 Data Preprocessing (Hotel booking).ipynb ☆

File Edit View Insert Runtime Tools Help [All changes saved](#)

+ Code + Text

```
[1] from google.colab import files
    uploaded = files.upload()
```

Choose Files hotel_bookings.csv

- hotel_bookings.csv(text/csv) - 16855599 bytes, last modified: 1/12/2023 - 100% done

Saving hotel_bookings.csv to hotel_bookings.csv

```
[3] import matplotlib.pyplot as plt
    import seaborn as sns
    import pandas as pd
```

```
[4] df= pd.read_csv("hotel_bookings.csv")
```

df

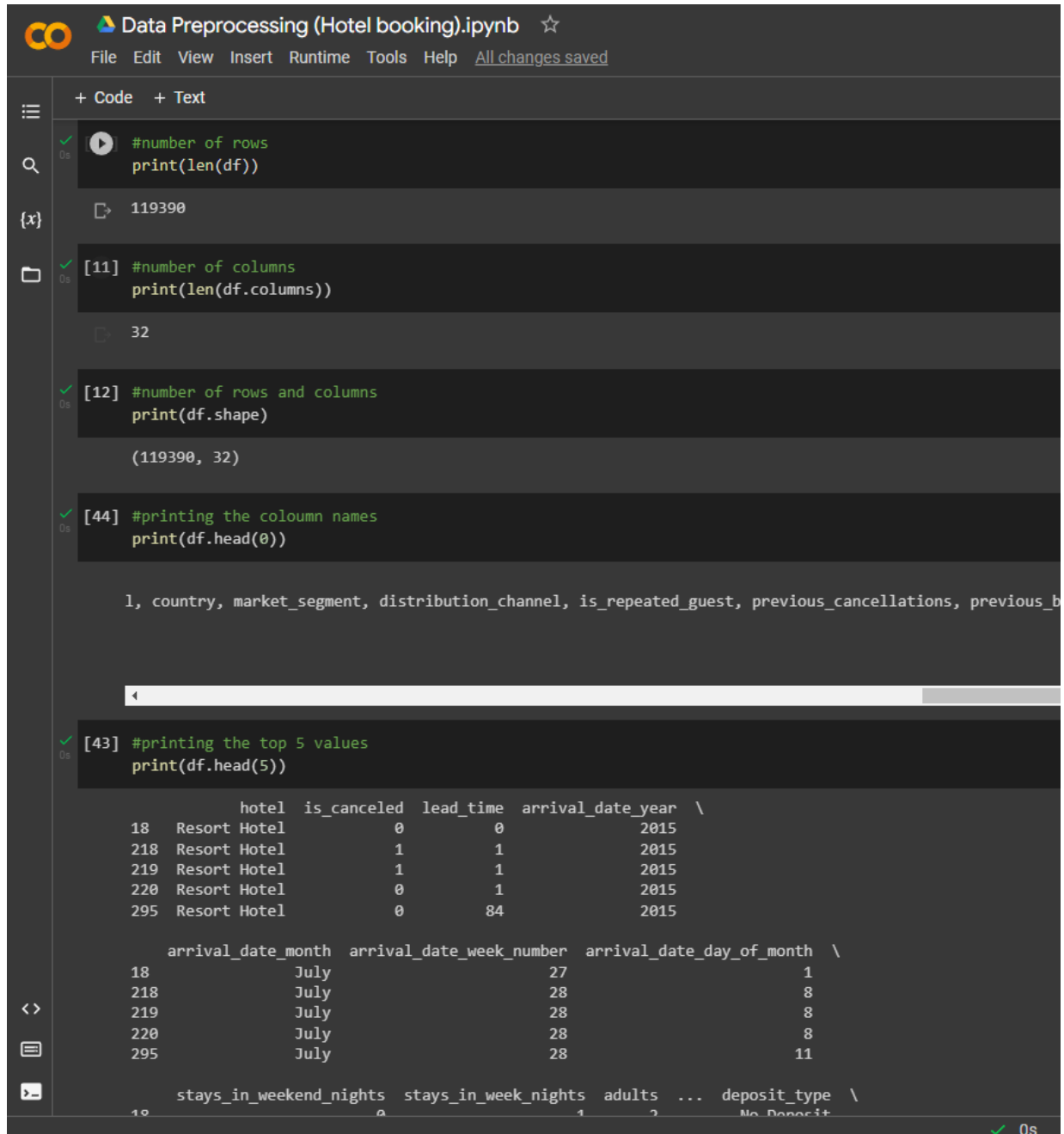
	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_w
0	Resort Hotel	0	342	2015	July	
1	Resort Hotel	0	737	2015	July	
2	Resort Hotel	0	7	2015	July	
3	Resort Hotel	0	13	2015	July	
4	Resort Hotel	0	14	2015	July	
...
119385	City Hotel	0	23	2017	August	
119386	City Hotel	0	102	2017	August	
119387	City Hotel	0	34	2017	August	
...

The screenshot shows a Jupyter Notebook titled "Data Preprocessing (Hotel booking).ipynb". The code cell contains the command `df.info()`, which has been executed. The output displays the DataFrame's structure, including the number of entries (119390), the total number of columns (32), and a detailed list of columns with their data types and non-null counts.

```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 119390 entries, 0 to 119389
Data columns (total 32 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
dtypes: float64(4), int64(16), object(12)
memory usage: 29.1+ MB

```



The screenshot shows a Jupyter Notebook interface with the following code and output:

```
#number of rows
print(len(df))
```

119390

```
#number of columns
print(len(df.columns))
```

32

```
#number of rows and columns
print(df.shape)
```

(119390, 32)

```
#printing the coloumn names
print(df.head(0))
```

1, country, market_segment, distribution_channel, is_repeated_guest, previous_cancellations, previous_b

```
#printing the top 5 values
print(df.head(5))
```

	hotel	is_canceled	lead_time	arrival_date_year	\
18	Resort Hotel	0	0	2015	
218	Resort Hotel	1	1	2015	
219	Resort Hotel	1	1	2015	
220	Resort Hotel	0	1	2015	
295	Resort Hotel	0	84	2015	

	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	\
18	July		27	1
218	July		28	8
219	July		28	8
220	July		28	8
295	July		28	11

	stays_in_weekend_nights	stays_in_week_nights	adults	...	deposit_type	\
18	0	1	2		No Deposit	

0s

0s

	stays_in_weekend_nights	stays_in_week_nights	adults	...	deposit_type	\
18	0	1	2	...	No Deposit	
218	0	1	2	...	No Deposit	
219	0	1	2	...	No Deposit	
220	0	2	2	...	No Deposit	
295	1	1	2	...	No Deposit	

	agent	company	days_in_waiting_list	customer_type	adr	\
18	NaN	110.0	0	Transient	107.42	
218	NaN	110.0	0	Transient	104.72	
219	NaN	110.0	0	Transient	104.72	
220	NaN	110.0	0	Transient	104.72	
295	NaN	113.0	0	Transient	100.00	

	required_car_parking_spaces	total_of_special_requests	\
18	0	0	
218	0	1	
219	0	1	
220	1	1	
295	1	0	

	reservation_status	reservation_status_date
18	Check-Out	2015-07-02
218	Canceled	2015-07-08
219	Canceled	2015-07-08
220	Check-Out	2015-07-10
295	Check-Out	2015-07-13

[5 rows x 32 columns]

0s [45] #print last 5 values
print(df.tail(5))

	hotel	is_canceled	lead_time	arrival_date_year	\
119119	City Hotel	0	40	2017	
119122	City Hotel	0	40	2017	
119123	City Hotel	0	40	2017	
119124	City Hotel	0	0	2017	
119248	City Hotel	0	22	2017	

	arrival_date_month	arrival_date_week_number	\
119119	August	35	
119122	August	35	
119123	August	35	
119124	August	35	
119248	August	35	

```
co Data Preprocessing (Hotel booking).ipynb ☆
File Edit View Insert Runtime Tools Help All changes saved

+ Code + Text

[74] #cleaning the dataset
      #finding null values
      print(df.isnull())

      220      False      False      False      False      False
      295      False      False      False      False      False
      ...      ...
      119119      False      False      False      False      False
      119122      False      False      False      False      False
      119123      False      False      False      False      False
      119124      False      False      False      False      False
      119248      False      False      False      False      False

      arrival_date_week_number  arrival_date_day_of_month \
      18                      False                      False
      218                      False                      False
      219                      False                      False
      220                      False                      False
      295                      False                      False
      ...                      ...
      119119                      False                      False
      119122                      False                      False
      119123                      False                      False
      119124                      False                      False
      119248                      False                      False

      stays_in_weekend_nights  stays_in_week_nights  adults  ... \
      18                      False                  False  False  ...
      218                      False                  False  False  ...
      219                      False                  False  False  ...
      220                      False                  False  False  ...
      295                      False                  False  False  ...
      ...                      ...
      119119                      False                  False  False  ...
      119122                      False                  False  False  ...
      119123                      False                  False  False  ...
      119124                      False                  False  False  ...
      119248                      False                  False  False  ...

      deposit_type  agent  company  days_in_waiting_list  customer_type \
      18           False    True    False                  False          False
      218           False    True    False                  False          False
      219           False    True    False                  False          False
      220           False    True    False                  False          False
      295           False    True    False                  False          False
      ...           ...
      119119         False    True    False                  False          False
      119122         False    True    False                  False          False
      119123         False    True    False                  False          False
```

```
#finding sum of all null values
print(df.isnull().sum())
```

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	0
babies	0
meal	0
country	174
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	6580
company	0
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
dtype:	int64

```
[62] #dropping the coloum company because it is mostly null
df.dropna(subset=['company'], inplace=True)
print (df)
```

```

#dropping the coloum company because it is mostly null
df.dropna(subset=['company'], inplace=True)
print (df)

```

295	Resort Hotel	0	84	2015
...
119119	City Hotel	0	40	2017
119122	City Hotel	0	40	2017
119123	City Hotel	0	40	2017
119124	City Hotel	0	0	2017
119248	City Hotel	0	22	2017

	arrival_date_month	arrival_date_week_number	\
18	July	27	
218	July	28	
219	July	28	
220	July	28	
295	July	28	
...	
119119	August	35	
119122	August	35	
119123	August	35	
119124	August	35	
119248	August	35	

	arrival_date_day_of_month	stays_in_weekend_nights	\
18	1	0	
218	8	0	
219	8	0	
220	8	0	
295	11	1	
...	
119119	29	0	
119122	29	0	
119123	29	0	
119124	29	0	
119248	29	0	

	stays_in_week_nights	adults	...	deposit_type	agent	company	\
18	1	2	...	No Deposit	NaN	110.0	
218	1	2	...	No Deposit	NaN	110.0	
219	1	2	...	No Deposit	NaN	110.0	
220	2	2	...	No Deposit	NaN	110.0	
295	1	2	...	No Deposit	NaN	113.0	
...	
119119	1	1	...	No Deposit	NaN	451.0	
119122	1	1	...	No Deposit	NaN	451.0	
119123	1	1	...	No Deposit	NaN	451.0	

```

[48] #checking the remaining null values
print(df.isnull().sum())

```

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	0
babies	0
meal	0
country	174
market segment	0

```
#checking the remaining null values
print(df.isnull().sum())
```

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	0
babies	0
meal	0
country	174
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	6580
company	0
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
dtype: int64	

```
[63] #checking the sum of remaining null values
print(df.isnull().values.sum())

6754

[64] #replacing the remaining null values with NAN
data= df.fillna('NAN')
data.isna().sum().sum()
```



```
[64] #replacing the remaining null values with NAN
data= df.fillna('NAN')
data.isna().sum().sum()

0

#checking the datatype
data.dtypes

hotel                object
is_canceled          int64
lead_time            int64
arrival_date_year     int64
arrival_date_month    object
arrival_date_week_number  int64
arrival_date_day_of_month  int64
stays_in_weekend_nights  int64
stays_in_week_nights  int64
adults              int64
children            float64
babies              int64
meal                object
country             object
market_segment       object
distribution_channel  object
is_repeated_guest    int64
previous_cancellations  int64
previous_bookings_not_canceled  int64
reserved_room_type    object
assigned_room_type     object
booking_changes       int64
deposit_type         object
agent               object
company             float64
days_in_waiting_list  int64
customer_type        object
adr                 float64
required_car_parking_spaces  int64
total_of_special_requests  int64
reservation_status    object
reservation_status_date  object
dtype: object

[72] #converting object into int64 for reservation_status_date
df['reservation_status_date'].str.replace('-', '').astype(int)
```

```
previous_bookings_not_canceled    int64
reserved_room_type                object
assigned_room_type                object
booking_changes                   int64
deposit_type                      object
agent                            object
company                           float64
days_in_waiting_list             int64
customer_type                     object
adr                               float64
required_car_parking_spaces       int64
total_of_special_requests         int64
reservation_status                object
reservation_status_date           object
dtype: object

[72] #converting object into int64 for reservation_status_date
df['reservation_status_date'].str.replace('-', '').astype(int)

18      20150702
218     20150708
219     20150708
220     20150710
295     20150713
...
119119  20170830
119122  20170830
119123  20170830
119124  20170830
119248  20170901
Name: reservation_status_date, Length: 6797, dtype: int64
```