Predicting Airline Customer Satisfaction

Group 1

Changa Fernando s3842381

Athan Katanos s3777056

Submission Mode: **Regular**

We hereby declare that this project report submission is an original research contribution.

Table of Contents

Introduction

Dataset Source

The dataset used in this project relates to data obtained from a US Airline passenger satisfaction survey (Klein, 2020). There are over one hundred thousand completed surveys in this dataset, from a wide range of people with a plethora of different airline experiences. This data set was sourced from kaggle.com and can be found at the following link: www.kaggle.com/teejmahal20/airline-passenger-satisfaction (Klein, 2020). The specific csv file being used is the train.csv file.

Dataset Details

The original dataset Airline Passenger Satisfaction (Klein, 2020) contains 103,903 rows and 25 columns. Each row contains information about each passenger's ratings of satisfaction as well as their personal and flight details. These entries were obtained in the form of a survey. Examples of ratings of satisfaction include inflight wifi service, food and drink, seat comfort. Examples of their personal and flight details include gender, age, class and flight distance. The ratings of satisfaction are spread across different degrees of service provided by the airline. As a result, the initial survey required the passengers to rate each degree of service into an ordinal scale from 0, being unreceived or poor, to 5, being well received or excellent.

Assumptions

The following assumptions have been made in order to maximise contextual understanding and mathematical computing to its fullest in order to help literary conclusions further on in this report.

- flight distance has been recorded in miles (given dataset is based in the United States)

Dataset Features

Below is a table outlining the features of the original dataset (train.csv).

#	NAME OF FEATURE	DATA TYPE	UNITS	BRIEF DESCRIPTION
1	NA	Numerical: Continuous	NA	Original dataset's first column contained row numbers, excluding the first row as white space and the second row starting as 0.
2	id	Numerical: Continuous	NA	Identification number of the completed survey or the passenger.
3	Gender	Categorical: Binary	NA	Identified gender of the passenger.
4	Customer Type	Categorical: Binary	NA	Loyalty of the passenger - either a loyal customer or disloyal customer.
5	Age	Numerical: Continuous	Years	Age of the passenger at the time of completing the survey.
6	Type of Travel	Categorical: Binary	NA	Purpose of the passenger's reason for flying - either personal travel or business travel.
7	Class	Categorical: Nominal	NA	Chosen flight class the passenger has booked - either Eco, Eco Plus or Business.
8	Flight Distance	Numerical: Continuous	Miles	Flight distance travelled in journey.
9	Inflight wifi service	Categorical: Nominal	0 to 5	Passenger's level of satisfaction with their flight's inflight wifi service - 0 being worst, 5 being best.
10	Departure/Arrival time convenient	Categorical: Nominal	0 to 5	Passenger's level of satisfaction with their flight's departure/arrival time convenience - 0 being worst, 5 being best.
11	Ease of Online booking	Categorical: Nominal	0 to 5	Passenger's level of satisfaction with their flight's ease of online booking - 0 being worst, 5 being best.
12	Gate location	Categorical: Nominal	0 to 5	Passenger's level of satisfaction with their flight's gate location - 0 being worst, 5 being best.
13	Food and drink	Categorical: Nominal	0 to 5	Passenger's level of satisfaction with their flight's food and drink quality - 0 being worst, 5 being best.
14	Online Boarding	Categorical: Nominal	0 to 5	Passenger's level of satisfaction with their flight's online boarding - 0 being worst, 5 being best.
15	Seat comfort	Categorical: Nominal	0 to 5	Passenger's level of satisfaction with their flight's seat comfort - 0 being worst, 5 being best.
16	Inflight entertainment	Categorical: Nominal	0 to 5	Passenger's level of satisfaction with their flight's inflight entertainment - 0 being worst, 5 being best.
17	On-board service	Categorical: Nominal	0 to 5	Passenger's level of satisfaction with their flight's on-board service - 0 being worst, 5 being best.
18	Leg room service	Categorical: Nominal	0 to 5	Passenger's level of satisfaction with their flight's leg room service - 0 being worst, 5 being best.

#	NAME OF FEATURE	DATA TYPE	UNITS	BRIEF DESCRIPTION
19	Baggage handling	Categorical: Nominal	0 to 5	Passenger's level of satisfaction with their flight's baggage handling - 0 being worst, 5 being best.
20	Checkin service	Categorical: Nominal	0 to 5	Passenger's level of satisfaction with their flight's checkin service - 0 being worst, 5 being best.
21	Inflight service	Categorical: Nominal	0 to 5	Passenger's level of satisfaction with their flight's inflight service - 0 being worst, 5 being best.
22	Cleanliness	Categorical: Nominal	0 to 5	Passenger's level of satisfaction with their flight's cleanliness - 0 being worst, 5 being best.
23	Departure Delay in Minutes	Numerical: Continuous	Minutes	Amount of time delayed since departure.
24	Arrival Delay in Minutes	Numerical: Continuous	Minutes	Amount of time delayed since arrival.
25	Satisfaction	Categorical: Binary	NA	Passenger's overall satisfaction of their flight and it's overall services - satisfied or neutral/dissatisfied.

Target Feature

The target feature of this report is the satisfaction of customers (last column). This feature is a binary feature with the two values being "satisfied" and "neutral or dissatisfied". Using all of the data available in this dataset, machine learning algorithms will be used to try and determine if a customer will be "satisfied" or "neutral or dissatisfied" with their experience. Some predictions can be made such as: people who fly business class are more likely to be "satisfied" with their experience compared to those who don't, whereas, people with an inflight Wi-Fi service of less than 2 will rate the experience as "neutral or dissatisfied".

Goals and Objectives

Using the largely descriptive and attribute-rich dataset of the satisfaction of airline passengers, we intend to draw conclusions and visualise relationships between multiple variables to extract deeper meanings.

Below are a list of objectives we hope to achieve:

- Visualise singular, pair and multiple variable relationships using Python and the Scikit-Learn module.
- Extract meaningful conclusions from the extended analysis of numerical and categorical ratings of the airline passengers' satisfaction levels.
- Illicit areas of improvement for the airline from the analysis of multi-relationship variables to better understand their passengers and improve their satisfaction ratings.
- Convert numerical conclusions to literary meanings.

Data Cleaning and Preprocessing

In [1]:

```python
import warnings
warnings.filterwarnings('ignore')
import pandas as pd
import numpy as np
data = pd.read_csv('Phase1_Group1.csv')
data.head()
```

Out[1]:

	Unnamed: 0	id	Gender	Customer Type	Age	Type of Travel	Class	Flight Distance	Inflight wifi service	Departure/Arrival time convenient	...	ent
0	0	70172	Male	Loyal Customer	13	Personal Travel	Eco Plus	460	3	4	...	
1	1	5047	Male	disloyal Customer	25	Business travel	Business	235	3	2	...	
2	2	110028	Female	Loyal Customer	26	Business travel	Business	1142	2	2	...	

	Unnamed: 0	id	Gender	Customer Type	Age	Type of Travel	Class	Flight Distance	Inflight wifi service	Departure/Arrival time convenient	...	ent
3	3	24026	Female	Loyal Customer	25	Business travel	Business	562	2	5	...	
4	4	119299	Male	Loyal Customer	61	Business travel	Business	214	3	3	...	

5 rows × 25 columns

From the table above the first two columns ('Unamed' and 'id') need to be removed as they are not descriptive features.

In [2]:

```
data.drop(['Unnamed: 0', 'id'], axis = 'columns', inplace = True)
data.head()
```

Out[2]:

	Gender	Customer Type	Age	Type of Travel	Class	Flight Distance	Inflight wifi service	Departure/Arrival time convenient	Ease of Online booking	Gate location	...	ente
0	Male	Loyal Customer	13	Personal Travel	Eco Plus	460	3	4	3	1	...	
1	Male	disloyal Customer	25	Business travel	Business	235	3	2	3	3	...	
2	Female	Loyal Customer	26	Business travel	Business	1142	2	2	2	2	...	
3	Female	Loyal Customer	25	Business travel	Business	562	2	5	5	5	...	
4	Male	Loyal Customer	61	Business travel	Business	214	3	3	3	3	...	

5 rows × 23 columns

Now that the data set has been cleaned, missing values are checked for.

In [3]:

```
data.isna().sum()
```

Out[3]:

```
Gender                                 0
Customer Type                          0
Age                                    0
Type of Travel                         0
Class                                  0
Flight Distance                        0
Inflight wifi service                  0
Departure/Arrival time convenient      0
Ease of Online booking                 0
Gate location                          0
Food and drink                         0
Online boarding                        0
Seat comfort                           0
Inflight entertainment                 0
On-board service                       0
Leg room service                       0
Baggage handling                       0
Checkin service                        0
Inflight service                       0
Cleanliness                            0
Departure Delay in Minutes             0
Arrival Delay in Minutes             310
satisfaction                           0
dtype: int64
```

'Arrival Delay in Minutes' column has 310 missing values. The rows containing these columns will be deleted from the dataset.

In [4]:

```
data_no_na = data.dropna()
data_no_na.isna().sum()
```

Out[4]:

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

Now that the dataset has been cleaned, data visualisation can be done.

Data Exploration and Visualisation

One Variable Plots

In [5]:

```
import random
import matplotlib.pyplot as plt
%matplotlib inline
%config InlineBackend.figure_format = 'retina'
import seaborn as sns
sns.set()
```

In [6]:

```
sns.set(rc = {'figure.figsize':(15,8)})
sns.set_palette("gist_ncar")
Fig1 = sns.countplot(x = 'satisfaction', data = data_no_na).set(title ='Histogram plot of S
```

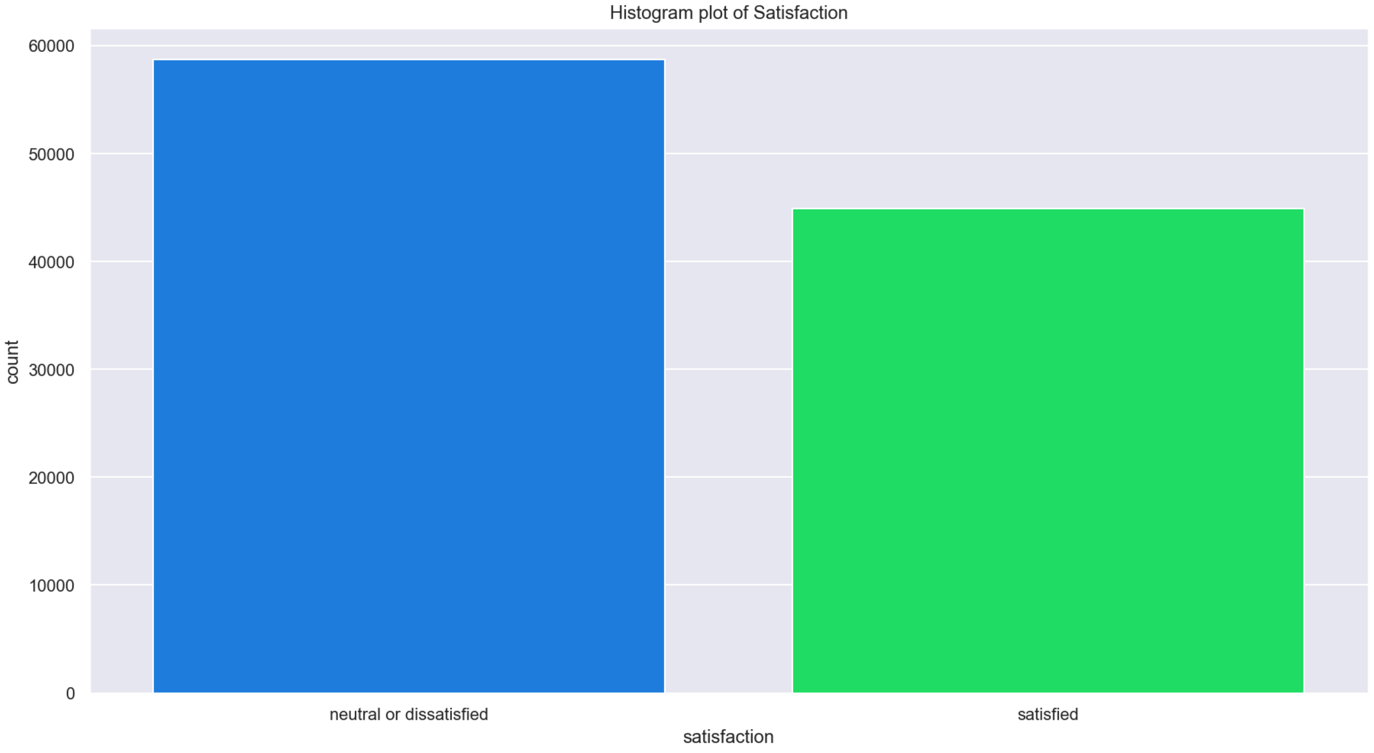

Figure 1. Histogram plot of Satisfaction

In the above, the histogram count plot of the singular variable *Satisfaction* has been plotted. The y-axis is the count of people, ranging from 0 to 60,000. The x-axis is the categories of *Satisfaction* being compared. On the left of the plot, we can see the selection of 'neutral or dissatisfied' as blue with 'satisfied' as green on the right.

As stated in the **Dataset Details** section, *Satisfaction* is a binary categorical variable. It has two categories, 'neutral and dissatisfied' and 'satisfied'. The number of passengers that identified with either were tallied as two individual bars on the histogram count plot as seen above.

From the plot, we can identify that the larger proportion of passengers in the survey felt 'neutral or dissatisfied' in regards to their overall satisfaction with their flight experience. To be specific, close to 59,000 passengers rated their flight experience as 'neutral or dissatisfied' and close to 45,000 as 'satisfied'. The reasoning behind the larger consensus of feeling 'neutral or dissatisfied' can be explained further by investigating each factor of the flight experience in Phase 2. For now, we can deduce that the flight experience will have many surveys critiquing and rating many parts of their flight experience as less than 3 stars - assuming that each passenger does not have high standards and the *Satisfaction* rating is proportional to the additive factors of their flight experience.

In [7]:

```
Fig2 = sns.distplot(data_no_na['Flight Distance'], kde = True, bins = 10, color='red').set(
```

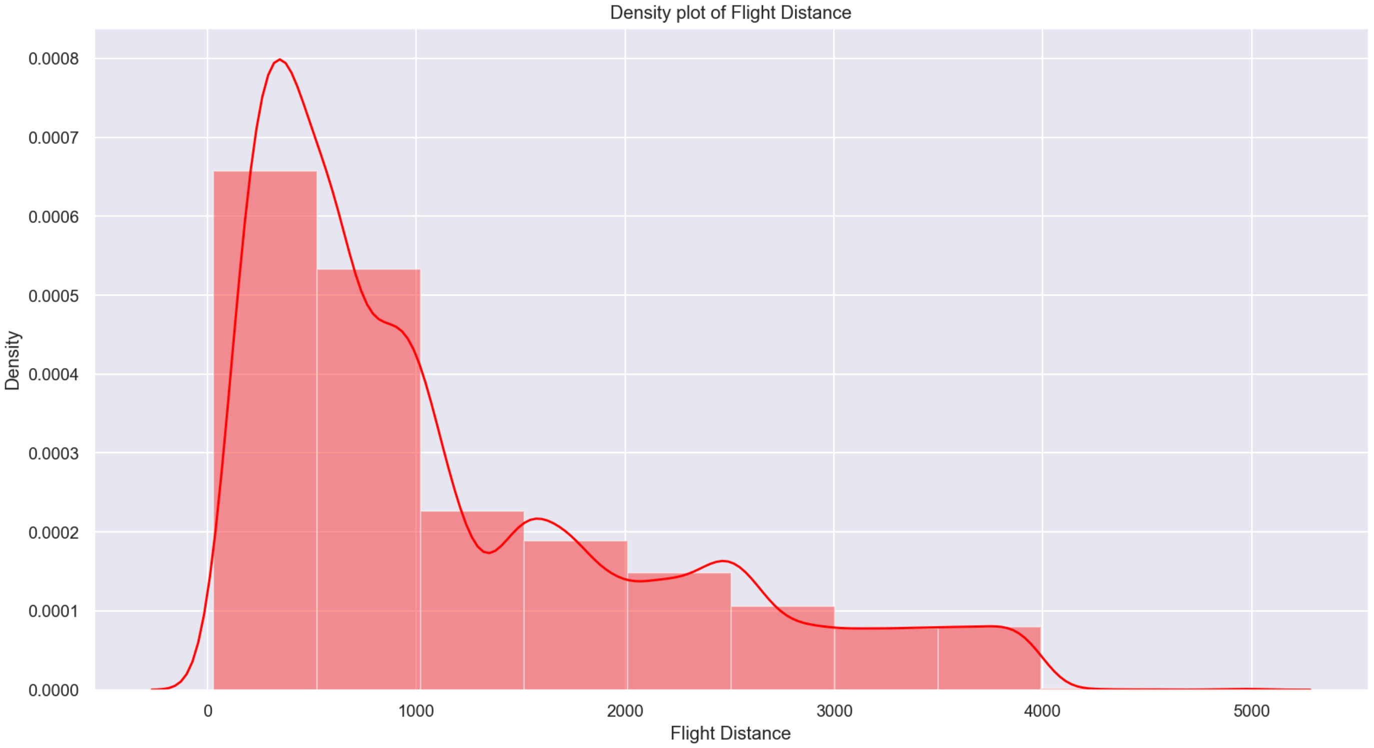

Figure 2. Density plot of Flight Distance

In the above, the density plot of the singular variable *Flight Distance* has been plotted. The y-axis is the continuous density. The x-axis is the *Flight Distance* in miles, ranging from 0 to 5,000. The *flight distance* has been split up into 10 equally-binned bars, indicating ranges for the traveled flight distance by each passenger. Infused with the bars is a singular line plot. The line provides a continuous visual of how densely distributed each 'flight distance' group is.

As stated in the **Dataset Details** section, *Flight Distance* is a continuous numerical variable. Its measuring unit, miles, cannot be easily counted and therefore is not classified as a discrete variable. As it's a continuous numerical variable, it can be useful for plotting line plots to demonstrate characteristics such as quantity, range and density - which can all be used to make conclusions.

For the plot, we can identify that a large proportion of the passengers' flights traveled 0-500 miles most frequently. This is detailed by the highest density proportion of the flight distances by each passenger. We can calculate the volume of passengers traveling a specific flight distance by multiplying the 'density' by 'flight distance'. For example, we can calculate the volume for the max of this plot: $volume = (500 - 0) \times 0.0008 = 0.4$. From this simple calculation, we can deduce that nearly half of the surveys completed describe the passengers to have traveled roughly 500 miles. This may mean the passengers have traveled from city to city, rather than from country to country. Calculating $volume = (4000 - 3500) \times 0.0001 = 0.05$, we can see that as few as 5% of the passengers traveled 3500-4000 miles in a single flight, or 10% of the passengers traveled 3000-4000 miles if we assume the right most visible bins are of the same size. This further adds to the conclusion of the passengers primarily traveling from city to city, rather than country to country.

Another observation we can make is that it is positively skewed, or a large density of the passengers traveled a smaller distance and a small density traveled a large distance. In other words, the graph is heavily populated to the left. The line plot helps indicate which side of the bins are more dense than the other.

In [8]:

```
Fig3 = sns.distplot(data_no_na['Age'], kde = True, bins=5, color='purple').set(title ='Dens
```

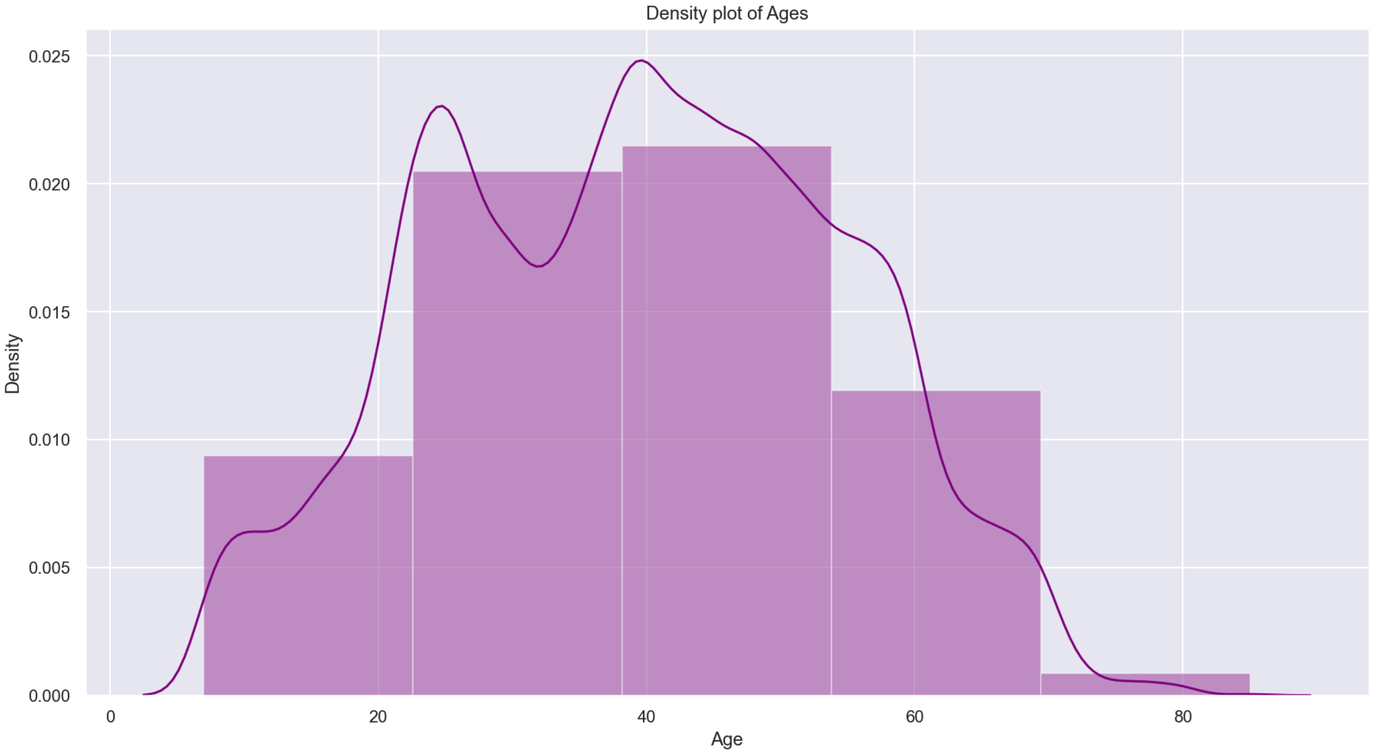

Figure 3. Density plot of Ages

In the above, the density plot of the singular variable *Ages* has been plotted. The y-axis is the continuous density. The x-axis is the age in years, ranging from 0 to 80. The age has been split up into 5 equally-binned bars, indicating ranges for the age group for each passenger. Infused with the bars is a singular line plot. The line provides a continuous visual of how densely distributed each age group is.

As stated in the **Dataset Details** section, *Age* is a continuous numerical variable. Its measuring unit, years, cannot be easily counted and therefore is not classified as a discrete variable. As it's a continuous numerical variable, it can be useful for plotting line plots to demonstrate characteristics such as quantity, range and density - which can all be used to make conclusions.

This plot details that the common passenger is most likely to be aged 39, given the max of the continuous line peaks at 39. This can provide insight towards the flight company's marketing team to tailor their current strategies to attract more of the same age demographic, or perhaps start a new marketing movement for a different age demographic.

Another detail is that there is a large fall off of passengers aged 60 and older. This might describe populations that might have settled or are no longer in need for various flights. However, it might indicate an area of concern being that the flight company isn't the first choice for passengers aged 60 and older.

The singular *Ages* variable may play a large part in identifying which other factors may influence the overall satisfaction of any passenger.

Two Variable Plots

In [9]:

```
sns.set_palette("gnuplot2_r", 8)
Fig4 = sns.violinplot(data = data_no_na, x = 'satisfaction', y = 'Flight Distance').set(tit
```

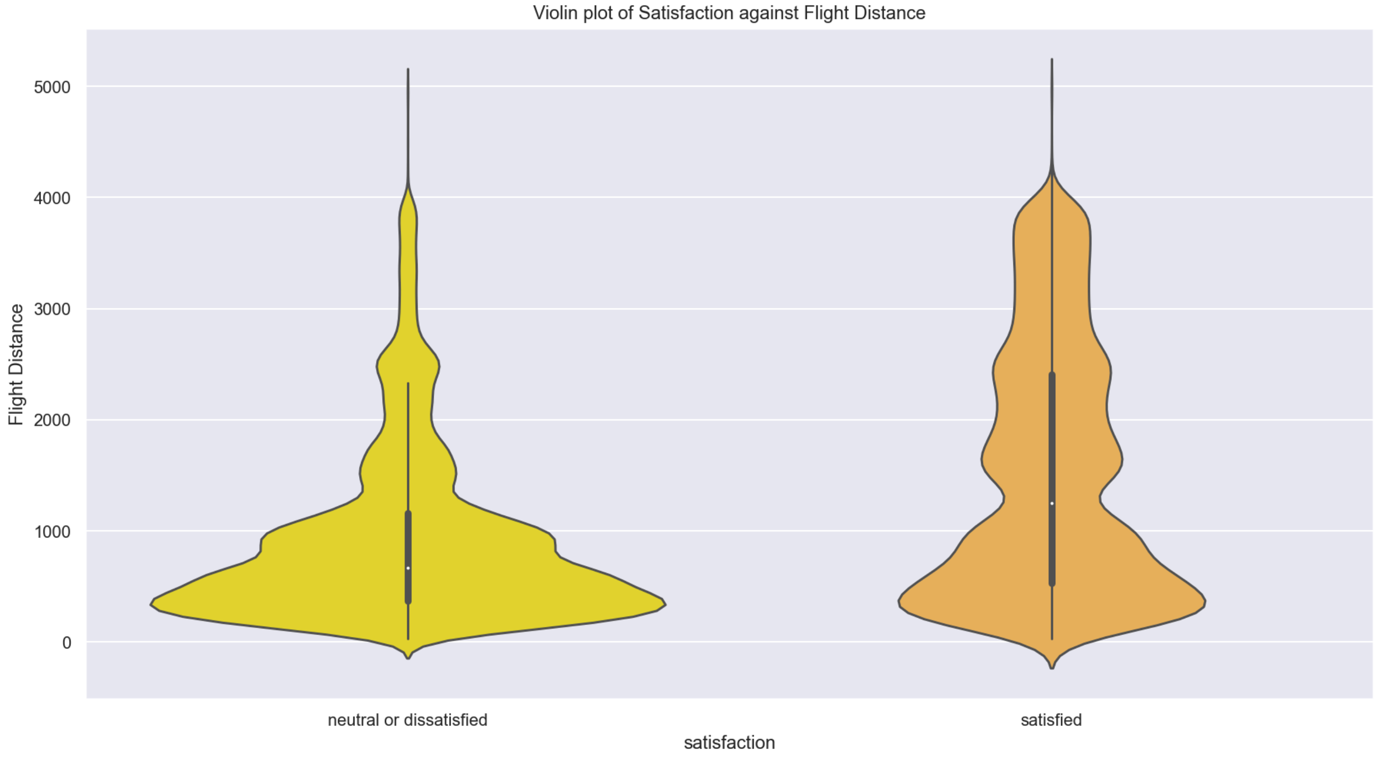

Figure 4. Violin plot of Satisfaction against Flight Distance

In the above, the violin density plot of the *Satisfaction* against *Flight Distance* has been plotted. The y-axis is the *Flight Distance* in miles, ranging from 0 to 5,250. The x-axis is the categories of *Satisfaction* being compared. On the left of the plot, we can see the selection of 'neutral or dissatisfied' as yellow with 'satisfied' as orange on the right.

As stated in the **Dataset Details** section, *Flight Distance* is a continuous numerical variable. Its measuring unit, miles, cannot be easily counted and therefore is not classified as a discrete variable. As it's a continuous numerical variable, it can be useful for plotting plots to demonstrate characteristics such as quantity, range and density - which can all be used to make conclusions. Whereas, *Satisfaction* is a binary categorical variable. It has two categories, 'neutral and dissatisfied' and 'satisfied'. The density of passengers that identified with either were tallied as two individual violins on the plot as seen above.

From this plot, we can identify that passengers were more likely to be satisfied with their flight over longer distances, and were more likely to be neutral or dissatisfied with their flight over short distances. This can indicate that the flight company excels in services and experiences that accompany passengers traveling larger distances. This can be proved by the frequently occurring satisfied rating of distances greater than 1,000 miles compared to the density of neutral or dissatisfied ratings of distances greater than 1,000 miles. This also includes the conclusion that the flight company doesn't excel in services and experiences that accompany passengers traveling shorter distances. This is indicated by the large majority of customers feeling neutral or dissatisfied when traveling distances lesser than 1,000 miles compared to the density of customers feeling satisfied when traveling distances lesser than 1,000 miles.

In [10]:

```
sns.set_palette("husl", 8)
Fig5 = sns.violinplot(data = data_no_na, x = 'satisfaction', y = 'Arrival Delay in Minutes'
    title = 'Violin plot of Satisfaction against Arrival Delay')
```

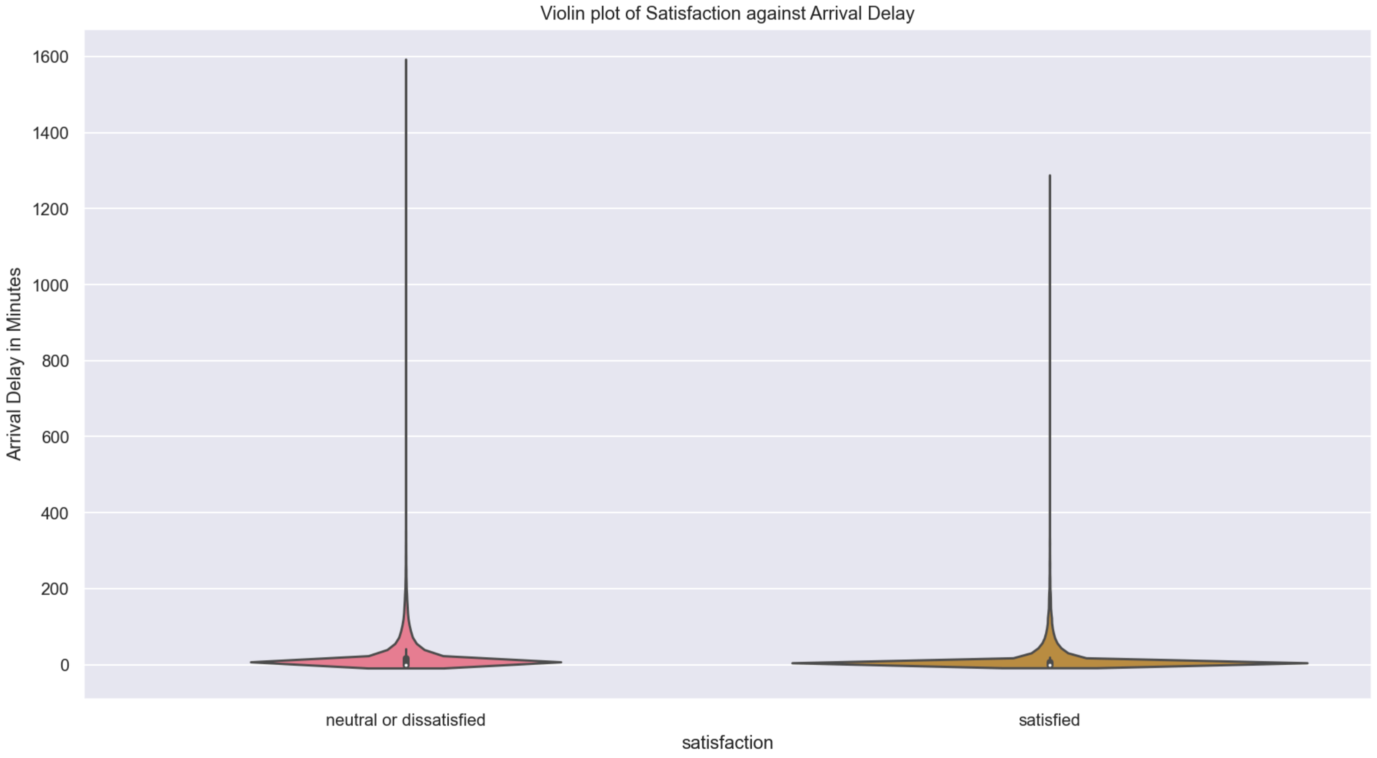

Figure 5. Violin plot of Satisfaction against Arrival Delay

In the above, the violin density plot of the *Satisfaction* against the *Arrival Delay in Minutes* has been plotted. The y-axis is the *Arrival Delay* in minutes, ranging from 0 to 1,600 minutes. The x-axis is the categories of *Satisfaction* being compared. On the left of the plot, we can see the selection of 'neutral or dissatisfied' as pink with 'satisfied' as brown on the right.

As stated in the **Dataset Details** section, *Arrival Delay* is a continuous numerical variable. Its measuring unit, minutes, cannot be easily counted and therefore is not classified as a discrete variable. As it's a continuous numerical variable, it can be useful for plotting plots to demonstrate characteristics such as time, range and density - which can all be used to make conclusions. Whereas, *Satisfaction* is a binary categorical variable. It has two categories, 'neutral and dissatisfied' and 'satisfied'. The density of passengers that identified with either were tallied as two individual violins on the plot as seen above.

From the plot, we can identify that the less delay a passenger has to experience for their flight to arrive at their destination, the passenger is more likely to be satisfied with their overall flight. This is evident when the right violin plot, 'satisfied', is much more populated and dense around 0-50 minutes compared to the left violin plot, 'neutral or dissatisfied'. However, the 'neutral or dissatisfied' plot is more concentrated around 50 minutes compared to the 'satisfied' plot. In addition to this, the 'neutral or dissatisfied' plot is elongated and peaks at 1,600 minutes, whereas the 'satisfied' peaks at 1,300 minutes. This must mean that there have been flights amounting to a 'neutral and dissatisfied' rating where a passenger has waiting excessively long for their arrival. The large delay can be responsible for the reasoning behind passengers feeling 'neutral or dissatisfied' about their flight, and vice versa.

In [11]:

```
sns.set_palette("Set2", 8)
Fig6 = sns.violinplot(data = data_no_na, x = 'satisfaction', y = 'Departure Delay in Minute
    title = 'Violin plot of Satisfaction against Departure Delay')
```

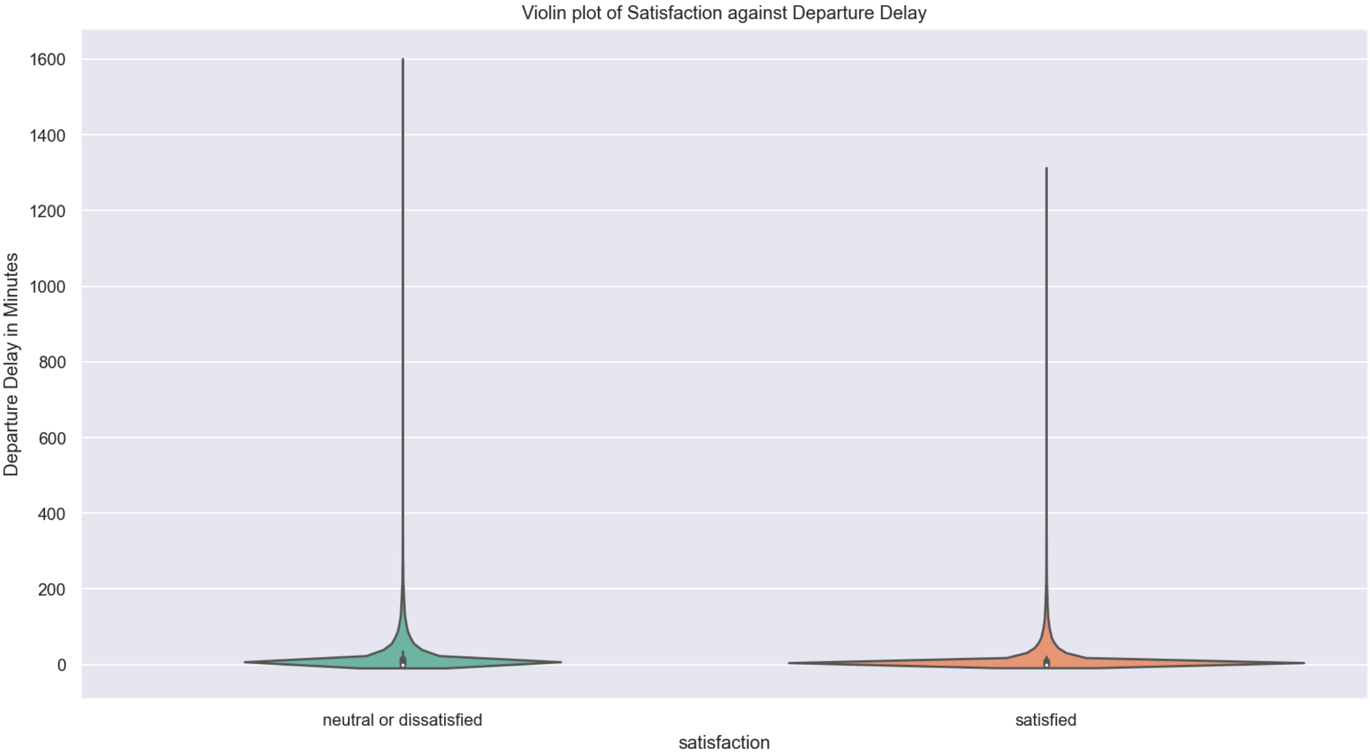

Figure 6. Violin plot of Satisfaction against Departure Delay

Much like the previous plot, the violin density plot of the *Satisfaction* against the *Departure Delay in Minutes* has been plotted. The y-axis is the *Departure Delay* in minutes, ranging from 0 to 1,600 minutes. The x-axis is the categories of *Satisfaction* being compared. On the left of the plot, we can see the selection of 'neutral or dissatisfied' as teal with 'satisfied' as orange on the right.

As stated in the **Dataset Details** section, *Departure Delay* is a continuous numerical variable. Its measuring unit, minutes, cannot be easily counted and therefore is not classified as a discrete variable. As it's a continuous numerical variable, it can be useful for plotting plots to demonstrate characteristics such as time, range and density - which can all be used to make conclusions. Whereas, *Satisfaction* is a binary categorical variable. It has two categories, 'neutral and dissatisfied' and 'satisfied'. The density of passengers that identified with either were tallied as two individual violins on the plot as seen above.

From the plot, we can identify that the less delay a passenger has to experience for their flight to departure from their destination, the passenger is more likely to be satisfied with their overall flight. This is evident when the right violin plot, 'satisfied', is much more populated and dense around 0-50 minutes compared to the left violin plot, 'neutral or dissatisfied'. However, the 'neutral or dissatisfied' plot is more concentrated around 50 minutes compared to the 'satisfied' plot. In addition to this, the 'neutral or dissatisfied' plot is elongated and peaks at 1,600 minutes, whereas the 'satisfied' peaks at 1,300 minutes. This must mean that there have been flights amounting to a 'neutral and dissatisfied' rating where a passenger has waiting excessively long for their departure. The large delay can be responsible for the reasoning behind passengers feeling 'neutral or dissatisfied' about their flight, and vice versa.

Three Variable Plots

In [12]:

```
Fig7 = sns.violinplot(data = data_no_na, x = 'Gender', y = 'Flight Distance', hue = 'satisf
        title = 'Violin plot of Gender against Flight Distance against Satisfaction')
```

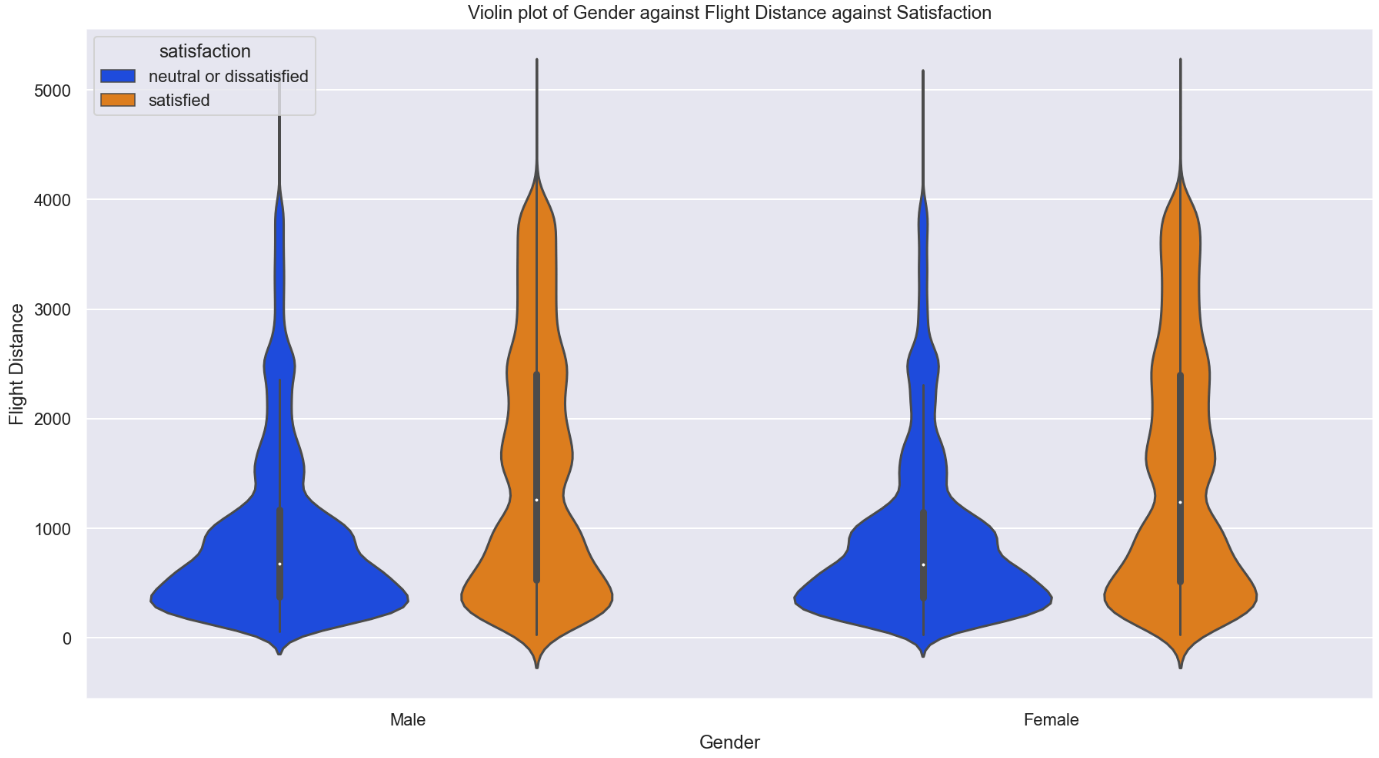

Figure 7. Violin plot of Gender against Flight Distance against Satisfaction

In the above plot, the violin density plot of *Gender* against *Flight Distance* against *Satisfaction* has been plotted. The y-axis is the *Flight Distance* in miles, ranging from 0 to 5,500. The x-axis is the identified *Gender* of the passengers, male or female. The colour is the categories of *Satisfaction* being compared, blue for 'neutral or dissatisfied' and orange for 'satisfied'. On the left side of the plot, the males' rating of *Satisfaction* is plotted. On the right side of the plot, the females' rating of *Satisfaction* is plotted.

As stated in the **Dataset Details** section, *Gender* is a binary categorical variable. It has two categories, 'male' and 'female'. The density of passengers that identified with either were tallied as two individual violins on the plot as seen above. *Flight Distance* is a continuous numerical variable. Its measuring unit, miles, cannot be easily counted and therefore is not classified as a discrete variable. As it's a continuous numerical variable, it can be useful for plotting plots to demonstrate characteristics such as quantity, range and density - which can all be used to make conclusions. *Satisfaction* is a binary categorical variable. It has two categories, 'neutral and dissatisfied' and 'satisfied'. The density of passengers that identified with either were tallied as two groups of violins on the plot as seen above.

From **Figure 4**, we can see that many of the same overall conclusions of *Flight Distance* against *Satisfaction* remain. That being, the flight company doesn't excel in services and experiences that accompany passengers traveling shorter distances. With the addition of the *Gender* variable, we can observe that there is little to no difference with the distribution of *Satisfaction* levels. Visually, both male and female plots do not differ with their responses in regards to feeling 'neutral or dissatisfied' or 'satisfied'. This indicates that *Gender* is not the primary cause of dissatisfaction nor is it related to satisfaction. Therefore, **Figure 4**'s conclusions stand.

</h3>

In [13]:

```
Fig8 = sns.boxplot(data = data_no_na, x = 'Ease of Online booking', y = 'Flight Distance',
        title = 'Box plot of Flight Distance against Ease of Online Booking against Satisfa
```

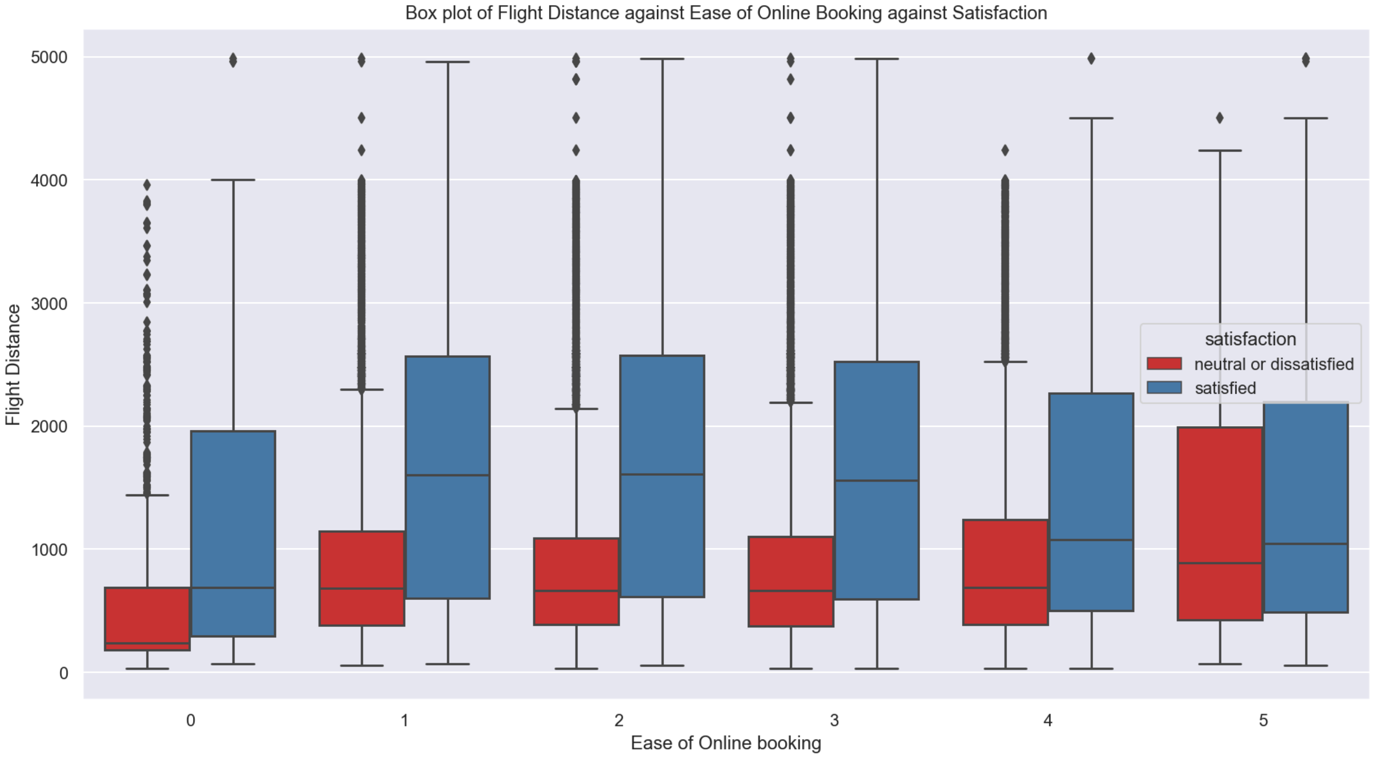

Figure 8. Box plot of Flight Distance against Ease of Online Booking against Satisfaction

In the above plot, the box plot of *Flight Distance* against *Ease of Online Booking* against *Satisfaction* has been plotted. The y-axis is the *Flight Distance* in miles, ranging from 0 to 5,000 miles. The x-axis is the *Ease of Online Booking* rating, ranging from 0 to 5 - 0, being unreceived or the worst; 5, being the best. The colour is the categories of *Satisfaction* being compared, red for 'neutral or dissatisfied' and blue for 'satisfied'.

As stated in the **Dataset Details** section, *Flight Distance* is a continuous numerical variable. Its measuring unit, miles, cannot be easily counted and therefore is not classified as a discrete variable. As it's a continuous numerical variable, it can be useful for plotting plots to demonstrate characteristics such as quantity, range and density - which can all be used to make conclusions. *Ease of Online Booking* is a nominal categorical variable. It has a rating system with integers indicating order and levels of satisfaction. *Satisfaction* is a binary categorical variable. It has two categories, 'neutral and dissatisfied' and 'satisfied'. The ranges of passengers that identified with either were colour coded as seen in the plot.

From the plot, we can deduce that the flight company resulted in many satisfied passengers when traveling larger distances despite having low ratings of *Ease of Online Booking*. An interesting observation is that many passengers that rated 5 stars for *Ease of Online Booking* and traveled a distance greater than 1,250 miles, felt 'neutral or dissatisfied' with their overall flight experience. This supports **Figure 4**'s conclusions.

As previously stated in **Figure 4**'s plot of *Flight Distance* against *Satisfaction*, the flight company doesn't excel in services and experiences that accompany passengers traveling shorter distances. This statement remains true as a large range of passengers do feel 'neutral or dissatisfied' with their overall satisfaction regardless of their *Ease of Online Booking* experience.

From this, we can identify that the involvement of the *Ease of Online Booking* experience towards a passenger's overall satisfaction is minimal. A concerning area is the fact that passengers still felt 'neutral or dissatisfied' despite having rated 5 stars for *Ease of Online Booking*. The reason why this may be the case is that *Ease of Online Booking* is not a large criteria of the passenger's additive conclusion of a satisfying flight experience.

```
In [14]:  Fig9 = sns.jointplot(data = data_no_na, x = 'Age', y = 'Flight Distance', hue = 'satisfacti
```

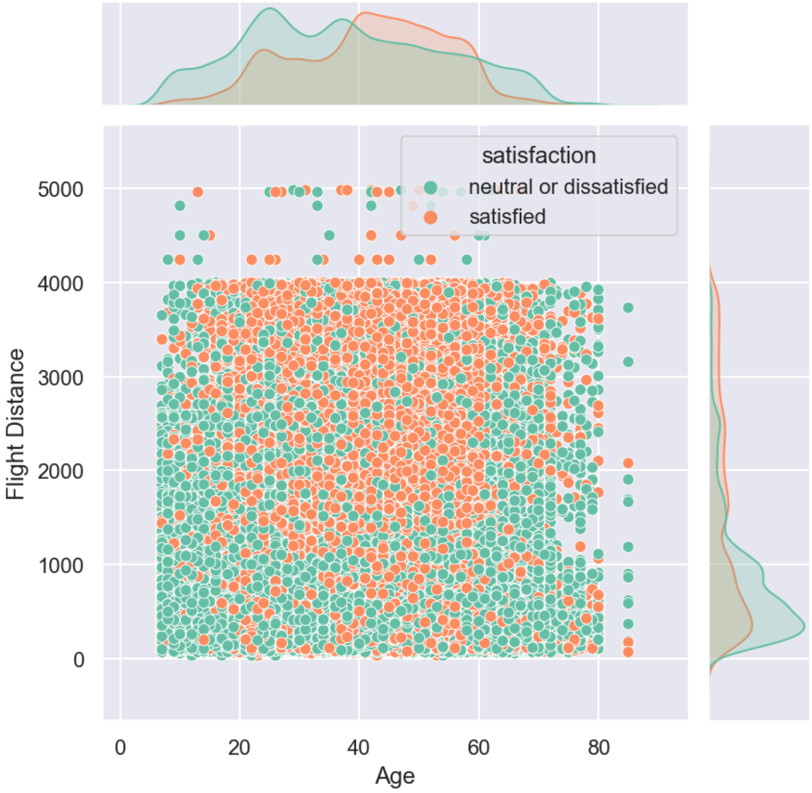

Figure 9. Scatter Plot of Age vs Flight Distance vs Satisfaction with kernal density estimates

In the above plot, the scatter plot of *Flight Distance* against *Age* against *Satisfaction* has been plotted, with the kernal density estimates above and to the right of the scatter plot. The y-axis is the *Flight Distance* in miles, ranging from 0 to 5,000 miles. The x-axis is the *Age*, ranging from 7 to 85 years of age. The colour is the categories of *Satisfaction* being compared, green for 'neutral or dissatisfied' and orange for 'satisfied'.

As stated in the **Dataset Details** section, *Flight Distance* is a continuous numerical variable. Its measuring unit, miles, cannot be easily counted and therefore is not classified as a discrete variable. As it's a continuous numerical variable, it can be useful for plotting plots to demonstrate characteristics such as quantity, range and density - which can all be used to make conclusions. *Age* is a numerical variable, with the same plotting benefits as *Flight Distance*. *Satisfaction* is a binary categorical variable. It has two categories, 'neutral and dissatisfied' and 'satisfied'. The ranges of passengers that identified with either were colour coded as seen in the plot.

From the plot, the kernal density estimates on the right provide information that customers who travel less than 1300 miles are more likely to be 'neutral or dissatisfied' rather than 'satisfied'. An interesting observation is that passengers that are aged between 40 and 60, are more likely to to be 'satisfied' compared to 'neutral or dissatisfied'.

As previously stated in **Figure 4**'s plot of *Flight Distance* against *Satisfaction*, the flight company doesn't excel in services and experiences that accompany passengers traveling shorter distances. It is not quite clear if *Flight Disance* and *Age* is enough to determine *Satisfaction* of a particular experience from a customer. It is also not clear if *Age* is a factor at all in determining a customers satisfaction.

Summary and Conclusion

Phase 1 started with finding a dataset. The dataset found was from www.kaggle.com and the specific data set relates the customer satisfaction surveys from airline passengers. The specific data set can be found at the following link: www.kaggle.com/teejmahal20/airline-passenger-satisfaction (Klein, 2020).

Next the dataset was inspected. The target feature was defined as the satisfaction. After that, the dataset was cleaned. To clean the dataset, the first two columns were removed, and secondly the dataset was scanned for missing values. It was found that there were 310 missing values in the *Arrival Delay in Minutes* column so the rows containing these observations were removed. Following the data cleaning, data visualisaton and data exploration were conducted.

For one variable plots, *Satisfaction*, *Flight Distance* and *Age* were plotted. For **Figure 1**'s *Satisfaction*, it was found that more people are likely to be 'neutral or dissatisfied' rather than 'satisfied'. For **Figure 2**'s *Flight distance*, it was recognised that most people travel less than 1000 miles. For **Figure 3**'s *Age*, it was identified that mean of ages was 39 and the mean of ages is almost normally distributed.

For two variable plots, there were three plots produced. **Figure 4** was a Violin plot of Satisfaction against Flight Distance. This plot showed that people who travel less than 1000 miles are more likely to be "neutral or dissatisfied", indicating that may affect the satisfaction of passengers. **Figure 5** was a Violin plot of Satisfaction against arrival delay and **Figure 6** was Violin plot of Satisfaction against Departure Delay. Both of these plots show that the less time passengers are delayed in arrival and departure, the more likely they are to be 'satisfied'. There may be a strong correlation between delay and satisfaction.

For 3 variable plots, there were three plots produced. **Figure 7** was a Violin plot of *Gender* against *Flight Distance* against *Satisfaction*. This plot indicates that gender does not appear to impact satisfaction as the violin plots for both 'Male' and 'Female' appear to be relatively the same. **Figure 8** shows a Box plot of *Flight Distance* against *Ease of Online Booking* against *Satisfaction*. This plot confirms what was found in **Figure 4** in that people who travel short distances are more likely to be 'neutral or dissatisfied'. *Ease of Online Booking* appears to no have no impact on *Satisfaction*. The last plot, **Figure 9**, shows a Scatter Plot of *Age* against *Flight Distance* against *Satisfaction* with kernel density estimates. While supporting what was found in **Figure 4**, *Age* was more difficult to analyse. Through the kernel density estimates, it appears that passengers aged between 40 to 60 are more likely to be 'satisfied' with their experience. The trend is unclear and it is unknown if *Age* has a noticeable impact on *Satisfaction*.

Overall *Flight Distance* and *Age* may be significant in determining the *Satisfaction* of passengers. It is highly unlikely that *Gender*, *Arrival Delay*, *Departure Delay* and *Ease of Online Booking* contribute to customer satisfaction.

References

- Klein, R. (2020, February). Airline Passenger Satisfaction Version 1. Retrieved August 6, 2021 from https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction/metadata