Wrangling

February 9, 2024

```
[ ]: import pandas as pd
     import plotly.graph_objects as go
     from sklearn.preprocessing import LabelEncoder, StandardScaler
```

```
[ ]: original_df = pd.read_csv('Original_Data.csv')
     original_df.head()
```

```
[ ]:   Booking_ID  no_of_adults  no_of_children  no_of_weekend_nights  \
     0   INN00001             2               0                     1
     1   INN00002             2               0                     2
     2   INN00003             1               0                     2
     3   INN00004             2               0                     0
     4   INN00005             2               0                     1

        no_of_week_nights type_of_meal_plan  required_car_parking_space  \
     0                  2       Meal Plan 1                           0
     1                  3      Not Selected                           0
     2                  1       Meal Plan 1                           0
     3                  2       Meal Plan 1                           0
     4                  1      Not Selected                           0

       room_type_reserved  lead_time  arrival_year  arrival_month  arrival_date  \
     0        Room_Type 1        224          2017             10             2
     1        Room_Type 1          5          2018             11             6
     2        Room_Type 1          1          2018              2            28
     3        Room_Type 1        211          2018              5            20
     4        Room_Type 1         48          2018              4            11

       market_segment_type  repeated_guest  no_of_previous_cancellations  \
     0             Offline               0                             0
     1              Online               0                             0
     2              Online               0                             0
     3              Online               0                             0
     4              Online               0                             0

        no_of_previous_bookings_not_canceled  avg_price_per_room  \
     0                                     0               65.00
     1                                     0              106.68
```

```
2                                     0            60.00
3                                     0           100.00
4                                     0            94.50

   no_of_special_requests booking_status
0                       0   Not_Canceled
1                       1   Not_Canceled
2                       0       Canceled
3                       0       Canceled
4                       0       Canceled
```

[ ]:
```
original_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 36275 entries, 0 to 36274
Data columns (total 19 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   Booking_ID                            36275 non-null  object
 1   no_of_adults                          36275 non-null  int64
 2   no_of_children                        36275 non-null  int64
 3   no_of_weekend_nights                  36275 non-null  int64
 4   no_of_week_nights                     36275 non-null  int64
 5   type_of_meal_plan                     36275 non-null  object
 6   required_car_parking_space            36275 non-null  int64
 7   room_type_reserved                    36275 non-null  object
 8   lead_time                             36275 non-null  int64
 9   arrival_year                          36275 non-null  int64
 10  arrival_month                         36275 non-null  int64
 11  arrival_date                          36275 non-null  int64
 12  market_segment_type                   36275 non-null  object
 13  repeated_guest                        36275 non-null  int64
 14  no_of_previous_cancellations          36275 non-null  int64
 15  no_of_previous_bookings_not_canceled  36275 non-null  int64
 16  avg_price_per_room                    36275 non-null  float64
 17  no_of_special_requests                36275 non-null  int64
 18  booking_status                        36275 non-null  object
dtypes: float64(1), int64(13), object(5)
memory usage: 5.3+ MB
```

1 Preprocessing

1.1 Heatmap

[ ]:
```
numerical_columns = original_df.select_dtypes(include=['int64', 'float64']).
  ↪columns
print(numerical_columns)
```

```
Index(['no_of_adults', 'no_of_children', 'no_of_weekend_nights',
       'no_of_week_nights', 'required_car_parking_space', 'lead_time',
       'arrival_year', 'arrival_month', 'arrival_date', 'repeated_guest',
       'no_of_previous_cancellations', 'no_of_previous_bookings_not_canceled',
       'avg_price_per_room', 'no_of_special_requests'],
      dtype='object')
```

[ ]:

```
fig = go.Figure(data=go.Heatmap(
    z=original_df.values,
    x=numerical_columns,
    y=numerical_columns,
))

fig.update_layout(
    title='<b>Heatmap'
)

fig.show()
```

1.2 Enconding Categorical Data

[ ]:

```
categorical_columns = ['type_of_meal_plan', 'room_type_reserved',␣
 ↪'market_segment_type', 'repeated_guest', 'booking_status']
labelencoder = LabelEncoder()
original_df[categorical_columns] = original_df[categorical_columns].
 ↪apply(labelencoder.fit_transform)

original_df.head()
```

[ ]:

```
  Booking_ID  no_of_adults  no_of_children  no_of_weekend_nights  \
0   INN00001             2               0                     1
1   INN00002             2               0                     2
2   INN00003             1               0                     2
3   INN00004             2               0                     0
4   INN00005             2               0                     1

   no_of_week_nights  type_of_meal_plan  required_car_parking_space  \
0                  2                  0                           0
1                  3                  3                           0
2                  1                  0                           0
3                  2                  0                           0
4                  1                  3                           0

   room_type_reserved  lead_time  arrival_year  arrival_month  arrival_date  \
0                   0        224          2017             10             2
1                   0          5          2018             11             6
2                   0          1          2018              2            28
```

```
3                   0        211          2018              5            20
4                   0         48          2018              4            11

   market_segment_type  repeated_guest  no_of_previous_cancellations  \
0                    3               0                             0
1                    4               0                             0
2                    4               0                             0
3                    4               0                             0
4                    4               0                             0

   no_of_previous_bookings_not_canceled  avg_price_per_room  \
0                                     0               65.00
1                                     0              106.68
2                                     0               60.00
3                                     0              100.00
4                                     0               94.50

   no_of_special_requests  booking_status
0                       0               1
1                       1               1
2                       0               0
3                       0               0
4                       0               0
```

1.3 Initial Analysis

1.3.1 Average Prices per Type of Room

```
[ ]: columns_new_df = ['avg_price_per_room', 'room_type_reserved']
     new_df = original_df[columns_new_df]
     new_df.head()
```

```
[ ]:    avg_price_per_room  room_type_reserved
     0               65.00                   0
     1              106.68                   0
     2               60.00                   0
     3              100.00                   0
     4               94.50                   0
```

```
[ ]: new_df.groupby(['room_type_reserved']).min()
```

```
[ ]:                     avg_price_per_room
     room_type_reserved
     0                                  0.0
     1                                  0.0
     2                                  0.0
     3                                  0.0
     4                                  0.0
```

```
5                                  0.0
6                                  0.0
```

[ ]:
```python
new_df.groupby(['room_type_reserved']).mean()
```

[ ]:
```
                    avg_price_per_room
room_type_reserved
0                            95.918532
1                            87.848555
2                            73.678571
3                           125.287317
4                           123.733623
5                           182.212836
6                           155.198291
```

[ ]:
```python
new_df.groupby(['room_type_reserved']).max()
```

[ ]:
```
                    avg_price_per_room
room_type_reserved
0                               540.00
1                               284.10
2                               130.00
3                               375.50
4                               250.00
5                               349.63
6                               306.00
```

1.3.2 Average of Adults and Children in 2017 and 2018

[ ]:
```python
columns_new_df = ['no_of_adults', 'no_of_children', 'arrival_year']
new_df = original_df[columns_new_df]
new_df.groupby(['arrival_year']).min()
```

[ ]:
```
              no_of_adults  no_of_children
arrival_year
2017                     0               0
2018                     0               0
```

[ ]:
```python
new_df.groupby(['arrival_year']).max()
```

[ ]:
```
              no_of_adults  no_of_children
arrival_year
2017                     3               9
2018                     4              10
```

[ ]:
```python
new_df.groupby(['arrival_year']).mean().astype(int)
```

```
[ ]:               no_of_adults  no_of_children
     arrival_year
     2017                     1               0
     2018                     1               0
```

1.4 Normalizing

```
[ ]: columns_to_normalize = ['lead_time', 'arrival_year', 'arrival_month',␣
      ↪'arrival_date', 'avg_price_per_room', 'no_of_special_requests']
     scaler = StandardScaler()
     fitting = scaler.fit(original_df[columns_to_normalize])
     original_df[columns_to_normalize] = fitting.
      ↪transform(original_df[columns_to_normalize])
     original_df.head()
```

```
[ ]:   Booking_ID  no_of_adults  no_of_children  no_of_weekend_nights  \
     0   INN00001             2               0                     1
     1   INN00002             2               0                     2
     2   INN00003             1               0                     2
     3   INN00004             2               0                     0
     4   INN00005             2               0                     1

        no_of_week_nights  type_of_meal_plan  required_car_parking_space  \
     0                  2                  0                           0
     1                  3                  3                           0
     2                  1                  0                           0
     3                  2                  0                           0
     4                  1                  3                           0

        room_type_reserved  lead_time  arrival_year  arrival_month  arrival_date  \
     0                   0   1.614896     -2.137469       0.839242     -1.555662
     1                   0  -0.933701      0.467843       1.164990     -1.098013
     2                   0  -0.980250      0.467843      -1.766747      1.419055
     3                   0   1.463610      0.467843      -0.789501      0.503757
     4                   0  -0.433291      0.467843      -1.115250     -0.525952

        market_segment_type  repeated_guest  no_of_previous_cancellations  \
     0                    3               0                             0
     1                    4               0                             0
     2                    4               0                             0
     3                    4               0                             0
     4                    4               0                             0

        no_of_previous_bookings_not_canceled  avg_price_per_room  \
     0                                     0           -1.095033
     1                                     0            0.092806
     2                                     0           -1.237528
```

```
3                                     0           -0.097567
4                                     0           -0.254312

   no_of_special_requests  booking_status
0                -0.78814               1
1                 0.48376               1
2                -0.78814               0
3                -0.78814               0
4                -0.78814               0
```