Title for your project

Measuring customer satisfaction with airports and airplanes of the air transportation industry
Name of first dataset: Airline customer Satisfaction

- **Link:** https://www.kaggle.com/datasets/sjleshrac/airlines-customer-satisfaction

Name of second dataset: Airport Quarterly Passenger Survey

- **Link:** https://catalog.data.gov/dataset/airport-quarterly-passenger-survey

Brief description/context

First dataset context:

Customers have given their feedback in a variety of airline trips. The problem we are having has to do with customer satisfaction, there are many users that have been a long time with the airline and have shown dissatisfaction and likewise those users that are not loyal have also showed dissatisfaction with the service provided. It would be interesting to understand the role some variables play in the satisfaction result for these trips.

Second dataset context:

Airport customers have given feedback on a variety of features of the airport. For each of the features customers provide their level of satisfaction and also give an overall satisfaction. The problem we have is that there are many features in which customers show a dissatisfaction with the value provided by the airport; it would be interesting to understand which variables are the most important to the overall customer satisfaction.

Description of the dataset

The **first dataset** has 129,880 records and has 23 variables that were measured per trip as detailed
below:

VARIABLES			
satisfaction	Gender	Customer Type	Age
Type of Travel	Class	Flight Distance	Seat comfort
Departure/Arrival time convenient	Food and drink	Gate location	Inflight wifi service
Inflight entertainment	Online support	Ease of Online booking	On-board service
Leg room service	Baggage handling	Checkin service	Cleanliness
Online boarding	Departure Delay in Minutes	Arrival Delay in Minutes	

The data was included in the source in 2020 and the information was anonymously given. The name for the airline used for this example is Invistico airlines. The rows represent the individual customer feedback, and the column names are the measures that capture the

customer related information as well as their experience of travelling with the company including how they have rated several airplane features. The columns also mention details about the flight that the customer has taken, as well as features of the particular airplane.

The **second dataset** has 3501 records and has 37 variables that were measured per customer as detailed below:

VARIABLES					
Quarter	Availability of baggage carts	Courtesy of security staff	Walking distance inside terminal	Shopping facilities	Customs inspection
Date recorded	Efficiency of check-in staff	Thoroughness of security inspection	Ease of making connections	Shopping facilities (value for money)	Overall satisfaction
Departure time	Check-in wait time	Wait time of security inspection	Courtesy of airport staff	Comfort of waiting/gate areas	
Ground transportation to/from airport	Courtesy of of check-in staff	Feeling of safety and security	Internet access	Cleanliness of airport terminal	
Parking facilities	Wait time at passport inspection	Restaurants	Business/executive lounges	Ambience of airport	
Parking facilities (value for money)	Ease of finding your way through the airport	Restaurants (value for money)	Availability of washrooms	Arrivals passport and visa inspection	
Courtesy of inspection staff	Flight information screens	Availability of banks/ATM/money changing	Cleanliness of washrooms	Speed of baggage delivery	

The data was first created for use in 2020. The information comes from customer surveys in Austin-Bergstrom airport. The surveys were conducted from 2015-2017 and are reported per quarters. The rows represent individual customer feedback and the columns detail the users overall satisfaction in many variables such as restaurants for example.

Initial idea of what to do with the data

First dataset:

- One business question we would be interested in is if a customer would be satisfied or dissatisfied in their trip taking into account their attributes and the attributes of the flight
- Another business question we would like to address is which type of users do we have. We would like to cluster our users by the type of customer they are and other features such as satisfaction to see if we can fit them into different groups depending on their characteristics and addressing them differently in the future

Second dataset:

- One business question we are interested in is which are the features that influence the most our users satisfaction so we may work on those which are the most relevant to achieving a greater customer satisfaction

Reading and cleaning your datasets

First dataset

Task a

As we can see in the image below, these are the number of missing values per column

```
satisfaction                          0
Gender                                0
Customer Type                         0
Age                                   0
Type of Travel                        0
Class                                 0
Flight Distance                       0
Seat comfort                          0
Departure/Arrival time convenient     0
Food and drink                        0
Gate location                         0
Inflight wifi service                 0
Inflight entertainment                0
Online support                        0
Ease of Online booking                0
On-board service                      0
Leg room service                      0
Baggage handling                      0
Checkin service                       0
Cleanliness                           0
Online boarding                       0
Departure Delay in Minutes            0
Arrival Delay in Minutes            393
dtype: int64
```

Given that the dataset is quite big and 393 ends up only being 0.3% of the total rows we have decided to drop the rows with this missing values.

Task b

The following image shows the count per class for the satisfaction variable

```
satisfied       70882
dissatisfied    58605
```

Task c

```
 #   Column                                Non-Null Count   Dtype
---  ------                                --------------   -----
 0   satisfaction                          129487 non-null  object
 1   Gender                                129487 non-null  object
 2   Customer Type                         129487 non-null  object
 3   Age                                   129487 non-null  int64
 4   Type of Travel                        129487 non-null  object
 5   Class                                 129487 non-null  object
 6   Flight Distance                       129487 non-null  int64
 7   Seat comfort                          129487 non-null  int64
 8   Departure/Arrival time convenient     129487 non-null  int64
 9   Food and drink                        129487 non-null  int64
 10  Gate location                         129487 non-null  int64
 11  Inflight wifi service                 129487 non-null  int64
 12  Inflight entertainment                129487 non-null  int64
 13  Online support                        129487 non-null  int64
 14  Ease of Online booking                129487 non-null  int64
 15  On-board service                      129487 non-null  int64
 16  Leg room service                      129487 non-null  int64
 17  Baggage handling                      129487 non-null  int64
 18  Checkin service                       129487 non-null  int64
 19  Cleanliness                           129487 non-null  int64
 20  Online boarding                       129487 non-null  int64
 21  Departure Delay in Minutes            129487 non-null  int64
 22  Arrival Delay in Minutes              129487 non-null  float64
dtypes: float64(1), int64(17), object(5)
```

After validating the data types, they are all the right type of variable. We have classes for our categorical values and ordinal values for our ratings.

Second dataset

Task a

Below we can see the number of missing values per column we have for the airport customer satisfaction dataset

Quarter	0
Date recorded	0
Departure time	0
Ground transportation to/from airport	54
Parking facilities	39
Parking facilities (value for money)	46
Availability of baggage carts	91
Efficiency of check-in staff	38
Check-in wait time	39
Courtesy of of check-in staff	52
Wait time at passport inspection	69
Courtesy of inspection staff	96
Courtesy of security staff	31
Thoroughness of security inspection	46
Wait time of security inspection	50
Feeling of safety and security	43
Ease of finding your way through the airport	36
Flight information screens	26
Walking distance inside terminal	37
Ease of making connections	83
Courtesy of airport staff	40
Restaurants	59
Restaurants (value for money)	60
Availability of banks/ATM/money changing	41
Shopping facilities	46
Shopping facilities (value for money)	57
Internet access	73
Business/executive lounges	91
Availability of washrooms	35
Cleanliness of washrooms	37
Comfort of waiting/gate areas	41
Cleanliness of airport terminal	32
Ambience of airport	54
Arrivals passport and visa inspection	143
Speed of baggage delivery	181
Customs inspection	201
Overall satisfaction	172

Since we want to use the Overall satisfaction column as our classifier, we will drop the 172 rows which don't have a value since they only make up a 5% of the data.

On the other hand we will impute the rest of the columns with the mode (given that as seen from the histograms, the data is very skewed) to make up for the missing values in each of the columns. We wish to keep them so as to not drop any more information that will come in handy for the classification task.

Task b

The following shows the count per label. It is important to note that users that gave a rating of 4 or 5 are classified as satisfied; all the other have been classified as not satisfied

```
not satisfied    2088
satisfied        1241
```

Task c

```
 #   Column                                    Non-Null Count  Dtype
---  ------                                    --------------  -----
 0   Quarter                                   3329 non-null   object
 1   Date recorded                             3329 non-null   object
 2   Departure time                            3329 non-null   object
 3   Ground transportation to/from airport     3329 non-null   float64
 4   Parking facilities                        3329 non-null   float64
 5   Parking facilities (value for money)      3329 non-null   float64
 6   Availability of baggage carts             3329 non-null   float64
 7   Efficiency of check-in staff              3329 non-null   float64
 8   Check-in wait time                        3329 non-null   float64
 9   Courtesy of of check-in staff             3329 non-null   float64
 10  Wait time at passport inspection          3329 non-null   float64
 11  Courtesy of inspection staff              3329 non-null   float64
 12  Courtesy of security staff                3329 non-null   float64
 13  Thoroughness of security inspection       3329 non-null   float64
 14  Wait time of security inspection          3329 non-null   float64
 15  Feeling of safety and security            3329 non-null   float64
 16  Ease of finding your way through the airport 3329 non-null float64
 17  Flight information screens                3329 non-null   float64
 18  Walking distance inside terminal          3329 non-null   float64
 19  Ease of making connections                3329 non-null   float64
 20  Courtesy of airport staff                 3329 non-null   float64
 21  Restaurants                               3329 non-null   float64
 22  Restaurants (value for money)             3329 non-null   float64
 23  Availability of banks/ATM/money changing  3329 non-null   float64
 24  Shopping facilities                       3329 non-null   float64
 25  Shopping facilities (value for money)     3329 non-null   float64
 26  Internet access                           3329 non-null   float64
 27  Business/executive lounges                3329 non-null   float64
 28  Availability of washrooms                 3329 non-null   float64
 29  Cleanliness of washrooms                  3329 non-null   float64
 30  Comfort of waiting/gate areas             3329 non-null   float64
 31  Cleanliness of airport terminal           3329 non-null   float64
 32  Ambience of airport                       3329 non-null   float64
 33  Arrivals passport and visa inspection     3329 non-null   float64
 34  Speed of baggage delivery                 3329 non-null   float64
 35  Customs inspection                        3329 non-null   float64
 36  Overall satisfaction                      3329 non-null   float64
 37  satisfied                                 3329 non-null   object
```

After observing all the variables we see that they have the correct type needed for our classification

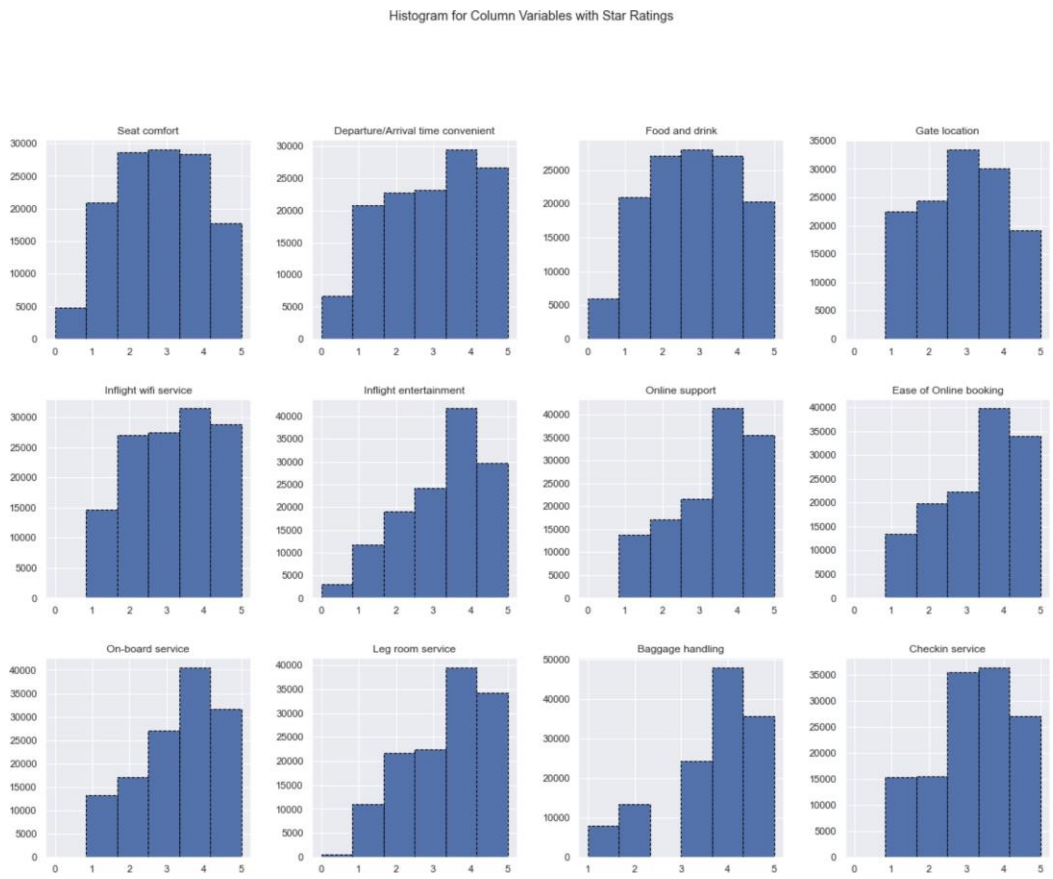
Decision on keeping the datasets separate

We have decided to keep both the datasets separate as the questions answered from each data set can be answered by using the data in each separately. Both have different numbers of rows and columns, which will make it harder for the information to be merged. For data set one, we will predict whether the customer will take another flight with the airline or not as well as clustering the customer into different groups. For the second data set, we will mainly deal with the attributes that can have an impact on customer satisfaction using binary classification. For this reason, merging the datasets is not required and it is better to treat them separately.

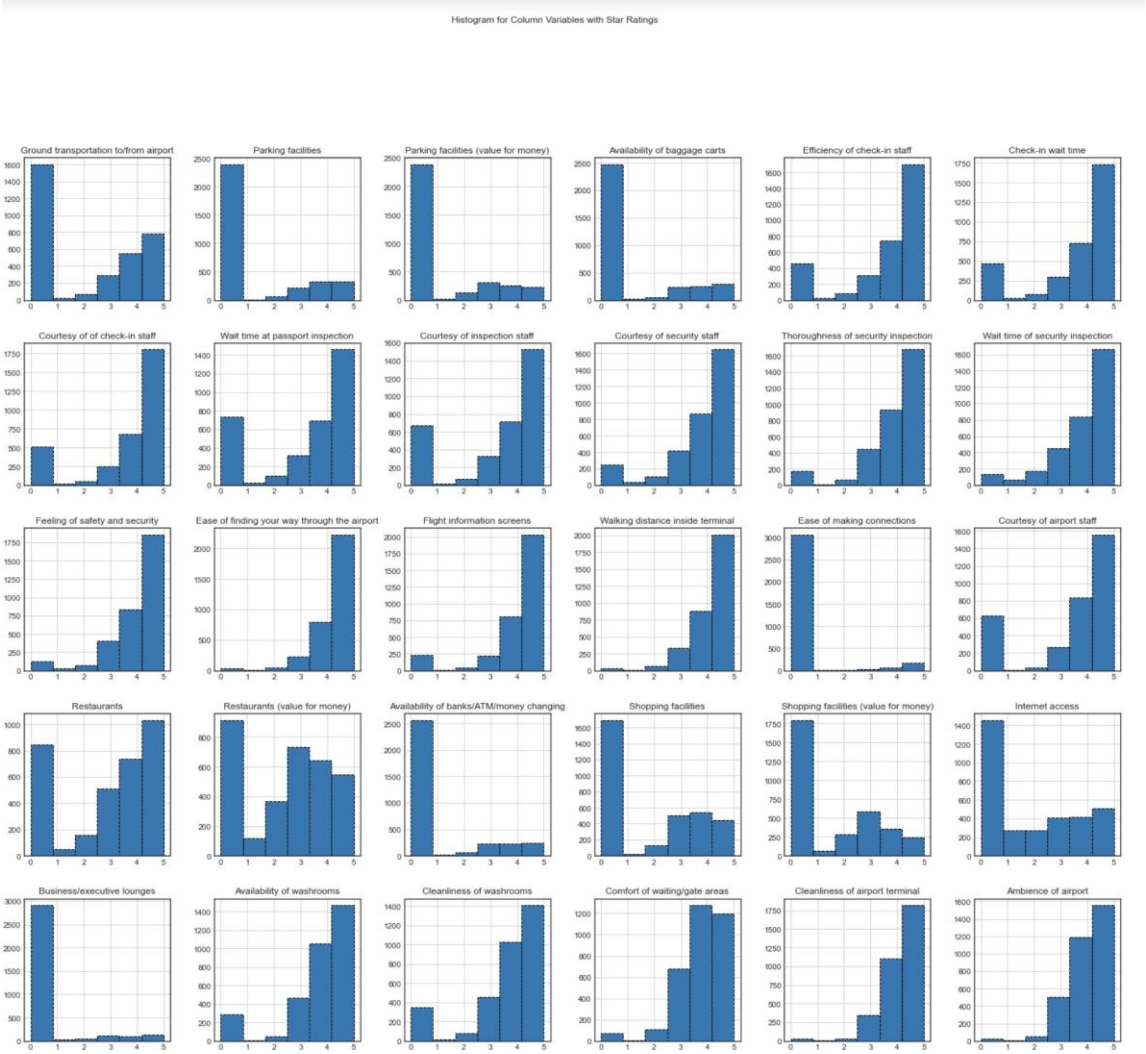
Data Visualization

For both the data sets, we made a histogram and a regplot as that goes with our problem statements. A histogram best explains that for the attributes in both data sets with ratings given from 0 to 5, how much is the count of each star in each attribute. This also gives an idea which attributes were ranked higher in stars than the others and which attributes did poorly. The histogram of these attributes are below:

Data set 1



Data set 2:

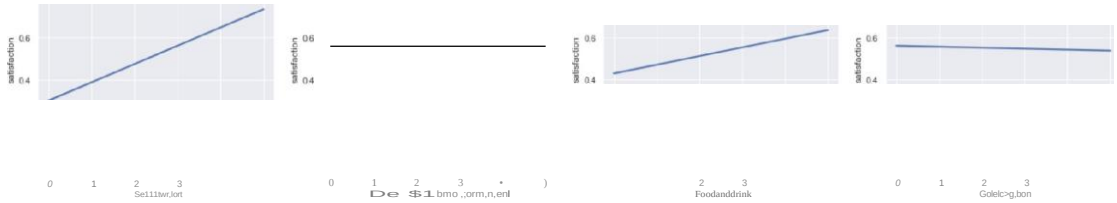


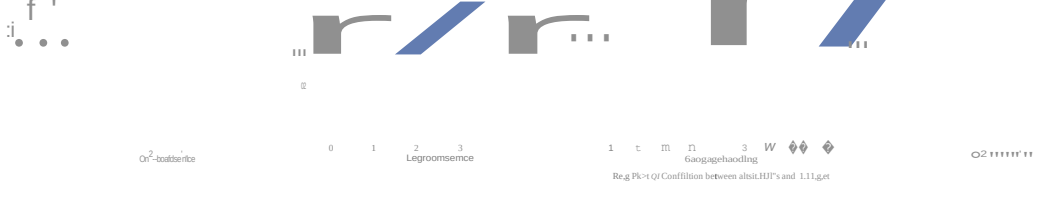
Secondly, we calculated the correlation score of these attributes with satisfaction and plotted a regplot to visualize this information as well as to see what attributes have most impacted the satisfaction score and what type of relationship do they share with the target variable.

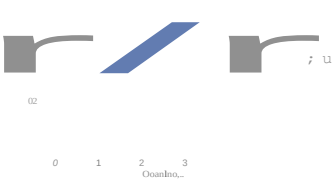
Data set 1:

Seat comfort	0.242371
Departure/Arrival time convenient	-0.015624
Food and drink	0.120568
Gate location	-0.012272
Inflight wifi service	0.227010
Inflight entertainment	0.523364
Online support	0.389890
Ease of Online booking	0.432017
On-board service	0.352283
Leg room service	0.305115
Baggage handling	0.260398
Checkin service	0.266089
Cleanliness	0.259504
Online boarding	0.338118

dtype: float64

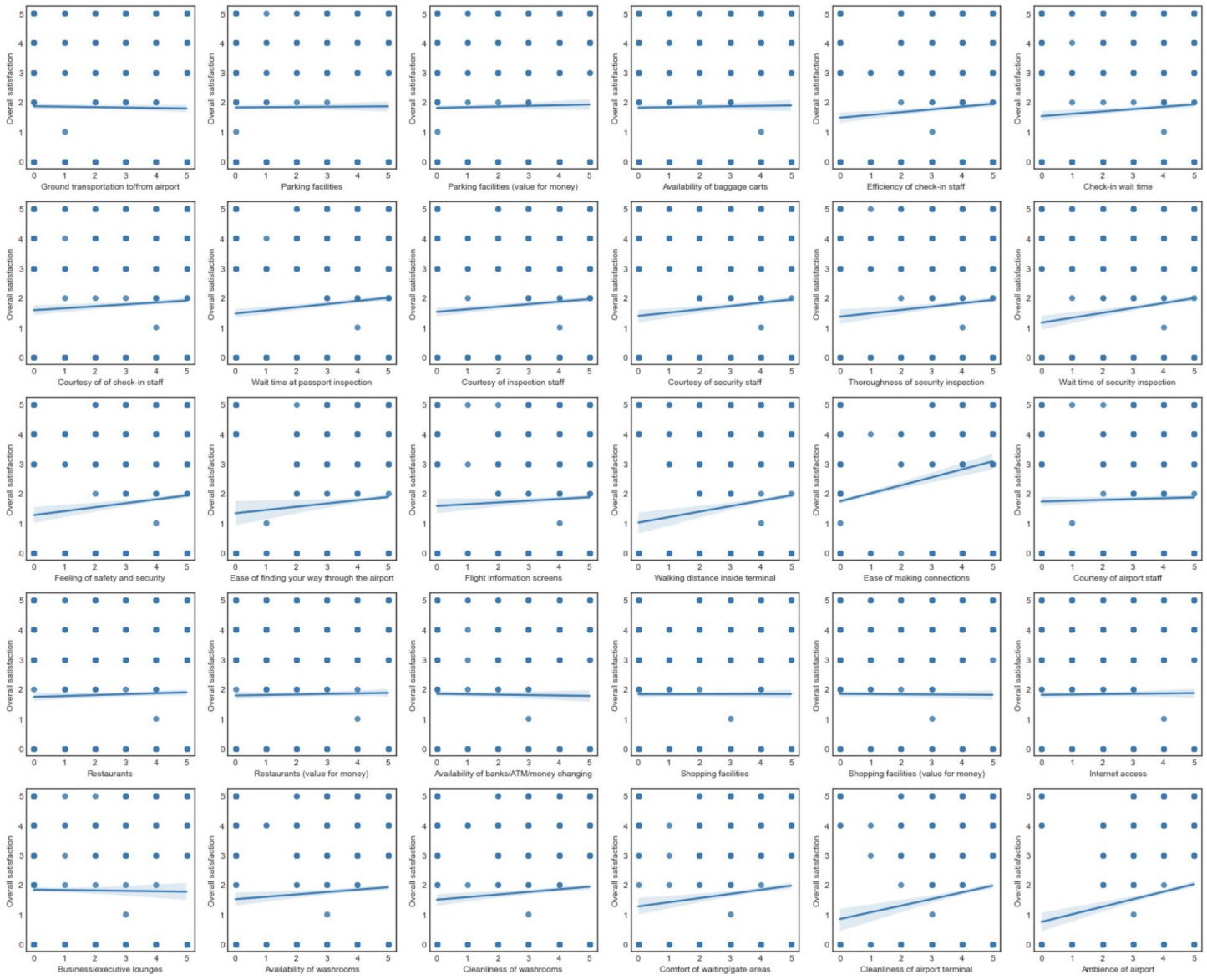






Data set 2:

```
Ground transportation to/from airport          -0.014919
Parking facilities                              0.006306
Parking facilities (value for money)            0.017704
Availability of baggage carts                   0.012070
Efficiency of check-in staff                    0.072653
Check-in wait time                              0.061926
Courtesy of of check-in staff                   0.052286
Wait time at passport inspection                0.094178
Courtesy of inspection staff                    0.074165
Courtesy of security staff                      0.071371
Thoroughness of security inspection             0.064964
Wait time of security inspection                0.096904
Feeling of safety and security                  0.071428
Ease of finding your way through the airport    0.042013
Flight information screens                      0.035525
Walking distance inside terminal                0.074720
Ease of making connections                      0.150044
Courtesy of airport staff                       0.023830
Restaurants                                     0.027231
Restaurants (value for money)                   0.014550
Availability of banks/ATM/money changing       -0.010806
Shopping facilities                             0.001176
Shopping facilities (value for money)          -0.005498
Internet access                                 0.010482
Business/executive lounges                     -0.008013
Availability of washrooms                       0.052074
Cleanliness of washrooms                        0.060536
Comfort of waiting/gate areas                   0.064303
Cleanliness of airport terminal                 0.084277
Ambience of airport                             0.101082
Arrivals passport and visa inspection          -0.877192
Speed of baggage delivery                       0.597898
Customs inspection                             -0.517455
dtype: float64
```



Define your classification or prediction variables

In the first data set, the classification variable is the first column with the name ‘satisfaction’. This column tells whether each customer was satisfied or not with the airline based on the attributes mentioned in the first dataset.

For the second dataset, our classification variable is the column ‘Overall satisfaction’ where we have two labels; satisfied and non satisfied. This column initially had rating values from 0 to 5. As mentioned above, users that gave a rating of 4 or 5 are classified as satisfied; all the others have been classified as not satisfied.