Team Members:  
Cara-Linn Reusch (Data Science Track)  
Julia Krebbers (Data Science Track)

Selinay Cengiz (Data Science Track)

Project on Github: <https://github.com/Catrex90/TechsLabTeam2>

# 

# **Title: HappyStay - Exploring Hotel Data for a Smile**

# **Abstract:**

Our project analysed guest data from two distinct hotels in Portugal: A city hotel in Lisbon and a resort hotel in the Algarve. During our analysis, we not only focussed on the data from the two hotels but also explored correlations between happiness ratings from diverse nations and travel preferences, focusing on visits to Portugal by combining a second dataset. This approach revealed valuable insights into guest behaviour and choices, aiding informed decision-making in the hospitality industry. Our findings emphasise notable disparities in guest origin and booking behaviour, especially regarding family stays. By grasping these differences, hotels can adjust their services to meet the needs of diverse clientele, optimising operations during peak booking periods.

# **Introduction:**

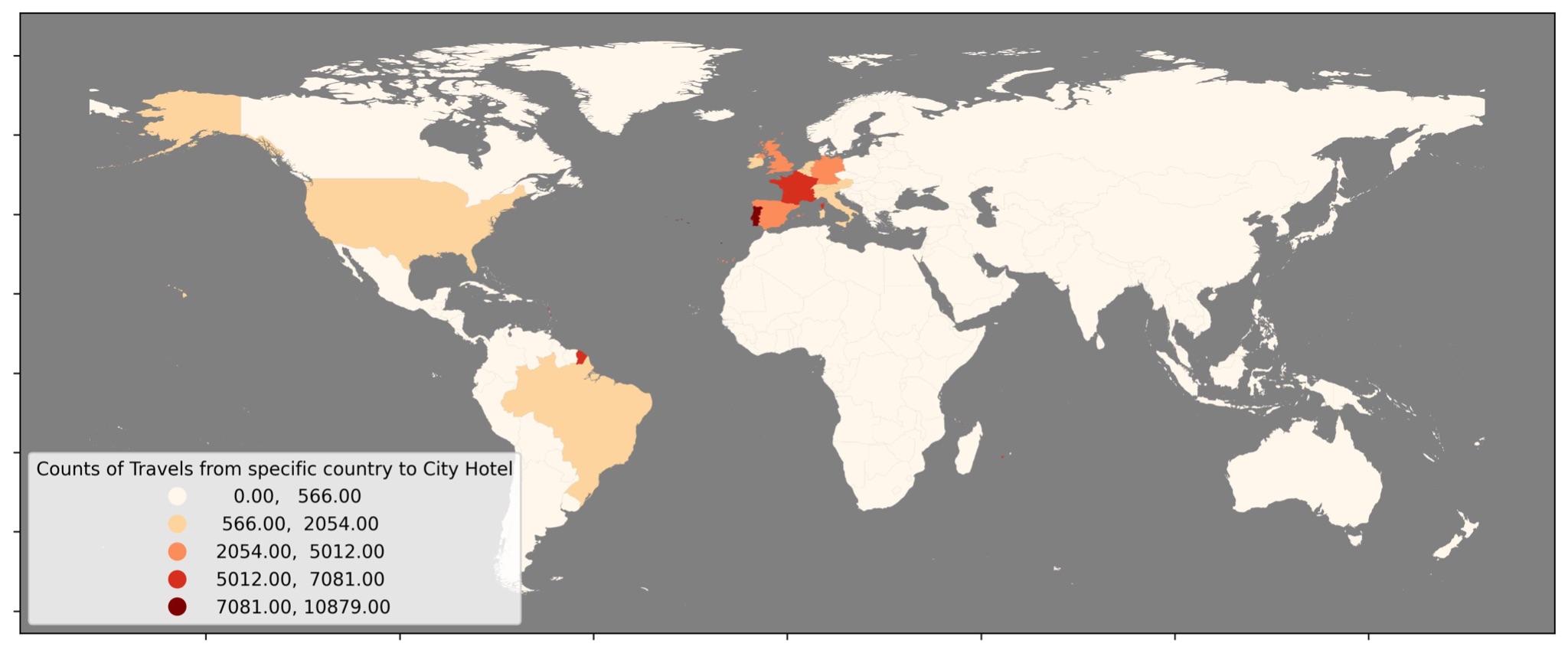
Understanding guest demographics and preferences is paramount for operational success in the dynamic hospitality industry landscape. The different nature of the two observed hotels in Portugal, which also suit different types of holidays or trips (relaxing holidays vs city breaks), led us to assume that the guest behaviour between the two hotels must also be different. With a meticulous focus on guest origins, family compositions, and national affiliations, we compared the hotel data. We focused on unveiling nuanced insights into the divergent guest profiles, shedding light on their preferences, behaviour and general dynamics shaping the hospitality experience. Furthermore, we explored the potential correlation between happiness ratings from diverse nations and their travel preferences, particularly on visits to Portugal. To achieve this, we merged two datasets: One containing hotel booking information and another encompassing happiness scores from various countries.

# **Methods:**

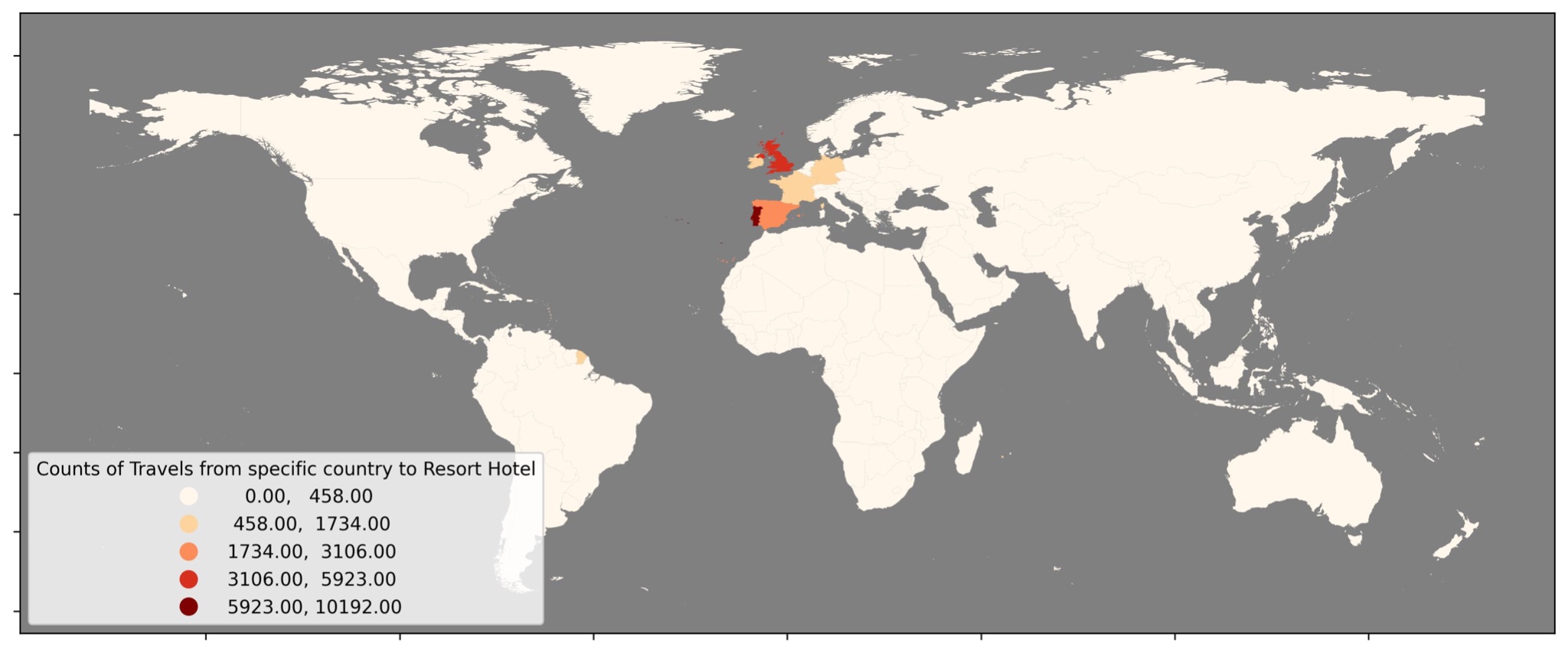
The project was based on the following Kaggle project (1); however, we conducted an expanded analysis and extended the evaluation to include another dataset (2). The data for the two hotels was originally obtained from external sources (3) and spanned approximately two years, from the middle of 2015 until the middle of 2017. It comprised various parameters, including the number of guests, booking period, day of arrival, and the countries from which the guests came. We have focused exclusively on bookings that were not cancelled and where at least one guest visited the hotel (some bookings did not contain any guests, as we considered this to be an error in the data and removed them in our data cleaning process). We used Python as the programming language and Jupyyter Notebook as a programming environment. The Pandas and Numpy libraries were used to analyse the data, and Matplotlib, Plotly, Seaborn, and Geopandas (4, 5) were used for the subsequent graphical presentation of our results. We chose GitHub for version control and Visual Studio Code for code editing and development. Additionally, GitHub has provided us with tools such as Kanban Boards and Wikis to help us better organise our project. We also recorded our data analyses and graphs in a dashboard on Miro so that we had a complete overview of our findings at all times and could compare partial results with each other simply.

# **Results:**

Overall, there were more bookings in the City Hotel than in the Resort Hotel over the same period. Still, the age distribution of guests (split between adults, children and babies) is relatively equal in both hotels. The average number of guests per booking also hardly differs (City Hotel: ∅ 1,95; Resort Hotel: ∅ 1,98 )

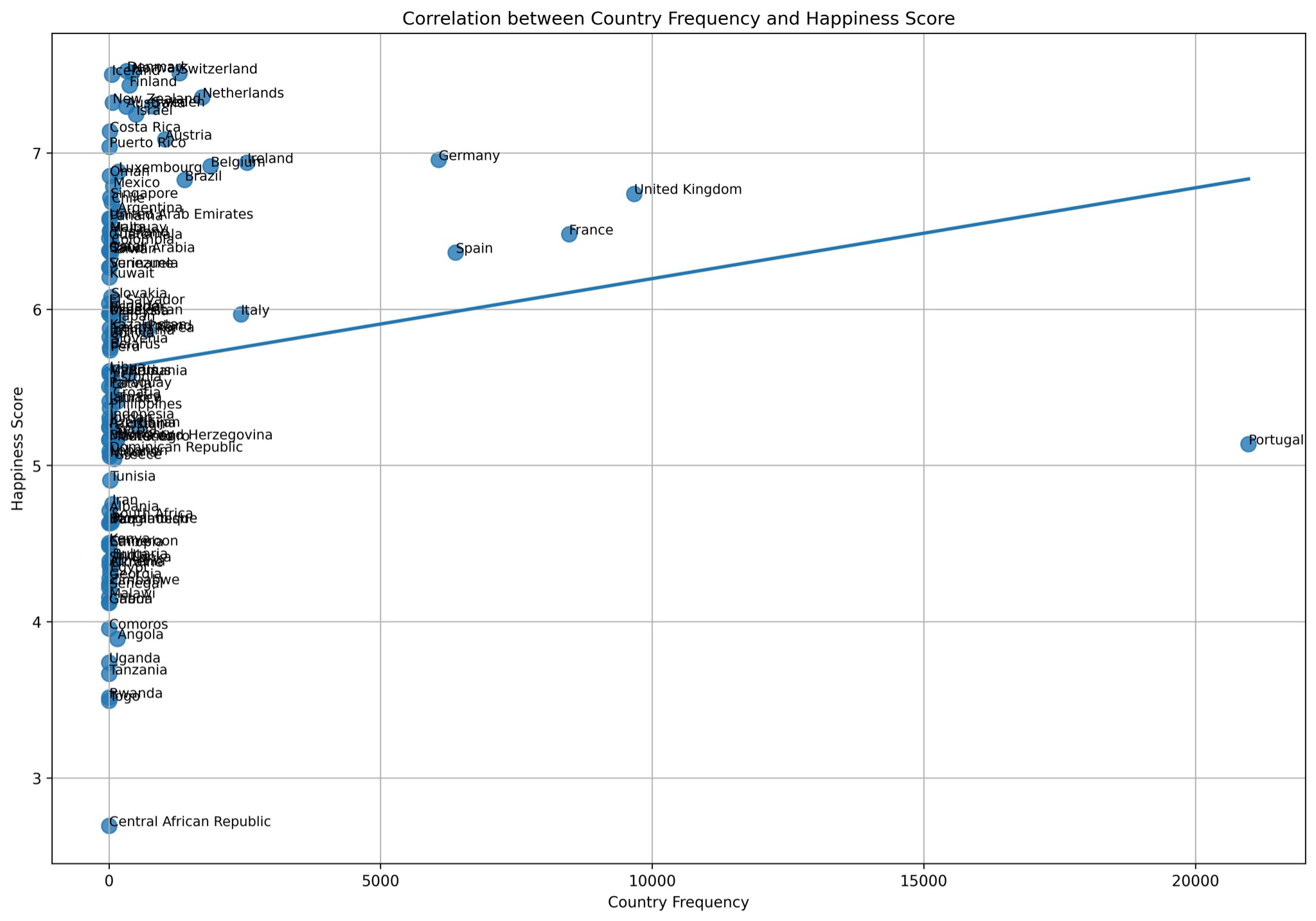


**Fig 1:** World map showing the origins of the City Hotel guests



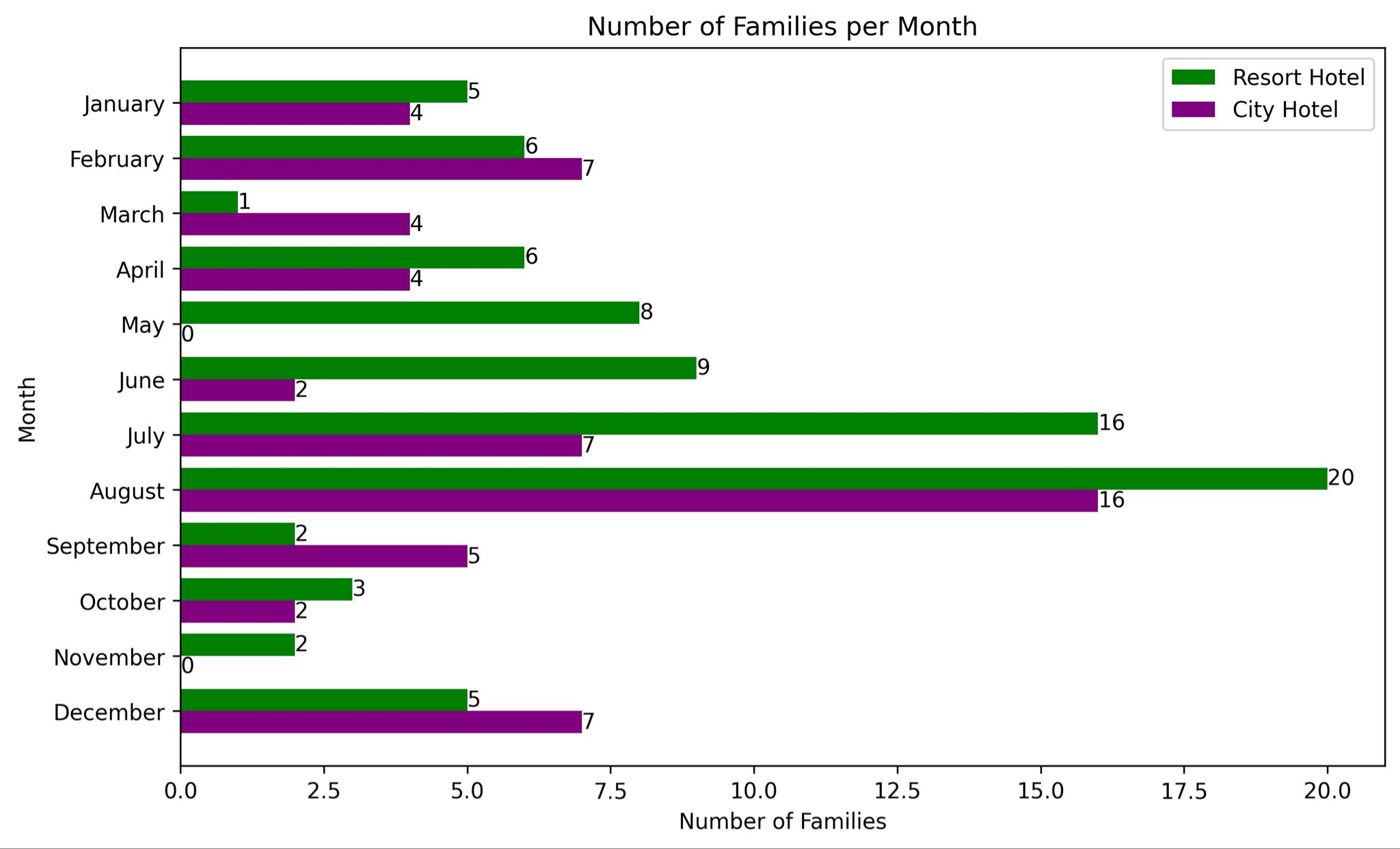
**Fig 2:** World map showing the origins of the Resort Hotel guests

The situation is different regarding guests' booking behaviour and origin. For these factors, we found apparent differences between the two hotel types. While the vast majority of guests in both hotels come from Europe, it is noticeable that there are guests from other countries in the City Hotel, especially from the USA and Brazil. On the other hand, the Resort Hotel has very few non-European guests.



**Fig. 3:** Diagram showing the correlation between country frequency and happiness score for both hotels

A Spearman correlation coefficient of 0.53 revealed a moderate to strong positive correlation between the frequency of countries and their happiness scores, with a statistically significant p-value < 0.001 (2.58e-09), indicating that countries with higher happiness scores tended to be more represented in bookings.



**Fig. 4:** Bar Diagram showing in which month families visit which hotel

There are also differences in the behaviour of families (bookings with at least one child or baby travelling with them) and other guests for both hotels. Most families stay around 1-4 days in the City hotel. In comparison, in the Resort Hotel, the most popular duration of stay is five or seven days (which does fit an average leisure holiday of a week). In both hotels, guests visit during the summer and spring when the weather is best. However, especially for the Resort Hotel, it is very noticeable that the families mainly arrive in July and August, during the summer holidays in Portugal. This also fits in with our analysis that most of the families in both hotels come from Portugal and the neighbouring European or Romance-speaking countries. There are very few non-European families.

# **Conclusion:**

Our investigation reveals distinct guest behaviour disparities between the two hotels, emphasising the importance of tailored approaches for a diverse clientele. Our findings empower hotels to optimise their services and precisely anticipate peak booking periods by discerning preferences among families and childless guests alongside international visitor trends.

# **References:**

1. Hotel project based on kaggle project from Nitesh Yadav, Title: Hotel Booking Prediction (99.5% acc): <https://www.kaggle.com/code/niteshyadav3103/hotel-booking-prediction-99-5-acc>, Licence: [Apache License 2.0](https://www.apache.org/licenses/LICENSE-2.0)
2. Happiness Dataset provided on kaggle by Sustainable Development Solutions Network (Owner) and Abigail Larion (Admin), Title: World Happiness Report - Happiness scored according to economic production, social support, etc.: <https://www.kaggle.com/datasets/unsdsn/world-happiness/data>, Licence: [CC0: Public Domain](https://creativecommons.org/licenses/by/4.0/)
3. Hotel dataset provided on kaggle by Jesse Mostipak, Title: Hotel booking demand:

<https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand>, Licence: [Attribution 4.0 International (CC BY 4.0)](https://creativecommons.org/licenses/by/4.0/)

Originating from the paper: Hotel Booking Demand Datasets: Nuno Antonio, Ana Almeida, Luis Nunes, Data in Brief, 2019. Downloaded from <https://www.sciencedirect.com/science/article/pii/S2352340918315191> and cleaned by Thomas Mock and Antoine Bichat as part of #TidyTuesday:<https://github.com/rfordatascience/tidytuesday/blob/master/data/2020/2020-02-11/readme.md>

1. Geopandas diagram based on Quisl, April 14, 2021, <https://quisl.de/b/wie-du-weltkarten-mit-python-erstellst-geopandas-in-vier-schritten/>
2. Map made with Natural Earth. Free vector and raster map data @ naturalearthdata.com.

Text for our Certificate:

Our project analysed guest data from two distinct hotels in Portugal: A city hotel in Lisbon and a resort hotel in the Algarve. We used various tools and technologies, including Python, Jupyter Notebook, Pandas, and Numpy, to conduct an in-depth analysis of guest data. Additionally, we employed a range of visualisation tools such as Matplotlib, Plotly, Seaborn, and Geopandas. During our analysis, we not only focused on the data from the two hotels but also explored correlations between happiness ratings from diverse nations and travel preferences, focusing on visits to Portugal by combining a second dataset. This approach revealed valuable insights into guest behaviour and choices, aiding informed decision-making in the hospitality industry. Our findings emphasise notable disparities in guest origin and booking behaviour, especially regarding family stays. By grasping these differences, hotels can adjust their services to meet the needs of diverse clientele, optimising operations during peak booking periods.