Pandas Project Exercise

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.

Acknowledgements The data is originally from the article Hotel Booking Demand Datasets, written by Nuno Antonio, Ana Almeida, and Luis Nunes for Data in Brief, Volume 22, February 2019.

NOTE: Names, Emails, Phone Numbers, and Credit Card numbers in the data are synthetic and not real information from people. The hotel data is real from the publication listed above.

Data Column Reference

Variable	Type	Description
ADR	Numeric	Average Daily Rate as defined by [5]
Adults	Integer	Number of adults
Agent	Categorical	ID of the travel agency that made the booking[a]
ArrivalDateDayOfMonth	Integer	Day of the month of the arrival date
ArrivalDateMonth	Categorical	Month of arrival date with 12 categories: “January” to “December”
ArrivalDateWeekNumber	Integer	Week number of the arrival date
ArrivalDateYear	Integer	Year of arrival date
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
Babies	Integer	Number of babies
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
Children	Integer	Number of 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
Country	Categorical	Country of origin. Categories are represented in the ISO 3155–3:2013 format [6]
CustomerType	Categorical	Type of booking, assuming one of four categories:
		Contract - when the booking has an allotment or other type of contract associated to it;
		Group – when the booking is associated to a group;

Variable	Type	Description
		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
DepositType	Categorical	Indication on if the customer made a deposit to guarantee the booking. This variable can assume three categories:
		No Deposit – no deposit was made;
		Non Refund – a deposit was made in the value of the total stay cost;
		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”
IsCanceled	Categorical	Value indicating if the booking was canceled (1) or not (0)
IsRepeatedGuest	Categorical	Value indicating if the booking name was from a repeated guest (1) or not (0)
LeadTime	Integer	Number of days that elapsed between the entering date of the booking into the PMS and the arrival date
MarketSegment	Categorical	Market segment designation. In categories, the term “TA” means “Travel Agents” and “TO” means “Tour Operators”
Meal	Categorical	Type of meal booked. Categories are presented in standard hospitality meal packages:
		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
PreviousCancellations	Integer	Number of previous bookings that were cancelled by the customer prior to the current booking
RequiredCardParkingSpaces	Integer	Number of car parking spaces required by the customer
ReservationStatus	Categorical	Reservation last status, assuming one of three categories:
		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-
ReservedRoomType	Categorical	Code of room type reserved. Code is presented instead of designation for anonymity reasons
StaysInWeekendNights	Integer	Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
StaysInWeekNights	Integer	Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel
TotalOfSpecialRequests	Integer	Number of special requests made by the customer (e.g. twin bed or high floor)

TASKS

Complete the tasks shown in bold below. The expected output is shown in a cell below. Be careful not to run the cell above the expected output, as it will clear the expected output. Try your best to solve these in one line of pandas code (not every single question can be solved in one line, but many can be!) Refer to solutions notebook and video to view possible solutions. NOTE: Many tasks have multiple correct solution methods!

TASK: Run the following code to read in the "hotel_booking_data.csv" file. Feel free to explore the file and understand the data statistics a littel bit before continuing with the rest of the exercise.

```
import pandas as pd
from datetime import datetime

hotels = pd.read_csv("hotel_booking_data.csv")
```

TASK: How many rows are there?

```
hotels.shape[0]
```

```
119390
```

TASK: Is there any missing data? If so, which column has the most missing data?

```
hotels.isnull().sum()
```

	0
hotel	0
is_canceled	0
lead_time	0
arrival_date_year	0
arrival_date_month	0
arrival_date_week_number	0
arrival_date_day_of_month	0
stays_in_weekend_nights	0
stays_in_week_nights	0
adults	0
children	4
babies	0
meal	0
country	488
market_segment	0
distribution_channel	0
is_repeated_guest	0
previous_cancellations	0
previous_bookings_not_canceled	0
reserved_room_type	0
assigned_room_type	0
booking_changes	0
deposit_type	0
agent	16340
company	112593
days_in_waiting_list	0
customer_type	0

```
                        adr    0
required_car_parking_spaces    0
  total_of_special_requests    0
         reservation_status    0
    reservation_status_date    0
                       name    0
                      email    0
               phone-number    0
                credit_card    0

dtype: int64
```

The most null --> Company

TASK: Drop the "company" column from the dataset.

```
hotels.drop('company', axis=1, inplace=True)
```

TASK: What are the top 5 most common country codes in the dataset?

```
hotels['country'].value_counts()[:5]
```

```
         count
country
  PRT    48590
  GBR    12129
  FRA    10415
  ESP     8568
  DEU     7287

dtype: int64
```

TASK: What is the name of the person who paid the highest ADR (average daily rate)? How much was their ADR?

```
hotels.sort_values('adr', ascending=False)[['adr', 'name']].iloc[0]
```

```
             48515
adr         5400.0
name  Daniel Walter

dtype: object
```

TASK: 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?

```
hotels['adr'].mean()
```

```
101.83112153446686
```

TASK: What is the average (mean) number of nights for a stay across the entire data set? Feel free to round this to 2 decimal points.

```
hotels['total_stay_nights'] = hotels['stays_in_weekend_nights'] + hotels['stays_in_week_nights']
hotels['total_stay_nights'].mean()
```

```
3.4279001591423066
```

TASK: 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). Feel free to round this to 2 decimal points.

```
hotels['total_paid'] = hotels['adr'] * hotels['total_stay_nights']
hotels['total_paid'].mean()
```

```
357.84820780634897
```

TASK: What are the names and emails of people who made exactly 5 "Special Requests"?

```
hotels[hotels['total_of_special_requests'] == 5][['name', 'email']]
```

	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

TASK: What percentage of hotel stays were classified as "repeat guests"? (Do not base this off the name of the person, but instead of the is_repeated_guest column)

```
hotels['is_repeated_guest'] = hotels['is_repeated_guest'].apply(lambda x: 'Yes' if x == 1 else 'No')
hotels['is_repeated_guest'].value_counts(normalize=True)
```

	proportion
is_repeated_guest	
No	1.0

dtype: float64

TASK: 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)

```
hotels['name'].apply(lambda name: name.split()[-1]).value_counts()[:5]
```

```
          count
name
Smith      2503
Johnson    1990
Williams   1618
Jones      1434
Brown      1423

dtype: int64
```

TASK: What are the names of the people who had booked the most number children and babies for their stay? (Don't worry if they canceled, only consider number of people reported at the time of their reservation)

```
hotels['total_kids'] = hotels['children'] + hotels['babies']
hotels.sort_values('total_kids', ascending=False)[['name', 'adults', 'children', 'babies', 'total_kids']][:3]
```

```
                  name  adults  children  babies  total_kids
328      Jamie Ramirez       2      10.0       0        10.0
46619  Nicholas Parker       2       0.0      10        10.0
78656    Marc Robinson       1       0.0       9         9.0
```

TASK: What are the top 3 most common area code in the phone numbers? (Area code is first 3 digits)

```
hotels['phone-number'].apply(lambda num: num[:3]).value_counts()[:3]
```

```
              count
phone-number
799             168
185             167
541             166

dtype: int64
```

TASK: 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?

```
hotels['arrival_date_day_of_month'].apply(lambda day: day in range(1, 16)).sum()
```

```
58152
```

HARD BONUS TASK: Create a table for counts for each day of the week that people arrived. (E.g. 5000 arrivals were on a Monday, 3000 were on a Tuesday, etc..)

```
import pandas as pd

df = pd.read_csv('hotel_booking_data.csv')

# Convert 'arrival_date' to datetime
df['arrival_date_day_of_month'] = pd.to_datetime(df['arrival_date_day_of_month'])

# Extract day of the week (0=Monday, 6=Sunday)
df['day_of_week'] = df['arrival_date_day_of_month'].dt.day_name()

# Count arrivals by day of week
arrival_counts = df['day_of_week'].value_counts().sort_index()

# Create a table
arrival_counts_table = pd.DataFrame(arrival_counts).reset_index()
arrival_counts_table.columns = ["Day of Week", "Arrival Count"]

# Display table
print(arrival_counts_table)
```

```
  Day of Week  Arrival Count
0    Thursday         119390
```