CS677_HW4_Pandas_questions.ipynb

October 5, 2024

1 Pandas Project

1.0.1 The Data

This data set 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.

All personally identifying information has been removed from the data.

Source: Antonio, Nuno, Ana de Almeida, and Luis Nunes. "Hotel booking demand datasets." Data in brief 22 (2019): 41-49.

##

Data Column Reference

Variable

Type

Description

Source/Engineering

ADR

Numeric

Average Daily Rate as defined by [5]

BO, BL and TR / Calculated by dividing the sum of all lodging transactions by the total number of staying nights

Adults

Integer

Number of adults

BO and BL

Agent

Categorical

ID of the travel agency that made the bookinga

BO and BL

ArrivalDateDayOfMonth

Integer

Day of the month of the arrival date

BO and BL

ArrivalDateMonth

Categorical

Month of arrival date with 12 categories: “January” to “December”

BO and BL

ArrivalDateWeekNumber

Integer

Week number of the arrival date

BO and BL

ArrivalDateYear

Integer

Year of arrival date

BO and BL

AssignedRoomType

Categorical

Code for the type of room assigned to the booking. Sometimes the assigned room type differs from the reserved room type due to hotel operation reasons (e.g. overbooking) or by customer request. Code is presented instead of designation for anonymity reasons

BO and BL

Babies

Integer

Number of babies

BO and BL

BookingChanges

Integer

Number of changes/amendments made to the booking from the moment the booking was entered on the PMS until the moment of check-in or cancellation

BO and BL/Calculated by adding the number of unique iterations that change some of the booking attributes, namely: persons, arrival date, nights, reserved room type or meal

Children

Integer

Number of children

BO and BL/Sum of both payable and non-payable children

Company

Categorical

ID of the company/entity that made the booking or responsible for paying the booking. ID is presented instead of designation for anonymity reasons

BO and BL.

Country

Categorical

Country of origin. Categories are represented in the ISO 3155–3:2013 format [6]

BO, BL and NT

CustomerType

Categorical

Type of booking, assuming one of four categories:

BO and BL

Contract - when the booking has an allotment or other type of contract associated to it;

Group – when the booking is associated to a group;

Transient – when the booking is not part of a group or contract, and is not associated to other transient booking;

Transient-party – when the booking is transient, but is associated to at least other transient booking

DaysInWaitingList

Integer

Number of days the booking was in the waiting list before it was confirmed to the customer

BO/Calculated by subtracting the date the booking was confirmed to the customer from the date the booking entered on the PMS

DepositType

Categorical

Indication on if the customer made a deposit to guarantee the booking. This variable can assume three categories:

BO and TR/Value calculated based on the payments identified for the booking in the transaction (TR) table before the booking s arrival or cancellation date.

No Deposit – no deposit was made;

In case no payments were found the value is “No Deposit”.

If the payment was equal or exceeded the total cost of stay, the value is set as “Non Refund”.

Non Refund – a deposit was made in the value of the total stay cost;

Otherwise the value is set as “Refundable”

Refundable – a deposit was made with a value under the total cost of stay.

DistributionChannel

Categorical

Booking distribution channel. The term “TA” means “Travel Agents” and “TO” means “Tour Operators”

BO, BL and DC

IsCanceled

Categorical

Value indicating if the booking was canceled (1) or not (0)

BO

IsRepeatedGuest

Categorical

Value indicating if the booking name was from a repeated guest (1) or not (0)

BO, BL and C/ Variable created by verifying if a profile was associated with the booking customer. If so, and if the customer profile creation date was prior to the creation date for the booking on the PMS database it was assumed the booking was from a repeated guest

LeadTime

Integer

Number of days that elapsed between the entering date of the booking into the PMS and the arrival date

BO and BL/ Subtraction of the entering date from the arrival date

MarketSegment

Categorical

Market segment designation. In categories, the term “TA” means “Travel Agents” and “TO” means “Tour Operators”

BO, BL and MS

Meal

Categorical

Type of meal booked. Categories are presented in standard hospitality meal packages:

BO, BL and ML

Undefined/SC – no meal package;

BB – Bed & Breakfast;

HB – Half board (breakfast and one other meal – usually dinner);

FB – Full board (breakfast, lunch and dinner)

PreviousBookingsNotCanceled

Integer

Number of previous bookings not cancelled by the customer prior to the current booking

BO and BL / In case there was no customer profile associated with the booking, the value is set to 0. Otherwise, the value is the number of bookings with the same customer profile created before the current booking and not canceled.

PreviousCancellations

Integer

Number of previous bookings that were cancelled by the customer prior to the current booking

BO and BL/ In case there was no customer profile associated with the booking, the value is set to 0. Otherwise, the value is the number of bookings with the same customer profile created before the current booking and canceled.

RequiredCardParkingSpaces

Integer

Number of car parking spaces required by the customer

BO and BL

ReservationStatus

Categorical

Reservation last status, assuming one of three categories:

BO

Canceled – booking was canceled by the customer;

Check-Out – customer has checked in but already departed;

No-Show – customer did not check-in and did inform the hotel of the reason why

ReservationStatusDate

Date

Date at which the last status was set. This variable can be used in conjunction with the ReservationStatus to understand when was the booking canceled or when did the customer checked-out of the hotel

BO

ReservedRoomType

Categorical

Code of room type reserved. Code is presented instead of designation for anonymity reasons

BO and BL

StaysInWeekendNights

Integer

Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel

BO and BL/ Calculated by counting the number of weekend nights from the total number of nights

StaysInWeekNights

Integer

Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel

BO and BL/Calculated by counting the number of week nights from the total number of nights

TotalOfSpecialRequests

Integer

Number of special requests made by the customer (e.g. twin bed or high floor)

BO and BL/Sum of all special requests

This Pandas project has twelve (12) questions. Answer all twelve.

Do not run the cells above the outputs!

1. Read the file "hotel_booking_data.csv" into a DataFrame. Show the first 4 rows of the DataFrame.

[40]:

```
```

[2]:

```
```

[2]:

```
          hotel  is_canceled  …  phone-number       credit_card
0  Resort Hotel            0  …  669-792-1661  ************4322
1  Resort Hotel            0  …  858-637-6955  ************9157
2  Resort Hotel            0  …  652-885-2745  ************3734
3  Resort Hotel            0  …  364-656-8427  ************5677

[4 rows x 36 columns]
```

2. How many data points and features does this data set have? List the features of the data set.

[41]:

```
```

```
This DataFrame has 119390 rows and 36 columns
```

[42]:

```
```

```
[42]: Index(['hotel', 'is_canceled', 'lead_time', 'arrival_date_year',
              'arrival_date_month', 'arrival_date_week_number',
              'arrival_date_day_of_month', 'stays_in_weekend_nights',
              'stays_in_week_nights', 'adults', 'children', 'babies', 'meal',
              'country', 'market_segment', 'distribution_channel',
              'is_repeated_guest', 'previous_cancellations',
              'previous_bookings_not_canceled', 'reserved_room_type',
              'assigned_room_type', 'booking_changes', 'deposit_type', 'agent',
              'company', 'days_in_waiting_list', 'customer_type', 'adr',
              'required_car_parking_spaces', 'total_of_special_requests',
              'reservation_status', 'reservation_status_date', 'name', 'email',
              'phone-number', 'credit_card'],
             dtype='object')
```

3. Is there any missing data? If so, which column has the most missing data?

```
[ ]:
```

```
[12]:
```

```
[12]: 'company'
```

4. Print the non-null count and data type of each column.

```
[ ]:
```

```
[9]:
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 119390 entries, 0 to 119389
Data columns (total 36 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   hotel                           119390 non-null  object
 1   is_canceled                     119390 non-null  int64
 2   lead_time                       119390 non-null  int64
 3   arrival_date_year               119390 non-null  int64
 4   arrival_date_month              119390 non-null  object
 5   arrival_date_week_number        119390 non-null  int64
 6   arrival_date_day_of_month       119390 non-null  int64
 7   stays_in_weekend_nights         119390 non-null  int64
 8   stays_in_week_nights            119390 non-null  int64
 9   adults                          119390 non-null  int64
 10  children                        119386 non-null  float64
 11  babies                          119390 non-null  int64
 12  meal                            119390 non-null  object
 13  country                         118902 non-null  object
 14  market_segment                  119390 non-null  object
 15  distribution_channel            119390 non-null  object
 16  is_repeated_guest               119390 non-null  int64
 17  previous_cancellations          119390 non-null  int64
 18  previous_bookings_not_canceled  119390 non-null  int64
 19  reserved_room_type              119390 non-null  object
 20  assigned_room_type              119390 non-null  object
 21  booking_changes                 119390 non-null  int64
 22  deposit_type                    119390 non-null  object
 23  agent                           103050 non-null  float64
 24  company                         6797 non-null    float64
 25  days_in_waiting_list            119390 non-null  int64
 26  customer_type                   119390 non-null  object
 27  adr                             119390 non-null  float64
 28  required_car_parking_spaces     119390 non-null  int64
 29  total_of_special_requests       119390 non-null  int64
 30  reservation_status              119390 non-null  object
 31  reservation_status_date         119390 non-null  object
 32  name                            119390 non-null  object
 33  email                           119390 non-null  object
 34  phone-number                    119390 non-null  object
 35  credit_card                     119390 non-null  object
dtypes: float64(4), int64(16), object(16)
memory usage: 32.8+ MB
```

5. Remove the ‘company’ column from the data set.

```
[ ]:
```

6. What are the top 5 most common country codes in the dataset? Answer using a single line of code.

```
[ ]:
```

```
[19]:
```

```
[19]: country
      PRT    48590
      GBR    12129
      FRA    10415
      ESP     8568
      DEU     7287
      Name: count, dtype: int64
```

7. The adr is the average daily rate for a person’s stay at the hotel. What is the mean adr across all the hotel stays in the dataset? Round this to 2 decimal points.

```
[ ]:
```

```
[43]:
```

[43]:
```
101.83
```

8. What is the average (mean) number of nights for a stay across the entire data set? Round this to 2 decimal points.

[22]:

[27]:

[27]:
```
3.43
```

9. What is the average *total cost* for a stay in the dataset? Not average daily cost, but *total stay* cost. (You will need to calculate total cost your self by using ADR and week day and weeknight stays). Round this to 2 decimal points.

[26]:

[26]:
```
357.85
```

10. What are the names and emails of people who made exactly 5 “Special Requests”?

[29]:

[29]:
```
                          name                          email
7860             Amanda Harper           Amanda.H66@yahoo.com
11125            Laura Sanders      Sanders_Laura@hotmail.com
14596              Tommy Ortiz            Tommy_O@hotmail.com
14921           Gilbert Miller         Miller.Gilbert@aol.com
14922           Timothy Torres         TTorres@protonmail.com
24630          Jennifer Weaver             Jennifer_W@aol.com
27288           Crystal Horton             Crystal.H@mail.com
27477           Brittney Burke       Burke_Brittney16@att.com
29906          Cynthia Cabrera    Cabrera.Cynthia@xfinity.com
29949              Sarah Floyd              Sarah_F@gmail.com
32267           Michelle Villa         Michelle.Villa@aol.com
39027           Nichole Hebert       Hebert.Nichole@gmail.com
39129         Lindsey Mckenzie       Lindsey.Mckenzie@att.com
39525           Ashley Edwards       Edwards.Ashley@yahoo.com
70114       Christopher Torres   Torres.Christopher@gmail.com
78819   Mrs. Tara Sullivan DVM           Mrs..DVM@xfinity.com
78820           Michaela Brown          MichaelaBrown@att.com
78822        Kurt Maldonado MD              KMD15@xfinity.com
97072         Jason Richardson               Jason.R@zoho.com
97099             Terri Hurley            THurley@xfinity.com
97261        Mrs. Caitlin Webb             Mrs._W@comcast.net
98410             Holly Arroyo          Arroyo_Holly@mail.com
98674          Denise Campbell             Denise_C@gmail.com
99887            Michael Smith            Michael.S42@aol.com
99888       Dr. Trevor Sellers                  Dr._S@aol.com
```

```
101569           Kayla Murphy         Kayla.Murphy@yahoo.com
102061        Taylor Martinez     Taylor.Martinez@hotmail.com
109511         Charles Wilson       Charles_Wilson@yahoo.com
109590          Tyler Allison         Tyler.A@protonmail.com
110082         Matthew Bailey         Matthew_Bailey@aol.com
110083      Charlotte Acevedo        Charlotte_A@verizon.com
111909        Darrell Brennan  Brennan_Darrell51@hotmail.com
111911         Melinda Jensen         MelindaJensen@zoho.com
113915           Terry Arnold          Arnold.Terry@zoho.com
114770            Mary Nguyen     Nguyen.Mary@protonmail.com
114909         Lindsay Cuevas       Lindsay.Cuevas40@mail.com
116455      Cynthia Hernandez   CynthiaHernandez@xfinity.com
116457         Angela Hawkins             Angela_H@gmail.com
118817             Sue Lawson             Sue.L52@comcast.net
119161        Alyssa Richards        Alyssa_Richards@aol.com
```

11. What are the top 5 most common last name in the dataset? Bonus: Can you figure this out in one line of pandas code?
    - For simplicity treat the a title such as MD as a last name, for example Caroline Conley MD can be said to have the last name MD.
    - You may have to revisit string methods and lambda functions for this question.

[34]:

```
[34]: name
      Smith       2503
      Johnson     1990
      Williams    1618
      Jones       1434
      Brown       1423
      Name: count, dtype: int64
```

12. How many arrivals took place between the 1st and the 15th of the month (inclusive of 1 and 15)? Bonus: Can you do this in one line of pandas code?

[37]:

```
[37]: 58152
```

[ ]: