**Project step:**

**Data analysis(bussiness problem)**

**1.Create a problem statements**

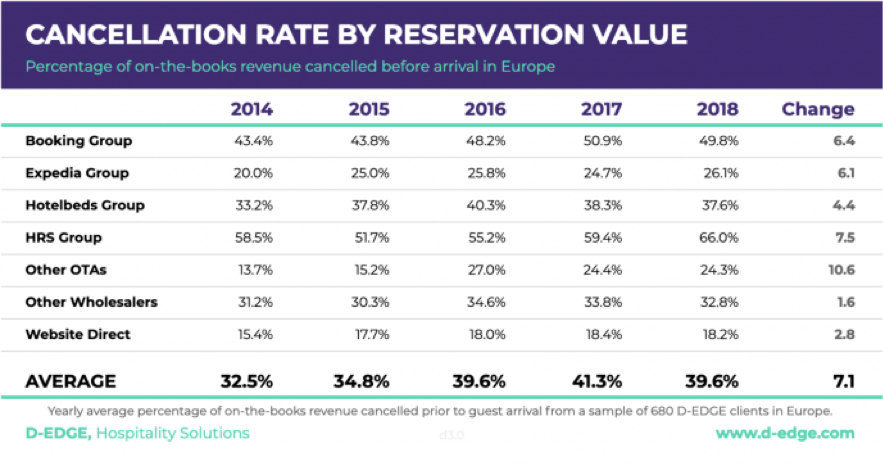
**2.Identify the data you want to analyse**

**3.explore and clean the data set**

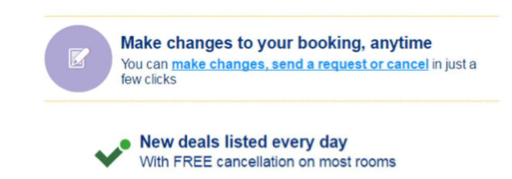
**4.analyze the data**

**5.present the data set in term to report &visualization in dashboard Business Problem**

Booking cancellations are undoubtedly one the biggest headaches of any revenue manager or hotel manager nowadays. If we look in to the data available of the European market we see that in 2018 49.8% of bookings reserved on the OTA Booking.com were cancelled:



 In the graph above we can clearly see this is a growing trend and generates a bigger problem when it comes to understand how many rooms a hotel should sell and which overbooking techniques it should apply.



OTA´s are encouraging customers to cancel when they actively encourage you to book now and cancel later, free of charge, whenever you want. This results in customers booking more than one hotel and decide later which one they will choose. Examples of the impact of cancellations on a hotel:

1.    Lost of revenue when they cannot resell the room

2.    Additional costs of distribution channels by increasing commissions or paying for publicity to help sell these rooms

3.    Lowering prices last minute, so they can resell a room, resulting in reducing profit margin

So, what can hotels do to reduce this uncertainty and maximize their product and revenue? A lot can be done with revenue management techniques when it comes to rates restrictions, like increasing the number of days until the arrival date that the customer can cancel without cost, giving you more time to resell the room. But nowadays you have to apply similar restrictions to those applied by your competitive set and hotels around you, so if you are going to be stricter, costumers will prefer other hotels that are more permissive.

Therefore it would seem that we have a complex problem and not a viable solution. However, thanks to data science and machine learning there are many things we can do to accurately predict which individual and specific reservations are going to cancel. In the next chapters I am going to take on a public dataset of hotel bookings and apply an EDA (Exploratory Data Analysis) to understand the data and use descriptive analysis techniques to get a full picture of its behavior. Then we will see how can we use machine learning and select the best models for our data. And last but not least, we will analyse the model results and apply it to unseen data (data not used to train the model) so we can verify how good our model is.

**All of the work done below was made in Python using Jupyter Notebook and only open sourced libraries.**

**Dataset**

Our dataset is available at Kaggle in the link above. Let's look at the description:

**Content**

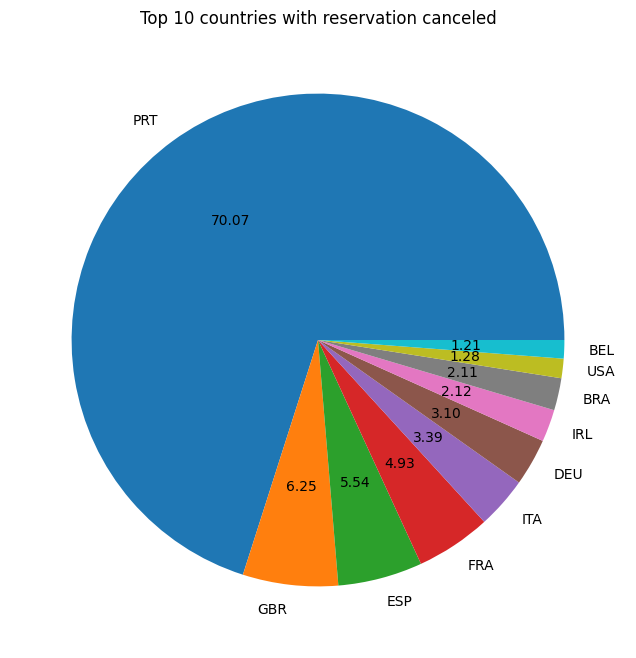
This data set contains data for a city hotel and a resort hotel and includes information such as when the booking was made, length of stay, the number of adults, children, and/or babies, and the number of available parking spaces, among other things.

All personally identifying information has been removed from the data.

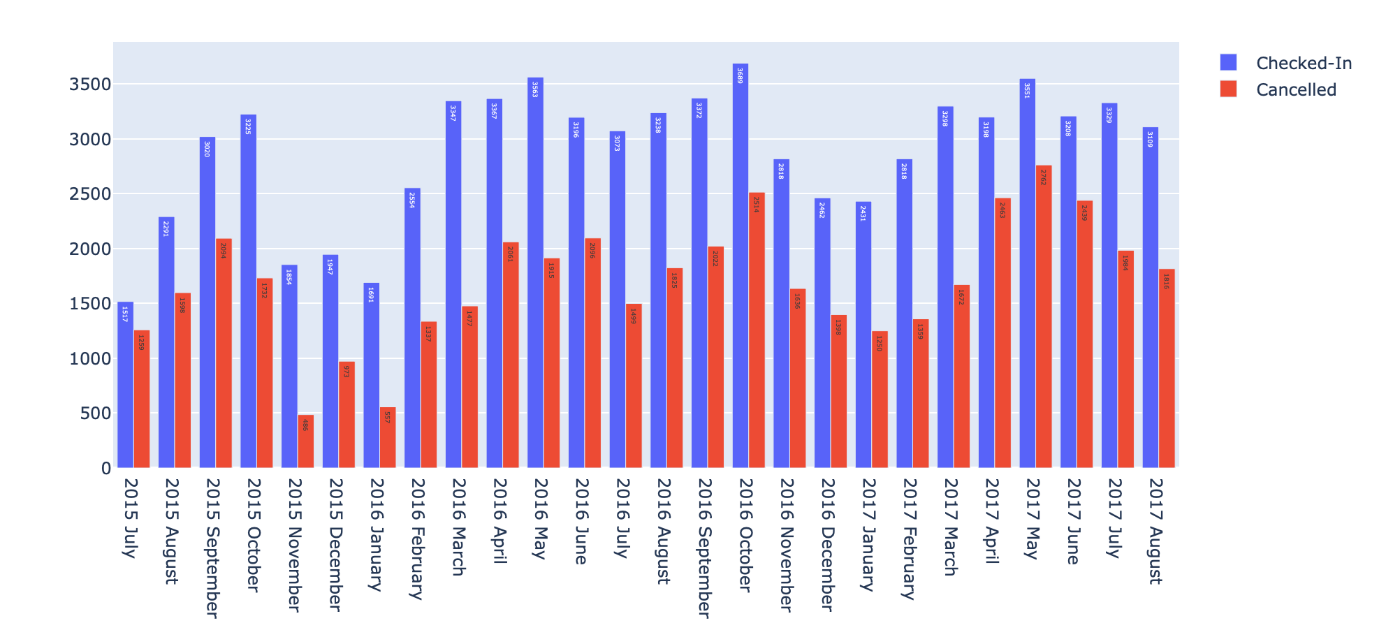
**Acknowledgements**

The data is originally from the article [Hotel Booking Demand Datasets](https://www.sciencedirect.com/science/article/pii/S2352340918315191), written by Nuno Antonio, Ana Almeida, and Luis Nunes for Data in Brief, Volume 22, February 2019.

**EDA – Exploratory Data Analysis**

The first thing we want to see is the distribution of reservations cancelled and not cancelled. In the image below we can see that we have data from a Resort Hotel and City hotel. In x axis we see the 0 and 1 representing Checked-In and Cancelled respectively. The first thing we conclude is that there are more cancellations in the City hotel. 

Then we plot this data through time to see if we find some trends:



We can see that there are some seasonality during the years, which is normal in hospitality, specially in Resort Hotel, which get more bookings in the summer.

Also cancellations ratio are bigger in Spring and Summer and have a lowest ratio in the winter. So we can clearly star to see some patterns. Let's look to the market segments and their distribution for Checked In and Cancelled reservations:

Most of the reservations Checked In are from online markets and the same thing happens for the cancelled ones. The interesting thing is to see that Groups are the second market segment that has the higher number of cancellations, but only the 4th in the ranking for the Checked In ones. Normally in Hospitality Groups segment have 3 (sometimes 4 if you count Inquiry) types os status, Tentative, Pending and Definitive. So it would be interesting the see if the tentative reservations count for the cancellations. Since groups normally have washing (when company reduce the initial number of rooms requested) it is expected that the segment gets the 2nd place when it comes to booking cancellations.

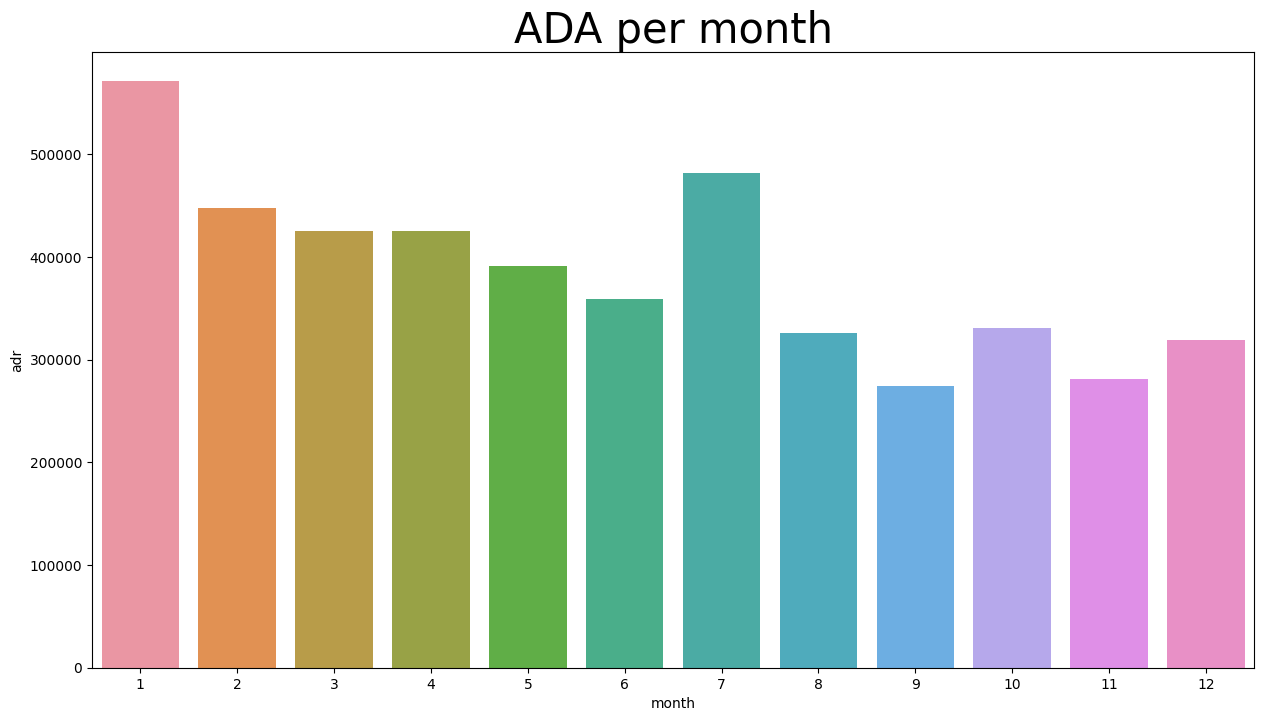
**Lead Time**

Lead time is one the most important metrics when it comes to analyse hotel revenue performance, so I felt I had to plot different visualizations to get a good picture of it. A lot of other visualization would also be good to analyse.

The last two we can see that the longer the lead time, the higher the probability of the booking would be cancelled. In the last plot we can see a correlation between lead time and the cancellations ratio.

**ADR**

ADR is the only variable related to revenue in this dataset. When it comes to it, I always find box plots are the perfect way to understand ADR behavior.



The mean is represented in the dashed horizontal line of each segment, and the median by the non dashed horizontal line. ie: The mean (average) and median of Direct are 115 and 105 respectively.

* Corporate, Aviation and Offline TA/TO have lowest variation of prices which is usually because these segments have contracts with flat prices and are not yieldable like other segments
* Direct and Online TA are the market segments with highest prices and higher variation, showing well how yield revenue management is a good thing

In the top Nationalities we only find European Countries and all from South or Central Europe. Portugal leads the race of reservations in both Transient and Transient-Party customer types.

**Introducing the Power Predict Score (PPS)**

Predict Power Score is a great way to see the correlations between our variables, not only the numeric but also the categorical ones. Some are pretty obvious like the reservation\_status and is\_canceled feature. To find more about it check this article by Florian Wetschoreck called [RIP correlation. Introducing the Predictive Power Score](https://towardsdatascience.com/rip-correlation-introducing-the-predictive-power-score-3d90808b9598).

Here are some findings that helps understand better our dataset:

* Market Segment and Distribution Channel are mostly affected by the agent feature. This helps understand that these hotels have different agents when it comes to specific segments of clients. For example Booking.com is an agent that is only represented in the Online market segment
* Customer type is more influenced by the company. Which is similar to the point above
* is\_repeated is also influenced by the company from which the customer comes. This may be impacted by Corporate and Aviation segments

**Modeling**

Now that we have our descriptive analysis it is time to create a model to predict cancellations. The first thing is to understand the problem that we have. In the data science field this is known as supervised classification problem, since we want to predict if a booking is cancelled or not, in other words it is a binary prediction.

**Data Cleaning**

The first thing to do is to check if we have missing values, then we will have to replace those missing values with the mean of that feature if that variable is numeric or the constant if it is a categorical feature.

In this data set we have the following missing values:

children                               4

country                              488

agent                              16340

company                           112593

Then we will have to remove the column reservation\_status due to is possible source of leakage and overfitting.

**Selecting the best models**

Before we start testing the best algorithms for our data, we will split our dataset in 80%-20%. This will helps use the 80% of our data to train a model, and leave 20% of unseen data to test our trained model. This has to be done, because if we use all of our data the model will memorize it and this will result in overfitting predictions. Then we select our target in this is case is the feature is\_canceled.

In order to understand which models were the best fit for our dataset, I had to execute a number of tasks to clean and adjust the models, however for the purpose of this article I will only show you the results for the model who had the better predict results.

**In this article we saw that we can predict reservations that will cancel with a confidence of:**

* **Accuracy - 90%**
* **Precision - 88%**
* **Recall - 84%**

**With this information hotels can, for example, contact clients that the model predicted will cancel in order to get a cancellation earlier - so they can have more time to resell the room. Or perhaps approach the client in a way to make them feel special and keep their reservation and therefore cancel the others he or she had made in other hotels in the same city.**

We are currently seeing changes in the customer behavior due to the COVID-19 pandemic, however is too early to make predictions, we already saw that the hotels are lowering the window of days that a customer can cancel, as a way of encouraging them to book more. So, if it was important to manage and predict cancelations before the pandemic, it is even more important in the future days that the industry has in their path.

We all know the importance of data and revenue management in the hotel industry, but we also saw the importance that data science has to maximize the product, distribution and profit of a hotel. So now is the perfect time for the industry to start investing in this area and particularly in machine learning, since the advantages are huge in all fields of hospitality. Not only in predicting cancellations, but using regression analysis to forecast revenue and costs, cluster their clients on similar behaviors for marketing campaign or many other techniques.

I hope you enjoyed this article and project and find it helpful in any way. Please feel free to comment, make questions or share it with your friends and colleagues.