Introduction:

The data set we have chosen represents real world data regarding hotel bookings (Antonio, de Almeida and Nunes, 2019). There are 36 Columns with 119,390 observations. The data contains specific information for each booking such as: country of customers, payment method, Agency used, number of adults/children etc.

We note that of the 119,390 observations there are 81,234 unique names, and 115,425 unique emails.

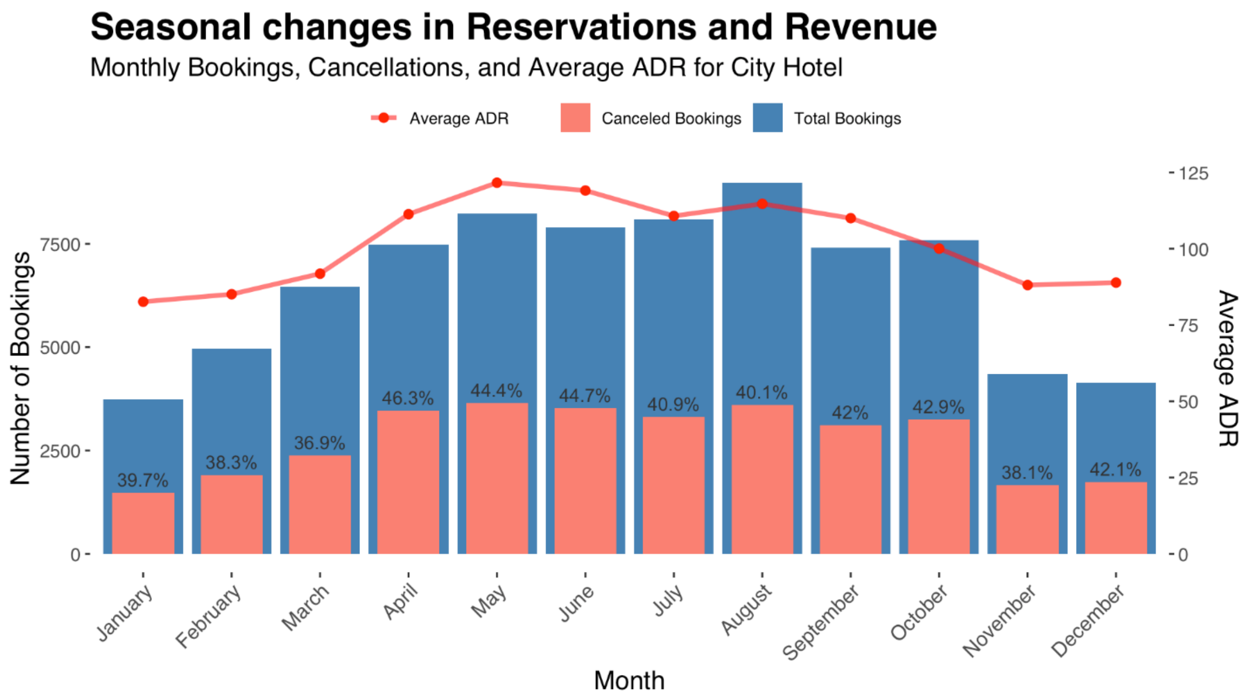
Business Problem

In this dataset, the City Hotel has been experiencing a large number of hotel booking cancellations, which has had an impact on the revenue stream of the business. The aim of this Exploratory Data Analysis is in the identification of trends amongst the people cancelling to better manage the Hotels expectations. We have decided to look at the City Hotel data purely in order to find the trends in cancellation.

Link to data: https://www.kaggle.com/datasets/mojtaba142/hotel-booking/code

Plot 1: Seasonal changes in Cancellations and average ADR

The graph shows monthly variations in booking numbers and average changes in the Average Daily Rate (ADR) for each month. This graph is based on a subset of the original dataset, focusing specifically on data from City Hotel. The booking counts for each month are aggregated from data spanning the years 2015, 2016, and 2017.



Graph Design:

This graph is purposed to provide a comprehensive view of monthly booking trends, cancellation patterns, and revenue performance (measured by Average Daily Rate, or ADR) for City Hotel. To achieve this, a **dual-axis bar and line chart** format was chosen for its effectiveness in displaying both absolute numbers (total and canceled bookings) and average rates (ADR) within a single, coherent visual. We utilized the 'Tufte' theme in the code across all plots to minimize visual clutter (Tufte, 2012), coupled with the sans serif font to ensure clear readability across the plots.

The **nested bar format** was selected to distinctly display total bookings and cancelled bookings for each month. Using color differentiation—steel blue for total bookings and salmon for cancelled bookings—enables an immediate, visual contrast, allowing viewers to quickly spot patterns and compare these two categories. This color choice also helps identify potential issues, such as months with proportionally high cancellations, which are essential to monitor and address for better revenue stability.

With the precise **cancellation percentage** displayed within each column, the graph highlights the ratio of canceled bookings to total bookings. These additional details make it easier to observe seasonal fluctuations in the cancellation rate, providing viewers with a clearer understanding of City Hotel's customers reservation behaviors.

An **Average ADR line** was added to represent revenue trends over the same period. The line format is effective here as it can visually show its relationship with the bars easily, creating a distinct focal point for ADR trends. By plotting ADR on a **secondary y-axis**, the graph maintains clarity and scale integrity for both the booking counts and the ADR data, ensuring that neither metric overshadows the other. **Month labels** on the **x-axis** and the tilt in axis text enhance readability by accommodating many categories in a compact space.

This design was chosen to allow viewers to quickly assess how bookings, cancellations, and revenue performance fluctuate month-by-month, aiding in data-driven decision-making for seasonal and operational strategies.

Findings and Interpretations:

Seasonal Trends in Bookings and Cancellations

One of the first visible insights from the graph is the seasonal pattern in total bookings. City Hotel experiences noticeable peaks in high-demand months, particularly during the summer (May to August). This is common in the hospitality industry, where leisure travel spikes during vacation seasons. The graph also shows that during these peak months, cancellation volumes tend to increase alongside total bookings. This pattern suggests that cancellations are partly influenced by high-demand periods when customers might make multiple bookings or hold reservations across different options, only to cancel as their plans solidify or other preferred options become available.

ADR Patterns and Their Relationship to Demand

The ADR line trend provides insights into pricing behavior of City Hotel across the months. ADR generally rises during high-demand periods like May and August, responding to increased demand by capitalizing on higher prices. This dynamic pricing strategy is effective in maximizing revenue during popular months. However, some high-booking months, such as August, show less of an increase in ADR relative to the booking volume. This observation suggests an opportunity to optimize ADR more aggressively in certain peak months to maximize revenue.

Cancellation Patterns in High ADR Months

Looking closer to the exact percentage of cancellations, the graph also reveals that months with elevated ADR correspond with significant cancellation rates (e.g., April, May and June). This correlation suggests that while high ADRs capture additional revenue, they might also pose a risk of deterring commitment from certain customer segments. Price-sensitive customers may initially book at a high ADR rate but later cancel if they find cheaper options.

Recommendation

To address the high cancellation rate and achieve greater booking stability, the following strategies are recommended:

Seasonal Discounts for Early Bookings:

- Provide discounted rates for guests who book and make an upfront financial commitment for high-demand months. This approach encourages guests to finalize their plans early, lowering the chance of speculative bookings that frequently result in cancellations.

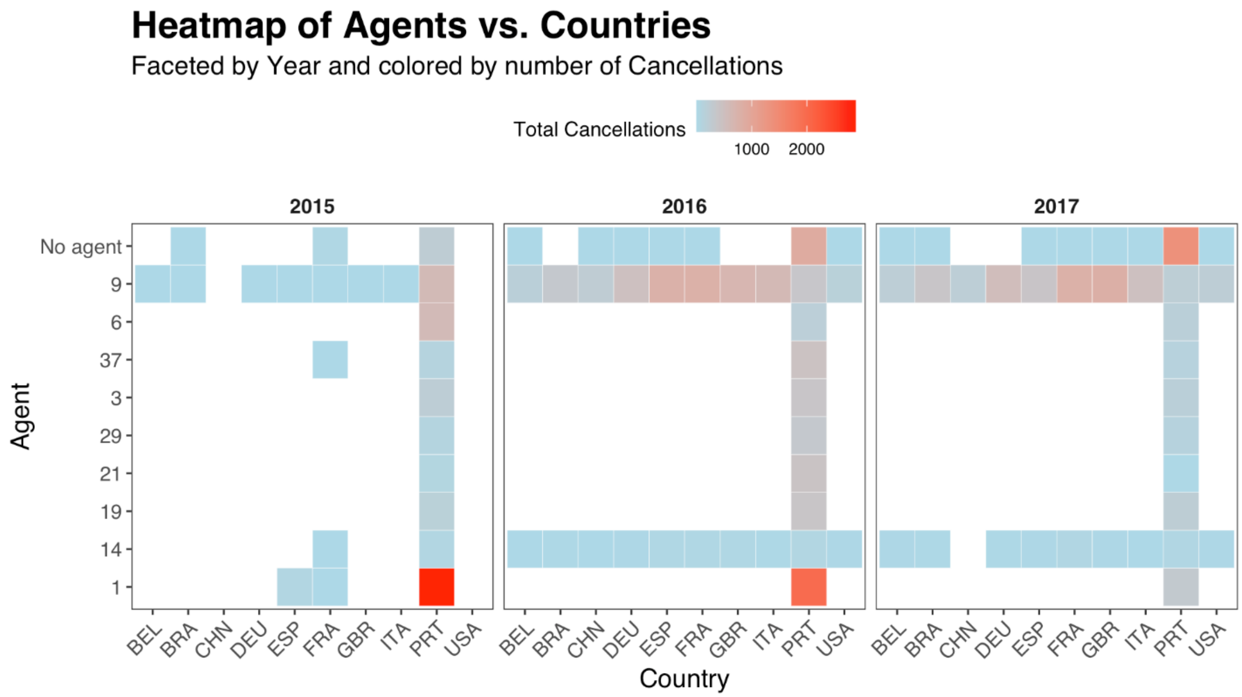
Non-Refundable Rates with Perks:

- Offer special perks, such as room upgrades, complimentary breakfast, or spa access, for guests who choose non-refundable bookings. These added benefits provide guests with more value, making them less likely to cancel.

Flexible Pricing for Longer Stays:

- In months with high cancellations, offering reduced rates for extended stays or complimentary nights can incentivize longer bookings. This can help increase occupancy rate and attract guests with high sensitivity to value over price ratio.
- By implementing these promotions, City Hotel can incentivize early bookings, increase booking commitment, and reduce overall cancellation rates.

Plot 2: Cancellations based on Agents and Countries



Graphic Design

The graph above shows a four-variate model, where variables year, agent used, country of origin, and number of cancellations are all considered.

This graph utilizes a heat map which is ideal for a tri-variate model or three cardinalities, the 2 dimensions of an x-y plot with the heat map displaying a colored spectrum across the surface. The major benefit of a heatmap is its ability to identify hotspots, or high correlations in a large dataset using a single visualization.

Heatmaps also allow for many variables whilst still being quite quick and easy to interpret, making it a powerful graph for early trend detection. This comes with the downside of lacking numeric inference and overlapping of values. The use of Faceting helps to split the heatmap and view change over a certain variable, which can relieve the overlapping of the data.

In terms of design choices, we utilized the colors “Light-blue” and “red” as they are different on the value scale allowing for anyone who is color-impaired to easily interpret the graph and drive the readers eyes to the areas of high cancellations. Moreover, the initial Tufte design didn’t separate the facets clearly, therefore we entered a border grid around the facet to ensure clear separation as the heatmap can be misinterpreted between years easily.

Findings and interpretations

The graph is Faceted by the 3 years in the data from 2015-2017 and has the x-axis illustrating the top 10 countries with the highest cancellations. The Y-axis illustrates the top 10 Agents with highest cancelation used by individuals and “No agent” represents the people who booked a room directly. Finally, the graph's coloring illustrates the number of people cancelling their booking where red indicates a high number of cancellations higher than 2000.

Firstly, facet helps segregate the data annually to help in identifying the source of recent cancellations. We chose to split this data into countries as we recognized that a substantial portion of the bookings came from Portugal, and thus wanted to better understand the number of people cancelling their booking from other regions. Finally, we plot against the top 10 agents with high cancellations to isolate these cases.

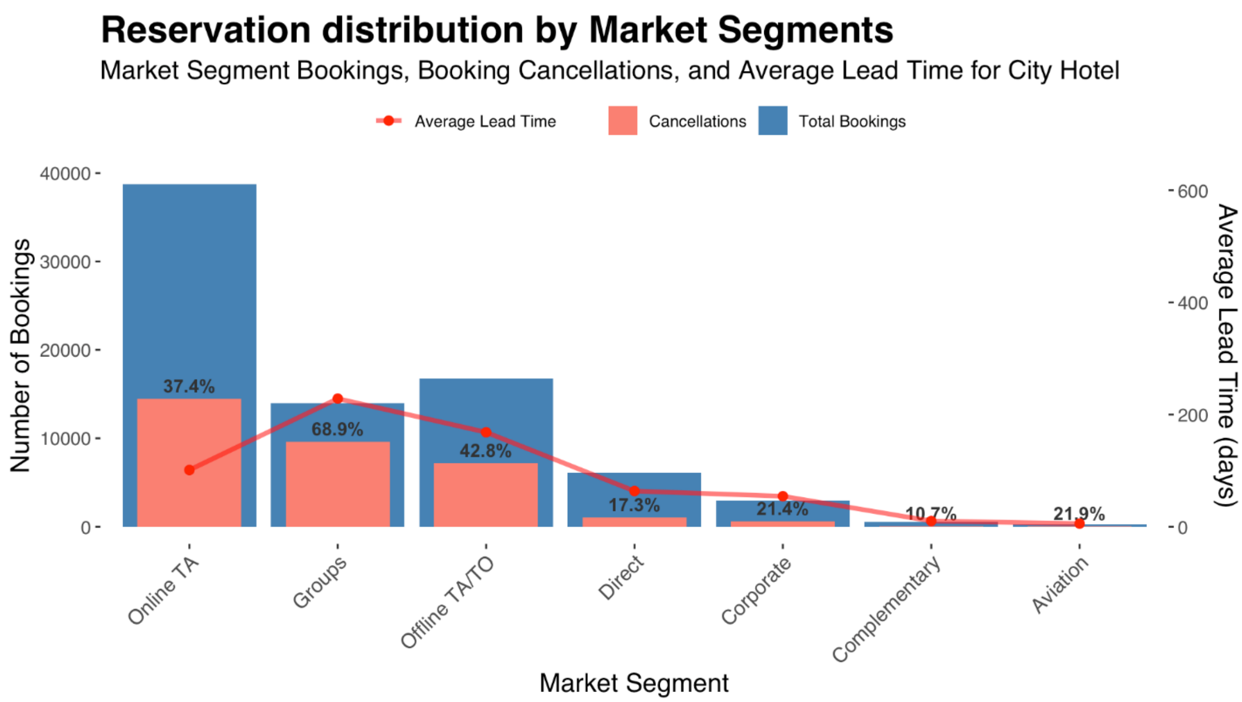
Looking at the data horizontally we identify 3 agent types that book for multiple countries of origin (No agent, 9, and 14). We see that agents 9 and 14 have booked for individuals across all 10 countries, whereas individuals who did not use an agent are mixed and the diversity of countries being booked directly are increasing each year. Among these we see that agent 14 has a strong track record with extremely low levels of cancellations across all countries. Agent 9 has a more balanced track record with higher cancellations especially amongst French people. Non-agents have a low cancellation number across all countries except Portugal which has seen a clear rise in cancellations over the last 3 years. The number of cancellations in 2017 was the highest in the subgroup of direct bookings from Portuguese guests. Agent 1 has seen a fall in the number of cancellations across the years whereas agent 9 has stagnated.

Recommendations

The most highlighted issue that this graph brings to light is the recent rise in Direct Portuguese cancellations. Some potential changes to the experience of booking directly with the hotel could be implemented:

- **Increased transparency and flexibility when booking directly:** To encourage those who book directly not to cancel, allowing guests to shift their dates and providing a more transparent user experience where price, room type, cancellation policies are all easily accessible to the website can help decrease cancellations.
- **Cancellations policies for direct users:** Tightening the cancellation policies by decreasing the flexibility once a deposit has been paid, or making deposits non-refundable after a certain period can act as a deterrent to anyone looking to book and cancel in the future. Although this may decrease the number of bookings, it will heavily decrease the cancellation rates and thus allow the business to plan more easily.
- **Further analysis:** The heatmaps strength is as a foundational graph to leverage in deeper analysis. An analysis of the market segments of the Portuguese cancellations could highlight strong relationships. Moreover, a deeper analysis of the success of agents 1 and 6 is required to understand the changes the agent has made over time which caused the cancellation rates to fall. This insight can help determine the best course of action the business should take in minimizing cancellation rates.

Plot 3: Cancellations based on Market Segmentation and Lead Time



Graphic Design

Plot 3 is a tri-variate visualization combining market segment bookings, lead time (days booked ahead), and cancellations. This plot uses a **dual-axis bar and line chart**—an ideal setup to represent both discrete and continuous variables in a clear and interpretable manner for a broad audience.

The **nested bar format** organizes information effectively. The wider blue bar represents total bookings per market segment, while the narrower red bar inside it shows cancellations. This nesting emphasizes cancellations as a subset of bookings, effectively portraying a binary variable (cancellations) without adding data-ink or sacrificing detail (Tufte, 2012). By positioning the cancellation bar within the booking bar, rather than side-by-side, we reinforce the subset relationship while reducing clutter. To improve differentiation, we chose high-contrast colors—steel blue for bookings and salmon for cancellations. Additionally, we display **cancellation rates** as bold, large-font percentages to emphasize these values as the plot's focal point.

Lead time was added because of its correlation with both market segments and cancellations, based on our exploratory data analysis. Average lead time per segment is displayed as a point on each booking bar and depicted with a secondary y-axis on the plot's right side. We chose a red line graph for lead time, both for simplicity and for clear contrast with the steel blue and salmon bars. The line is semi-transparent (alpha 0.6) to ensure it does not obscure the cancellation percentage values. While adding exact lead time values per segment was considered, we avoided it to prevent oversaturation.

While averaging values across the three years for the City Hotel provides a broad overview, faceting by year would enhance temporal detail and reveal trends over time. However, we opted against this to avoid overcrowding the visualization; it would be more appropriate for a deeper analysis.

This design aims to enable viewers to quickly assess fluctuations in bookings, cancellations, and lead time across market segments, supporting insights for strategic adjustments to reduce cancellations.

Findings and Interpretations

Our analysis shows that a significant portion of City Hotel's bookings originate from online travel agencies (TA), with this segment showing the highest absolute number of cancellations at 14,490, representing 37.4% of all bookings. While this volume warrants further investigation, group bookings are even more concerning. Group bookings experience 9,623 cancellations, equating to a higher cancellation rate of 68.9%. This elevated rate is likely due to the complexities of coordinating group plans, which increases the likelihood of changes. Consequently, group bookings are a critical area to address due to their higher likelihood of cancellation.

Offline travel agency and tour operator (TA/TO) bookings also show a notable cancellation rate, with 7,158 cancellations or 42.8% of bookings. Although not the highest, this rate suggests that offline TA/TO bookings would benefit from closer analysis to reduce cancellations.

Direct and corporate bookings, by contrast, show lower cancellation rates of 17.3% and 21.4%, respectively. These segments appear more stable, presenting opportunities for focused marketing to reduce overall cancellations.

Finally, our analysis highlights that canceled bookings tend to have significantly longer lead times than non-cancelled ones. Group bookings have the longest average lead time at 228.5 days, indicating that extended planning time correlates with higher cancellation risk. These findings underline lead time as a factor to address in future policies.

Recommendations

Our primary recommendation is to revise booking policies, especially concerning lead time, as it significantly impacts cancellation rates, particularly for group bookings. Suggested policy adjustments include:

- **Limiting booking lead time**: Setting a cap on how far in advance customers can book, such as a 1-year limit (or tailored limits by segment), could help reduce cancellations, especially among groups. Balancing this restriction is essential to prevent discouraging group bookings and to manage potential opportunity costs.
- **Adjusting the deposit structure**: Requiring a higher, non-refundable deposit for bookings made more than a year in advance may discourage cancellations without reducing total bookings. In cases of cancellations, City Hotel retains the deposit, potentially increasing revenue.

A secondary recommendation is to redirect some marketing toward direct and corporate bookings, which have lower cancellation rates. Potential strategies include:

- **Offering corporate booking contracts**: City Hotel could offer discounted, bulk booking contracts for corporate clients, promoting higher booking consistency.
- **Incentivizing direct bookings**: City Hotel could introduce exclusive discounts or a membership program for customers booking directly with the hotel, encouraging loyalty and reducing cancellations by creating a more stable customer base.

These recommendations are aimed at lowering cancellation rates and creating steadier demand across reliable customer segments.

Bibliography

Antonio, N., de Almeida, A. and Nunes, L. (2019). Hotel booking demand datasets. *Data in Brief*, 22, pp.41–49. doi:https://doi.org/10.1016/j.dib.2018.11.126.

Holtz, Y. (2018). *Data visualization with R and ggplot2 | the R Graph Gallery*. [online] r-graph-gallery.com. Available at: https://r-graph-gallery.com/ggplot2-package.html.

Tufte, E.R. (2012). *Visual explanations : images and quantities, evidence and narrative*. Cheshire, Connecticut: Graphics Press.

Wickham, H. (2019). *Create Elegant Data Visualisations Using the Grammar of Graphics*. [online] Tidyverse.org. Available at: https://ggplot2.tidyverse.org/.