NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Analyzing attendance at performing art shows

Understanding profiles, motivations and barriers

Ana Sofia da Silva Mendes

Master Thesis

presented as partial requirement for obtaining a Master's Degree in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Analyzing attendance at performing art shows

Understanding profiles, motivations and barriers

by

Ana Sofia da Silva Mendes

Master Thesis presented as partial requirement for obtaining the Masters' degree in Data Science and Advanced Analytics, with a specialization in Business Analytics

Supervised by

Roberto Henriques, Ph.D. in Gestão da Informação, NOVA IMS

July, 2024

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Lisbon, June 25, 2024

ABSTRACT

The cultural and performing arts areas are not very discussed, which increases people's unawareness of art. Besides that, COVID-19 has had a severe impact on this field. To increase affluence in the performing arts, this project aims to analyze patterns of people's behavior and their perception of the performing arts, understanding what influences people to attend performing art shows and what could prevent them from going. Some work has already been done regarding social, economic, and demographic features, motivations and barriers to attending a performing art show. Still, this study adds a new feature regarding whether or not the person follows theatres or ticket-selling companies on social media, receives newsletters by email or checks their website. The data was collected by an online survey to 290 participants, which was then adequately analyzed to create clusters defining the three patterns found. Based on that, we present some suggestions to improve the affluence in the performing arts.

KEYWORDS

Performing art shows; audience segmentation; art shows attendance; motivations and barriers; strategies to improve attendance; cluster analysis

Sustainable Development Goals (SDG):

TABLE OF CONTENTS

1. Introduction 1
2. Literature Review 3
3. Methodology 8
    3.1. Objectives 8
    3.2. Survey Design 8
    3.3. Sampling 9
    3.4. Survey Administration 9
    3.5. Ethical Considerations 9
    3.6. Data Collection 10
    3.7. Data Analysis 10
        3.7.1. Data Understanding 11
        3.7.2. Data Preparation 13
        3.7.3. Modelling 18
        3.7.4. Evaluation 19
        3.7.5. Deployment 19
4. Results and Discussion 21
5. Conclusions and Future Work 24
Bibliographical References 26
Appendix A – Ethics Committee 28
Appendix B – Survey 29

LIST OF FIGURES

Figure 1 - CRISP-DM methodology phases 10
Figure 2 – Visualizations of all features 12
Figure 3 - Visualizations with and without having attended a performing art show in the past six months 13
Figure 4 - Correlations Matrix 16
Figure 5 - Feature Importance using Mutual Information Method 17
Figure 6 - Inertia plot on the left and Average Silhouette plot on the right 18
Figure 7 - Means and Distribution of the Clusters 21

LIST OF ABBREVIATIONS AND ACRONYMS

CRISP-DM CRoss-Industry Standard Process for Data Mining.

1. INTRODUCTION

Art is a way of communication that can shape people's everyday lives and inspire them. It helps understand history, culture and the experience of others while simply appreciating its beauty. This being said, studying more about it and promoting its awareness is very relevant. For this reason, with this project, we aim to identify the motivations and limitations of people to attend performing art shows to bring art to more people ultimately.

The performing arts can bring so many benefits that can help attendees and practitioners achieve success in the future by improving academic achievement, improving memory, stimulating their brains, enhancing creativity, improving social skills, language and reading skills, as well as critical thinking skills, building confidence, creating new connections with people with the same interests and learning commitment and discipline, crucial to their future (Bishop, 2021).

Performing art has a significant influence on the Portuguese economy. It plays several roles in the nation's well-being, including education, tourism, and job opportunities for actors, musicians, technicians, dancers, and many other professionals. Moreover, in Portugal, performing art was one of the most affected industries in the world by COVID-19 since the government imposed some measures to restrict the operation of the industry (Fresh Essays, 2023).

The cultural and performing arts areas are not very discussed topics, which has only worsened with the appearance of COVID-19, as mentioned above. Besides this problem with people's unawareness of their culture, many studies about this subject are unpublished (Andreasen & Belk, 1980), and many occur abroad. This project will analyse attendance at performing arts shows to reduce this gap, with the goal of increasing affluence in the performing arts, bringing more art to people and improving artists' employability. The objectives are:

1) Analyzing patterns of people's behaviour and their perception of the performing arts;

2) Understanding what influences people to attend performing art shows and what could be preventing them from attending;

3) Develop better strategies to improve affluence in performing art shows;

4) Understanding which features weigh more on the decision to attend performing art shows.

This project will follow the CRISP-DM (Cross-Industry Standard Process for Data Mining), including business understanding, data understanding, data preparation, modelling, evaluation and deployment. First, data exploration is essential to understand the data and context better; good data preparation is halfway to better results. Thus, some methods will be used regarding feature selection, such as correlations and mutual information, and a model

will be tested to create clusters, K-Means Clustering. It will be evaluated using R-squared, Silhouette Score, Davies-Bouldin Score and Calinski and Harabasz Score.

This project will make it possible to analyze motivations to attend performing art shows and understand which kind of shows people are more likely to participate in. This will make it easier for marketers of these events to improve their business strategy (Mitchell & SRI International, 1983) to become more profitable, improve customer satisfaction and consequently get customer loyalty. Besides that, this study can also conclude on the limitations of going to such events so those companies can mitigate them.

The rest of the master thesis will be structured as follows: Chapter 2 will cover the Literature Review, providing a comprehensive overview of existing research and theoretical frameworks relevant to the study. Chapter 3 will detail the Methodology, outlining the research design, data collection, and analysis procedures employed in the investigation. Chapter 4 will present the Results and Discussion, highlighting the key findings and interpreting their implications in the context of the research questions. Finally, Chapter 5 will offer Conclusions and Future Work, summarizing the main conclusions drawn from the study and suggesting directions for future research.

2. LITERATURE REVIEW

This chapter will explore in-depth previous studies about the topic in question, their main differences, methodologies and conclusions.

By analyzing attendance at the performing arts, (Andreasen & Belk, 1980) mentioned that up until then, studies have concluded that attendance at performing arts is strongly positively associated with income, educational and occupational attainment. Moreover, past studies had little success linking attendance or non-attendance to individual perceptions of the performing arts. So, to improve previous work, besides the usual socioeconomic variables, their study adds two features: one about the extent to which respondents were interested in classical music or live theatre when they were growing up and the other about the extent to which their parents were also interested in the same performing arts. So, with data collected by telephone interviews of 14 years old or older and analyzing it with an attitude model, correlations and multiple regression, (Andreasen & Belk, 1980) results showed that sex and race are not significantly related to attendance likelihood; education, membership in the highest income category, years living in the area and the number of cars owned are positively correlated and age, number of children over 14 years old and being a home-maker are negatively correlated with attendance likelihood.

These authors concluded that life-style, attitude and developmental experiences are more important in understanding consumer behaviour regarding performing arts than socioeconomic variables and that the previous studies are shortsighted. (Andreasen & Belk, 1980)

With the same line of thought but some different conclusions, (Manolika & Baltzis, 2022) also agree that socioeconomic variables are not enough, as previous studies showed. Individual attendance decisions result from multiple determinants since people have a broad array of distinct needs that are fulfilled through passive participation in various cultural activities. By conducting a survey of 480 attendees at several venues in Thessaloniki (Greece) and analyzing the data, the authors used ordinal logistic regression for predicting arts attendance, as well as the Chi-square test to find significant gender differences, the One-way ANOVA to analyze age, Kruskal-Wallis test for income, among others. The results showed that gender, age and education are not significantly related to attendance likelihood. Perhaps if art education were available, it would have a different impact. Income, entertainment, and art interest positively correlate to attendance likelihood and the need to expand intellectual horizons. Moreover, people who want to experience positive emotions, such as joy and obtain pleasurable emotional experiences are more likely to attend arts performances. It is also worth mentioning that different events attract visitors with different motivations.

This being said, (Manolika & Baltzis, 2022) concluded that the events people select reflect their personal characteristics and associated needs and, therefore, they suggest that arts

managers should pay attention to the needs and preferences of attendees at specific performing art events to increase profitability of arts attendance. (Manolika & Baltzis, 2022)

(Seaman, 2005) also mentions education as a stronger determinant than income in arts attendance. In particular, arts training is distinguished from formal education. Econometric evidence has shown that this factor and family socialization was essential in explaining arts attendance variations. Moreover, it is worth noting that theatre critics' opinions matter since they can affect attendance, although sometimes negatively, at least for some audience segments. (Seaman, 2005)

Still regarding socioeconomic and demographic variables, *Bone J, Bu F, Fluharty M et al.* used cross-sectional data from the General Social Survey (GSS) in the US from 1993 to 2016, analyzed it with a logistic regression and concluded that demographic, socioeconomic, residential, and health factors may influence engagement in the arts throughout the life course. Children of parents with more education may be more likely to receive arts education in childhood and, therefore, more likely to attend art classes in the future. Furthermore, attendance at arts events was less likely for those with lower income and social class, poorer health, and fewer urban areas, as expected, and these factors would be considered barriers to interested non-attendees. Moreover, being female, compared to male, was also consistently associated with higher levels of art engagement and more years of education in general. (Bone et al., 2021) These authors mention that there is growing evidence that engagement with the arts can lead to a range of health benefits, independent of demographic and socioeconomic factors, so given its importance, individuals should be provided with equal opportunities, art activities and creative group membership should be encouraged as a way of increasing participation and attendance in the arts.

Regarding the analysis of motivations and barriers to attending performing art events, the National Arts Administration and Policy Publications Database (NAAPPD) was analyzed, and (Mitchell & SRI International, 1983) concluded that the primary motivations for arts attendance are to be entertained and to see a particular show, performer, or group; for sociability and business purposes or to fulfil themselves.

On the other hand, some barriers to attending art shows were found in (Mitchell & SRI International, 1983) study, namely the lack of leisure time and preference for other leisure time activities. It was probably not the most intuitive barrier since it first came to mind to cut ticket prices. However, interpretation of the data suggests that cutting prices would not make much of a difference (perhaps it would attract a few in the middle-income ranges, but not a significant portion).

Based on the 2012 U.S. General Social Survey and statistical analysis, (Hill Strategies, 2016) study concluded that the top motivations for attending arts performances are to socialize with family or friends, see a specific artist or performer and see a performance at a specific location. In particular, parents with young children cited socializing with family or friends, learning new

things, and celebrating cultural heritage. In contrast, people without children wanted to experience high-quality art, visit the event's location or venue, support the community, and celebrate cultural heritage. Finally, people in rural areas seem to be more motivated by supporting the community than urban residents, who may have other motivations as a priority. On the contrary, the main barriers to attending are lack of time, cost, and difficulty getting to a venue. (Hill Strategies, 2016)

To conclude, using the 2012 GSS Arts survey, *NEA Office of Research & Analysis* mentions that organizations should provide opportunities for attendees to socialize, meet new people, and experience new art forms in a flexible format that combines the arts with other activities these individuals enjoy. This way, art marketers could attract and retain audience members. (NEA Office of Research & Analysis, 2015)

Still, on the same line of thought, (Carter, 2020) analyzed some statistics from two different surveys, the 2017 Survey of Public Participation in the Arts and the 2016 General Social Survey and also agreed with some of the motivations referred to above, adding some new ones, such as feeling creatively inspired and experiencing high-quality art. (Carter, 2020)

It is worth noting that across all generations, Generation Z is the most interested in attending arts events, being the one generation who participated in the most in-person arts events, followed by the Millennials, who are mainly motivated by socializing with friends and family. Furthermore, income was not a significant predictor of attendance. Low-income people were more likely to identify celebrating their heritage and supporting their communities as reasons for arts attendance than low-cost.

As for barriers, it is safe to say that these do not suffer many changes among different studies, except adding the inability to find someone to go with. This factor may not be the main barrier to attendance, but it is an important one since although it was not a frequently cited barrier to arts attendance, nearly 30 percent of those who cited it at all said it was the only barrier or the most important one.

With this in mind, (Carter, 2020) suggests that to attract people who lack time, it is essential to provide them with clear and accessible advanced information that includes when an event may be, how long a performance will be, directions to the venue, the availability of parking, options for food and drink at and near the venue so that it will help with decision-making and supports a positive arts experience. Besides that, sharing information about free and lower-cost events is critical to attracting possible attendees with the lowest income. Finally, for people seeking socializing, ticketing discounts for groups, customized tours for families or groups with unique interests, hosting events around a particular topic, showcasing images of families and groups interacting at their programs, emphasizing the arts as a fun, creating materials and messaging about enjoying the arts. These suggestions could also help people who cannot find someone to go with to mitigate this problem since art is being promoted as

fun on its own, and the campaigns about it might encourage them to go alone and meet new people. (Carter, 2020)

Once again, with the statistical analysis that emerged from the NEA's 2017 Survey of Public Participation in the Arts, (Klickstein, 2022) also came to the same conclusion regarding motivations and barriers to arts attendance, with socializing with friends or family, learning new things, experiencing high-quality arts and supporting the community as the main motivations and time, cost, access and no one to go with as the top barriers. (Klickstein, 2022)

NEA has also shown the same conclusions regarding motivations and barriers to arts attendance by using the 2012 GSS Arts survey. As for motivations, low cost, wanting to experience high-quality art, socialize with family or friends, learn about or celebrate your or your family's cultural heritage, support a community organization or (community) event, gain knowledge or learn something new, experience a performance or see an exhibit at this location, see a specific, individual performer were the most mentioned. On the other hand, cost, access (too difficult to get there), social reasons (could not find anyone to go with), lack of time, program (programs or events were not of interest) and venue (did not want to go to that location) were the main barriers, not forgetting work commitments, sickness or disability, and not being able to find childcare. These were the greatest findings in this research. (NEA, 2015)

A new approach is introduced by (Walmsley & Ba, 2011) regarding why people go to the theatre, with a comprehensive qualitative study, containing qualitative techniques, including responsive depth interviews and participant observation. This study on motivations for theatre-goers rose some new conclusions, namely that emotion is a very powerful one that keeps theatre audiences coming back for more, so emotion-based marketing would be a good option on promotion. Furthermore, another finding was that post-show discussions enhanced the audience experience by providing a forum to share ideas, meaning this is a motivation for attendance. Finally, despite the latest developments in creative technology, consumers are ultimately human beings who are essentially motivated by human interaction and insight. The collective experience is what is important here, so marketers should highlight in their imagery the interactive privilege of watching actors perform, and facilitate deeper relationship-building opportunities with actors and creative teams, both on and off the stage. This research has clearly indicated the overriding importance of emotional impact in audiences' motivations to go and see a play, so far that respondents actually mentioned that theatre is live, dynamic and about people, and that is what makes them go: "It's done live, in front of you, every night, and you're part of it, you're part of the experience with the actors, dancers, orchestras, whoever it is.". (Walmsley & Ba, 2011)

In summary, research on performing arts attendance reveals that while socioeconomic factors like income and education are influential, they do not fully explain audience behaviour. (Andreasen & Belk, 1980) and (Manolika & Baltzis, 2022) emphasize the importance of developmental experiences, lifestyle, and personal motivations such as intellectual

enrichment and emotional satisfaction. (Seaman, 2005) highlights the critical role of arts-specific education and family socialization, while (Bone et al., 2021) point to structural barriers faced by lower-income and socially disadvantaged groups. Studies by (Mitchell & SRI International, 1983) and (Hill Strategies, 2016) identify practical barriers like time, cost, and access, suggesting that logistical improvements could enhance attendance. Recommendations from (NEA Office of Research & Analysis, 2015) and (Carter, 2020) focus on integrating social opportunities and clear information to mitigate these barriers, while (Walmsley & Ba, 2011) stress the emotional and communal aspects of arts experiences.

Despite these insights, there remains a gap in research on specific cultural contexts, such as Portuguese audiences, since few studies exist, and many have not been published yet. Additionally, the impact of social media and modern technology on arts attendance has not been properly explored.

This project will build on the existing literature by incorporating the features used by the authors mentioned above and introducing new variables, to help Portuguese art marketers understand their customers and better strategize campaigns to promote upcoming shows. A survey with targeted questions will be conducted, and the data will be analyzed to create meaningful clusters, thereby providing a more comprehensive understanding of audience behavior.

3. METHODOLOGY

The methodology used in this study was to first create a survey, based on the Literature Review, distribute it and then analyze the data that results from it.

3.1. OBJECTIVES

To collect data, a survey was built and conducted since there are not many open datasets available on this topic, and some of the questions made were very specific to the purpose of this study. The more data, the better, the broader the purview of the problem, and the results will be more trustable, so the prior goal was to reach as many people as possible.

3.2. SURVEY DESIGN

The survey was made based on the combination of several previous studies already named in the Literature Review chapter, with one additional feature: whether or not the person follows theaters or ticket selling companies on social media, receives newsletters by email or checks their website. The new feature brings a different perspective, more adequate to the present and that will make it possible to see if digital marketing is a good option for promotion by understanding if whether or not people engage with theaters and ticket selling companies online.

Thus, the survey (see full version in Appendix B) consisted of the following questions:

1. Age
2. Sex
3. Income
4. Occupation
5. Did you practice a performing art when you were young?
6. How many children do you have?
7. Is at least one of your children younger than 6 years old?
8. Do your children practice a performing art?
9. Do you live in a urban or rural area?
10. Do you have a car?
11. Do you have easy acessability to theaters?
12. Considering that performing art shows are live presentations to an audience, do you have interest in attending them?
13. Which type of performing art do you prefer to watch?
14. Did you attend a performing art show in the past six months?
15. What type of performing art was the last show you attended?
16. What is your main motivation to attend a performing art show?
17. Which main barrier can prevent you from going?

18. What makes you select performing art shows instead of others (like cinema, exposition, etc)?
19. Do you follow theaters or ticket selling companies on social media / receive newsletters by email / check their websites?

Regarding socioeconomic and demographic features, it was transversal to the literature its importance, so it was included in the survey. As for more specific questions, such as emotion (question 18), it was based on (Walmsley & Ba, 2011) study. On the other hand, the inclusion of the social media feature (question 19) was encouraged by the natural progression and evolution of technology. Regarding motivations and barriers (questions 16 and 17), several studies have mentioned it, such as (Klickstein, 2022), (NEA Office of Research & Analysis, 2015), (Hill Strategies, 2016) and (Carter, 2020). Finally, as for art's training, (Seaman, 2005) was the main source of question 5.

3.3. SAMPLING

The sampling method used was convenience sampling, which means people are selected based on their accessibility and availability, once the survey was distributed online via *Instagram* and *Whatsapp* on my own network. The survey was conducted to 290 participants, which considering the Portuguese population and according to Slovin's Formula, it is a sample of the population with an error of approximately 6%.

3.4. SURVEY ADMINISTRATION

Some assumptions were made *prior to* conducting this study. First, all participants are 18 years old or older and, since the survey was in English, it is expected that this might be a limitation for the ones that do not speak this language. Moreover, 290 people is just a sample, meaning conducting research with these participants may lead to biased data.

3.5. ETHICAL CONSIDERATIONS

At the beginning of the survey, it is stated that it is completely anonymous, with no questions asking for name, email, or any other Personal Identifying Information; also that the responses will remain confidential and will be utilized solely for academic purposes. Besides that, it is also mentioned that the participation in this research study is entirely voluntary, and people may opt out if they choose to.

Before diving into the survey itself, the participants need to consent that they have read and understood the above information, they voluntarily consent to participate and they're at least 18 years old.

This study was approved by the Ethics Committee of NOVA IMS, as it can be found the document in Appendix A.

3.6. DATA COLLECTION

Since the survey was shared online, on my own network, it might limit the range of participants that had access to it, which may cause some error. However, it can bring some insights anyway.

The survey resulted in a dataset with 290 rows, each representing a record from a participant and 26 columns, that denote the various features. Thus, the first step was to import the dataset to a *jupyter notebook* in *python* using *pandas*, where the data analysis will be conducted.

3.7. DATA ANALYSIS

The data analysis follows the CRISP-DM method, which includes business understanding, data understanding, data preparation, modelling, evaluation and deployment. This method was not built theoretically or academically, and it worked from technical principles. Thus, it succeeds because it is soundly based on the practical, real-world experience of how people conduct data mining projects. (Chapman et al., 2000)

Figure 1 - CRISP-DM methodology phases

Figure 1 shows the steps regarding CRISP-DM methodology. The initial phase focuses on understanding the project objectives and requirements from a business perspective, which was already described in the introduction. The data understanding phase starts with initial data collection and proceeds with activities that enable you to become familiar with the data, identify data quality problems and discover first insights into the data. The data preparation phase covers all activities needed to construct the final dataset (data that will be used in the modelling phase) from the initial raw data. Various modelling techniques are selected and applied in the modelling phase, and their parameters are calibrated to optimal values. At the evaluation stage of the project, it is important to thoroughly evaluate it and review the steps executed to create it to be sure the model properly achieves the business objectives. A key aim is to determine if some critical business issue has not been sufficiently considered. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise, but its purpose is to maintain the achieved solution (Chapman et al., 2000).

3.7.1. Data Understanding

Before diving into data preparation, it is crucial to explore and understand the current dataset in detail to know what it contains, how it is structured, and whether there are any issues with it, so that the adjustments needed can be done later.

To begin with, the columns were renamed in order to become easily read and then better understand the dataset, so instead of the questions of the survey, the features have adequate names, such as *'Consent', 'Age', 'Sex', 'Income', 'Occupation', 'PracticeYoung', 'NumberOfChildren', 'ChildrenUnderSix', 'ChildUnderSix', 'ChildrenPractice', 'UrbanArea', 'Car', 'EasyAccess', 'Interest', 'TypePreference', 'AttendancePastSixMonths', 'TypeLastShow', 'Motivation', 'Barrier', 'PerformingArtDifferentiator', 'SocialMedia'.*

By default, Microsoft Forms creates automatically the columns *Beginning Time, End Time, Email and Name*, filled with the date and time of beginning and conclusion, *anonymous* and *NaN* values, since this is not part of the defined survey, so these features will not be considered. This will be corrected in the data preparation phase.

Afterwards, the data types of the variables and missing values were checked. Some missing values were found regarding children related variables, *TypePreference, AttendancePastSixMonths, TypeLastShow, Motivation, Barrier, PerformingArtDifferentiator, SocialMedia*. However, this is due to the fact that once people answered they had no interest in art at all or prefer other types of art rather than performing arts, the questions who followed did not appear, so it was expected that those had missing values.

Regarding the main descriptive statistics for the categorical variables, it can be seen that the majority of the sample are middle class, employed and have no children. Besides that, most of the participants live in urban areas, have a car, and have easy access to theaters. They are also interested in art, namely performing art shows. It is also worth mentioning that the majority of the population surveyed answered 'It's more dynamic and realistic' to "What makes you select performing art shows instead of others (like cinema, exposition, etc)?".

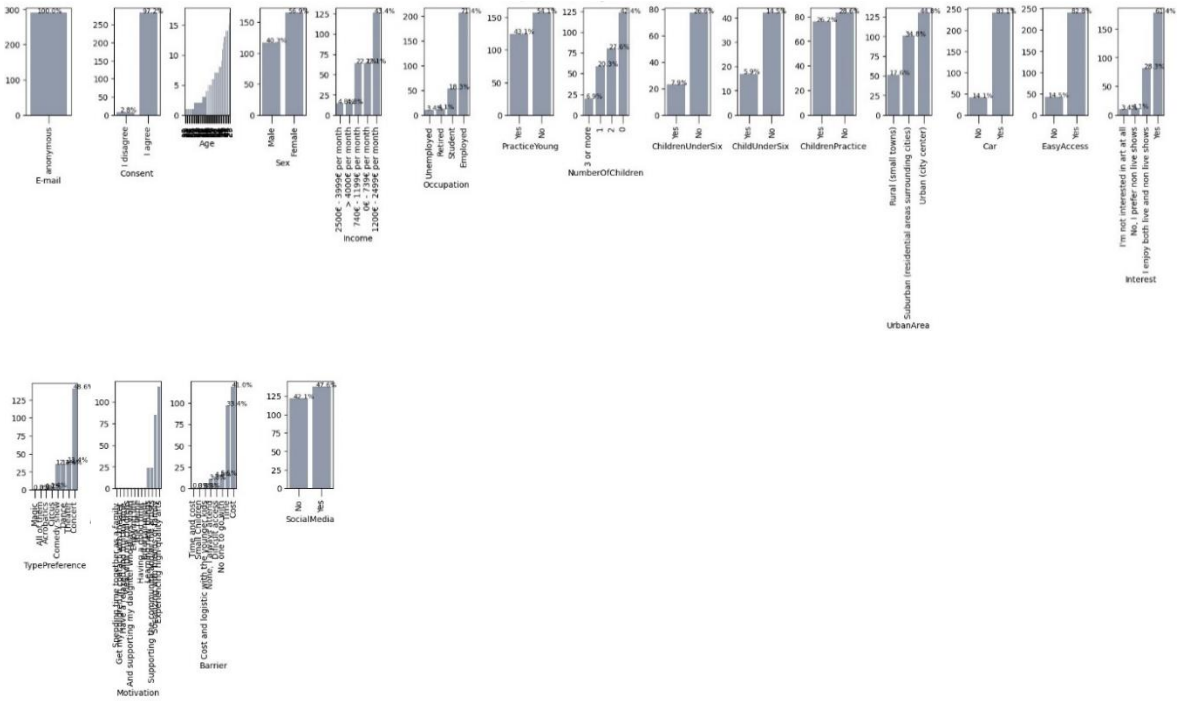

Figure 2 – Visualizations of all features

Figure 2 shows that 56.9% are female, 83% have a car, and approximately the same percentage have easy access to theaters. It is also evident that the majority, more specifically 48.6%, prefer "Concert" over other performing art types. Moreover, the main motivation for attending performing art shows is *"Experiencing high-quality arts"* and the main barrier is *"Cost"* (41%), followed by *"Time"* (33.4%). Notice that, even though it was a small percentage, it is still important to mention that some people pointed out that having young children and not having no one to care for them is a barrier to attending. Furthermore, it is worth noting that the second primary motivation is *"Socializing with friends and family"*, which had more people attending, but also had a lot not attending in the past six months, so perhaps this could be a good point to explore, in order to bring more people to performing art shows.

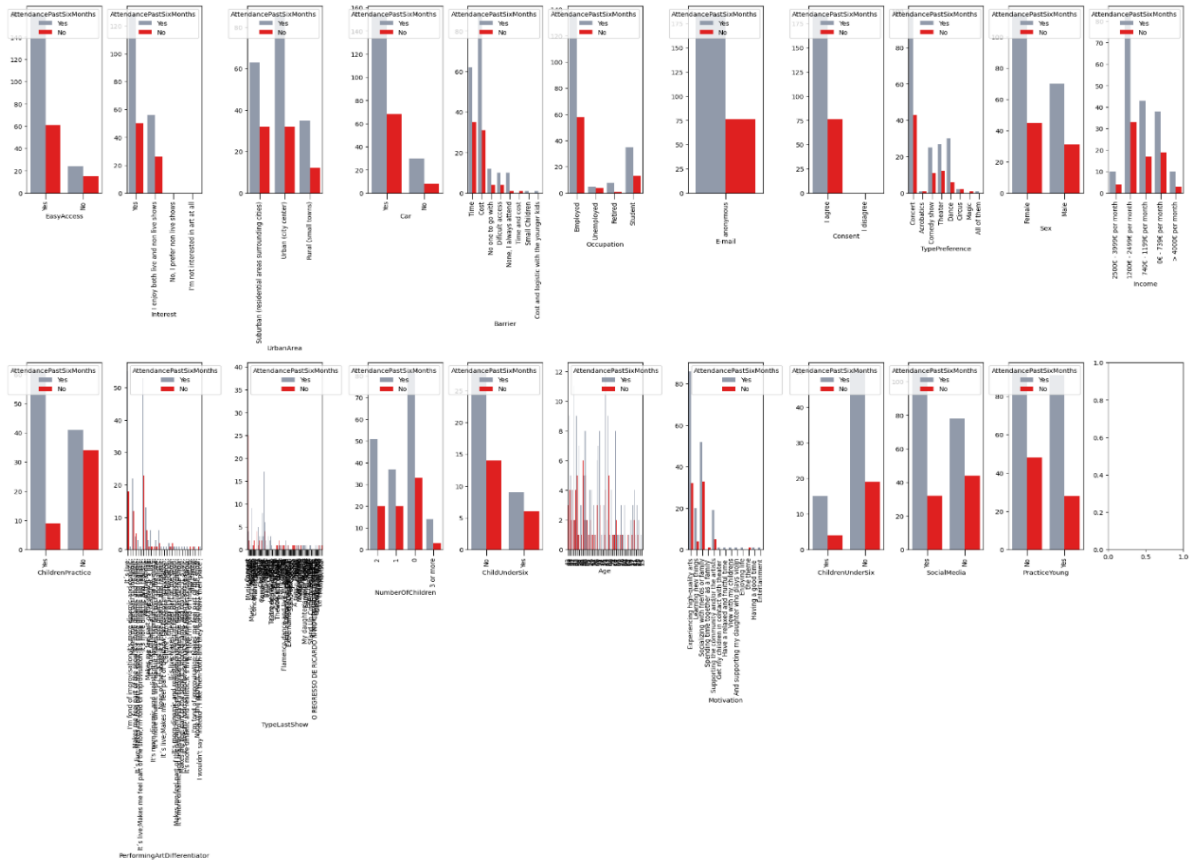

Figure 3 - Visualizations with and without having attended a performing art show in the past six months

As shown in Figure 3, by analyzing bar plots of all features to compare the behavior of variables with or without having attended a performing art show in the past six months, it can be seen that people with no children are more likely to attend, but even the ones who have are more likely to attend if they do not have children under six years old. Furthermore, employed ones are more likely, although students have a good proportion as well. It is important to mention that people who follow the ticket selling companies on social media are more likely to attend and the ones that don't are more likely not to attend, which rises the relevance and the role of technology in the performing arts. Moreover, it can be seen that people who did not practice a performing art when they were young are more likely not to attend. Finally, the ones with children who practice a performing art are very likely to attend, whereas those who don't practice it doesn't make much of a difference.

3.7.2. Data Preparation

The data preparation process is a crucial step in any data mining project, which consists of transforming raw data into a clean, organized, and suitable format for modelling.

First, some columns were removed, namely *Beginning Time, End Time, Name* and *Email*, since the first ones were not needed for the analysis and the remaining were filled with one unique value only, and the *ID* column was set as the index of the dataset. Furthermore, the variable '*Age'* was converted to numeric values.

To normalize the data, namely the metric features, which in this case is only the variable *'Age'*, the MinMax Scaler was used, since it will preserve the shape of the dataset (no distortion) and also because it is the least disruptive to the information in the original data. With this method, the features normalized have a minimum value of 0 and a maximum value of 1.

Regarding duplicates and missing values, the data was checked. However, no tasks were conducted for the first one, since there were no duplicates in the dataset. On the other hand, some missing values existed on the children related variables, because the people who answered that had no children did not answer the remaining questions about them, and also the ones that had no interest in art did not answer the questions bellow that. Although the solutions used for the problem of missing values are deleting records or filling in with a measure of central tendency, in this particular case, since the missing values of the children related variables are due to not having children, they will be filled in with 0, meaning not having children under six years old and not having children who practice a performing art. As for the 22 missing values from the last 7 variables, those occur because the individuals do not have interest in art at all or because they prefer non live shows, so the last 7 questions did not apply. For this reason, the missing values will not be removed or filled in with a measure of central tendency. Since these 22 participants do not have interest in the performing arts, two different datasets were created: one with the people interested in the performing arts (*data_interested*) and another with those who don't (*data_not_interested*). Once divided into the two datasets, the feature *'Interest'* is no longer necessary in any of them, so it will be removed. Regarding the *data_not_interested* dataset*,* it will not be conducted any analysis, since people who do not have interest in the performing arts area, are not relevant to this project. This would correspond to around 8% of the sample used. So, from now on, all the analysis will be focused on the *data_interested* dataset.

As for data types, the mistakes seen in the data understanding phase were corrected. Thus, *'Age'* was converted to integer. Moreover, *'Sex'* was renamed as *'Male'* and converted to a binary variable (0 or 1) instead of remaining in the Yes/No format, as well as *'PracticeYoung', 'ChildrenUnderSix', 'ChildUnderSix', 'ChildrenPractice', 'Car', 'EasyAccess', 'AttendancePastSixMonths'* and *'SocialMedia'.* Finally, the feature *'NumberOfChildren'*, which had three categories ("0", "1", "2" and "3 or more") was first replaced the value "3 or more" with "3", since it does not make much of a difference in this context, and then converted the feature to integer, since it represents the number of children of the participant.

Once the survey is voluntary and participants need to consent the use of the data, although it was not asked for any Personal Identifying Information, and be at least 18 years old, there was also a question on the survey with this in mind, that was renamed into *'Consent'*. So, it was

removed from the dataset the observations with *"I disagree"* on this feature, which deleted 8 rows.

Regarding new features, one was created to simplify the reading and understanding of the dataset. Since the survey had two questions asking the same thing, one for people with only one child and another for the ones with more, this created two different features: *'ChildrenUnderSix'* and *'ChildUnderSix'*, which will then be merged to one feature only named *'ChildrenUnderSixYears'*, that has the information of both of them and represents whether or not the children (one or more) of the participant is under six years old. With this step completed, the previous features were removed. This was performed on both the datasets created.

At last, one-hot encoding was applied to the categorical variables (*'Income', 'Occupation', 'NumberOfChildren', 'UrbanArea', 'TypePreference', 'Motivation'* and *'Barrier'*), creating new binary columns indicating the presence of each category of the variable. For each row in the dataset, the column corresponding to the category for that row is set to 1, while all others are set to 0. The use of this encoding has some drawbacks because creating new columns for each category can lead to the curse of dimensionality and sparsity. This increased the number of features as expected, so feature selection is necessary as the next step. The features *'TypeLastShow'* and *'PerformingArtDifferentiator'* had a lot of unique values, so they were not considered, since it would increase too much the number of features after encoding.

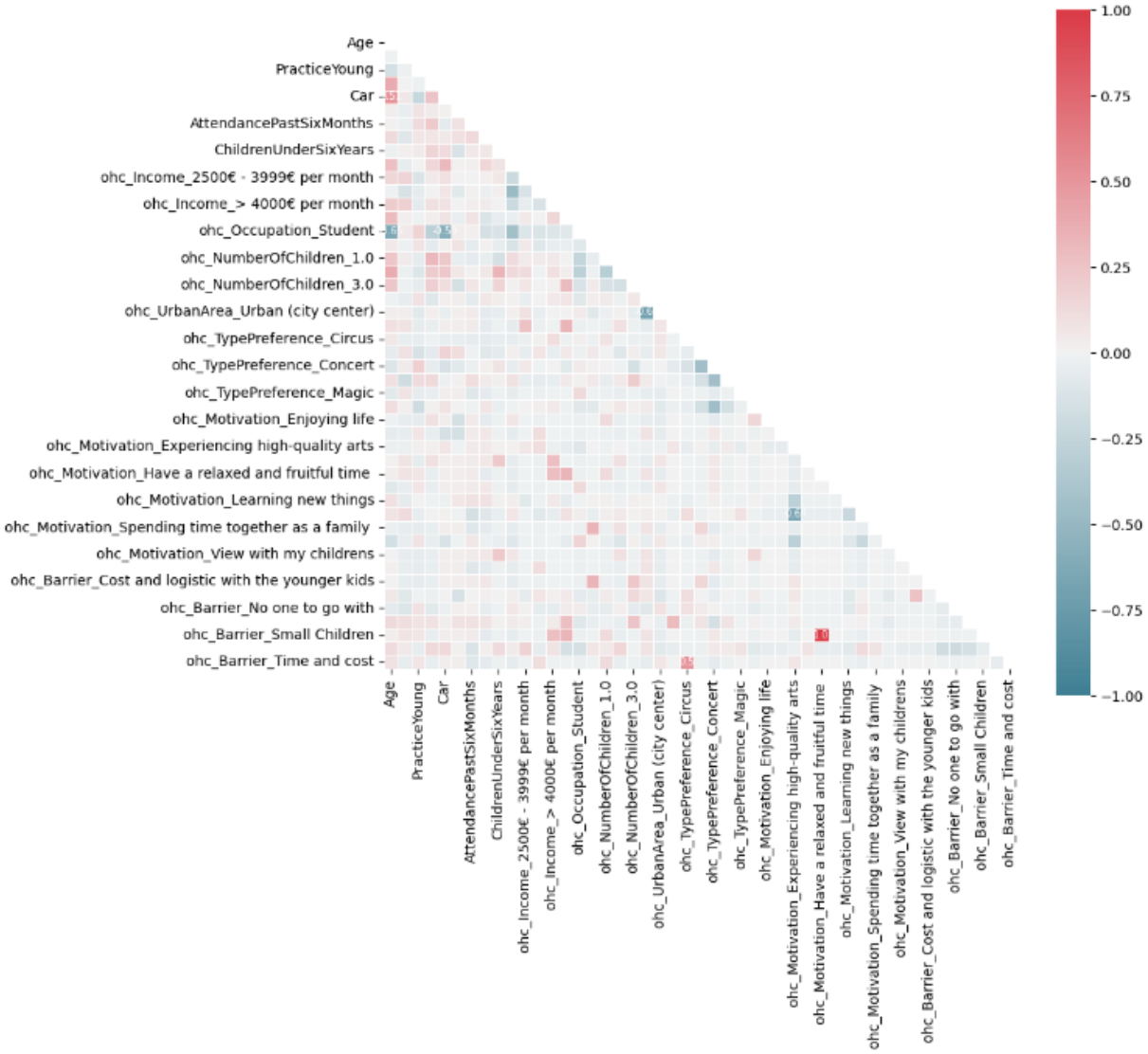

Figure 4 - Correlations Matrix

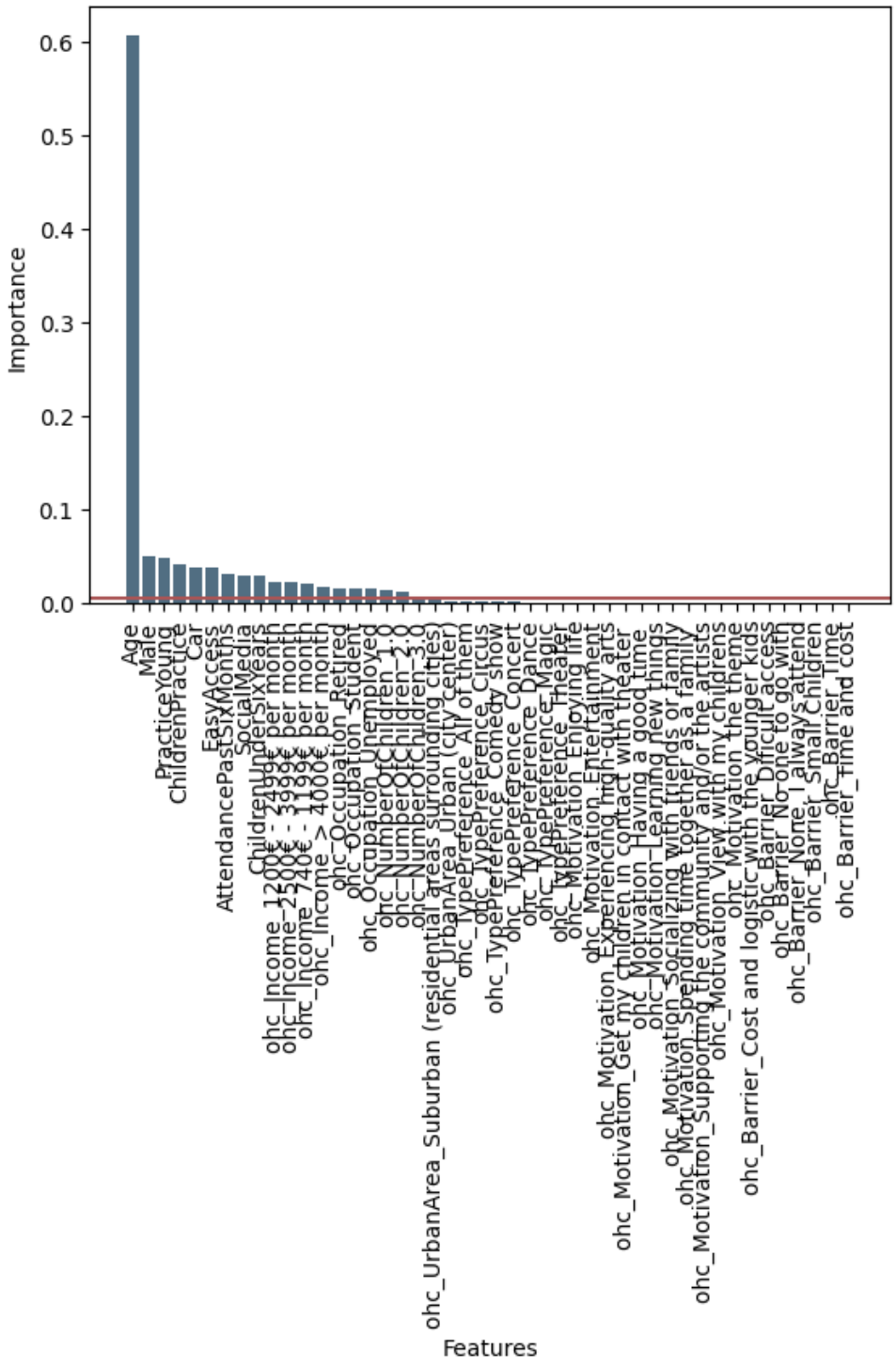

Figure 5 - Feature Importance using Mutual Information Method

Regarding feature importance, some methods were used, namely correlations, represented in Figure 4, and mutual information, illustrated by Figure 5. First, it was confirmed that the variable *'ohc_Motivation_Have a relaxed and fruitful time'* had a correlation of 1.0 with *'Barrier_Small Children'*, so the first one had to be deleted. Then, using the Mutual Information

Method, it was possible to identify other variables with almost zero importance, which led to removing several features. Finally, the dataset is ready for the modelling phase.

3.7.3. Modelling

K-means clustering is a partition-based algorithm that divides a dataset into K clusters based on similarity, making it ideal for customer segmentation, which is a goal of this study. It is particularly useful when there is no specific outcome variable to predict; instead, it uses a set of features to identify groups of observations that share similar characteristics. K-means algorithm only handles numeric features, however, since the one-hot encoding has already been applied to the categorical features, this should not be a problem.

The algorithm works by first specifying the number of clusters K to create. Then, the K data points are randomly initialized in the data space, and each data point is assigned to the cluster with the nearest centroid, based on the Euclidean distance. The mean of all data points assigned to each centroid is calculated and moved to that mean. This process is repeated until the centroids no longer move or change only minimally.

Regarding the parameter settings, the number of clusters to use is an important parameter to set *apriori*, because if the number of clusters is not specified correctly, the algorithm may not be able to effectively group the data points into meaningful clusters, and the resulting clusters may not accurately reflect the underlying structure of the data. Two methods were used to decide on that number: plotting the distortion (which measures the compactness of each cluster) and computing the silhouette score (which measures how similar an observation is to its own cluster compared to other clusters).

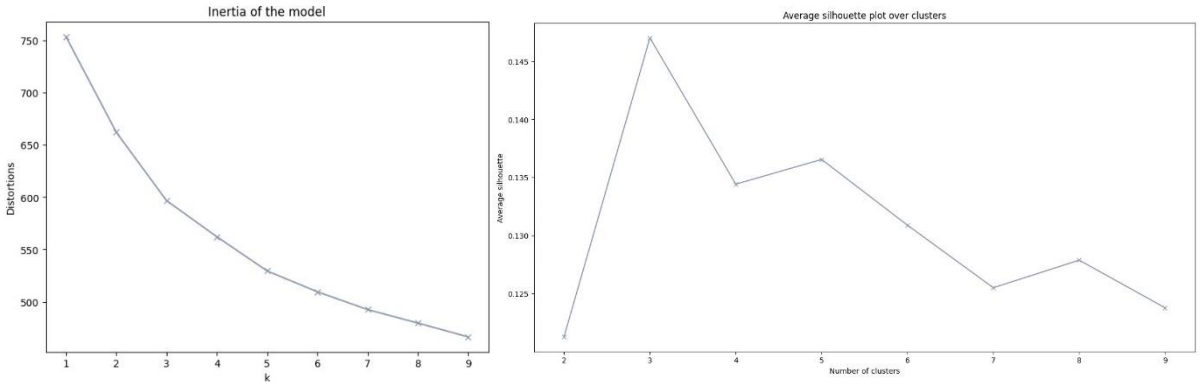

Figure 6 - Inertia plot on the left and Average Silhouette plot on the right

Figure 6 shows on the left the inertia plot, which measures how well K-Means clustered a dataset. The point is to find the *elbow* of the plot, the point where the decrease in inertia begins to slow, that is the optimal number of clusters. On the right side, the image represents

the average silhouette plot, which evaluates the quality of the clusters. The purpose is to look for the number of clusters that have the highest silhouette score.

Thus, the best solution according to the inertia plot was three clusters and the silhouette score, so the final decision for the number of clusters was indeed three.

The results of the model were the creation of a new column in the dataset with the clustering labels, so it was possible to know to which cluster a specific observation belonged. Since the data was still scaled and encoded and had some missing features, the created labels were added to the original dataset to interpret and understand each cluster's different characteristics.

3.7.4. Evaluation

The model used was a K-Means algorithm with K = 3 clusters. To evaluate the clustering solution, four metrics were used: R-Squared, Silhouette Score, Davies-Bouldin Score, and Calinski and Harabasz Score. R-Squared measures how well the clustering solution explains the proportion of variance in the data. The Silhouette Score assesses how well each observation fits within its assigned cluster relative to other clusters. The Davies-Bouldin Score evaluates the quality of the clustering solution by computing the average similarity between each cluster and its most similar cluster, considering cluster size and dispersion. The Calinski and Harabasz Score measures the ratio of between-cluster variance to within-cluster variance.

These metrics were chosen to provide a comprehensive assessment of the clustering quality from different perspectives, including variance explanation, cluster cohesion, separation, and overall cluster structure.

The evaluation results were not very satisfactory, with scores of 0.3417 for R-Squared, 0.2167 for Silhouette Score, 1.8970 for Davies-Bouldin Score, and 66.6972 for Calinski and Harabasz Score.

This not-so-good outcome may be due to the method used to collect the data. Since it was a survey, conducted to only 260 participants that consented to it and have interest in the performing arts, it was not a very large sample, as well as not very diversified. Thus, the clusters might be biased. However, it was possible to get some insights from the analysis done, which will be discussed further on the Results and Discussion chapter.

3.7.5. Deployment

Regarding the potential deployment of this model, when new people answer the survey, it is necessary to assign them to a new cluster, and this can be done using the model created, as it is scalable (can handle large datasets with many variables), fast (can easily calculate the cluster

labels), and easy to interpret and to visualize (so, it is easy to understand and communicate the clustering results). So, using K-Means to predict the cluster of a new person will include all the data preparation steps that were applied to the data corresponding to the new person and the calculation of the distances of the new observation to all the cluster's centroids. This will then become a classification problem, and classification models can be used, such as a decision tree, as it is a valuable model for interpreting the decision rules that were made for assigning a particular customer to a specific cluster.

The final clusters with the new people would be re-evaluated and a new solution with the correspondent metrics would rise.

4. RESULTS AND DISCUSSION

The final clustering solution included 3 clusters with distinct behaviours. Overall, the main motivations and barriers support the literature reviewed: (Klickstein, 2022), (NEA Office of Research & Analysis, 2015), (Hill Strategies, 2016) and (Carter, 2020). Furthermore, it was transversal to all clusters that the majority of the participants answered *"It's more dinamic and realistic"* and *"Makes me feel part of the show"* to the question of *"What makes you select performing art shows instead of others (like cinema, exposition, etc)?"*, as also concluded by (Walmsley & Ba, 2011).

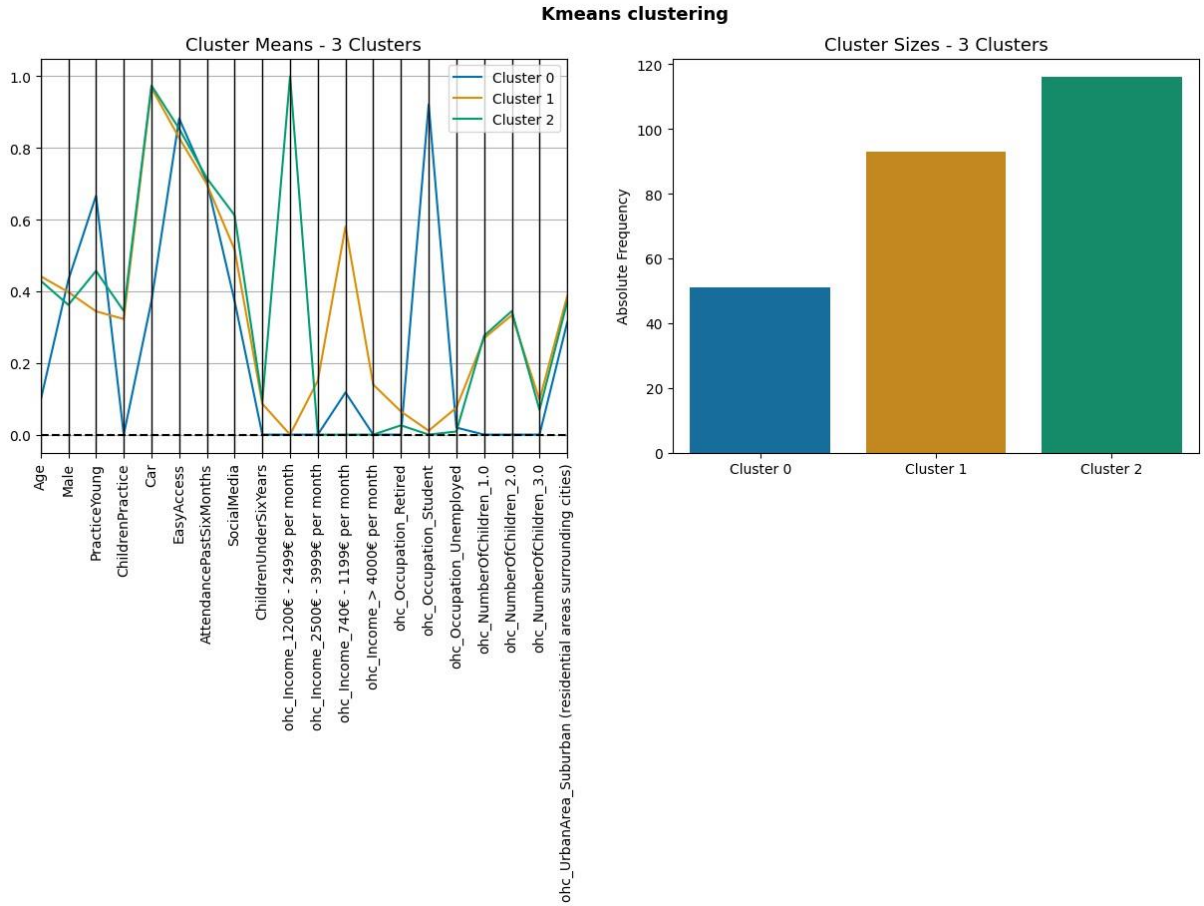

Figure 7 - Means and Distribution of the Clusters

Figure 7 shows the results of the clusters. The image on the left reflects the main characteristics of the clusters, whereas the one on the right represents the clusters' sizes.

In particular, the specific clusters of this study can be described as:

- *Bronze*: students who, even if they do not have a car, have easy access to theatres and prefer to attend concerts or dance performances. Their primary motivations are to experience high-quality arts and to socialize with friends and family. Most of them

practised a performing art when they were young, as also seen by authors (Andreasen & Belk, 1980). Their main barriers are cost and not having anyone to go with.

- *Silver*: employed people with younger children and have a car. They prefer to attend concerts and theatres. Their main barriers are time and having small children, whereas their main motivation is to experience high-quality arts.
- *Gold*: middle-class employed people with children who have a car. The majority of these are the ones who follow theatres on social media. Their primary motivations are to experience high-quality arts, socialize with friends and family and to learn new things. Furthermore, their main barrier is cost. These people attended a performing art show more frequently in the past six months.

It is important to note that these results might be biased, because of the data collection method used. As it was a survey designed by me, based on the existing literature and the evolution of technology, namely social media, with only 260 valid participants, the conclusions may not be quite accurate, as it is not a large sample.

However, based on these results, it is still possible to discuss the practical implications of the findings for performing arts organizations, addressed to the identified clusters:

- *Bronze*: promote events where people can socialize with others with the same interests in arts. These are activities before or after the show where everyone can go alone, meet new people, and eventually make new friends, as (NEA Office of Research & Analysis, 2015) purposed. It could also provide a discount for students with student cards.
- *Silver*: offer a type of "childcare" where the parents can leave their children while attending the performing art show. This could be free when one adult ticket was purchased, for example.
- *Gold*: offer an exclusive discount on social media since they follow the theaters online, but their main barrier is cost.

On the other hand, as shown by cluster *Gold* and confirmed in the data understanding phase, it can be seen that people who follow theaters and companies selling tickets on social media or receive newsletters are more likely to attend performing art shows. Thus, it is important to invest more in digital marketing. Some actions to be taken could be to create partnerships with influencers or other brands since many people follow celebrities on social media, especially young ones. Moreover, since the ones who engage less online with theatres are students, perhaps invest more in *Instagram*, the most popular app among them. (Marktest, 2023)

These actions aim to improve affluence in theaters, by encouraging people who are already interested, try to mitigate their barriers and difficulties and invest on their motivations.

Finally, since most people mentioned their main motivation was to experience high quality arts, this could also be an important point. Promoting performing art shows as being of the highest quality they actually are is crucial because it is known that people are attracted to good publicity and something that catches their attention. So, to eventually find a high-quality performance, they must first be drawn to the publicity. Artists deserve the opportunity to be seen, supported and appreciated. With this in mind, good publicity may be a good way to get more people to attend performing art shows.

5. CONCLUSIONS AND FUTURE WORK

This study aims to analyze attendance at performing art shows, improve it, bring art to more people, and increase people's awareness of the performing arts. The main objectives include: 1) Analyzing patterns of people's behaviour and their perception of the performing arts; 2) Understanding what influences people to attend performing art shows and what could be preventing them from attending; 3) Develop better strategies to improve affluence in performing art shows; 4) Understanding which features have more weight on the decision of attending performing art shows.

After all data analysis, the first objective was reached by creating clusters that combine people's behaviour patterns and gather them by similarity. The conclusion was that there were three patterns:

- *Bronze:* Students who have easy access to theaters but do not own cars. Many were involved in performing arts as children. They are inspired by interacting with friends and family and expericiencing high-quality arts. Some barriers include not having anyone to go with and cost.
- *Silver:* Working adults with small children who own a car. They are motivated by experiencing high-quality arts, but time and having young children rise as barriers.
- *Gold:* People in the middle-class who work, have cars, kids, and are active on social media. Motivated by experiencing high-quality arts, socializing with friends and family, and learning new things, having cost as main barrier. Attendance at performing art shows on a more regular basis in the previous six months.

The second one was reached by analyzing the frequency of the responses to the motivations and barriers questions, which raised the main conclusion that the greatest motivation to attend performing art shows is experiencing high-quality arts, followed by socializing with friends, and the most voted barrier is cost, followed by time.

The third objective was addressed by suggesting some actions that could be taken to improve affluence in performing art shows, either in general or specified for each cluster in the solution, stated in Chapter 4 – Results and Discussion.

Finally, the fourth objective was reached by performing feature selection, which made it possible to eliminate some features that were not so relevant to the purpose of the project. The features that have more impact on the decision to attend performing art shows are age, sex, whether or not people used to practice performing art when younger or if their children practice and the accessibility to theatres. Moreover, it was possible to see that the new feature created impact on the attendance, since people who follow theaters or ticket selling companies on social media or receive newsletters are more likely to attend a performing art show.

This project included some limitations. The first is data, in that it only gathered 290 participants, 18 years old or older, which probably did not include very diversified people in terms of age, geography, occupation and income. This could have affected the clusters and, consequently, the suggestions for improving attendance at performing art shows.

Regarding future work, the main goal is to expand this project and gather more data to get more accurate results and help theatres with more confidence to increase their profit, so more artists can work in what they like to do. Besides this, it would be a good idea to do a kind of *business case* with one theater and work directly with them to analyze only people who attend that specific venue *ceteris paribus*, so that the focus is on what makes people attend and if they usually go but do not attend some of them interviewing them to ask the reason.

Another suggestion could be to explore more the social media feature, in a way that it could be interesting to dive into what kind of content people are more drawn to. That can include the definition of KPI's to better analyze and understand how many *clicks* or *conversions* the website is getting by type of event or artist, so that the theater can address their strategy to that specific audience and increase its attendance.

BIBLIOGRAPHICAL REFERENCES

Andreasen, A. R., & Belk, R. W. (1980). Predictors of Attendance at the Performing Arts. *Journal of Consumer Research*, *7*(2), 112–120. https://doi.org/10.1086/208800

Bishop, T. (2021). *10 Reasons why engagement in the Performing Arts is critical for future success. Sunmarke School Blog*. https://www.sunmarke.com/blog/art/12145/engagement-in-the-performing/

Bone, J. K., Bu, F., Fluharty, M. E., Paul, E., Sonke, J. K., & Fancourt, D. (2021). Who engages in the arts in the United States? A comparison of several types of engagement using data from The General Social Survey. *BMC Public Health*, *21*(1), 1–13. https://doi.org/10.1186/S12889-021-11263-0

Carter, M. A. (2020). *WHY WE ENGAGE: ATTENDING, CREATING, AND PERFORMING ART*. https://www.arts.gov/sites/default/files/Why-We-Engage-0920_0.pdf

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *Step-by-step data mining guide*. DaimlerChrysler.

Fresh Essays. (2023). *Current Issues in Portuguese Performing Art | Free Essay Examples*. https://samples.freshessays.com/current-issues-in-portuguese-performing-art.html

Hill Strategies. (2016). *When Going Gets Tough: Barriers and Motivations Affecting Arts Attendance - Hill Strategies Research Inc*. https://hillstrategies.com/2016/02/24/when-going-gets-tough-barriers-and-motivations-affecting-arts-attendance/

Klickstein, G. (2022). *Motivations & Barriers to Arts Attendance | MusiciansWay.com*. https://www.musiciansway.com/blog/2022/08/motivations-and-barriers-to-arts-attendance/

Manolika, M., & Baltzis, A. (2022). Concert Hall, Museum, Cinema, and Theater Attendance: What Difference Do Audience Motivations and Demographics Make? *Empirical Studies of the Arts*, *40*(1), 37–56. https://doi.org/10.1177/0276237420979569/ASSET/IMAGES/LARGE/10.1177_0276237420979569-FIG1.JPEG

Marktest. (2023). *Os Portugueses e as Redes Sociais 2023 : Estudos & Serviços*. https://www.marktest.com/wap/a/grp/p~96.aspx

Mitchell, A., & SRI International. (1983). *The Professional Performing Arts: Attendance Patterns, Preferences and Motives | Americans for the Arts*. https://www.americansforthearts.org/by-program/reports-and-data/legislation-policy/naappd/the-professional-performing-arts-attendance-patterns-preferences-and-motives

NEA. (2015). *Defining Interested Non-Attendance and the Barriers to Attendance | National Endowment for the Arts*. https://www.arts.gov/impact/research/arts-data-profile-series/adp-4/defining-interested-non-attendance

NEA Office of Research & Analysis. (2015). *When Going Gets Tough: Barriers and Motivations Affecting Arts Attendance*.

Seaman, B. A. (2005). Attendance and Public Participation in the Performing Arts: A Review of the Empirical Literature. *SSRN Electronic Journal*. https://doi.org/10.2139/SSRN.895099

Walmsley, & Ba. (2011). *Why people go to the theatre: a qualitative study of audience motivation*. https://doi.org/10.1362/147539211X13210329822545

APPENDIX A – ETHICS COMMITTEE

This is to certify that
Project No.: **DSCI2024-5-305727**
Project Title: **Analyzing attendance at performing art shows**
Principal Researcher: **Ana Sofia da Silva Mendes**

according to the regulations of the Ethics Committee of NOVA IMS and MagIC Research Center this project was considered to meet the requirements of the NOVA IMS Internal Review Board, being considered **APPROVED** on 5/30/2024.

It is the Principal Researcher's responsibility to ensure that all researchers and stakeholders associated with this project are aware of the conditions of approval and which documents have been approved.

The Principal Researcher is required to notify the Ethics Committee, via amendment or progress report, of
- Any significant change to the project and the reason for that change;
- Any unforeseen events or unexpected developments that merit notification;
- The inability of the Principal Researcher to continue in that role or any other change in research personnel involved in the project.

Lisbon, 5/30/2024

NOVA IMS Ethics Committee
ethicscommittee@novaims.unl.pt

APPENDIX B – SURVEY

Analyzing attendance at performing art shows

Hello! I'm currently pursuing my Data Science Master's Thesis and would really appreciate your help in completing a brief survey, which would only take a maximum of 10 minutes of your time.

Your input in this survey is vital for my research, which aims to analyze attendance patterns at performing art shows. By understanding what motivates people to attend and what obstacles they face, I hope to enhance audience engagement in these events.

It's important to know that this survey is completely anonymous, as I won't be asking for your name, email, or any other Personal Identifying Information. Your responses will remain confidential and will be utilized solely for academic purposes, ensuring you can provide honest feedback without any reservations.

This survey is open to individuals aged 18 and above, and there are no right or wrong answers. Feel free to share it with others; the more data I gather, the more insightful the findings will be.

Thank you for your participation!

CONSENT

The purpose of this research project is to analyze attendance at performing art shows, exploring both motivations and deterrents, with the aim of enhancing audience turnout.

Your participation in this research study is entirely voluntary, and you may opt out if you choose.

Participation involves completing a brief online survey lasting approximately 10 minutes. No Personal Identifying Information will be requested, ensuring the confidentiality and anonymity of your responses. The findings of this study will be used strictly for academic purposes.

Should you have any inquiries regarding the research study, please feel free to contact Ana Sofia Mendes at 20220687@novaims.unl.pt.

ELECTRONIC CONSENT:

By clicking the "I agree" button below, you confirm that:

- You have read and understood the above information.
- You voluntarily consent to participate.
- You are at least 18 years old.

If you prefer not to participate in the research study, kindly decline by clicking on the "I disagree" button.

* Obrigatória

1. Please select your choice bellow: *

- ◯ I agree
- ◯ I disagree

2. Age *

3. Sex *

- ◯ Male
- ◯ Female
- ◯ Outro

4. Income *

- ◯ 0€ - 739€ per month
- ◯ 740€ - 1199€ per month
- ◯ 1200€ - 2499€ per month
- ◯ 2500€ - 3999€ per month
- ◯ > 4000€ per month

5. Occupation *

- ◯ Student
- ◯ Employed
- ◯ Unemployed
- ◯ Retired

6. Did you practice a performing art when you were young? *

- ◯ Yes
- ◯ No

7. How many children do you have? *

- ◯ 0
- ◯ 1
- ◯ 2
- ◯ 3 or more

8. Is at least one of your children younger than 6 years old? *

- ◯ Yes
- ◯ No

9. Is your child younger than 6 years old? *

- ◯ Yes
- ◯ No

10. Do your children practice a performing art? *

- ◯ Yes
- ◯ No

11. Do you live in a urban or rural area? *

- ◯ Urban (city center)
- ◯ Suburban (residential areas surrounding cities)
- ◯ Rural (small towns)

12. Do you have a car? *

- ◯ Yes
- ◯ No

13. Do you have easy acessability to theaters? *

- ◯ Yes
- ◯ No

14. Considering that performing art shows are live presentations to an audience, do you have interest in attending them? *

- ◯ Yes
- ◯ No, I prefer non live shows
- ◯ I enjoy both live and non live shows
- ◯ I'm not interested in art at all

15. Which type of performing art do you prefer to watch? *

- ◯ Dance
- ◯ Theater
- ◯ Concert
- ◯ Comedy show
- ◯ Circus
- ◯ Magic
- ◯ Acrobatics
- ◯ Outro

16. Did you attend a performing art show in the past six months? *

- ◯ Yes
- ◯ No

17. What type of performing art was the last show you attended? *

18. What is your main motivation to attend a performing art show? *

- ◯ Socializing with friends or family
- ◯ Learning new things
- ◯ Experiencing high-quality arts
- ◯ Supporting the community and/or the artists
- ◯ Outro

19. Which main barrier can prevent you from going? *

- ◯ Time
- ◯ Cost
- ◯ Dificult access
- ◯ No one to go with
- ◯ None, I always attend
- ◯ Outro

20. What makes you select performing art shows instead of others (like cinema, exposition, etc)? *

- ☐ It´s live
- ☐ Makes me feel part of the show
- ☐ I'm fond of improvisation
- ☐ It's more dinamic and realistic
- ☐ None of the above
- ☐ Outro

21. Do you follow theaters or ticket selling companies on social media / receive newsletters by email / check their websites? *

- ◯ Yes
- ◯ No

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa