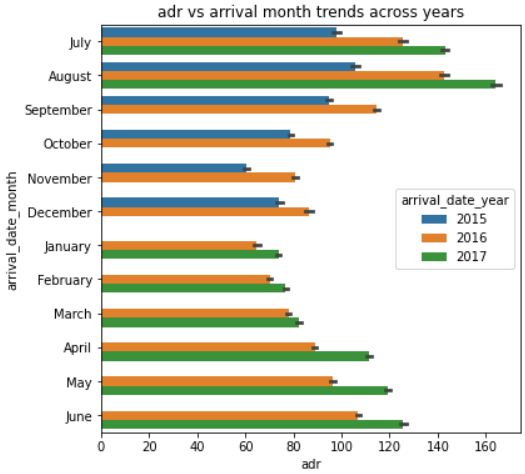
Hotel Booking Data Analysis

Context: As a hotel manager, I would be interested to observe the performance of the hotel. In the first part of the analysis let us understand the overview of ADR (Average Daily Rate)

"What are the factors leading to significant impact on ADR of a hotel?"

Approach

Starting with the process of identifying anomalies in the data corresponding to ADR (ADR values are non-negative), we understand the different time frames corresponding to the varying ADR. **Month of August** saw significantly **higher** ADR, maintaining the trend across the span of 3 years.



As we analyse the ADR feature, we treat the categorical and numerical features separately. There are some interesting trends to observe, customers getting the room type as according to their reservation specifications, lead to higher ADR compared to the cases where the assigned and reserved room types differ.

reserve_assigned	count	mean	std	min	25%	50%	75%	max
0	14916.0	83.363412	47.805168	0.0	49.00	75.8	109.8	510.0
1	104472.0	104.418173	47.624410	0.0	72.25	96.0	127.8	451.5

Reserve_assigned 0 represents that the assigned room type was different from the reserved while 1 represents that they were same

Certain **market segment types** such as **Online TA** (Travel Agents) **Aviation, Direct** contributed significantly more to the ADR as compared to other segments. *What market share is constituted by the three segment types based on the data?*

```
[18] #What share of market is constituted by the Aviation, Online TA and Direct?
     t1 = data.loc[data.market_segment.isin(['Aviation','Online TA','Direct'])]
     print(t1.shape[0]/data.shape[0] * 100)
```

```
58.06181422229667
```

```
[19] #what percentage of ADR is constituted by the above three?
     print(t1['adr'].sum()/data['adr'].sum() * 100)
```

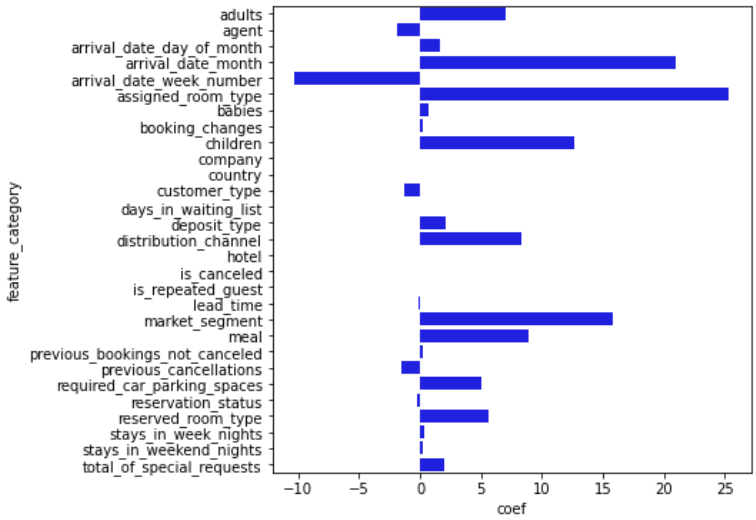
```
66.60823153378341
```

```
[17] #what percentage of bookings are constituted by Transient customers?
     t1 = data.loc[data.customer_type=='Transient']
     print(t1.shape[0]/data.shape[0] * 100)

     75.05905017170618

[20] #what percentage of ADR is contributed by the Transient customers?
     print(t1['adr'].sum() / data['adr'].sum() * 100)

     66.60823153378341
```

Transient customer comprised a large share in the entire bookings while contributing a significant portion to the ADR

Next we further try to understand the feature importance. We deal with the missing values as per the Hotel Booking Dataset paper. As per the paper the missing values are on purpose and hence we experiment by generating an additional category.

Thereafter, we one hot encode the categorical fields and train a linear regressor to learn the coefficients representing the ADR.



'arrival_date_month' forms an important feature with higher coefficient drawing evidence to our initial analysis, accompanied with 'assigned_room_type' and 'market_segment'

The above analysis is informatory providing direction to the further in-depth analysis to mark key identifiers and transform business decisions. We can form discretionary insights into the factors influencing the ADR

Based on the data, it was safe to assume that the data is collected in Portugal. One of the major festivals celebrated in Portugal is the Carnival festival around the month of February and March. However, the event is not a major contributor to the hotel business based on the collected data (probably due to the location of hotels from the location of festival).

The months of July and August, with the summer heat, sunseekers pack the resorts and beaches, with prices peaking in the two months. This can be seen reflecting in the data we have, hinting the location of the hotels closer to beaches.

(reference: https://www.lonelyplanet.com/portugal/narratives/planning/month-by-month)

With the improvement in Portugal's economy from the impact of Eurozone's debt crisis, the country's hospitality segment saw a growth after 2016, which was reflected in year-on-year growth.

(reference: https://www.hotelmanagement.net/development/why-portugal-a-winning-streak)