Module 11, Exploratory Data Analysis

Author: Jessica Cervi

Expected time = 2 hours

Total points = 70 points

Assignment Overview

In this assignment you will perform exploratory data analysis on a dataset from the travel industry. You will explore different ways of visualizing the data to better understand relationships between variables. You'll examine the descriptive statistics and plots to draw new insights.

This assignment is designed to build your familiarity and comfort coding in Python while also helping you review key topics from each module. As you progress through the assignment, answers will get increasingly complex. It is important that you adopt a data scientist's mindset when completing this assignment. **Remember to run your code from each cell before submitting your assignment.** Running your code beforehand will notify you of errors and give you a chance to fix your errors before submitting. You should view your Vocareum submission as if you are delivering a final project to your manager or client.

Vocareum Tips

- Do not add arguments or options to functions unless you are specifically asked to. This will cause an error in Vocareum.
- Do not use a library unless you are expicitly asked to in the question.
- You can download the Grading Report after submitting the assignment. This will include feedback and hints on incorrect questions.

Learning Objectives

- Visualize data with matplotlib and probe for insights.
- Use exploratory data analysis to describe data.

IMPORTANT INSTRUCTIONS:

- To be able to test for this module, you will be asked to save your figures as PNG into a folder called "results". Please don't change the name we ask you to give to the plots so you are able to get all the points in every question.
- Don't add any customization you're not asked to in the plots.

Index:

Module 11: Exploratory Data Analysis

In this assignment you will work with the pandas concepts you learned in Module 11 to examine a dataset from the travel industry and visualize relationships between variables. You will begin by familiarizing yourself with the columns of the dataset, then explore their relationships.

Inspecting your Data

The dataset that we will be using in this assignment contains booking information 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. More detailed information about the dataset can be found here.

We will begin by importing the necessary libraries for this assignment and by reading the dataset.

In [ ]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn import datasets
import scipy.stats as sp
import seaborn as sns

# Avoid warnings
import warnings
warnings.filterwarnings("ignore")

df = pd.read_csv("data/hotel_bookings.csv")
```

For convenience, we will use the command `.head()` to visualize the first 10 rows of our DataFrame

In [ ]:
```
df.head(10)
```

In [ ]:
```
df.describe()
```

Back to top

Question 1

5 points

We'll begin by exploring the arrival dates, to see when people begin a trip. Use `.value_counts()` on the column `arrival_date_year`. Assign the result to `ans1`.

In [ ]:
```
### GRADED

### YOUR SOLUTION HERE
ans1 = df['arrival_date_year'].value_counts()
### END SOLUTION
```

Back to top

Question 2

5 points

What data type the attribute `.value_counts()` return?

- a) A list
- b) A series
- c) A DataFrame.
- d) An object

Assign the character corresponding to your choice as a string to `ans2`.

In [ ]:
```
### GRADED

### YOUR SOLUTION HERE
ans2 = "b"
### END SOLUTION
```

Back to top

Question 3

5 points

Next, we will examine the lead time. How far in advance do people book travel? We can compute this with the median of the column `lead_time` by ignoring the NaN values. Assign the result to `ans3`.

In [ ]:
```
### GRADED

### YOUR SOLUTION HERE
ans3 = np.nanmedian(df["lead_time"])
### END SOLUTION
```

Back to top

Question 4

10 points

Now, we will create a heatmap which will help us explore relationships between the different variables, or columns, in the travel dataset. What relationships have we missed? This process will help us find out.

To begin, use `.figsize()` to set the figure size to `(15,15)`. Next, produce a heatmap with the correlation between the different columns of `df`. Specify the parameter `annot= True`. DO NOT specify any other parameter. Save your plot as a png with the name "plot4.png" in the folder "results".

In [ ]:
```
### GRADED

### YOUR SOLUTION HERE
plt.figure(figsize=(15,15))
sns.heatmap(df.corr(), annot= True)
plt.savefig("results/plot4.png")
### END SOLUTION
```

Back to top

Question 5

5 points

The hotel wants to know how many parking spaces to be filled with the number of adults booking rooms. We will use feature engineering to create a new measure by dividing `adults` by `required_car_parking_spaces`.

Assign the result of this to a new column created in `df` called `parking_spaces_per_adult`.

In [ ]:
```
### GRADED

### YOUR SOLUTION HERE
df['parking_spaces_per_adult'] = df['adults']/df['required_car_parking_spaces']
### END SOLUTION
```

Back to top

Question 6

10 points

Next, we'll examine the habits of travelers. Are people more likely to stay in on week nights or weekend nights?

Produce a jointplot that compares the relationship between `stays_in_week_nights` and `stays_in_weekend_nights` of `df`. DO NOT specify any parameter. Save your plot as a png with the name "plot6.png" in the folder "results".

In [ ]:
```
### GRADED

### YOUR SOLUTION HERE
sns.jointplot(df.stays_in_week_nights,df.stays_in_weekend_nights)
plt.savefig("results/plot6.png")
### END SOLUTION
```

Back to top

Question 7

5 points

Let's take a closer look at the graph you created to examine travelers who stay in on week nights versus weekend nights.

From the graph produced in question 6, what can you say about the relationship between `stays_in_week_nights` and `stays_in_weekend_nights`?

- a) The two variables have a correlation value close to one (high correlation)
- b) The two variables are not correlated with a value close to zero
- c) The two variables have a correlation value close to -0.5
- d) None of the above

Assign the character corresponding to your choice as a string to `ans7`.

In [ ]:
```
### GRADED

### YOUR SOLUTION HERE
ans7 = "a"
### END SOLUTION
```

Back to top

Question 8

10 points

Could the size of the group impact whether travelers are likely to stay in on weekend nights? Let's explore this.

To begin, use `.figsize()` to set the figure size to `(5,5)`. Next, produce a boxplot that compares the relationship between `adults` and `stays_in_weekend_nights` of `df` and set the x limits equal to `(-1,5)`. Save your plot as a png with the name "plot8.png" in the folder "results".

In [ ]:
```
### GRADED

### YOUR SOLUTION HERE
plt.figsize=(5,5)
sns.boxplot(df['adults'], df['stays_in_weekend_nights'])
plt.xlim(-1,5)
plt.savefig("results/plot8.png")
### END SOLUTION
```

Back to top

Question 9

10 points

Next, we want to examine the reservation date more closely. We will split the entries of this column into three new columns to see the year, month, and day.

Using the appropriate string method, split the column `reservation_status_date` at every occurrence of "-". Next, add three new columns to `df`: `year`, `month` and `day`.

In [ ]:
```
### GRADED

### YOUR SOLUTION HERE
new= df["reservation_status_date"].str.split("-", n = 2, expand = True)
df["year"] = new[0]
df["month"] = new[1]
df["day"] = new[2]
### END SOLUTION
```

Back to top

Question 10

5 points

To conclude, we will examine requests for required car parking spaces. Does a traveler need a parking space, yes or no?

Use hot encoding to create dummy categorical variables for modeling on the column `required_car_parking_spaces`. Make sure you drop the first level by using `drop_first= True`. Save this to a new DataFrame called `df1`.

In [ ]:
```
### GRADED

### YOUR SOLUTION HERE
df1 = pd.get_dummies(df, columns = ['required_car_parking_spaces'], drop_first=True)
### END SOLUTION
```