# Proposal: Hotel Review Sentiment Analysis and Rating Prediction

DSC 511 Group Project

Maria Tsilidou
Anastasios Nikodimou
Ioannis Demetriou

## Abstract

In the era of digital transformation, online hotel reviews have become a significant factor in shaping travelers' booking decisions. Reviews provide firsthand experiences from guests, offering insights into hotel quality, customer satisfaction, and areas of improvement. Using data collection, aggregation and processing techniques we seek to formulate and answer questions related to hotel ratings, customer sentiment, and prediction modeling.

## Dataset

The analysis will use the "515k Hotel Reviews Data in Europe" dataset obtained from here at the time of writing. It contains over 500k reviews, scraped from Booking.com, scoring 1493 luxury hotels across Europe. Each review includes 17 features, including textual feedback and scoring, reviewer nationality, hotel location, providing diverse data types for analysis.

We acknowledge that while the dataset contains actual reviews, there is no integrity protection mechanism to check against fake reviews, biased opinions, or spam.

## Objectives

Hotel reviews reflect genuine opinions and individual experiences from real people. Using the selected dataset, we seek to formulate and answer questions utilizing domain knowledge from the hospitality industry, such as bad reviews influence the average score of each hotel, seasonality affects the occupancy, causes of dissatisfaction, while also attempting to predict a hotel’s effective rating based of our review data.

In particular,

- Identify key factors leading to customer dissatisfaction.
- Examine how negative reviews impact the average score of a hotel.
- Understand traveler sentiment toward various aspects and help hotels improve service quality by identifying common complaints
- Identify reasons for customer satisfaction and dissatisfaction.
- Investigate the effects of seasonality in the review rating.
- Develop a predictive model to estimate a hotel's average rating based on review data.

# Methodology

To achieve our objectives, we will apply the following techniques on our review dataset:

## Machine Learning for Rating Prediction

We will apply regression models to predict the average hotel rating based on textual sentiment, review metadata, and other influential factors. The most suitable model will be selected based on accuracy and performance metrics.

## Sentiment Analysis

Natural Language Processing (NLP) techniques will be used to analyze textual reviews and extract sentiment-related insights. We expect that sentiment analysis will allow us to classify reviews as positive or negative indicating high or low customer satisfaction.

# Big Data Perspective: The 5 V's

This dataset aligns with four of the five key Big Data characteristics:

**Volume.** The dataset consists of 515,000 reviews for 1,493 luxury hotels across Europe. The large volume of data allows for meaningful analysis of customer sentiment and hotel quality.

**Variety.** The dataset includes multiple types of data, such as textual reviews, numerical ratings, and metadata (hotel names, locations, review timestamps). This diversity enables comprehensive analysis from different perspectives.

**Veracity.** While the reviews originate from a reliable source (Booking.com), there is no built-in fraud detection mechanism to filter out fake, biased, or spam reviews. This may impact the reliability of insights, and additional measures may be required to verify data authenticity.

**Value.** The dataset holds high business and research value. Hotels can leverage the insights to improve service quality and guest satisfaction, while travelers can use the findings for better decision-making when selecting accommodations.

**Velocity.** This dataset does not meet the velocity criterion, as hotel reviews are continuously generated in real time across multiple platforms. However, the dataset itself is static and does not update dynamically.

## Conclusion

This study aims to leverage sentiment analysis and machine learning to gain deeper insights into traveler experiences. By identifying factors that drive satisfaction and dissatisfaction, and predicting hotel ratings, we seek to provide actionable insights for hotels to improve service quality and customers to make informed booking decisions.