Airline Passenger Satisfaction

Sohila Abdallah 20108874

Hager Ashraf 20108609

Presented to:

Eng. Zeyad Abdelhady

Contents:

Table of Figure:

Chapter.1

1.1 Introduction

In today's fiercely competitive aviation industry, delivering outstanding services is key to gaining a competitive edge. It is imperative for aviation companies to comprehend how their services meet the needs and desires of customers to achieve passenger satisfaction. This research investigates airline passenger satisfaction through the application of data mining techniques. The study focuses on analyzing various service attributes, ultimately identifying the top areas that require improvement by airlines to enhance passenger satisfaction.

1.2 Data Contents

- Gender: Gender of the passengers (Female, Male)
- Customer Type: The customer type (Loyal customer, disloyal customer)
- Age: The actual age of the passengers
- Type of Travel: Purpose of the flight of the passengers (Personal Travel, Business Travel)
- Class: Travel class in the plane of the passengers (Business, Eco, Eco Plus)
- Flight distance: The flight distance of this journey
- Inflight Wi-Fi service: Satisfaction level of the inflight Wi-Fi service (0: Not Applicable)
- Departure/Arrival time convenient: Satisfaction level.
- Ease of Online booking: Satisfaction level of online booking
- Gate location: Satisfaction level of Gate location
- Food and drink: Satisfaction level of Food and drink
- Online boarding: Satisfaction level of online boarding
- Seat comfort: Satisfaction level of Seat comfort
- Inflight entertainment: Satisfaction level of inflight entertainment
- On-board service: Satisfaction level of On-board service
- Leg room service: Satisfaction level of Leg room service
- Baggage handling: Satisfaction level of baggage handling
- Check-in service: Satisfaction level of Check-in service
- Inflight service: Satisfaction level of inflight service
- Cleanliness: Satisfaction level of Cleanliness
- Departure Delay in Minutes: Minutes delayed when departure.
- Arrival Delay in Minutes: Minutes delayed when Arrival.

Chapter.2

Collecting Data

2.1 Import libraries.

This imports key libraries for data analysis and visualization, including 'Pandas' for data manipulation, 'NumPy' for numerical computations, 'Seaborn' for enhanced data visualization, and 'Missingno' for identifying missing values. These libraries offer valuable tools for conducting data analysis and visualization tasks efficiently.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import missingno as mn
```

Figure 1

2.2 Show data.

The '.head()' function in Python, commonly used with the Pandas library, provides a concise way to display a small subset of the beginning rows of a dataset. By default, it shows the first five rows, allowing for a quick overview of the data's structure and content. This function is useful for initial data inspection and gaining a brief understanding of the dataset before further analysis.

```
DataFile =  pd.read_csv("test.csv")
```

Figure 2

2.3 displays information.

The '.info()' function in Python, used with the Pandas library, provides a concise summary of a dataset's structure. It displays information such as the number of rows, columns, and data types of each column. Additionally, it provides an overview of memory usage. This function is useful for quickly understanding the composition and characteristics of a dataset, making it a valuable tool for initial data assessment.

```
DataFile.head()
```

Figure 3

Output of information:

	Unnamed: 0	id	Gender	Customer Type	Age	Type of Travel	Class	Flight Distance	Inflight wifi service	Departure/Arrival time convenient	...	Inflight entertainment	On-board service	Leg room service	Baggage handling	Checkin service	Inflight service	Cleanliness	Departure Delay in Minutes	Arrival Delay in Minutes	satisfaction
0	0	19556	Female	Loyal Customer	52	Business travel	Eco	160	5	4	...	5	5	5	5	2	5	5	50	44.0	satisfied
1	1	90035	Female	Loyal Customer	36	Business travel	Business	2863	1	1	...	4	4	4	4	3	4	5	0	0.0	satisfied
2	2	12360	Male	disloyal Customer	20	Business travel	Eco	192	2	0	...	2	4	1	3	2	2	2	0	0.0	neutral or dissatisfied
3	3	77959	Male	Loyal Customer	44	Business travel	Business	3377	0	0	...	1	1	1	1	3	1	4	0	6.0	satisfied
4	4	36875	Female	Loyal Customer	49	Business travel	Eco	1182	2	3	...	2	2	2	2	4	2	4	0	20.0	satisfied

Figure 4

Chapter.3

3.1 Preparing

Data Conversion:

The code snippet uses the replace() function in Python to convert categorical text values in the 'satisfaction' column of the 'DataFile' dataset to numeric format. 'satisfied' is replaced.

with 1, and 'neutral or dissatisfied' is replaced with 0. This conversion enables easier analysis and processing of the data, particularly in machine learning applications that rely on numeric inputs.

```
DataFile['satisfaction'].replace({'satisfied':1,'neutral or dissatisfied':0}, inplace=True)
```

Figure 5

Data set after converting:

	Unnamed: 0	id	Gender	Customer Type	Age	Type of Travel	Class	Flight Distance	Inflight wifi service	Departure/Arrival time convenient	...	Inflight entertainment	On-board service	Leg room service	Baggage handling	Checkin service	Inflight service	Cleanliness	Departure Delay in Minutes	Arrival Delay in Minutes	satisfaction
0	0	19556	Female	Loyal Customer	52	Business travel	Eco	160	5	4	...	5	5	5	5	2	5	5	50	44.0	1
1	1	90035	Female	Loyal Customer	36	Business travel	Business	2863	1	1	...	4	4	4	4	3	4	5	0	0.0	1
2	2	12360	Male	disloyal Customer	20	Business travel	Eco	192	2	0	...	2	4	1	3	2	2	2	0	0.0	0
3	3	77959	Male	Loyal Customer	44	Business travel	Business	3377	0	0	...	1	1	1	1	3	1	4	0	6.0	1
4	4	36875	Female	Loyal Customer	49	Business travel	Eco	1182	2	3	...	2	2	2	2	4	2	4	0	20.0	1

Figure 6

Information:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 25 columns):
 #   Column                            Non-Null Count  Dtype
---  ------                            --------------  -----
 0   Unnamed: 0                        10000 non-null  int64
 1   id                                10000 non-null  int64
 2   Gender                            10000 non-null  object
 3   Customer Type                     10000 non-null  object
 4   Age                               10000 non-null  int64
 5   Type of Travel                    10000 non-null  object
 6   Class                             10000 non-null  object
 7   Flight Distance                   10000 non-null  int64
 8   Inflight wifi service             10000 non-null  int64
 9   Departure/Arrival time convenient 10000 non-null  int64
 10  Ease of Online booking            10000 non-null  int64
 11  Gate location                     10000 non-null  int64
 12  Food and drink                    10000 non-null  int64
 13  Online boarding                   10000 non-null  int64
 14  Seat comfort                      10000 non-null  int64
 15  Inflight entertainment            10000 non-null  int64
 16  On-board service                  10000 non-null  int64
 17  Leg room service                  10000 non-null  int64
 18  Baggage handling                  10000 non-null  int64
 19  Checkin service                   10000 non-null  int64
 20  Inflight service                  10000 non-null  int64
 21  Cleanliness                       10000 non-null  int64
 22  Departure Delay in Minutes        10000 non-null  int64
 23  Arrival Delay in Minutes          9970 non-null   float64
 24  satisfaction                      10000 non-null  int64
dtypes: float64(1), int64(20), object(4)
memory usage: 1.9+ MB
```

Figure 7

3.2 Data Cleaning

3.2.1 Delete columns:

This process enables the reduction of the dataset's dimensionality and focuses the analysis on the remaining relevant columns. Deleting unnecessary columns can help streamline data processing, improve computational efficiency, and enhance the clarity of subsequent analyses.

```
del DataFile['Departure/Arrival time convenient']
del DataFile['Gate location']
del DataFile['Food and drink']
del DataFile['Leg room service']
del DataFile['Inflight service']
del DataFile['Inflight entertainment']
del DataFile['Flight Distance']
```

Figure 8

3.2.2 Handling missing data:

The code snippet `DataFile.isnull(). sum()` calculates the number of missing values in each column of the 'DataFile' dataset in Python. By using the `.isnull()` function on the dataset, it identifies missing values and represents them as True, while non-missing values are represented as False. The subsequent `.sum()` operation sums up the True values (indicating missing values) for each column. This provides a concise summary of the count of missing values in each column of the dataset, allowing for a quick assessment of data completeness and the need for any further data cleaning or imputation steps.

```
Unnamed: 0                      0
id                              0
Gender                          0
Customer Type                   0
Age                             0
Type of Travel                  0
Class                           0
Inflight wifi service           0
Ease of Online booking          0
Online boarding                 0
Seat comfort                    0
On-board service                0
Baggage handling                0
Checkin service                 0
Cleanliness                     0
Departure Delay in Minutes      0
Arrival Delay in Minutes       30
satisfaction                    0
dtype: int64
```

Figure 9

Code to handle missing data:

```
DataFile['Arrival Delay in Minutes'] = DataFile['Arrival Delay in Minutes'].interpolate(method ='linear', limit_direction='forward')

DataFile.isnull().sum()
```

Figure 10

```
Unnamed: 0                         0
id                                 0
Gender                             0
Customer Type                      0
Age                                0
Type of Travel                     0
Class                              0
Inflight wifi service              0
Ease of Online booking             0
Online boarding                    0
Seat comfort                       0
On-board service                   0
Baggage handling                   0
Checkin service                    0
Cleanliness                        0
Departure Delay in Minutes         0
Arrival Delay in Minutes           0
satisfaction                       0
dtype: int64
```

Figure 11

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 18 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Unnamed: 0                 10000 non-null  int64
 1   id                         10000 non-null  int64
 2   Gender                     10000 non-null  object
 3   Customer Type              10000 non-null  object
 4   Age                        10000 non-null  int64
 5   Type of Travel             10000 non-null  object
 6   Class                      10000 non-null  object
 7   Inflight wifi service      10000 non-null  int64
 8   Ease of Online booking     10000 non-null  int64
 9   Online boarding            10000 non-null  int64
 10  Seat comfort               10000 non-null  int64
 11  On-board service           10000 non-null  int64
 12  Baggage handling           10000 non-null  int64
 13  Checkin service            10000 non-null  int64
 14  Cleanliness                10000 non-null  int64
 15  Departure Delay in Minutes 10000 non-null  int64
 16  Arrival Delay in Minutes   10000 non-null  float64
 17  satisfaction               10000 non-null  int64
dtypes: float64(1), int64(13), object(4)
memory usage: 1.4+ MB
```

Figure 12

Data after handling missing value:

	Unnamed: 0	id	Age	Inflight wifi service	Ease of Online booking	Online boarding	Seat comfort	On-board service	Baggage handling	Checkin service	Cleanliness	Departure Delay in Minutes	Arrival Delay in Minutes	satisfaction
count	10000.00000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.00000	10000.00000	10000.000000	10000.000000	10000.000000
mean	4999.50000	64661.636500	39.740000	2.730800	2.758500	3.257600	3.450800	3.381300	3.627500	3.31230	3.28150	14.530900	14.826150	0.435500
std	2886.89568	37766.224308	15.188373	1.334507	1.409176	1.351601	1.318087	1.281825	1.171785	1.26741	1.31933	36.115123	36.135889	0.495847
min	0.00000	69.000000	7.000000	0.000000	0.000000	0.000000	1.000000	0.000000	1.000000	1.00000	0.00000	0.000000	0.000000	0.000000
25%	2499.75000	31595.500000	27.000000	2.000000	2.000000	2.000000	2.000000	2.000000	3.000000	3.00000	2.00000	0.000000	0.000000	0.000000
50%	4999.50000	64809.000000	40.000000	3.000000	3.000000	3.000000	4.000000	4.000000	4.000000	3.00000	3.00000	0.000000	0.000000	0.000000
75%	7499.25000	97405.750000	51.000000	4.000000	4.000000	4.000000	5.000000	4.000000	5.000000	4.00000	4.00000	13.000000	13.000000	1.000000
max	9999.00000	129876.000000	85.000000	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000	5.00000	5.00000	951.000000	940.000000	1.000000

Figure 13

Chapter.4

Visualization

Data visualization in Python utilizes libraries like Matplotlib, Seaborn, and Plotly to create graphical representations of data. These visualizations aid in understanding patterns, trends, and relationships within the data, enabling effective communication of insights. Python's versatility and customization options make it a preferred choice for data visualization tasks.

```
s=DataFile['satisfaction']
plt.title('Histogram For satisfaction')
plt.hist(s,bins=50)
plt.show()
```

Figure 15

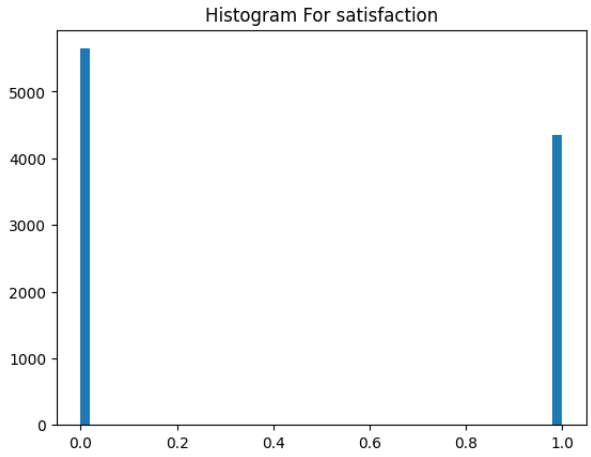

Figure 14

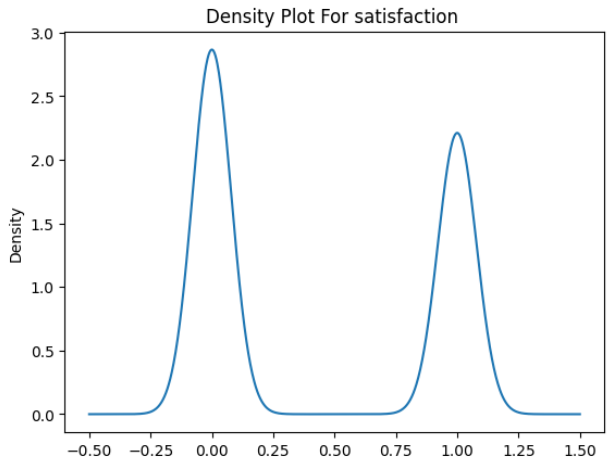

Figure 17

```
plt.title('Density Plot For satisfaction ')
s.plot.kde()
plt.show()
```

Figure 16

```
b=DataFile['Baggage handling']
plt.title('boxplot Plot For Baggage handling')
plt.boxplot(b, showmeans=True)
plt.show()
```

Figure 19

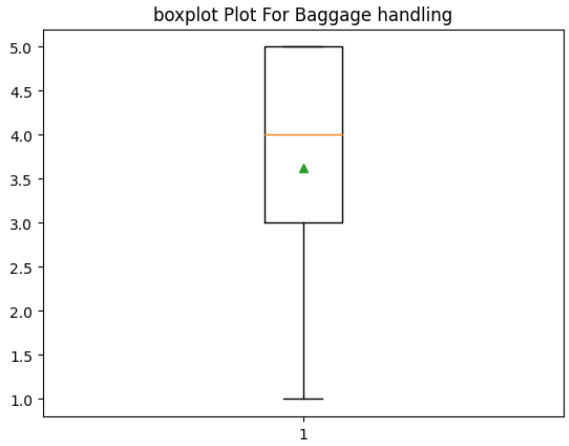

Figure 18

```
DataFile['Customer Type'].value_counts()
```

Figure 21

```
Loyal Customer       8155
disloyal Customer    1845
Name: Customer Type, dtype: int64
```

Figure 20

```
sns.countplot(data=DataFile,x=DataFile['Customer Type'])
```

Figure 22

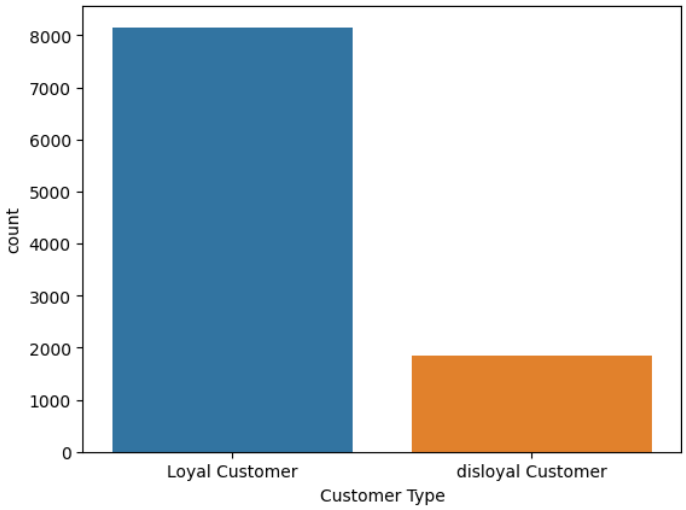

Figure 23

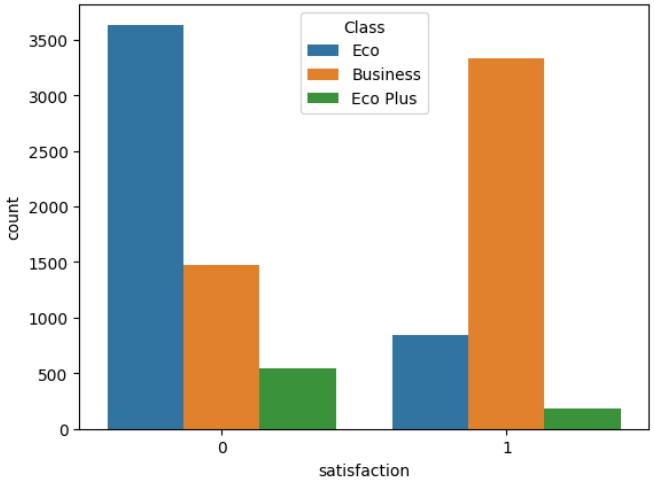

Figure 25

```
sns.countplot(
            x=DataFile['satisfaction'],
            hue=DataFile['Class'],
            data=DataFile
            )
```

Figure 24

```
plt.figure(figsize=(25,10))
sns.heatmap(DataFile.corr(),annot=True)
```

Figure 26

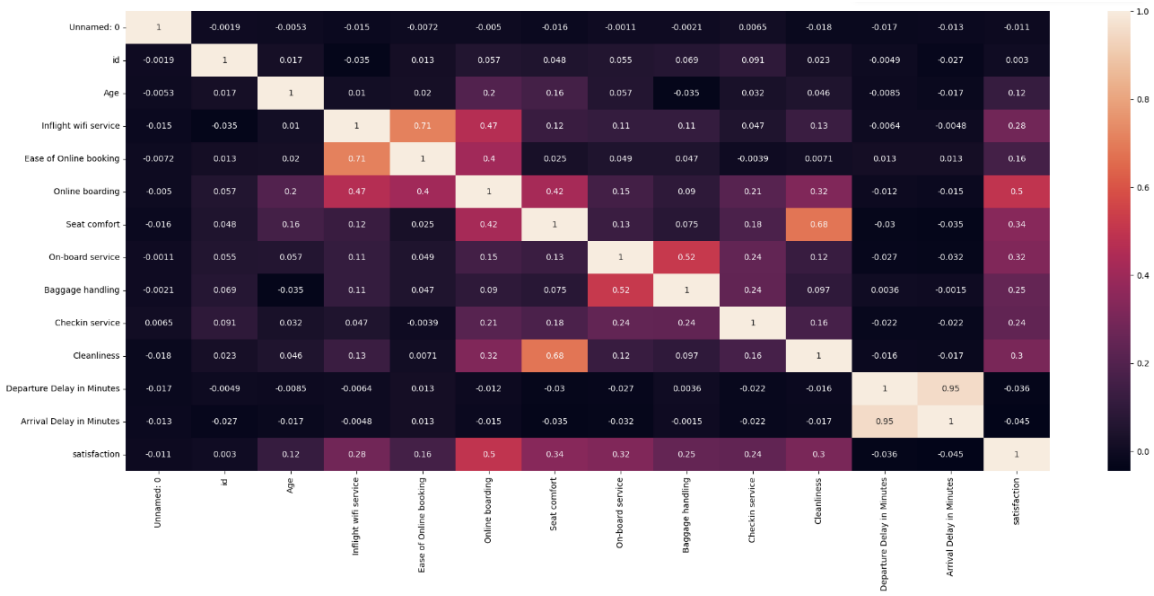

Figure 27

```
sns.countplot(x="Online boarding", hue="satisfaction", data=DataFile)
```

Figure 28

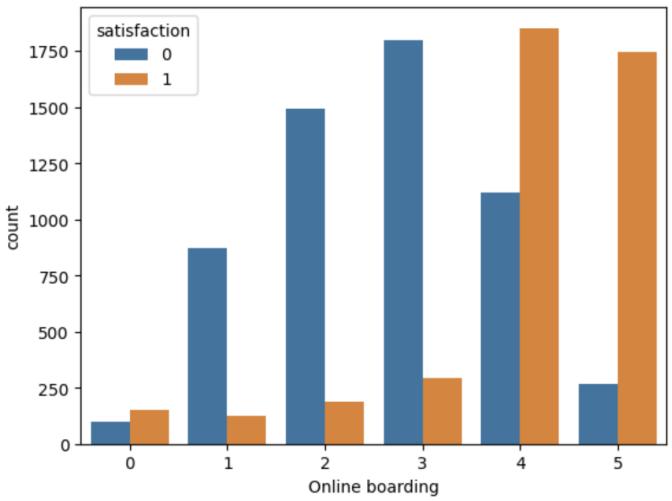

Figure 29

```
sns.countplot(x="Class", hue="satisfaction", data=DataFile)
```

Figure 30

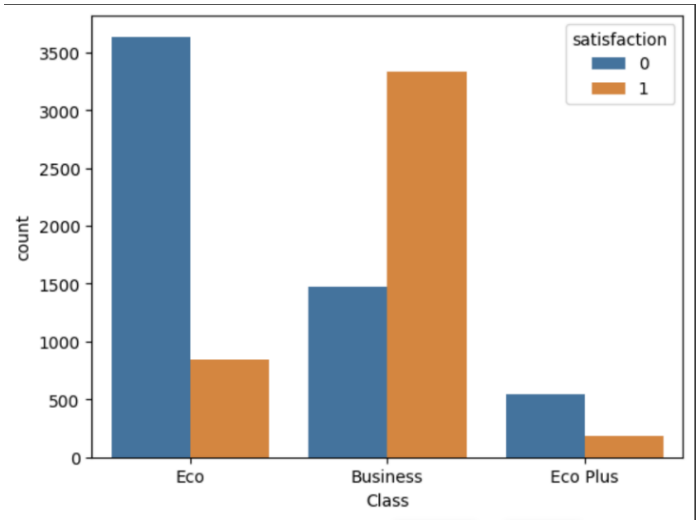

Figure 31

```
sns.countplot(x="Seat comfort", hue="satisfaction", data=DataFile)
```

Figure 33

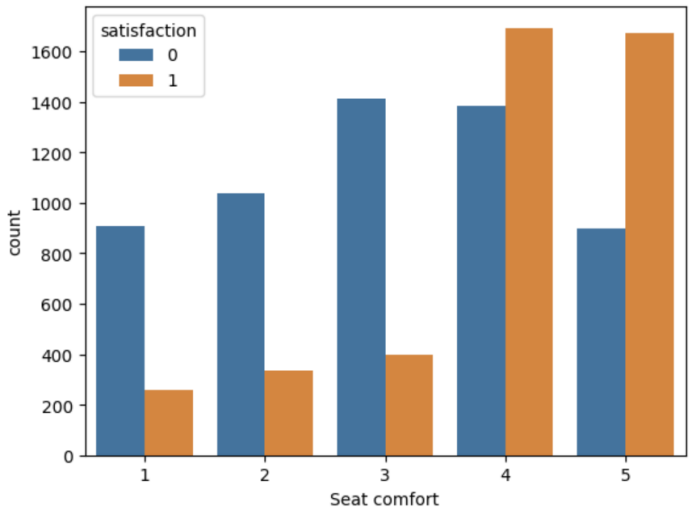

Figure 32

```
sns.countplot(x="Type of Travel", hue="satisfaction", data=DataFile)
```

Figure 34

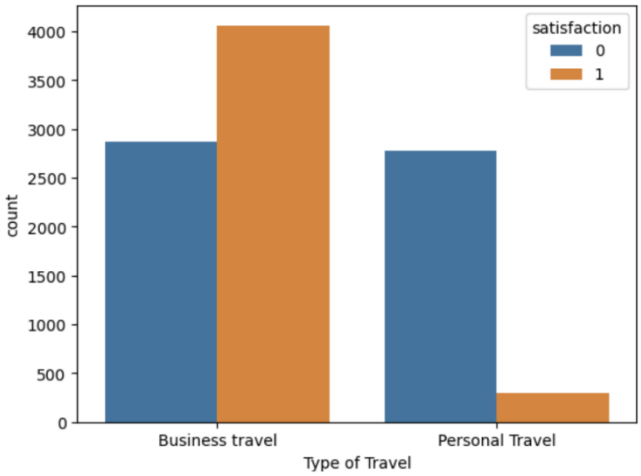

Figure 36

References

https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction