In [1]:
```
import os
os.getcwd()
```

Out[1]:
```
'C:\\Users\\guy74\\Documents\\NU Stuff\\ANA500\\Kasey Ryan ANA500 Micro Project'
```

In [3]:
```
import pandas as pd
import numpy as np

#Load in the data
df = pd.read_csv("airline.csv")

#There is a column which is presumably actually not named in the source but at some
#That variable is presumably not useful here analytically, and same goes for the "i
#The rest could serve useful and so we don't want to drop anything else yet at this
df = df.drop(columns=["Unnamed: 0", "id"])
#print(df)

#Loop through each column and calculate some data integrity metrics for each one to
column_checks = []
total_rows = df.shape[0]
for col in df.columns:
    col_data = df[col]

    populated_count = col_data.notnull().sum()
    missing_count = col_data.isnull().sum()
    distinct_count = col_data.nunique(dropna=True)
    data_type = col_data.dtype

    column_checks.append({
        "Column": col,
        "Populated Values": populated_count,
        "Missing Values": missing_count,
        "Distinct Values": distinct_count,
        "Data Type": str(data_type)
    })

integrity_df = pd.DataFrame(column_checks)

#This can sort it by missing values if we want the highest of those to rise to the
#integrity_df = integrity_df.sort_values(by="Missing Values", ascending=False)


print(integrity_df.to_string(index=False))
```

	Column	Populated Values	Missing Values	Distinct Values	Data Type
	Gender	129880	0	2	object
	Customer Type	129880	0	2	object
	Age	129880	0	75	int64
	Type of Travel	129880	0	2	object
	Class	129880	0	3	object
	Flight Distance	129880	0	3821	int64
	Inflight wifi service	129880	0	6	int64
	Departure/Arrival time convenient	129880	0	6	int64
	Ease of Online booking	129880	0	6	int64
	Gate location	129880	0	6	int64
	Food and drink	129880	0	6	int64
	Online boarding	129880	0	6	int64
	Seat comfort	129880	0	6	int64
	Inflight entertainment	129880	0	6	int64
	On-board service	129880	0	6	int64
	Leg room service	129880	0	6	int64
	Baggage handling	129880	0	5	int64
	Checkin service	129880	0	6	int64
	Inflight service	129880	0	6	int64
	Cleanliness	129880	0	6	int64
	Departure Delay in Minutes	129880	0	466	int64
	Arrival Delay in Minutes	129487	393	472	float64
	satisfaction	129880	0	2	object

In [5]:
```
#Let's also check for any duplicate rows where every single variable is exactly the
duplicate_count = df.duplicated().sum()
print(f"Total number of exact duplicate rows (excluding the two dropped columns): {
```

```
Total number of exact duplicate rows (excluding the two dropped columns): 0
```

In [7]:
```
#Swap out the column names so if any columns have spaces we change them to undersco
# Replace all spaces in column names with underscores
```

```
df.columns = df.columns.str.replace(" ", "_")
print(df)
```

```
        Gender      Customer_Type  Age   Type_of_Travel     Class  \
0         Male     Loyal Customer   13  Personal Travel  Eco Plus
1         Male  disloyal Customer   25  Business travel  Business
2       Female     Loyal Customer   26  Business travel  Business
3       Female     Loyal Customer   25  Business travel  Business
4         Male     Loyal Customer   61  Business travel  Business
...        ...                ...  ...              ...       ...
129875    Male  disloyal Customer   34  Business travel  Business
129876    Male     Loyal Customer   23  Business travel  Business
129877  Female     Loyal Customer   17  Personal Travel       Eco
129878    Male     Loyal Customer   14  Business travel  Business
129879  Female     Loyal Customer   42  Personal Travel       Eco

        Flight_Distance  Inflight_wifi_service  \
0                   460                      3
1                   235                      3
2                  1142                      2
3                   562                      2
4                   214                      3
...                 ...                    ...
129875              526                      3
129876              646                      4
129877              828                      2
129878             1127                      3
129879              264                      2

        Departure/Arrival_time_convenient  Ease_of_Online_booking  \
0                                       4                       3
1                                       2                       3
2                                       2                       2
3                                       5                       5
4                                       3                       3
...                                   ...                     ...
129875                                  3                       3
129876                                  4                       4
129877                                  5                       1
129878                                  3                       3
129879                                  5                       2

        Gate_location  ...  Inflight_entertainment  On-board_service  \
0                   1  ...                       5                 4
1                   3  ...                       1                 1
2                   2  ...                       5                 4
3                   5  ...                       2                 2
4                   3  ...                       3                 3
...               ...  ...                     ...               ...
129875              1  ...                       4                 3
129876              4  ...                       4                 4
129877              5  ...                       2                 4
129878              3  ...                       4                 3
129879              5  ...                       1                 1

        Leg_room_service  Baggage_handling  Checkin_service  Inflight_service  \
0                      3                 4                4                 5
1                      5                 3                1                 4
2                      3                 4                4                 4
```

```
3                    5                  3                  1                  4
4                    4                  4                  3                  3
...                ...                ...                ...                ...
129875               2                  4                  4                  5
129876               5                  5                  5                  5
129877               3                  4                  5                  4
129878               2                  5                  4                  5
129879               2                  1                  1                  1

        Cleanliness  Departure_Delay_in_Minutes  Arrival_Delay_in_Minutes  \
0                 5                         25                      18.0
1                 1                          1                       6.0
2                 5                          0                       0.0
3                 2                         11                       9.0
4                 3                          0                       0.0
...             ...                        ...                       ...
129875            4                          0                       0.0
129876            4                          0                       0.0
129877            2                          0                       0.0
129878            4                          0                       0.0
129879            1                          0                       0.0

                  satisfaction
0       neutral or dissatisfied
1       neutral or dissatisfied
2                     satisfied
3       neutral or dissatisfied
4                     satisfied
...                         ...
129875  neutral or dissatisfied
129876                satisfied
129877  neutral or dissatisfied
129878                satisfied
129879  neutral or dissatisfied

[129880 rows x 23 columns]
```

In [9]:
```
#If/when needed, export the new version of the file to another named .csv for furth
df.to_csv('AirlineSatisfaction_Transformed.csv', index=False)
```

In [ ]:
```

```