Washington University in St. Louis

OLIN BUSINESS SCHOOL

DAT 561 – Introduction to Python and Data Science

Final Project:
Application of Data Science for Developing a Hotel Selection System

Motivation:

In today's data-driven world, several e-commerce businesses have emerged that use data to help customers find their desired items which match to their personal preferences. These businesses provide customers with an online selection system in which they can filter items based upon particular information and features. In fact, selection systems collect classified information about a group of products or services and filter them based on the specific needs of individual users. For example, online travel agencies, such as Expedia and Priceline, help travelers with finding available hotels/properties that meet their expectations regarding price, star rating, number of rooms, amenities, location, etc.

Introduction to the Project:

In this project, you are going to help with building a "Hotel Selection System", using two datasets, i.e., Property_details and Order_details. All the columns in these two datasets are explained in detail as below:

Property_details:
propertyid: The ID of the property.
propertyname: The name of the property.
address: The address of the property.
city: The city where the property is located.
country: The country where the property is located.
zipcode: The zipcode of the property.
propertytype: The type of the property, i.e., Hotels, Inns, Motels, etc. (8 types in total).
starrating: The star rating of the property, from 1 to 5 (e.g., 4 indicates a 4-star property).
latitude: The latitude of the property.
longitude: The longitude of the property.
url: The website link to book the property (not necessarily the official website).

Order_details:
id: The id of the order.
PropertyCode: The ID of the property.
dtcollected: The date of collecting the information of the property.
reservationdate: The date that the order was received.
los: The length of stay (i.e., the number of nights that a guest stays in the property).
guests: Number of guests who stay in the property.
roomtype: The room type that is booked.
onsiteprice: The price a customer would pay for the order.
ratedescription: The booked room description, which may include room size, number of bedrooms and beds, etc.
ratetype: Booking information, including cancellation policy, possible pay in hotel option, and extra low-price reminding.
sourceurl: The website link to book the property (not necessarily the official website).
roomamenities: The amenities available of the booked room.
maxoccupancy: The maximum number of guests allowed to stay in the booked room.
ispromo: Whether promotion is used when the order is placed (Y means in promotion, N means not in promotion).
closed: Whether the hotel is closed when the order is placed (Y means close, N means not close).
discount: The monetary value of the discount that an order received (already been deducted from the "*onsiteprice*").
promoname: The name of the promotion that the order participated in.
proxyused: The proxy used to retrieve the order data.
mealinclusiontype: The type of meal included in the order.
hotelblock: Whether the hotel was fully booked (Sold Out means no rooms are available for an order, Blank means rooms are available for an order).
input_dtcollected: The date inputting the collected data of the hotel.

You are required to use the Pandas and Numpy packages in order to read the above csv files and analyze their data. You can find all the files related to the final project in the "Final Project" assignment on Canvas. If you have a RAM size problem and you have difficulty with using Jupyter, you can use google Colab (https://colab.research.google.com/notebooks/) instead.

Tasks: This project consists of the following two parts:

Part 1: This part contains 6 data science questions, each of which has two parts (i.e., part a and part b). The first 5 questions are required to be answered, while answering the last question (i.e., "Bonus Question") is optional. If you solve the Bonus Question, you will get an extra 2 points (i.e., your final project score will be out of 38 points). You can find all the questions in FL21_FinalProejct.ipynb. You need to write your code in the code cell provided after each question.

Part 2: In this part, you are required to make 5 interesting business questions based on the two datasets, i.e., Property_details and Order_details, write your code in Python to answer those questions, and visualize your results. Then, you need to write an executive summary (a short report

up to 300 words) about your most important findings and business insights from your answers to those 5 questions, and clearly explain the significance of your results.

Submission Instructions:

1) You need to submit your completed "DAT561_Final Project – FL21" Jupyter notebook (on the zip file) Canvas. You need to provide your code for answering the questions as well as your answers. Also, you need to provide your visualization, executive summary, and business insights using Markdown in the Jupyter notebook.
2) Please do not submit the data files with us.
3) **Only one person in each group needs to submit the final project.**
4) Please name your Jupyter notebook by including the student IDs of all teammates. For example, "14325_34672_12345_ FL21_FinalProejct.ipynb"

Grading:

Total points: 34 points (36 points with the extra credits in the Bonus Question)

Part 1: 25 points (27 points with the extra credits in the Bonus Question)
Question 1: 3 points (2 points for part (a) and 1 point for part (b))
Question 2: 5 points (3 points for part (a) and 2 points for part (b))
Question 3: 4 points (3 points for part (a) and 1 point for part (b))
Question 4: 7 points (3 points for part (a) and 4 points for part (b))
Question 5: 6 points (3 points for part (a) and 3 points for part (b))
Bonus Question: 2 points (extra credit): (1.5 points for part (a) and 0.5 point for part (b))

Part 2: 9 points
You need to ask five business-related questions (2 points).
You need to answer these five questions using Python and the two datasets (2 points).
You need to have at least "5" graphs to visualize your insights (2 points).
Your executive summary should be well-written (2 points).
Your results and business insights should be interesting and meaningful (1 point).

Notes:

- We will look at your logic and whether your code is working.
- We will look at the clarity of your explanations in the executive summary and whether your insights make sense.
- We will look at whether you submitted your file correctly.

Instructor and TAs for Help:

Please feel free to reach out for help, we are ready to help you!

Instructors:

Professor Salih Tutun
Email: salihtutun@wustl.edu
Office Hours: Sunday and Wednesday, 7:00 pm – 8:00 pm
Link: https://wustl.zoom.us/j/93938486334?pwd=aHp5UnhpRGJQUXc0Sm9YQXVPbndEQT09
Extra office hours will be announced on the canvas!
***TAs for help*:**

c.zhijie@wustl.edu (Zhijie Chen),
h.yixian@wustl.edu (Yixian He),
j.zhang1@wustl.edu (Jie Zhang)
Office Hours: Monday, Tuesday, Thursday, Friday, Saturday, Sunday 07:00 pm – 08:00 pm
Link: https://wustl.zoom.us/j/97560073414?pwd=eTRsRG5wcTBYMWtzT0hTK3d3czVEZz09
Extra office hours will be announced on the canvas!

If you have any questions, please contact me at salihtutun@wustl.edu

Good Luck, Guys… :D

Salih Tutun, PhD