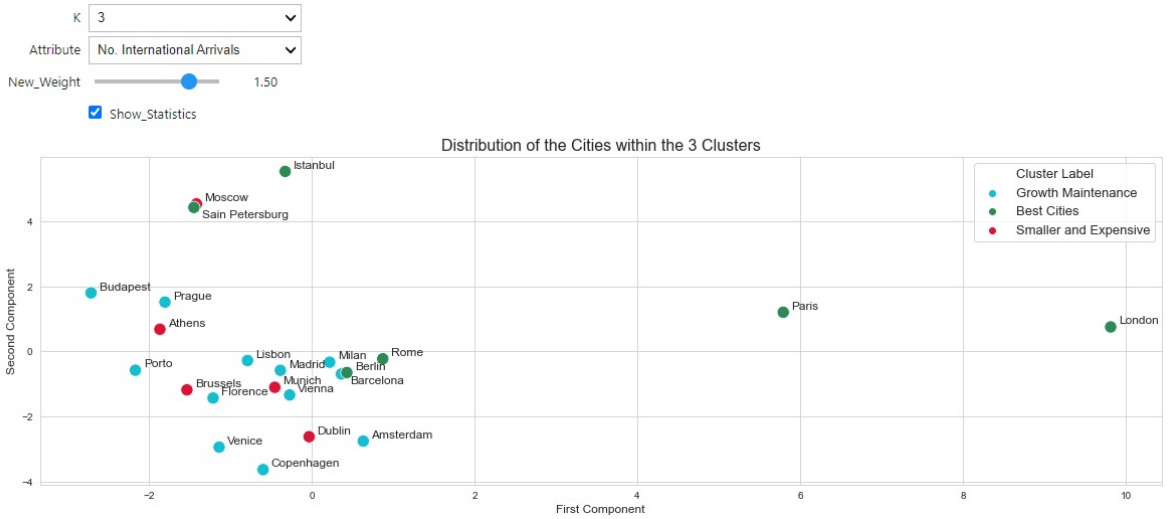

Figure 18 – Interactive Property in the Weighted K-Means