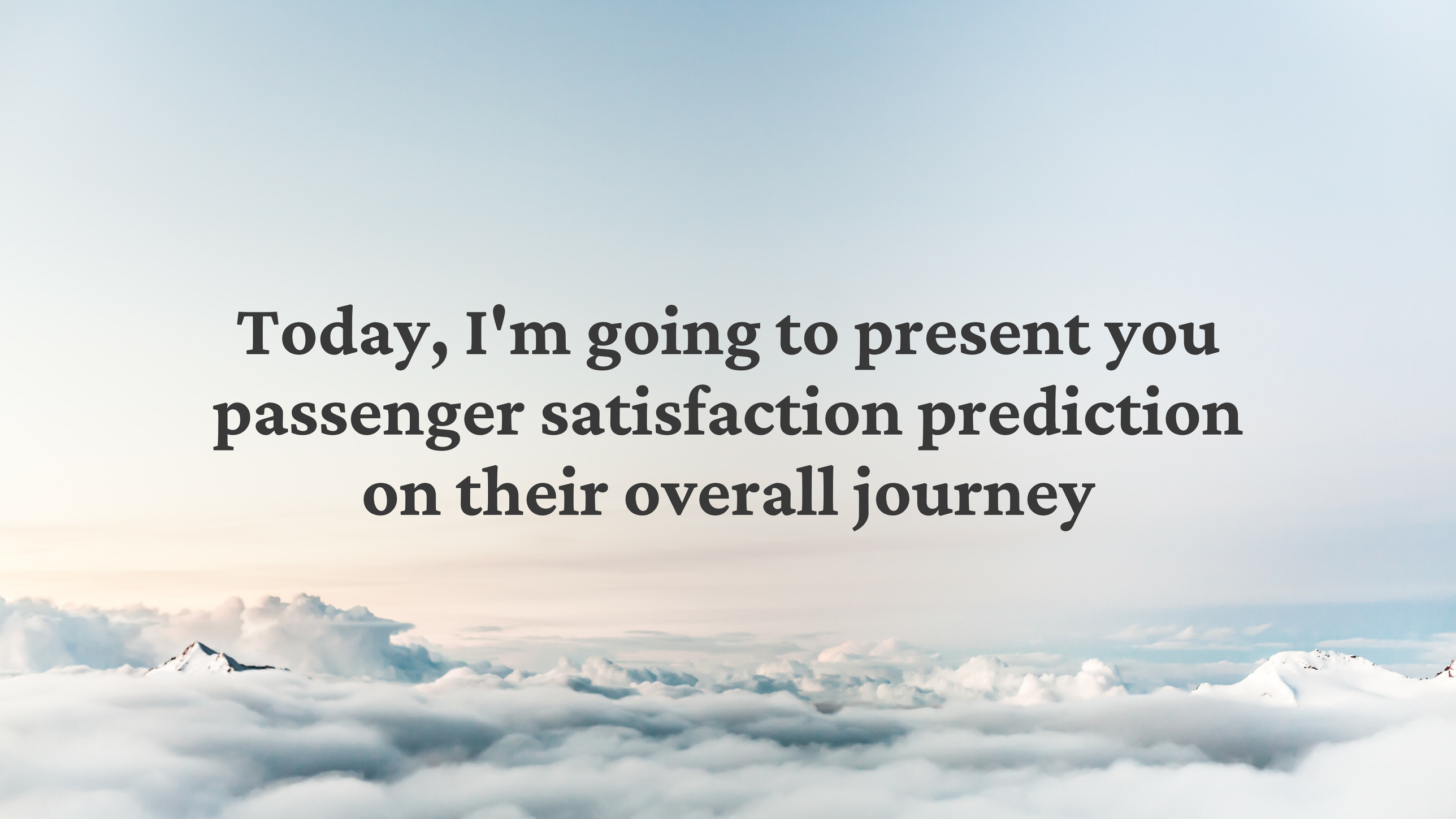
PASSENGER SATISFACTION


Today, I'm going to present you
passenger satisfaction prediction
on their overall journey

Summary :

1) Context of the dataset
2) Cleaning the dataset
3) Tableau visualizations
4) SQL analyse
5) Results

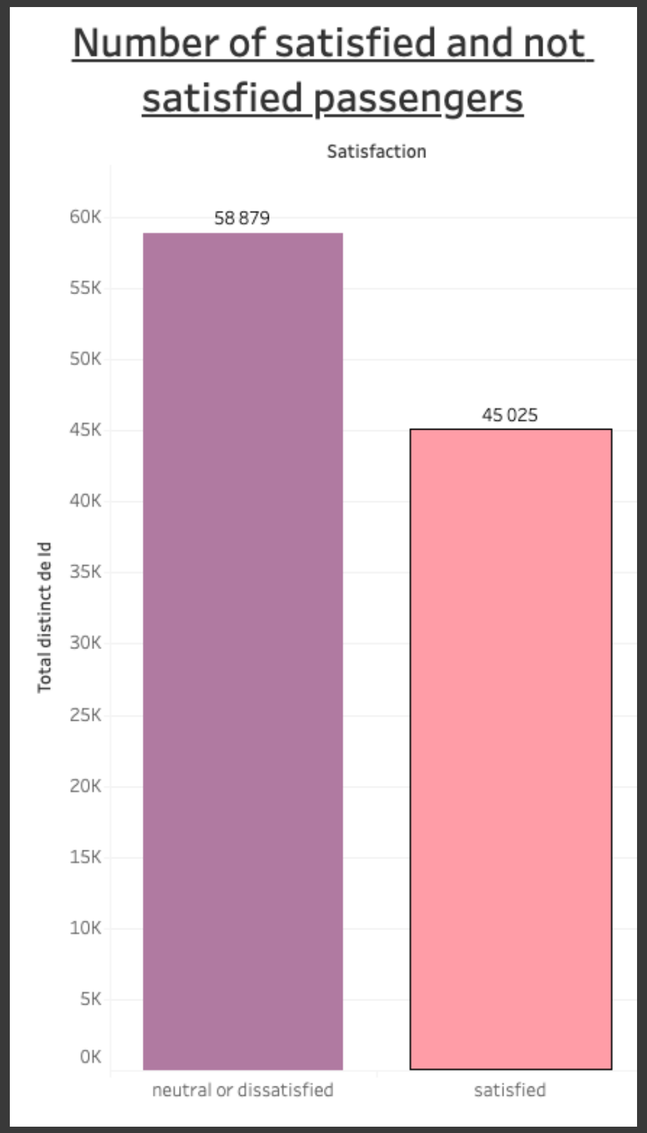
1) Context of the dataset

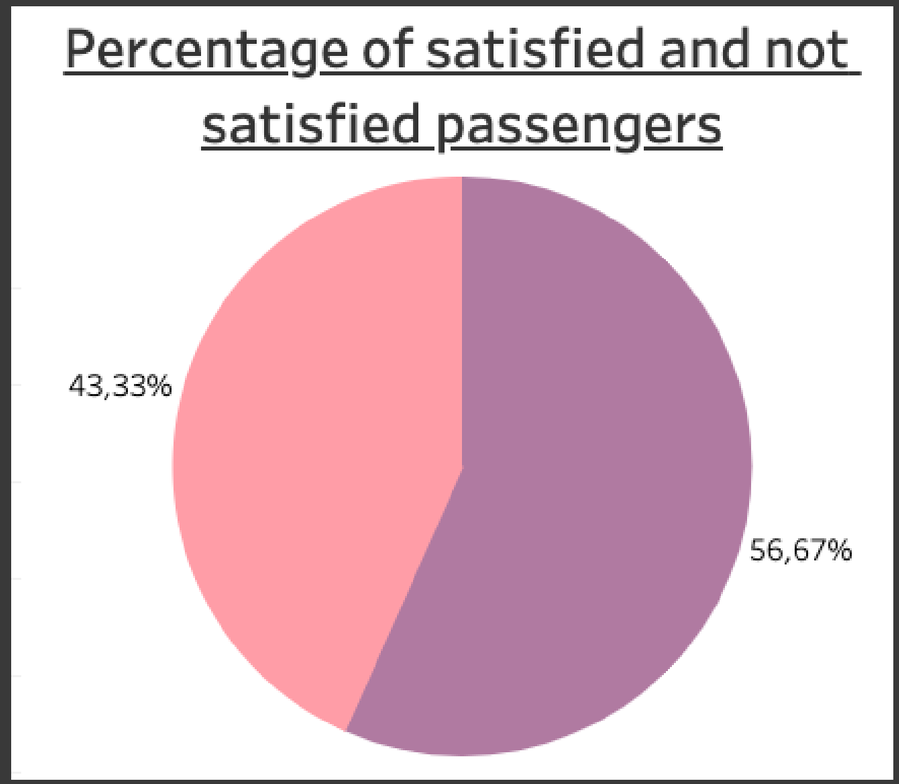
- US Airline 2015 passengers satisfaction with their flight
- The number of rows and columns :
  - before cleaning: 103 904 rows and 25 columns
  - after cleaning: 75 119 rows and 24 columns

2) Cleaning the dataset

- Normalizing the columns' name
- Cleaning rows from 1 column : "arrival_delay_in_minutes "
- Checking for duplicates
- Removing outliers from the following columns :
  - flight_distance
  - departure_delay_in_minutes
  - arrival_delay_in_minutes
- Saving the cleaned dataset to a new CSV file

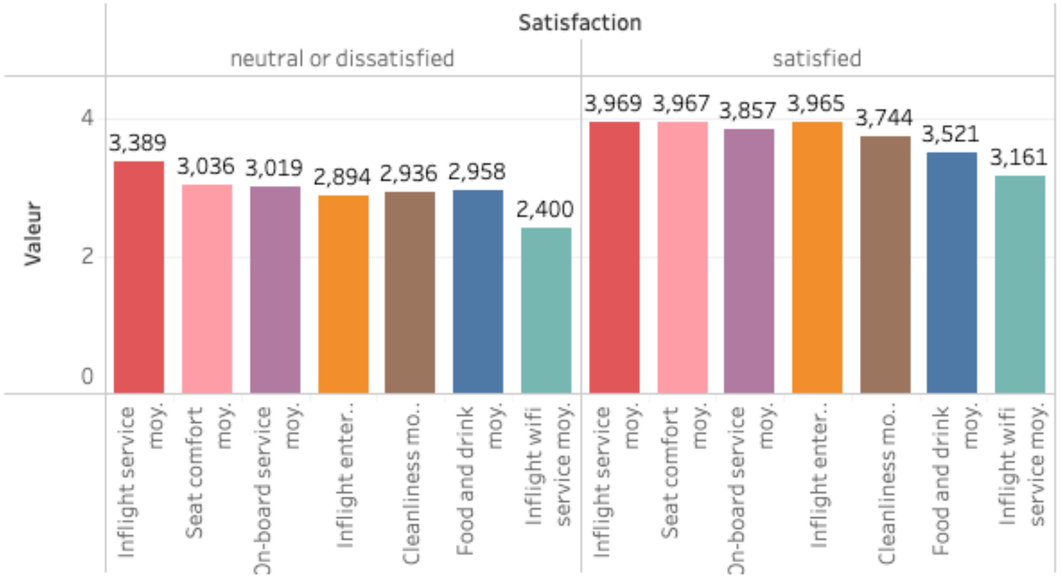
3) Tableau visualizations



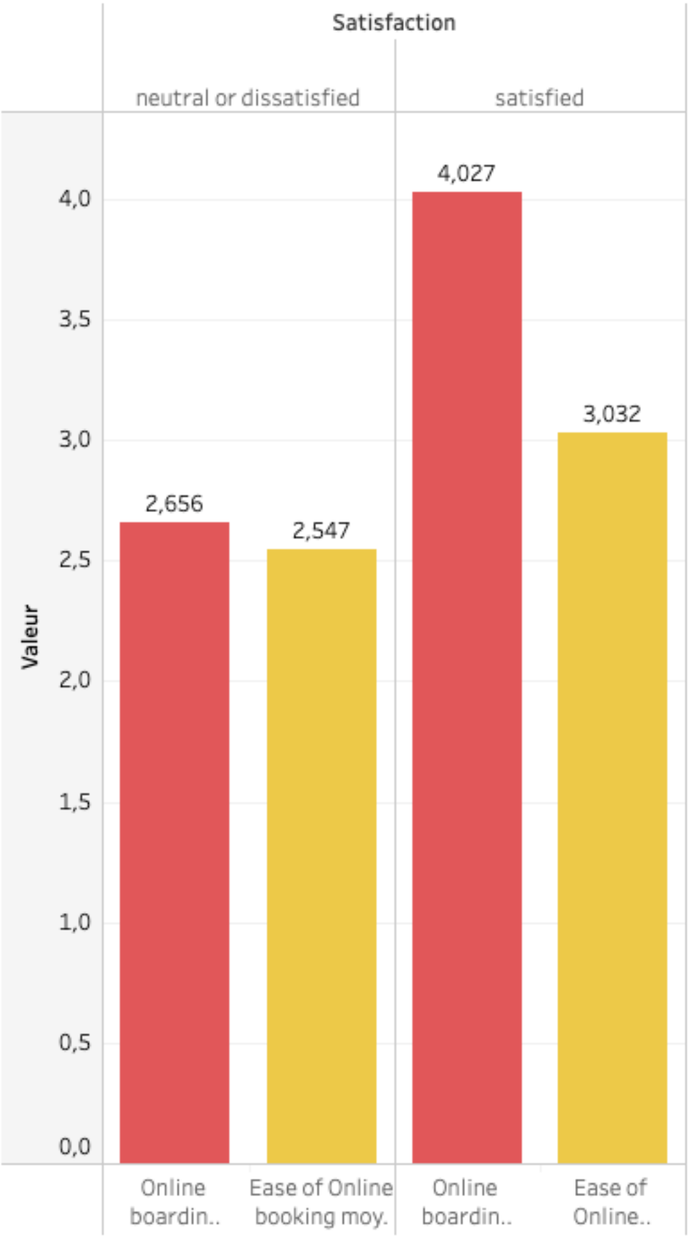


Average score regarding to the satisfaction

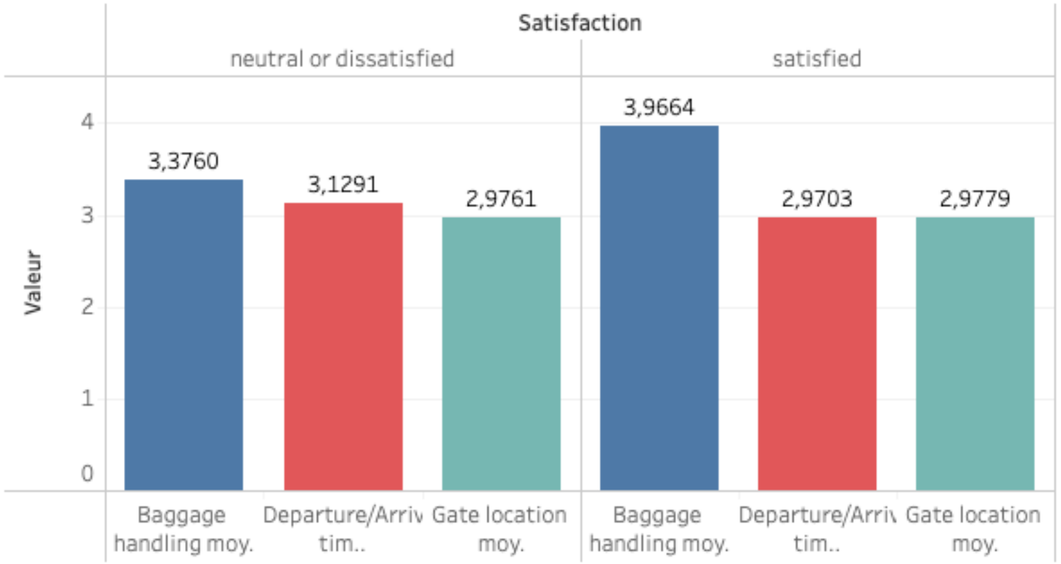
Average scores about plane's services



Average scores about online services

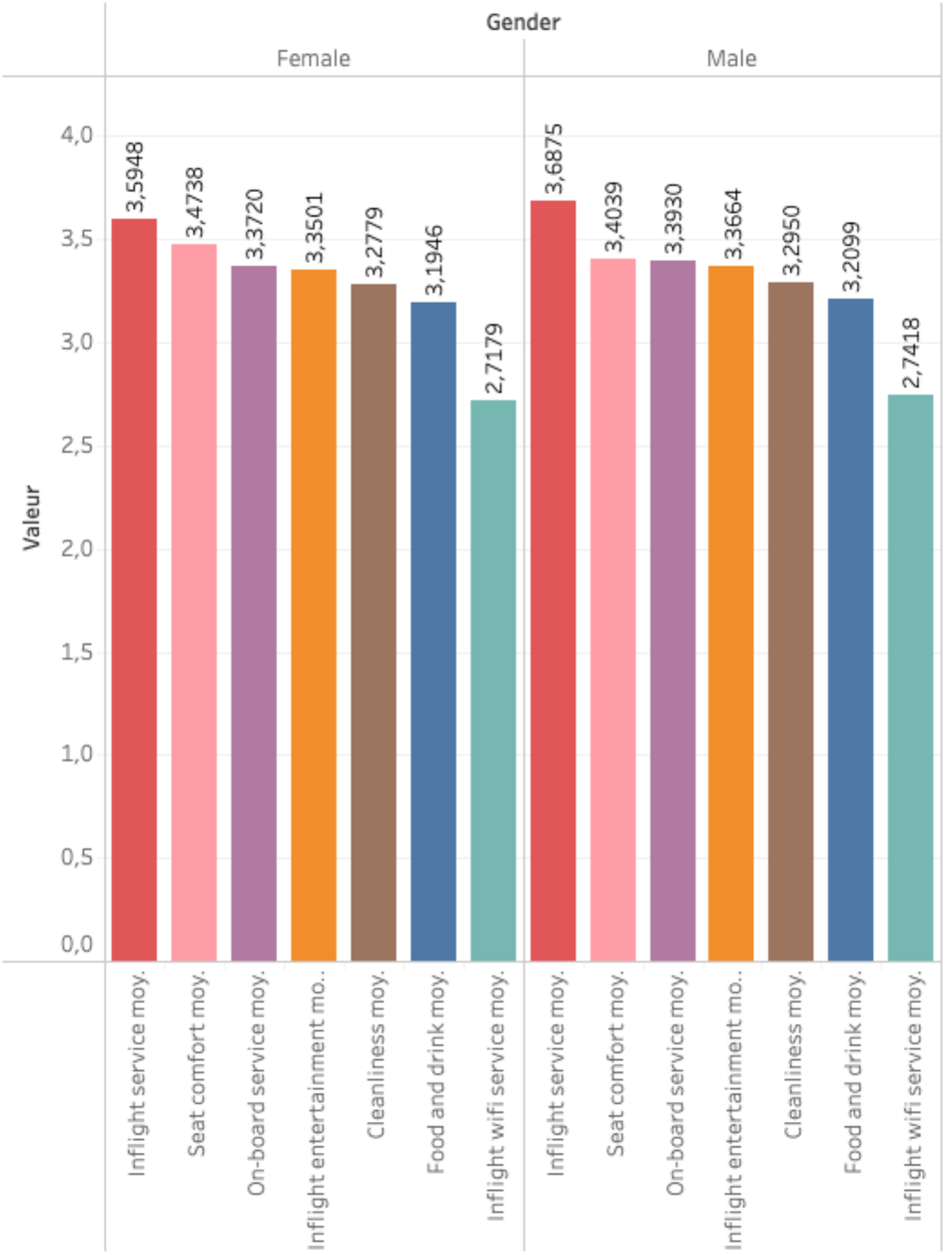


Average scores about airport's services

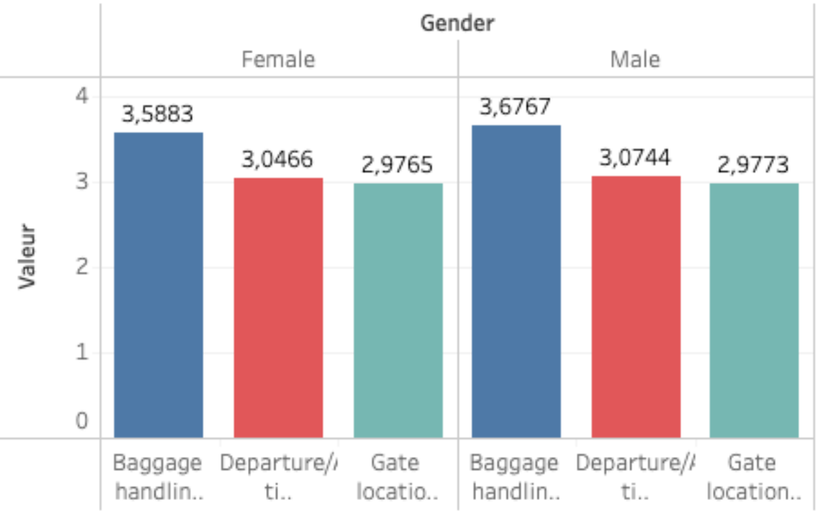


Average score regarding to passenger's gender

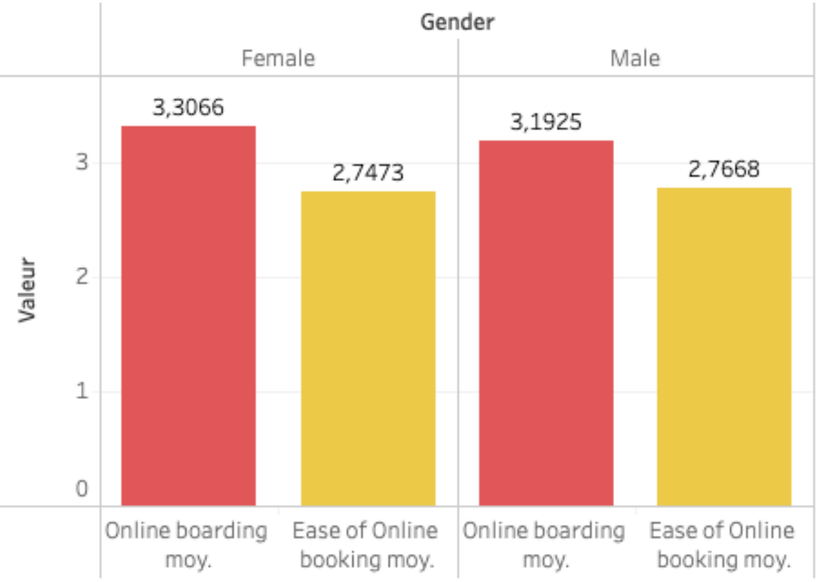
Average scores about plane's services



Average scores about airport's services



Average scores about online services



4) SQL analyse

```
query = '''
        SELECT satisfaction, gender, count(id)
        FROM data
        group by satisfaction, gender
        '''

data = pd.read_sql_query(query, engine)
data.head()
```

	satisfaction	gender	count(id)
0	neutral or dissatisfied	Male	20024
1	satisfied	Female	16960
2	neutral or dissatisfied	Female	21137
3	satisfied	Male	16998

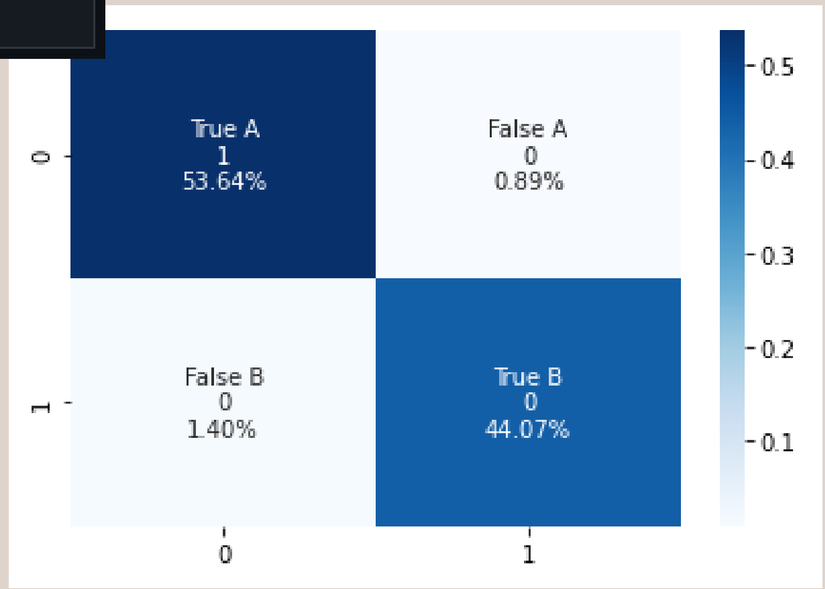
```
query = '''
        SELECT class,gender, avg(l.Total_Score)
        FROM (
        SELECT class, gender, (inflight_wifi_service+"departure/arrival_time_convenient"+ease_of_online_booking+
        FROM data) l
        group by class, gender
        '''

data = pd.read_sql_query(query, engine)
data.head()
```

	class	gender	avg(l.Total_Score)
0	Business	Male	41.730610
1	Business	Female	41.586083
2	Eco	Female	36.676786
3	Eco	Male	36.938678
4	Eco Plus	Female	36.790294
5	Eco Plus	Male	37.154412

The scores are on 70.

5) Results

Model	Score	Results
Logistic Regression	0,97	As we can see, the score is really high. I checked through a matrix, to understand better this score. This matrix shows that 97,7% of the preditions are "True" and 2,3% are "False".
Random Forest	0,97	This is obtained with a max_depth = 1 as a parameter. The main paramters are the following : random_state=0, max_depth=1, max_features='sqrt', min_samples_leaf=5, min_samples_split=5, n_estimators=250
	0,99	This is obtained with a max_depth = 2 as a parameter.
	1	This is obtained with a max_depth = 3 as a parameter. After 3, the score is always 1.



I would like to thank Abhi and his team for this incredible experience full of learning and enjoy !