POST-PANDEMIC TOURISM TRENDS IN PORTUGAL: INSIGHTS FROM SOCIAL MEDIA ANALYTICS AND COMPARATIVE ANALYSIS WITH KEY GLOBAL COMPETITORS

Data Science for Marketing

Bernardo Abreu – 20230062

Fernanda Trindade – 20230070

Muhammet Emin Imir – 20231378

Tausif Ahmad – 20231030

January 2024

Executive Summary

This project aims to provide valuable insights for the National Tourism Board Organization (NTBO) in a challenging post-pandemic period. Using an approach based on the CRISP-DM model, we divided our analysis into four main notebooks. The first notebook covers Business Understanding, Data Understanding, and Data Preparation, laying a solid foundation for the subsequent analyses. In the second notebook, we focus on Market Basket analysis of tourist attractions, exploring rule-based patterns of Support, Confidence, and Lift. The third notebook is dedicated to RFM analysis, focusing mainly on users of segment 144, identified as the most promising. Finally, the fourth notebook performs a comparative assessment, analyzing changes in visitor patterns before and after the pandemic and highlighting Portugal's main tourism competitors. Together, these findings aim to help the NTBO better understand visitor patterns and formulate effective strategies to revitalize tourism in the country.

It was observed that the pandemic caused a sharp drop in visitor numbers in Europe in general during the implementation of this project. Regarding Portugal, the number of tourists from the United States, the United Kingdom and Canada decreased by almost 100%, while the number of tourists from Australia reached zero. These are the countries from which most tourists come.

Likewise, the presence of tourist attractions that are considered icons of Portugal was observed, appearing alongside other equally famous places, for example: Mosteiro dos Jerónimos, Park and National Palace of Pena, Casa Milà - La Pedrera, and Parc Guell.

RFM analysis provided deep insight into visitor behavior patterns, allowing us to identify groups with similar characteristics for targeted marketing efforts. Overall, RFM analysis serves as a data-driven framework to enhance tourism marketing efforts and maximize return on investment in attracting and retaining tourists.

In conclusion, this project aims to provide concrete data to support NTBO's strategic marketing decisions and contribute to Portugal's return among the most visited destinations in Europe.

Index

1. Introduction

The National Tourism Board Organization (NTBO) asked us to study user-generated content related to various tourist attractions in Europe. The goal was to understand visitors' patterns and whether those patterns changed because of the pandemic.

Despite the impact of the pandemic, the tourism sector in Portugal continues to play an important role in the country's economy, accounting for 5.2% of Portuguese GDP in 2020/21, but has already reached a share of almost 15% in 2018/19.

According to data from the *Tripadvisor* platform, Portugal ranks fourth among the countries that have received the most reviews and are therefore the most visited. And this result shows that Portugal can compete for the top spots among the most visited destinations with countries like the United Kingdom, Spain, and Italy.

It must be considered that the pandemic has strongly influenced the analyzed results and that it will take some time to return to the previous normality. However, this means that the difficult path also presents itself to a greater or lesser extent to other European countries, it is true, but the difficulty exists for everyone.

During the analysis, the presence of notable tourist attractions in Portugal was verified, along with other places that are very popular among travelers, confirming the potential of this country as one of the most important European tourist destinations. After all, it has been voted the world's best destination for three years running by the World Travel Awards. In addition to other important categories such as the best city break, and island and adventure destination in the world.

In short, it's about studying the numbers and developing a strategic campaign to bring visitors back by encouraging the different categories of tourism that the country has to offer.

2. Methodology

The aim of this project is to uncover patterns and changes in Portugal's tourism landscape before and after the Covid-19 pandemic by analyzing user reviews from the social media platform TripAdvisor. This analysis is centered on the dataset "EuropeTop100Attractions_ENG_20190101_20210821," provided by the NTBO, to extract valuable insights from these user-generated reviews.

The analyses for this project were predominantly carried out using the Python programming language, specifically within Jupyter Notebooks facilitated by the Anaconda Navigator system. This choice of tools allowed for a flexible and powerful analysis environment. Alongside this, Microsoft Excel was also utilized for some aspects of data exploration. Excel's user-friendly interface provided an accessible means for initial data examination and simpler analytical tasks, complementing the more complex analyses performed in Python. This combination of Python's robust capabilities and Excel's intuitive layout ensured a comprehensive and effective approach to data analysis.

Given the extensive nature of the dataset, the methodology adopted for structuring this project was the CRISP-DM methodology. CRISP-DM, which stands for Cross-Industry Standard Process for Data Mining, is designed with practical application at its core. Through iterative refinements, this process has evolved, integrating practical, real-world experiences of data scientists in executing data mining projects (Chapman et al., 2000). Therefore, CRISP- DM is a structured approach for conducting data mining projects. It follows six distinctive steps. The process begins with the Business Understanding phase, where the objectives and requirements of the project are defined. This is followed by the Data Understanding phase, involving initial data collection and familiarization. Next is the Data Preparation phase, where data is cleaned and formatted for analysis. The Modeling phase then involves selecting and applying various techniques to the prepared data. In the Evaluation phase, the results are assessed to ensure they meet the project objectives. Finally, the Deployment phase involves implementing the findings into the organization's operations. This process is iterative, often requiring revisiting earlier steps as new insights emerge (Chapman et al., 2000).

In this project, navigating through most of the CRISP-DM steps, it was dedicated a significant portion of our development time to the data preparation phase. This stage was particularly crucial, as the dataset contained numerous entries with missing or invalid data. Addressing these issues was a meticulous process, requiring careful attention and prompt corrections to ensure the integrity and reliability of the analysis. The time invested in this phase was essential for laying a solid foundation for the subsequent stages.

During the modeling phase, tailored to the specific goals of the data project, it was implemented two key techniques: Market Basket Analysis and the Recency, Frequency, and Monetary (RFM) Model. These methodologies were chosen for their relevance and effectiveness in addressing the unique data objectives.

Market Basket Analysis is a technique used to uncover associations between items within a business context, revealing how certain objects frequently co-occur (Loshin, 2013). While its name originates from the typical supermarket setting, the application of Market Basket Analysis extends far beyond this scenario. It involves examining a variety of item collections to discover affinities or patterns that can be strategically leveraged for business benefits (Loshin, 2013). This method was employed to uncover the relationships between different tourist attractions and determine which ones are frequently visited in conjunction, providing insights into common travel patterns.

Additionally, the RFM Model is a behavior-based model used to analyze the behavior of a customer and then make predictions based on the behavior in the database (Hughes, 1994). Furthermore, "Recency" measures the time elapsed since the last purchase, "Frequency" indicates the number of purchases within a set time frame, and "Monetary" refers to the total spending in that period. These three metrics are behavioral variables, serving as segmentation tools by analyzing customer attitudes towards products, brands, benefits, and even loyalty (Hughes, 1994).

In the case for this project, the adaptation of these three variables to fit the data objective and problem:

- **Recency:** the recency of reviews for each user, calculated as the difference in days between the maximum global review date and the most recent review date for the user. Higher values indicate that the user has made more recent reviews.
- **Frequency:** the number of reviews each user has written.
- **Monetary:** the average of the ratings given by the user in their reviews. The average is calculated from individual review ratings.

3. Data Overview

This chapter of the report delves into the Data Understanding and Data Preparation phases, providing a detailed examination of these crucial stages in the data analysis process.

General analysis of the received data showed that much of it was not in a condition for processing at that time. Since *Tripadvisor* does not have criteria for filling out the review form, people entered the information they wanted in the field related to the user's location, for example. In addition, many people did not fill in the trip type they were taking, leaving this column with many missing values.

An additional observation noted in the study was that two attractions had their identification codes replaced by unrelated words, namely: “genis” and “u”, representing “Staromestske namesti” and Edinburgh Castle respectively (shown in Appendix 3.1. and 3.2., respectively).

Another noteworthy observation concerns duplicate information. Although it was possible to identify these lines by looking directly at the datasheet, they simply did not appear when we looked for them through coding (shown in Appendix 3.3. and 3.4., respectively).

The reason for this is that there may be cases where duplicates are considered different entries because of spaces or upper-case and lower-case letters. The solution to this problem was to separate the duplicate lines from the rest of the data set to work on all the others.

The next step was to adjust the information in the "User Location" column and to correct a few errors in the spreadsheet regarding the tourist attractions. However, it took more time and effort to customize the User Location column.

When the Attractions spreadsheet was analyzed, it was discovered that there were several errors, ranging from typos to mistakes in the location of some attractions. For example, there was a country called “Curaçao” which is considered to belong to the Netherlands, although is situated in the Lesser Antilles, Caribbean. As this destination is likely to be competing with other islands in the Caribbean, we consider it best to remove it from the analysis as it is not part of the region being analyzed.

As previously mentioned, users filled in the User Location field as they saw fit. So instead of identifying your origin place by the country’s name and its acronym, there was a variety of different information in this regard. Most of the data were corrected by coding, separating the column into two distinct pieces of information: one for country and one for ISO code. However, the remaining data had to be corrected manually in the Excel environment.

Meanwhile, it is important to note that some locations in the User Location column were numbers only, when they should be representing locations. Unfortunately, these rows have been deleted from the dataset.

Finally, after all these data understanding and preparation steps were completed, it was possible to merge the user review and tourist attraction data sets. This way it is possible to correctly identify the origin of the users, which attractions they have visited and in which country they have been.

This information will provide the answers that the National Tourism Board needs to promote the tourism in Portugal.

4. Key Findings

Market Basket Analysis (MBA), traditionally a tool for discovering product pairing trends in retail and e-commerce, has been applied in this project to delve into tourist behaviors. By using MBA, we aim to unravel the interconnectedness of various tourist attractions and understand which ones are commonly visited together, offering a fresh perspective on travel patterns. This analytical approach serves as a powerful means to decode the sequence and combination in which tourists prefer to visit attractions.

When analyzing tourist attraction data, various methods were used to uncover interesting patterns. The initial approach focuses on evaluating the reliability of these patterns, sorting them based on their level of confidence. For instance, a notable rule discovered indicates a strong link between visitors who toured Real Alcazar de Sevilla (Seville, Spain), Casa Milà - La Pedrera (Barcelona, Spain), and the Palace of Catalan Music (Barcelona, Spain), as they were also likely to visit Casa Batlló (Barcelona, Spain). This highlights a significant connection between the attractions and Casa Batlló.

To enhance our understanding, we looked at how often certain pairs of famous places in Portugal are visited together. For example, the Torre de Belém and the Mosteiro dos Jerónimos in Lisbon often appear as a popular duo, with about 0.46% of tourists visiting both. We then further analyzed these patterns to see which places are commonly visited one after the other. We found that visitors who go to the Mosteiro dos Jerónimos often also visit famous sites in Barcelona, Spain, like Casa Milà - La Pedrera and Parc Guell. This pattern is strong, suggesting that tourists tend to explore these attractions in a similar sequence.

In our final review, we noticed that tourists often visit Mosteiro dos Jerónimos and the Park and National Palace of Pena in Lisbon before heading to Parc Guell. This information, gathered from our pattern analysis, offers useful insights for planning trips, showcasing which famous landmarks are likely to be visited in sequence.

RFM (Recency, Frequency, Monetization) analysis is a sophisticated technique that yields critical insights into tourist preferences regarding specific attractions, facilitating the creation of marketing campaigns and the efficient allocation of companies' resources. This method segments tourists based on their engagement patterns, enabling the identification of distinct groups that exhibit similar characteristics for targeted promotional strategies.

Tourists with high RFM scores are typically deemed more valuable due to their higher engagement and interest levels, making them prime targets for focused marketing initiatives. Conversely, those with lower RFM scores might need alternative engagement strategies to enhance their interaction with the attractions.

In this analysis, the following key metrics were evaluated: the recency of tourists' visits, the frequency of these visits, and their overall engagement or expenditure levels. A lower recency score is indicative of recent or frequent visits, suggesting active engagement with the attraction. In contrast, higher frequency and monetization scores denote more frequent visits and higher levels of spending.

This analysis centers on the characteristics of the "144" visitor segment, which exhibits a notably low average recency. This suggests that the segment comprises either new or highly active visitors. Furthermore, their average frequency and spending patterns indicate a robust level of engagement and expenditure. Intriguingly, certain tourist attractions, such as Don Luis I Bridge (Porto, Portugal), the Park and National Palace of Pena, and Cais da Ribeira (Porto, Portugal), show minimal presence of visitors from this "144" segment.

However, these attractions should be viewed not as shortcomings, but as untapped opportunities. Even a modest number of visitors from this segment to these attractions could signify a keen interest.

Overall, this study illuminates the intricate patterns and connections among various tourist attractions, offering invaluable insights for strategists aiming to optimize tourism marketing. In this context, a thoughtful approach to setting goals, leveraging customer insights, and assessing feasibility becomes paramount in making well-informed decisions and enhancing marketing initiatives. This careful balance of analysis and strategy can lead to more effective engagement with potential tourists and a better understanding of the dynamics within the tourism sector.

5. Discussion and Recommendations

The findings from the Market Basket Analysis (MBA) and RFM analysis in tourism reveal several key insights.

Firstly, the interconnectedness of attractions suggests a pattern in tourist itineraries, highlighting opportunities for collaborative marketing between these attractions. The preference for visiting attractions within the same city or region, such as Mosteiro dos Jerónimos and Torre de Belém, indicates the influence of geographical and cultural proximity on tourist choices.

Additionally, the RFM analysis uncovers varying levels of engagement among tourists, with some showing higher frequency and recency in their visits, which can inform tailored marketing strategies.

Interestingly, the minimal presence of certain visitor segments in some attractions, like Don Luis I Bridge and Cais da Ribeira, suggests untapped potential in these areas. This variation in visitor engagement and attraction popularity provides a rich ground for developing nuanced and effective marketing strategies in the tourism sector.

This analysis provides great insights into the overall tourism sector and in specific for the NTBO of Portugal. Some of the main possible actionable marketing strategies could be:

- **Collaborative Marketing Campaigns:** Develop joint marketing strategies for attractions that are frequently visited together. For example, package deals or joint advertising for Mosteiro dos Jerónimos and nearby attractions in Lisbon. This approach could yield even more favorable results if these packages were marketed at competitive price points, strategically positioned against the main competitors in Portugal's tourism sector, such as United Kingdom, Spain, and Italy.

- **Segment-Specific Engagement Strategies:** Tailor marketing campaigns to different visitor segments identified in the RFM analysis. High-value visitors (with high RFM scores) could be targeted with premium experiences or loyalty programs, while strategies to engage lower-score tourists might include discounts or special events.

- **Boosting Lesser-Visited Attractions:** Design specific campaigns to increase the visibility and appeal of lesser-visited attractions. This could involve storytelling, highlighting unique aspects of these attractions, or bundling them with more popular attractions.

- **Digital Marketing and Social Media Engagement:** Using digital marketing and social media to target specific tourist segments, leveraging data insights to craft compelling content that resonates with the interests and preferences of those segments. Additionally, considering that the foundation of this project is the dataset derived from *Tripadvisor* reviews, it would be advantageous for the NTBO to boost engagement on this digital platform. This approach would not only provide a richer dataset for future analysis but also enable the organization to stay ahead of rapid shifts in tourist behavior.

6. Conclusion

In this project, we conducted a comprehensive analysis of visitor patterns to Portuguese attractions, focusing on the impact of the COVID-19 pandemic. We employed the CRISP-DM process model and used various data cleansing techniques to ensure the accuracy and reliability of our findings. Our journey began with a treasure trove of data and a map to guide us. But like any map, it needed to be deciphered. We found hidden gaps, duplicates, and outliers, much like hidden traps in a treasure hunt. But with our trusty tools, we navigated through, ensuring our map was accurate and reliable. We discovered patterns, like footprints in the sand, and generated association rules based on these patterns. We calculated all possible combinations of pairs of Portuguese landmarks, like joining pieces of a puzzle. The pairs were then sorted in descending order by support, revealing the most popular combinations.

We filtered association rules, like sieving through the sand for precious gems. We analyzed the lines in which Portuguese tourist attractions appear and other tourist attractions, revealing fascinating relationships. This was like viewing our map from a different perspective, revealing new information about each user, including the recency of the reviews, the frequency (number of reviews), and the average review rating. We calculated RFM (recency, frequency, and monetary) scores for each entry in DataFrame X, like decoding a secret language. These scores revealed groups of customers with similar characteristics, like tribes in an unexplored land.

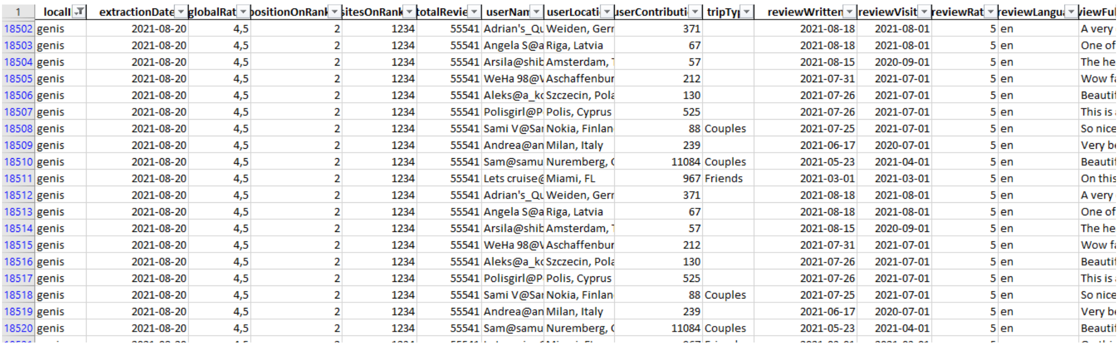
Finally, our adventure through the realm of data provided valuable insights into visitor patterns at Portuguese attractions during the pandemic. These findings can serve as a guide for the Portuguese National Tourism Board organizations, helping them navigate the post-pandemic world. However, it's important to remember that our map is based on available data, and certain areas remain unexplored due to missing values and outliers. Future explorations could involve refining our map and exploring additional data sources for a more comprehensive understanding of visitor patterns.

Bibliography

- Boosting Portugal, promoting the tourism destination. (2022). Turismodeportugal.pt. https://www.turismodeportugal.pt/en/O%20que%20fazemos/PromoverDestinoPortugal/Pages/default.aspx
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRIPS-DM 1.0: Step-by-step data mining guide*. CRISP-DM consortium.
- Hughes, A. M. (1994). *Strategic Database Marketing: The Masterplan for Starting and Managinng a Profitable, Customer-Based Marketing Program* (1st Edition). Probus Publishing.
- Loshin, D. (2013). Chapter 17 – Knowledge Discovery and Data Mining for Predictive Analytics. In *Business Intelligence: The Savvy Manager's Guide* (2nd Edition) (pp. 271-286). Massachusetts: Morgan Kaufmann.
- Riqueza criada no comércio e turismo em 2021 ficou ainda 12,5% abaixo do valor pré-pandemia. (n.d.). Jornal Expresso. https://expresso.pt/economia/2022-02-28-riqueza-criada-no-comercio-e-turismo-em-2021-ficou-ainda-125-abaixo-do-valor-pre-pandemia
- Turismo e viagem para Portugal 2023 - Férias em Portugal. (n.d.). Tripadvisor. https://www.tripadvisor.com.br/Tourism-g189100-Portugal-Vacations.html

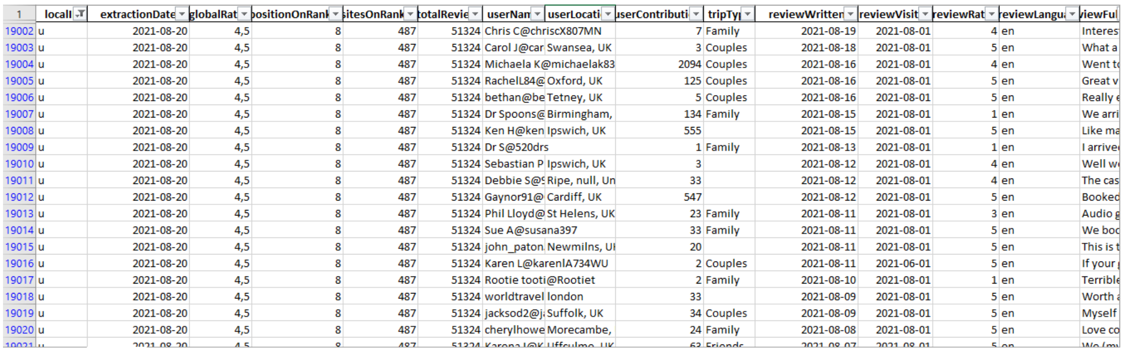
Appendix

3.1. Data Preparation – genis



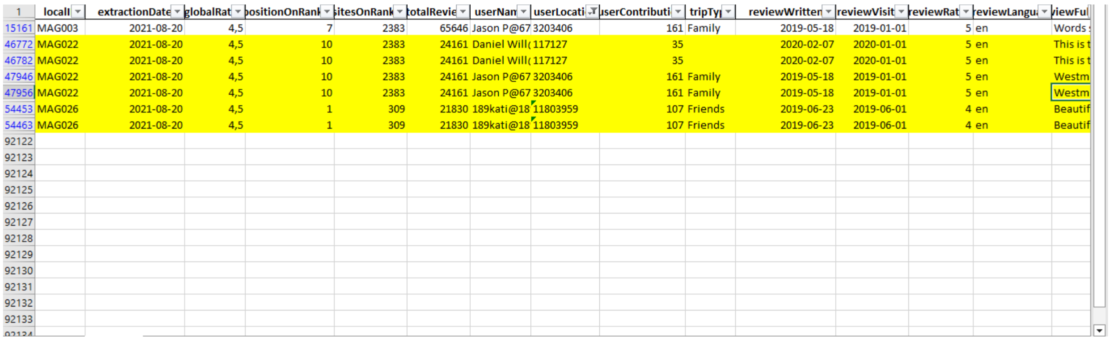
1	localI	extractionDate	globalRat	positionOnRank	sitesOnRank	totalRevie	userNam	userLocati	userContributi	tripTyp	reviewWritten	reviewVisit	reviewRat	reviewLangua	viewFul
18502	genis	2021-08-20	4,5	2	1234	55541	Adrian's_Qu	Weiden, Gerr	371		2021-08-18	2021-08-01	5	en	A very
18503	genis	2021-08-20	4,5	2	1234	55541	Angela S@a	Riga, Latvia	67		2021-08-18	2021-08-01	5	en	One of
18504	genis	2021-08-20	4,5	2	1234	55541	Arsila@shib	Amsterdam, 1	57		2021-08-15	2020-09-01	5	en	The he
18505	genis	2021-08-20	4,5	2	1234	55541	WeHa 98@\	Aschaffenbur	212		2021-07-31	2021-07-01	5	en	Wow fa
18506	genis	2021-08-20	4,5	2	1234	55541	Aleks@a_kc	Szczecin, Pola	130		2021-07-26	2021-07-01	5	en	Beautif
18507	genis	2021-08-20	4,5	2	1234	55541	Polisgirl@P	Polis, Cyprus	525		2021-07-26	2021-07-01	5	en	This is a
18508	genis	2021-08-20	4,5	2	1234	55541	Sami V@Sar	Nokia, Finlan	88	Couples	2021-07-25	2021-07-01	5	en	So nice
18509	genis	2021-08-20	4,5	2	1234	55541	Andrea@an	Milan, Italy	239		2021-06-17	2020-07-01	5	en	Very be
18510	genis	2021-08-20	4,5	2	1234	55541	Sam@samu	Nuremberg, (	11084	Couples	2021-05-23	2021-04-01	5	en	Beautif
18511	genis	2021-08-20	4,5	2	1234	55541	Lets cruise@	Miami, FL	967	Friends	2021-03-01	2021-03-01	5	en	On this
18512	genis	2021-08-20	4,5	2	1234	55541	Adrian's_Qu	Weiden, Gerr	371		2021-08-18	2021-08-01	5	en	A very
18513	genis	2021-08-20	4,5	2	1234	55541	Angela S@a	Riga, Latvia	67		2021-08-18	2021-08-01	5	en	One of
18514	genis	2021-08-20	4,5	2	1234	55541	Arsila@shib	Amsterdam, 1	57		2021-08-15	2020-09-01	5	en	The he
18515	genis	2021-08-20	4,5	2	1234	55541	WeHa 98@\	Aschaffenbur	212		2021-07-31	2021-07-01	5	en	Wow fa
18516	genis	2021-08-20	4,5	2	1234	55541	Aleks@a_kc	Szczecin, Pola	130		2021-07-26	2021-07-01	5	en	Beautif
18517	genis	2021-08-20	4,5	2	1234	55541	Polisgirl@P	Polis, Cyprus	525		2021-07-26	2021-07-01	5	en	This is a
18518	genis	2021-08-20	4,5	2	1234	55541	Sami V@Sar	Nokia, Finlan	88	Couples	2021-07-25	2021-07-01	5	en	So nice
18519	genis	2021-08-20	4,5	2	1234	55541	Andrea@an	Milan, Italy	239		2021-06-17	2020-07-01	5	en	Very be
18520	genis	2021-08-20	4,5	2	1234	55541	Sam@samu	Nuremberg, (	11084	Couples	2021-05-23	2021-04-01	5	en	Beautif

3.2. Data Preparation – u



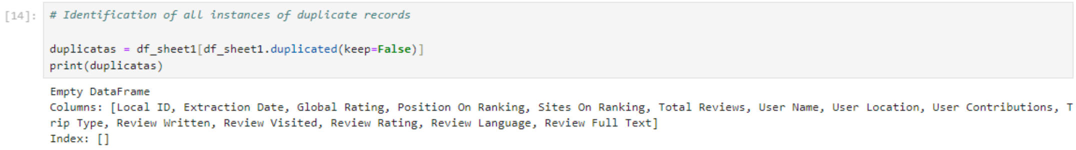
1	localI	extractionDate	globalRat	positionOnRank	sitesOnRank	totalRevie	userNam	userLocati	userContributi	tripTyp	reviewWritten	reviewVisit	reviewRat	reviewLangua	viewFul
19002	u	2021-08-20	4,5	8	487	51324	Chris C@chriscX807MN		7	Family	2021-08-19	2021-08-01	4	en	Interes
19003	u	2021-08-20	4,5	8	487	51324	Carol J@car	Swansea, UK	3	Couples	2021-08-18	2021-08-01	5	en	What a
19004	u	2021-08-20	4,5	8	487	51324	Michaela K@michaelak83		2094	Couples	2021-08-16	2021-08-01	4	en	Went to
19005	u	2021-08-20	4,5	8	487	51324	RachelL84@	Oxford, UK	125	Couples	2021-08-16	2021-08-01	5	en	Great v
19006	u	2021-08-20	4,5	8	487	51324	bethan@be	Tetney, UK	5	Couples	2021-08-16	2021-08-01	5	en	Really e
19007	u	2021-08-20	4,5	8	487	51324	Dr Spoons@	Birmingham,	134	Family	2021-08-15	2021-08-01	1	en	We arri
19008	u	2021-08-20	4,5	8	487	51324	Ken H@ken	Ipswich, UK	555		2021-08-15	2021-08-01	5	en	Like ma
19009	u	2021-08-20	4,5	8	487	51324	Dr S@520drs		1	Family	2021-08-13	2021-08-01	1	en	I arrived
19010	u	2021-08-20	4,5	8	487	51324	Sebastian P	Ipswich, UK	3		2021-08-12	2021-08-01	4	en	Well we
19011	u	2021-08-20	4,5	8	487	51324	Debbie S@!	Ripe, null, Un	33		2021-08-12	2021-08-01	4	en	The cas
19012	u	2021-08-20	4,5	8	487	51324	Gaynor91@	Cardiff, UK	547		2021-08-12	2021-08-01	5	en	Booked
19013	u	2021-08-20	4,5	8	487	51324	Phil Lloyd@	St Helens, UK	23	Family	2021-08-11	2021-08-01	3	en	Audio g
19014	u	2021-08-20	4,5	8	487	51324	Sue A@susana397		33	Family	2021-08-11	2021-08-01	5	en	We boo
19015	u	2021-08-20	4,5	8	487	51324	john_paton	Newmilns, UK	20		2021-08-11	2021-08-01	5	en	This is t
19016	u	2021-08-20	4,5	8	487	51324	Karen L@karenlA734WU		2	Couples	2021-08-11	2021-06-01	5	en	If your
19017	u	2021-08-20	4,5	8	487	51324	Rootie tooti@Rootiet		2	Family	2021-08-10	2021-08-01	1	en	Terrible
19018	u	2021-08-20	4,5	8	487	51324	worldtravel	london	33		2021-08-09	2021-08-01	5	en	Worth a
19019	u	2021-08-20	4,5	8	487	51324	jacksod2@j	Suffolk, UK	34	Couples	2021-08-09	2021-08-01	5	en	Myself
19020	u	2021-08-20	4,5	8	487	51324	cherylhowe	Morecambe,	24	Family	2021-08-08	2021-08-01	5	en	Love co

3.3. Data Preparation – Data Duplication



1	localI	extractionDate	globalRat	positionOnRank	sitesOnRank	totalRevie	userNam	userLocati	userContributi	tripTyp	reviewWritten	reviewVisit	reviewRat	reviewLangua	viewFul
15161	MAG003	2021-08-20	4,5	7	2383	65646	Jason P@67	3203406	161	Family	2019-05-18	2019-01-01	5	en	Words
46772	MAG022	2021-08-20	4,5	10	2383	24161	Daniel Will(	117127	35		2020-02-07	2020-01-01	5	en	This is t
46782	MAG022	2021-08-20	4,5	10	2383	24161	Daniel Will(	117127	35		2020-02-07	2020-01-01	5	en	This is t
47946	MAG022	2021-08-20	4,5	10	2383	24161	Jason P@67	3203406	161	Family	2019-05-18	2019-01-01	5	en	Westm
47956	MAG022	2021-08-20	4,5	10	2383	24161	Jason P@67	3203406	161	Family	2019-05-18	2019-01-01	5	en	Westm
54453	MAG026	2021-08-20	4,5	1	309	21830	189kati@18	11803959	107	Friends	2019-06-23	2019-06-01	4	en	Beautif
54463	MAG026	2021-08-20	4,5	1	309	21830	189kati@18	11803959	107	Friends	2019-06-23	2019-06-01	4	en	Beautif
92122															
92123															
92124															
92125															
92126															
92127															
92128															
92129															
92130															
92131															
92132															
92133															

3.4. Data Preparation – Invisible Duplicates through Python



```
[14]: # Identification of all instances of duplicate records

      duplicatas = df_sheet1[df_sheet1.duplicated(keep=False)]
      print(duplicatas)

      Empty DataFrame
      Columns: [Local ID, Extraction Date, Global Rating, Position On Ranking, Sites On Ranking, Total Reviews, User Name, User Location, User Contributions, T
      rip Type, Review Written, Review Visited, Review Rating, Review Language, Review Full Text]
      Index: []
```