Data Science Case Study

This notebook contains a series of exercises designed to explore a range of data science, python scripting, and quantitative reasoning skills. You can, in principle. solve these exercises using a number of different programming languages/environments, but it will likely be easiest for you to simply fill out this notebook with your solutions in the relevant sections following each exercise.

The data used in this case study come from the following academic work:
https://www.sciencedirect.com/science/article/pii/S2352340918315191

Feel free to peruse that paper in order to familiarize yourself with the data.

Part 1: A First Look at the Data

Exercise 1

The data are in two separate csv files in this directory. Bring them into this notebook as either a table or dataframe. Do the two datasets have the same column names (features)? If they do, combine the data into a single table or dataframe to work with for the rest of the notebook. How large is the dataset?

Exercise 2

What are the names of the features of the dataset and what are their data types (i.e. are they numerical data? categorical? strings? If numerical, are they floats, ints, ...?)

Exercise 3

At this point, what other steps would you take to get a better understanding of the data? Does it need to be cleaned or manipulated in any way? Explain and execute.

Part 2: Diving Into Some Details

Now that we have cleaned the data, put it in a friendly format, and have a basic understanding of it, lets dive deeper into some details.

Exercise 1

What is the distribution of the average daily rate (ADR)? Do you notice anything peculiar about this distribution? If so, and if you think something should be done about it, state it and do that. Also, approximate the mean and standard deviation of this distribution by eye, and then compare with the actual values.

Exercise 2

What is the distribution of values of the total length of stay? Do you notice anything noteworthy about it?

Exercise 3

What is the distribution of countries which people booking stays are coming from? Report this answer as percentages. What does this tell you about the probable location of the hotels?

Exercise 4

How would you investigate the relationships between the various features of the data? Explain and demonstrate.

Exercise 5

Suppose we are interested in the ADR feature. Further investigate the relationship between the ADR and some of the other features that you think might be relevant. What do you notice? Feel free to construct new features or modify existing ones if you think it would enhance the analysis.

Part 3: Machine Learning and Data Modeling

In this section, we will touch on some advanced data science tools to solve some realistic problems.

Exercise 1

Deploy a linear machine learning model that predicts the ADR as a function of whatever single feature you think is most strongly correlated. Are you able to get a good prediction stats on your validation set? Why or why not?

Exercise 2

Deploy a machine learning model that predicts the ADR as a function of whatever set of features you think are important. Explain your choices for the features and the type of ML model. If you use any, how can you handle the categorical features?

How does your model perform?

Exercise 3

Are you happy with the performance and efficiency of your model? What could you do to improve it? Explain and execute.

Part 4: Your Turn

In this final section, you are encouraged to further demonstrate your skills and knowledge using the provided dataset. Feel free to take things in whatever direction you find interesting.