Introduction:

Hotels have it difficult to manage bookings correctly, overbooking has been a common practice in hotel management with the objective of “improve the expected profit by selling the same room several times” (Birkenheuer, 2009 as cited by Zhechev, Todorov 2010). Although overbooking has a lot of advantages for revenue in the short term it can affect revenue in long term with impact on the reputation and customer loyalty of the hotel (Selmi, 2007). The correct application of overbooking is crucial to the hotel industry and if not applied correctly it can be harmful for the company (Zhechev, Todorov 2010). Cancellation policies have a great impact on revenue and although strict cancellation policies secure revenue of no shows and cancellations (M. Velten, 2017) they can affect bookings. Post covid most free cancellation policies have converted on average 4.2 times better (according to rentalscaleup.com) “Free cancellation bookings have surpassed all of the other policies combined” When free cancellation policies are just 30% of their total offerings. Having this in mind we can see the importance of managing correctly bookings. The aim of this project is to try to understand main trends in the hotel industry and build predictive models to help hotel companies forecast better and manage bookings correctly. We will try to predict average daily rate in order to adjust offers to segments and boost revenues and focus mainly on predicting if a booking will be cancelled. Lastly, we will use PCA to see if we can get better predictions with the available features. Dimensionality is not large given the size of the dataset, but I believe that to achieve better predictive results more features should be added to the dataset. In that case PCA could be beneficial preparing data for further exploration.

Dataset:

We will be working with the hotel booking demand data set provided by Jesse Mostipak in Kaggle (<https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand>). This dataset was previously used to write an article about hotel prediction in 2019 by Nuno Antonio, Ana de Almeida and Luis Nunes (<https://www.sciencedirect.com/science/article/pii/S2352340918315191>). This article will be used to compare results as a benchmark.

For word restrictions the data set features won’t be completely presented and explained in detail in this report but useful information can be found in the Jupyter notebook. We have 32 features with 11 categorical variables, 20 continuous variables and 1 date and time variable with 119390 rows.

The structure of the report is the following:  
We will first load the data and conduct some basic data preparation and data cleaning. Second, we will make a data exploration analysis leading to a second data preparation step in which we will do feature selection. After that we will do a regression analysis to predict average daily rate (referred from now on as adr) with different models, classification to predict booking cancellation and lastly PCA to do dimensionality reduction and rerunning previously used models to compare the performance.

Data Loading and Cleaning:

We first load the dataset (it was downloaded from Kaggle and loaded locally to the Jupyter notebook) The dataset is also saved in the GitHub repository for precaution as to being delated from Kaggle for any reason. We first check what the data types are and if we have any missing values. We find a lot of categorical variables. At first glance we have 12 object, 4 float64, 16 int64. A further look into the data we can see that most of the variables are discrete and even categorical. We can detect as continual adr, lead time, stays in weekend nights, stays in weeknights and days in waiting list as continuous with almost all of the rest as either categorical or discrete with little possible values that might be considered categorical.

In the missing values area, we find that most of the missing values come from two columns: company (94% of values missing) and agent (13% of values missing) so we will continue to drop both columns. Once these features are dropped, we continue dropping rows with missing values in the rest of the dataset (0.4% of the rows).

Then we search for outliers that need to be eliminated. Our data is very skewed in general, but I believe we need to remove only a few outliers that are wrongly imputed since relevant information is available in the outliers. The only outliers removed are in the adr column where we have negative values (considered impossible) and a single value that is 50 times larger than the mean. Zero values for adr are considered since they are most complementary and could be gifts or promotions. We also find great variability in the lead time variable which is the variable that counts the number of days between the day the booking is made and the first day of the booking. In this case there is no need to remove any outlier since the data seems to be consistent. With these minor tweaks we still have 118896 of our original 119390 rows and 30 of our original 32 features. We now pass to the exploratory data analysis in order to get better understanding of the data set and see if we need to remove or clean any further.

Exploratory Data Analysis and Visualization:

We first look at the cancelation rate which is high. We have that 37% of the reservations are cancelled. We also look at the number of cancelations by deposit type. We find here very counterintuitive results. We find that most of the non-refund type of deposits (99%) are cancelled while we might expect this number to be smaller than the no deposit type (71%). I don’t know what might be causing this, it might be an error in the data or a problem with the hotel policy. Further investigation should be made to confirm validity of this data.

We continue to plot some of the variables to see patterns in the data.

References:

Birkenheuer, G.; Brinkmann, A.; Karl H. (2009), The Gain of Overbooking. Job Scheduling Strategies for Parallel Processing, Volume 5798/2009, pp. 80-100

Selmi, N. (2007), Yield Management, a Technological Innovation in Services: Impacts on Hotels and Customers. Tourism, Mobility and Technology TTRA , Europe Conference 22 to 25 April 2007.

Todorov, Andrey and Zhechev, Vladimir (2010), The impact of overbooking on hotels’ operation management; <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1698103>

Velten (2017), Cancellation Policies in Combination With Scarcity and Social Proof Appeals: A study into the effects of cancellation policies and persuasion cues on consumer responses within the online booking industry; https://essay.utwente.nl/72502/1/Velten\_MA\_BMS.pdf

https://www.rentalscaleup.com/fully-refundable-cancellation-policies-see-booking-demand-at-almost-the-same-levels-as-pre-pandemic/