Project Title

COSC 4337

Saim Ali
Robert Duque
Muhaimin Badar

Data Description:

The *Airline Passenger Satisfaction*[1] dataset on Kaggle contains information on airline passenger satisfaction based on an airline satisfaction survey. The dataset consists of 24 attributes, including demographic information about the passengers, such as age, gender, and type of travel, as well as various attributes related to the airline service, such as the inflight entertainment, seat comfort, and onboard service. The target variable is the "satisfaction" attribute, which indicates whether the passenger was satisfied or not satisfied with their overall experience. The dataset contains approximately 129,880 rows, with 24 attributes in each row. The data types for the attributes include float, integer, and object. Some of the attributes have missing values, which will need to be addressed during data cleaning. This dataset is suitable for exploratory data analysis, classification, and regression tasks, and can provide insights for airlines to improve their service and customer satisfaction.

```
Data columns (total 25 columns):
 #   Column                            Non-Null Count   Dtype  
---  ------                            --------------   -----  
 0   Unnamed: 0                        103904 non-null  int64  
 1   id                                103904 non-null  int64  
 2   Gender                            103904 non-null  object 
 3   Customer Type                     103904 non-null  object 
 4   Age                               103904 non-null  int64  
 5   Type of Travel                    103904 non-null  object 
 6   Class                             103904 non-null  object 
 7   Flight Distance                   103904 non-null  int64  
 8   Inflight wifi service             103904 non-null  int64  
 9   Departure/Arrival time convenient 103904 non-null  int64  
 10  Ease of Online booking            103904 non-null  int64  
 11  Gate location                     103904 non-null  int64  
 12  Food and drink                    103904 non-null  int64  
 13  Online boarding                   103904 non-null  int64  
 14  Seat comfort                      103904 non-null  int64  
 15  Inflight entertainment            103904 non-null  int64  
 16  On-board service                  103904 non-null  int64  
 17  Leg room service                  103904 non-null  int64  
 18  Baggage handling                  103904 non-null  int64  
 19  Checkin service                   103904 non-null  int64  
 20  Inflight service                  103904 non-null  int64  
 21  Cleanliness                       103904 non-null  int64  
 22  Departure Delay in Minutes        103904 non-null  int64  
 23  Arrival Delay in Minutes          103594 non-null  float64
 24  satisfaction                      103904 non-null  object 
dtypes: float64(1), int64(19), object(5)
memory usage: 19.8+ MB
None
```

[1] Link to dataset: https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction

Data Cleaning:

The importance of data cleaning is magnified in this given dataset where we attempt to predict customer satisfaction based on various different variables from our data. Data cleaning is a crucial step in data analysis, as incorrect data can lead to unreliable or incorrect results, and can even mislead decision-making. In short, without examining and cleaning our dataset we would be prone to creating prediction models which are unable to perform to the best of their abilities when predicting customer satisfaction on an airline. Data cleaning as a whole in this instance required us to handle missing or null values, correct formatting issues, and remove outliers after examination of the dataset. This step in our data exploratory data analysis was time consuming but nonetheless was an important task that we are sure will pay off further down the line in this project's development.

Removal of irrelevant and duplicate data:

To start our data cleaning process of the Airline Passenger Satisfaction dataset we analyzed the dataset after loading it into our work environment and discussed which of our attributes were needed and which were not useful in our analysis. We deemed a few attributes were irrelevant as the removal of said attributes from our dataset would help simplify the analysis process and improve the accuracy of our later results. Two notable attributes were not useful for our purposes: 'id' and 'Unnamed: 0'. The 'id' attribute was used in the original survey to keep track of different participants' survey ids to differentiate and pull a specific survey if needed. In our case of creating prediction models for the dataset we do not need to pull survey results using the survey id. Since we have no need for the id in our analysis we removed the column 'id' from our dataset using the "df.drop()" function from the pandas library. Secondly, the original dataset we downloaded from Kaggle contained a column with no name which we refer to as 'Unnamed: 0' in our coding environment. As we looked into our dataset we found that this column was being used to track the number or index of the row entry in our dataset of surveys. As we were using our Deepnote environment with the pandas library we found no need for an indexing column as there were already built in functions which

could be used in its place. We decided to remove this column to reduce redundancy and establish one definitive way to refer to the index of surveys in our dataset.

Fix structural errors:

Our dataset had few “structural” errors as we came across no typos or incorrectly named attributes. We did however optimize our data’s attribute names in a way we could access and optimize our code in a more efficient manner. Using a function we coded we were able to turn attribute names from “Flight Distance” to “flightDistance”, to satisfy the camel-case variable name standard widely used in the industry. This overall would help us reference certain columns in our dataset more easily since everything will be following a uniform rule for naming of variables. We also found this dataset to be relatively big with over 100,000 entries for our environment to filter through. Our runtimes were high with even minimal tasks so we looked into reducing the memory usage of our dataset. One way to do so is by changing the original int64 datatype for variables to a less byte intensive int32 to reduce memory cost in our environment. We were able to do so since most of the numbers in our dataset were not fully utilizing the intensive int64 datatype and only really required an int32 to hold all of the numerical information we needed. When calculating the memory cost from before to after our datatype transformation we found a decreased memory usage of about 70% which was huge. As we were hosting our co-op python environment on Deepnote this was a huge reduction in runtime for just about every function in our dataset.

```
Before data transform 15.48
After data transform 3.74
The above transform has decreased memory usage by 75.84%
```

Handle missing data:

Null values in our datasets can occur for a variety of reasons. One of the most common reasons for many datasets containing null values is data entry errors. For example, when data is being entered manually for each person’s survey, the person entering the data may accidentally leave a field blank, resulting in a null value. Using “print(trainData.isnull().sum())” we were able to display any null values in our dataset added up based on their variable/columns name. We found that the ‘Arrival Delay in

Minutes' attribute was the only one in our dataset with null values and had 310 null values. We found that when participants were filling in their individual surveys, in the instance their flight had no arrival delays, they filled in a 'none' or 'none' value for their arrival delay input. Thus, in our dataset those instances show up as null values in the 'Arrival Delay in Minutes'. To tackle this issue we set null values for this attribute = 0 (int) as the arrival delay in minutes in theory would be 0 if it is null regarding this situation.

Conversion of categorical variables:

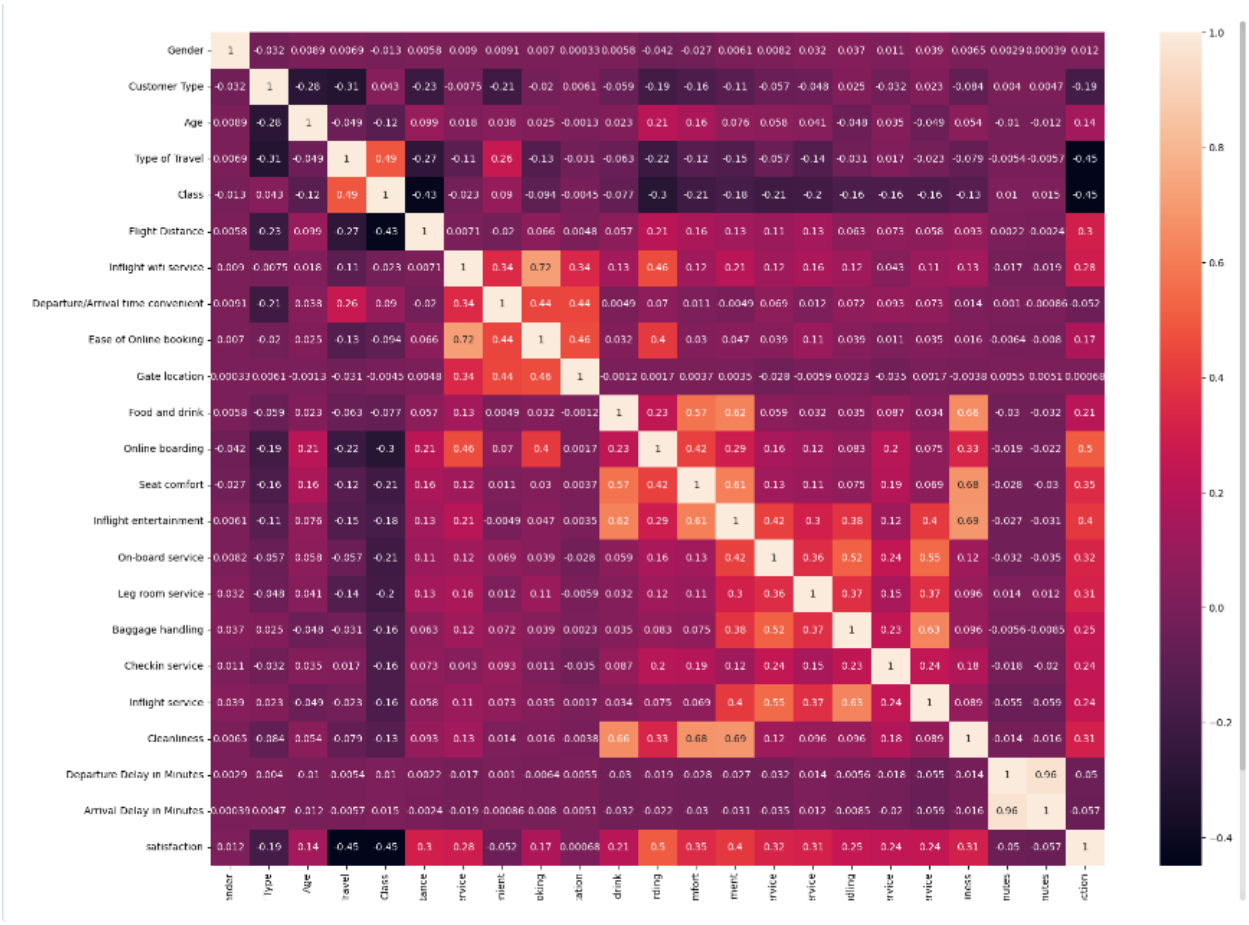
For our dataset there were various categorical variables which help in classification of the surveyor. We had to convert the following five variables: Satisfaction- determines if person is satisfied or not satisfied, Gender- Gender of the passengers (Female, Male), Customer Type- The customer type (Loyal customer, disloyal customer), Class:Travel class in the plane of the passengers (Business, Eco, Eco Plus), Type of Travel-Purpose of the flight of the passengers (Personal Travel, Business Travel). We converted all of these categorical variables into integers depending on the number of outcomes possible for a given variable. (Ex. Satisfaction = 0 - not satisfied; 1 - satisfied). In the end after converting all of these variables, we were able to make our data usable for functions we will create in the future as well as prediction models.

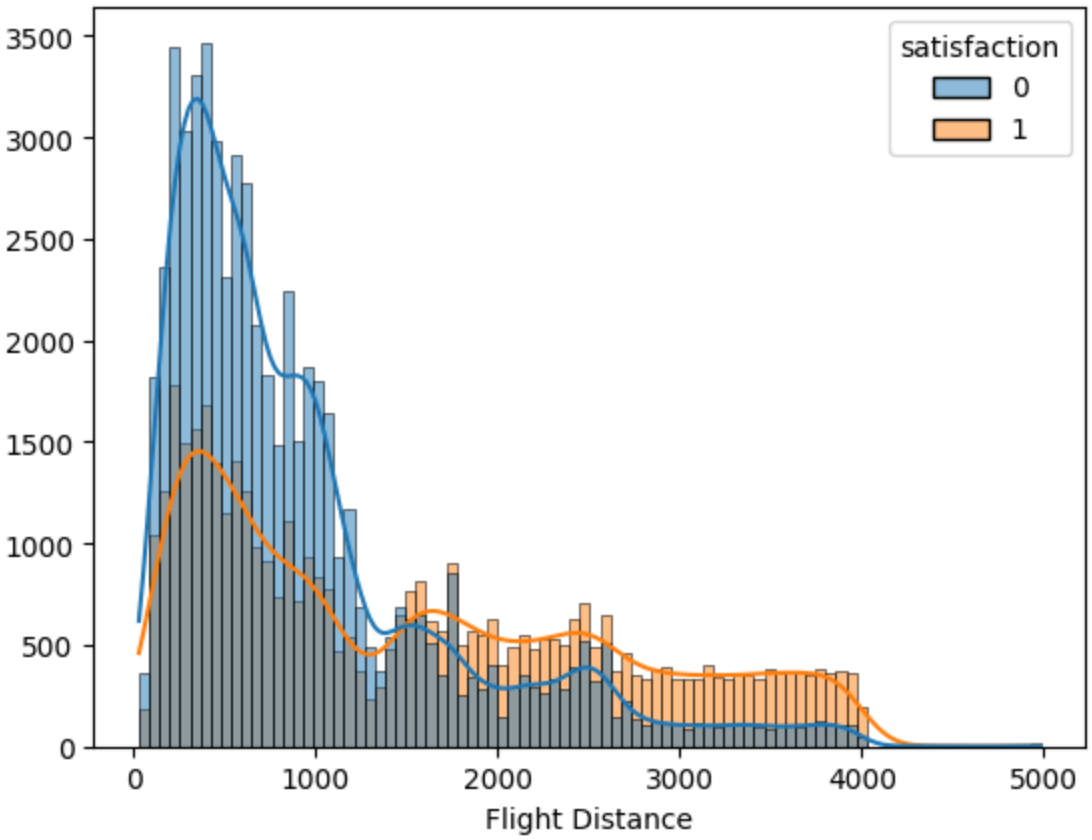
Data Analysis:

Data analysis is important for gaining insights and making informed decisions from a dataset. We must understand the structure and patterns of our data, identify problems and inconsistencies, and gain valuable insight to make informed decisions. Data analysis is also essential in detecting errors and inconsistencies in datasets, which can lead to incorrect conclusions if not properly addressed. The following are our findings regarding most attributes and their patterns in out dataset:

- There is almost an even number of males and females who took the survey
- The average customer age of participants is 39 years
- The flight distance varied greatly in this dataset with the mean being 1189 and the std benign 997

- Most categorical variables that were taken from participants in the 1-5 range had the average score of 3/5
- From looking at the initial correlation matrix we found that: **Best features for prediction of satisfaction** - Online Booking, Class, and Type of Travel; **Worst features for prediction of satisfaction** - Gate location, Gender, and Departure/Arrival Time Convenient


Gender
Customer Type
Age
Type of Travel
Class
Flight Distance
Inflight wifi service
Departure/Arrival time convenient
Ease of Online booking
Gate location
Food and drink
Online boarding
Seat comfort
Inflight entertainment
On-board service
Leg room service
Baggage handling
Checkin service
Inflight service
Cleanliness
Departure Delay in Minutes
Arrival Delay in Minutes
satisfaction
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4



Feature Selection:

To effectively pick and choose attributes that have a larger impact on the overall satisfaction rating, we need to take the data that we have and feed it into a method of feature selection. We chose 2 different ways of selecting prominent features, Mutual information regression and decision tree. Mutual information regression takes into consideration the joint and marginal probabilities of each attribute when compared to the satisfaction rating. For each attribute in the dataset, we will take all of the inputs and divide them into x rows (x being the amount of unique values in the dataset for that particular attribute) and y columns (amount of unique values in the dataset for satisfaction). Think of this table as a similar representation of a confusion matrix but with probabilities rather than predicted and actual values. All of the values in these boxes will represent the joint probability of the respective attribute value and satisfaction value. The final mutual information value will be the summation of the joint probabilities times

the log of the joint probabilities divided by the product of each marginal probability. The formula below shows a clear representation of the calculations.

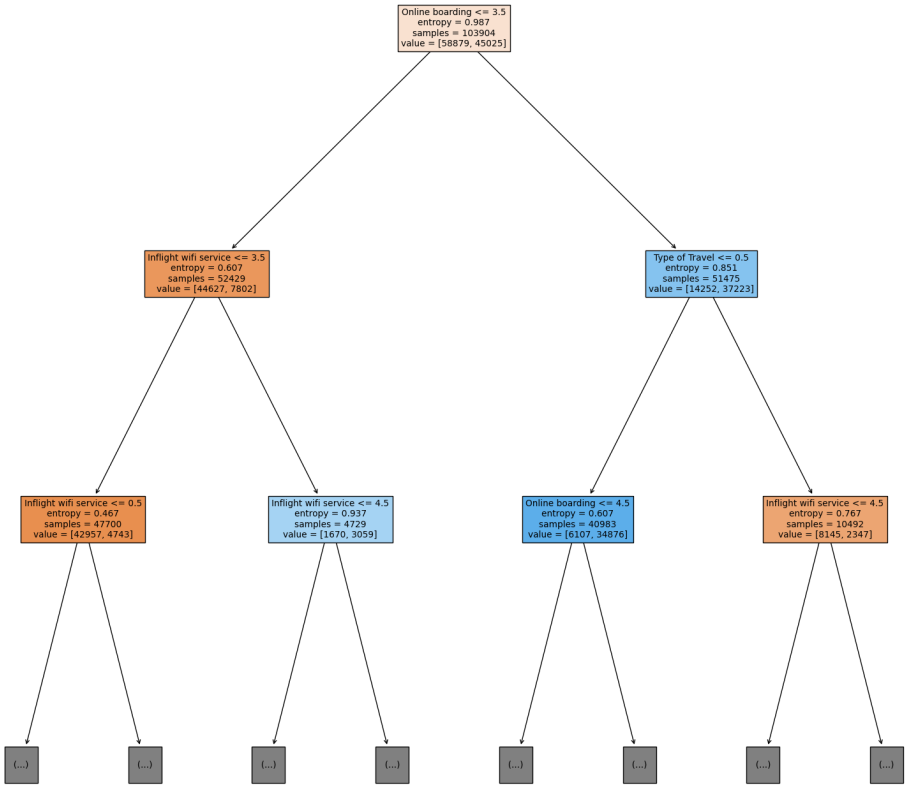
$$\sum_{j=1}^{m}\sum_{i=1}^{n} p_{YX}(y_j,x_i)\log\left(\frac{p_{YX}(y_j,x_i)}{p_Y(y_j)p_X(x_i)}\right)$$

Passing the dataset into this function results in the following dictionary that we sorted to show the highest relevant features first and vice versa for the last, the higher the mutual information value the more important this feature is for the satisfaction variable.

To actually select the features we opted for the SelectKBest method, which will take the K best features with the highest scores. For now we have chosen k=18 because there is a seemingly significant cutoff at the 18th feature's score, however we will using a tuning method such as GridSearch to select the best K and perhaps even a different method like VarianceThreshold or Recursive Feature Elimination.

The scores of the features can be found on the figure below.

```
{'Online boarding': 0.20982,
 'Inflight wifi service': 0.163689,
 'Class': 0.129764,
 'Type of Travel': 0.113483,
 'Inflight entertainment': 0.091286,
 'Seat comfort': 0.079906,
 'Leg room service': 0.062321,
 'Flight Distance': 0.061555,
 'On-board service': 0.055799,
 'Cleanliness': 0.049448,
 'Ease of Online booking': 0.048965,
 'Age': 0.048538,
 'Inflight service': 0.042985,
 'Baggage handling': 0.041984,
 'Checkin service': 0.031431,
 'Food and drink': 0.026445,
 'Customer Type': 0.020923,
 'Gate location': 0.013355,
 'Arrival Delay in Minutes': 0.004131,
 'Gender': 0.0,
 'Departure/Arrival time convenient': 0.0,
 'Departure Delay in Minutes': 0.0}
```



Conclusion:

In conclusion, the Airline Passenger Satisfaction dataset on Kaggle is a valuable resource for exploring airline passenger satisfaction and providing insights for airlines to improve their service and customer satisfaction. The dataset required extensive data cleaning to ensure accurate and reliable results, including the removal of irrelevant and duplicate data, fixing structural errors, and handling missing data. The steps followed in cleaning and reducing the dimensionality of the data improved the accuracy of the results and reduced the memory cost in the environment. This dataset is suitable for exploratory data analysis, classification, and regression tasks, and can provide valuable insights for airlines to enhance their customers' experience. We can now continue to utilize this dataset in our further studies in this group project in models to come.