May 9, 2025

Airline Customer Satisfaction – What Really Matters?

Customer Data Analytics – Term Project

Alyssa Schiera, Shagun Modi, Hariyat Andargachew

Professor Kihyun Hannah Kim

1. Objectives and Key Questions
2. Data Collection & Preprocessing
3. Exploratory Data Analysis
4. Analytical Modeling & Insights
5. Strategic Recommendations

Objectives and Key Questions

With the COVID-19 pandemic having a significant impact on the travel industry - aviation specifically - and a recent string of airplane crashes increasing travel anxiety, our team was interested in **analyzing customer data of an airline company to understand how different factors affect customer satisfaction** and determine how the airlines can improve customer retention

Key Research Questions

1. What are the variables that affect (most to least) the satisfaction for a customer?
2. Do loyal customers tend to be more satisfied with the airline offerings ?
3. Does the type of travel (Business v.s. Personal) influence how customers rate their satisfaction?
4. How does customer travel class (Business, Eco Plus, and Economy) influence satisfaction, and can this inform segmentation strategy?

Data Collection and Preprocessing

We'll use customer data from a real airline company from Kaggle to run our analysis

Data Acquisition Strategy

1. We used the customer satisfaction data from Kaggle at this link https://www.kaggle.com/datasets/sjleshrac/airlines-customer-satisfaction
2. This data is from a real airline company, which has been given the pseudonym Investico for privacy reasons
3. The dataset consists of information on customers who have already flown with the company and feedback they provided on their satisfaction level
4. While the data is well-structured, some challenges are that there is no price information or transactional data

Exploratory Data Analysis

We will discuss some common themes and key insights at a high view before taking a deep dive to unravel patterns.

Key Trends

1. **Travel Type:** Mostly Business Travel over Personal Travel (Figure 1)
2. **Travel Class:** Business is the largest class, followed by Economy and Economy Plus. (Figure 2)
3. **Gender Split:** Pretty balanced between Male and Female travelers(Figure 3)

Figure 1

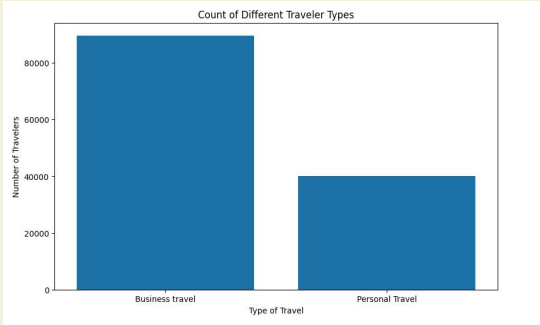


Figure 2

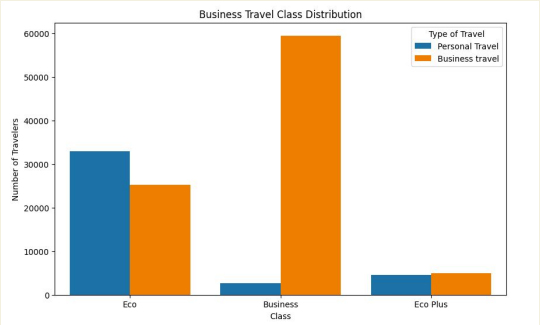
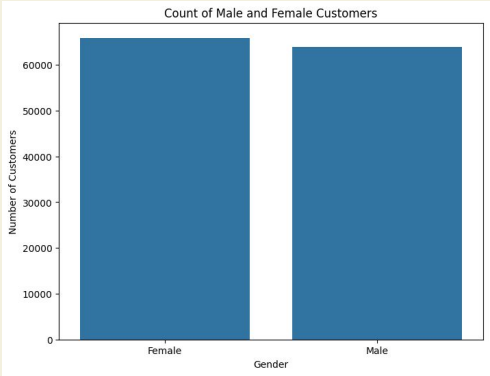


Figure 3

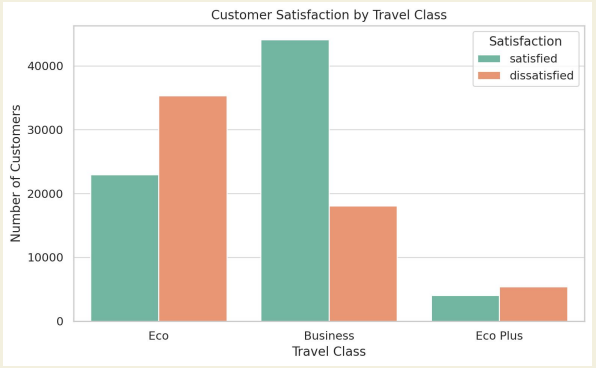


Customer Satisfaction – At a glance

1. **Customer Type**
   - Loyal customers are much more likely to be satisfied
   - Disloyal ones seem to be more dissatisfied (Figure 1)
2. **Business travelers**
   - Satisfied > personal travelers (Figure 2)
3. **Higher travel class** (Business > Eco Plus > Eco) correlates with higher satisfaction (Figure 2)

Figure 1

Figure 2



Analytical Modeling & Insights

Using Statistical techniques, particularly Pearson's correlation, OLS and K-Means clustering for exacting customer behaviour

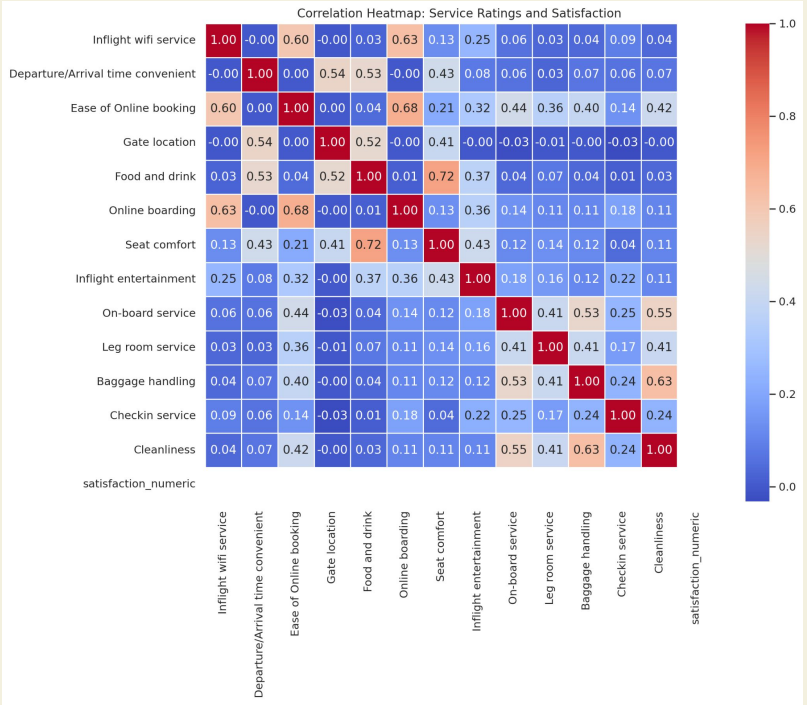
What Factors Are Correlated with Customer Satisfaction? -Approach

1. To understand how strongly satisfaction correlates with all the numerical service rating columns (seat comfort, food, cleanliness) used **pearson correlation coefficients**
2. The corr() method calculates Pearson correlations between all selected columns.
3. sns.heatmap() visualizes those correlations with colors and numeric values.
4. annot=True makes the correlation values visible on the map.
5. cmap='coolwarm' gives it the red-blue gradient.

```python
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Load the dataset
df = pd.read_csv('/mnt/data/Invistico_Airline.csv')

# Convert satisfaction column to numeric: Satisfied = 1, Neutral or Dissatisfied = 0
df['satisfaction_numeric'] = df['satisfaction'].map({
    'Satisfied': 1,
    'Neutral or Dissatisfied': 0
})

# Select relevant numerical service columns + satisfaction
service_cols = [
    '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', 'Cleanliness', 'satisfaction_numeric'
]

# Create correlation matrix
corr = df[service_cols].corr()

# Plot heatmap
plt.figure(figsize=(12, 10))
sns.heatmap(corr, annot=True, cmap='coolwarm', fmt=".2f", linewidths=0.5)
plt.title('Correlation Heatmap: Service Ratings and Satisfaction')
plt.tight_layout()
plt.show()
```

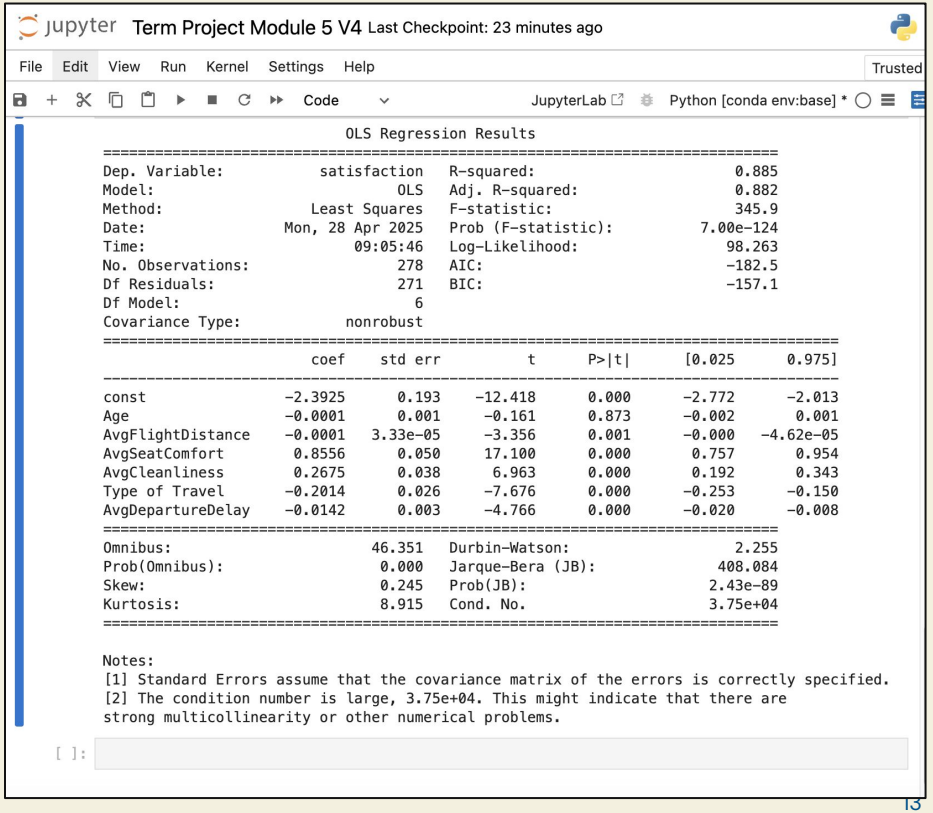
What Factors Are Correlated with Customer Satisfaction? -Insights

1. Key services that strongly correlate with overall satisfaction:
   a. **Seat Comfort**
   b. Inflight Entertainment
   c. **Cleanliness**
   d. Onboard Service
   e. **Ease of Online Booking**
2. Less important factors: Inflight Wifi and Gate Location seem to matter less overall



Statistical Techniques – OLS

1. Used the Ordinary Least Squares (OLS) Regression technique to help determine which independent variables have the most influence on the dependent variable (Customer Satisfaction).
2. Transformed Customer Satisfaction and Type of Travel variables from categorical to binary
    a. Satisfaction: from satisfied or dissatisfied to 1 or 0
    b. Type of Travel: from Personal Travel or Business Travel to 1 or 0



OLS Regression Results

Dep. Variable:	satisfaction	R-squared:	0.885
Model:	OLS	Adj. R-squared:	0.882
Method:	Least Squares	F-statistic:	345.9
Date:	Mon, 28 Apr 2025	Prob (F-statistic):	7.00e-124
Time:	09:05:46	Log-Likelihood:	98.263
No. Observations:	278	AIC:	-182.5
Df Residuals:	271	BIC:	-157.1
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-2.3925	0.193	-12.418	0.000	-2.772	-2.013
Age	-0.0001	0.001	-0.161	0.873	-0.002	0.001
AvgFlightDistance	-0.0001	3.33e-05	-3.356	0.001	-0.000	-4.62e-05
AvgSeatComfort	0.8556	0.050	17.100	0.000	0.757	0.954
AvgCleanliness	0.2675	0.038	6.963	0.000	0.192	0.343
Type of Travel	-0.2014	0.026	-7.676	0.000	-0.253	-0.150
AvgDepartureDelay	-0.0142	0.003	-4.766	0.000	-0.020	-0.008

Omnibus:	46.351	Durbin-Watson:	2.255
Prob(Omnibus):	0.000	Jarque-Bera (JB):	408.084
Skew:	0.245	Prob(JB):	2.43e-89
Kurtosis:	8.915	Cond. No.	3.75e+04

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.75e+04. This might indicate that there are strong multicollinearity or other numerical problems.

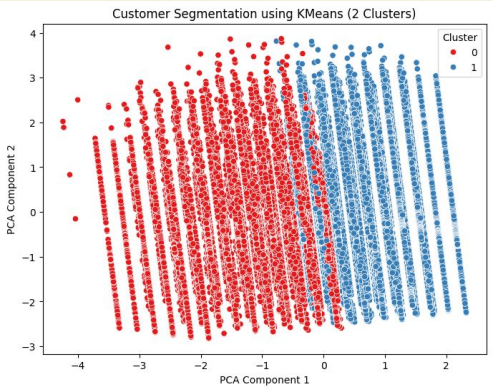
[ ]:

Analytical Insights

1. Overall the model explains 89% of the variation of Customer Satisfaction
2. Variable Interpretation
    a. Age: No statistical significance as p > 0.05
    b. Seat Comfort: Has a strong positive impact on Customer Satisfaction
    c. Cleanliness: Has a positive impact on Customer Satisfaction
    d. Type of Travel: Travel type has an impact on Customer Satisfaction
    e. Departure Delay: Has a slight impact on Customer Satisfaction

	coef	std err	t	P>|t|
const	−2.3925	0.193	−12.418	0.000
Age	−0.0001	0.001	−0.161	0.873
AvgFlightDistance	−0.0001	3.33e−05	−3.356	0.001
AvgSeatComfort	0.8556	0.050	17.100	0.000
AvgCleanliness	0.2675	0.038	6.963	0.000
Type of Travel	−0.2014	0.026	−7.676	0.000
AvgDepartureDelay	−0.0142	0.003	−4.766	0.000

Statistical Techniques - Clustering




```
features = ['Type of Travel', 'Flight Distance', 'Seat comfort', 'Ease of Online booking', 'Cleanliness']
X = data[features]

scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
kmeans = KMeans(n_clusters=2, random_state=42)
data['cluster'] = kmeans.fit_predict(X_scaled)

pca = PCA(n_components=2)
components = pca.fit_transform(X_scaled)

data['cluster'] = data['cluster'].astype(int)

plt.figure(figsize=(8,6))
sns.scatterplot(
    x=components[:, 0],
    y=components[:, 1],
    hue=data['cluster'],
    palette='Set1',
    legend='full'
)

plt.xlabel('PCA Component 1')
plt.ylabel('PCA Component 2')
plt.title('Customer Segmentation using KMeans (2 Clusters)')
plt.legend(title='Cluster')
plt.show()
```

2 clusters (K = 2)

PCA: Principal Component Analysis

- Used to compress many different features into few components
- Helps to visualise data in 2d despite the different components
- Shows essential patterns without loosing on the information

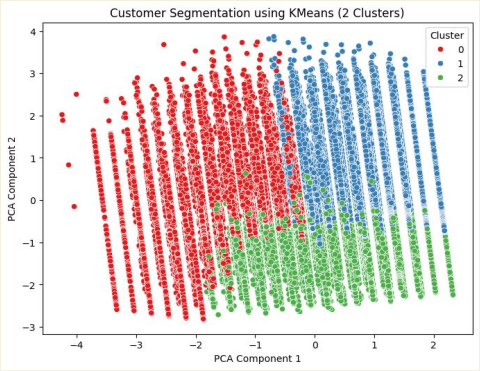
1. Using insights from the factors that mainly determine customer satisfaction **i.e Seat Comfort, Cleanliness and Ease of Online Booking**
2. Distinct clusters showing the segmentation using different colours, marking clusters as discrete categories, and a labelled legend

```
[74] data['Type of Travel'] = data['Type of Travel'].replace({'Personal Travel': 1, 'Business travel': 0})
```

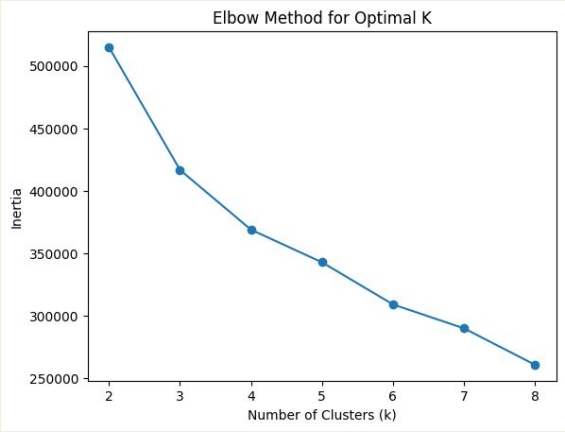
K-Means Clustering for Customer Segmentation

Elbow Method for optimal K :

- Choosing the bend (slow change in Inertia)
- Good model = Low Inertia & Low Clusters (K)
- At K = 3, there is an elbow bend



3 clusters (K = 3)




```
[84] cluster_profile = data.groupby('cluster')[features].mean()
     print(cluster_profile)

                Type of Travel  Flight Distance  Seat comfort  \
cluster
0                     0.075938      2039.285436      2.269903
1                     0.000000      2072.262584      3.159533
2                     1.000000      1789.210857      2.925380

                Ease of Online booking  Cleanliness
cluster
0                             2.201448     2.608369
1                             4.259088     4.256799
2                             3.562232     3.974556
```

Cluster 0 =
Mostly Personal Travellers
significantly long flights
Lowest seat comfort, cleanliness, & Ease of Online Booking
[Likely personal travellers in economy]

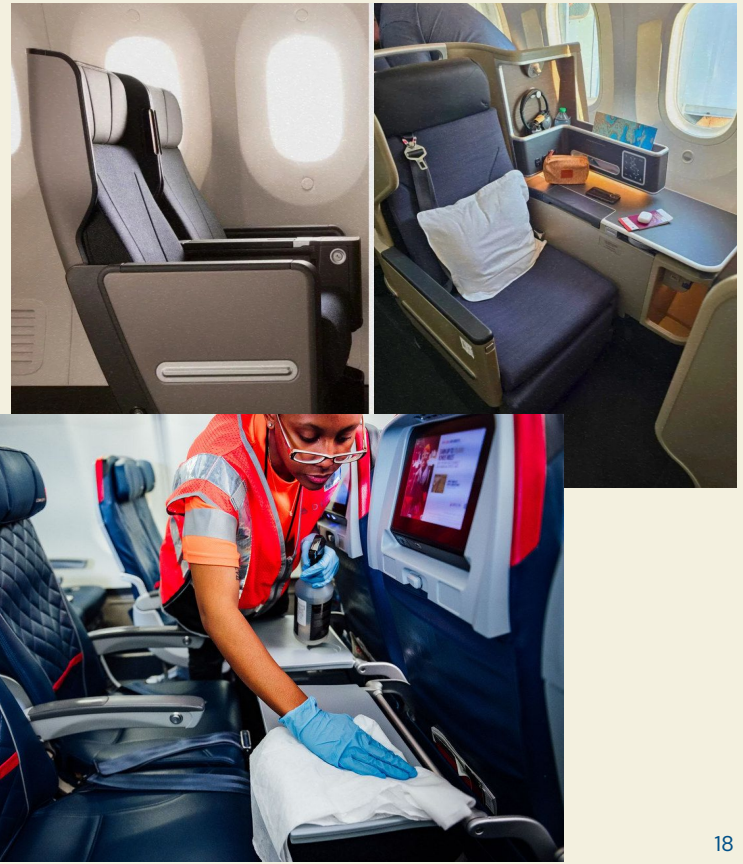
Cluster 1 =
All Business travellers
Highest avg flight distance
Seat comfort, Ease of Online Booking & Cleanliness
[Business Travellers travelling business class, most satisfied segment]

Cluster 2 =
All personal travellers
shortest avg distance
mediocre seat comfort, good cleanliness, and ease of online booking
[Personal travellers travelling short distance in ecoplus]

Strategic Recommendations

Key Long-Term Recommendation

1. **Consider removal of EcoPlus altogether**
    a. It is the small source of revenue for the airline
    b. Would help them better allocate resources and focus on providing better services in the economy seats
    c. Customers with the higher income band likely to move to business, and help the airline with higher capacity utilization



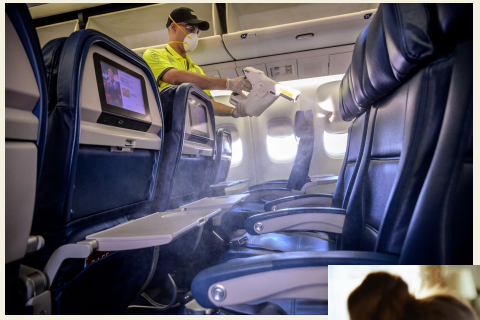
*Assumption: The revenue increases as the class goes higher

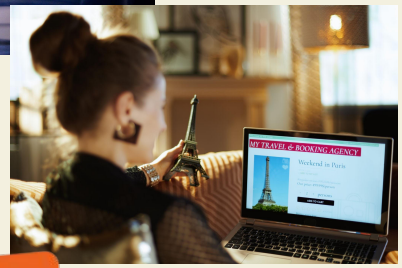
Short-Term Business Strategies (Part 1)

1. **Upgrade Seat Comfort in Economy Class**
    a. Improve cushioning, leg room, and seat ergonomics to close the comfort gap with Business Class.
2. **Invest in Better Inflight Entertainment**
    a. Update movie/music options frequently and ensure systems are easy to use and reliable across all cabins.
3. **Enhance On-board Service Training**
    a. Focus on friendly, efficient, and proactive customer service across all flights, especially for Economy.

Short-Term Business Strategies (Part 2)

4. **Prioritize Aircraft Cleanliness**
    a. Implement stricter cleaning protocols between flights to ensure consistently high cleanliness scores.
5. **Streamline and Improve Online Booking Experience**
    a. Make the website and mobile app faster, clearer, and more intuitive, especially for seat selection and upgrades.
6. **Target Disloyal Customers with Personalized Offers**
    a. Use targeted emails or loyalty programs offering upgrades, priority boarding, or bonus miles to turn neutral travelers into loyal ones.
    b. Create special offers for unsatisfied customers that fly economy to improve overall travel class satisfaction





Thank You


Merci
bedankt
DANKE SCHÖN
obrigado
Gracias
thank you
BEDANKT
Bedankt
danke schön
Obrigado
tack
gracias
THANK YOU
Danke Schön
OBRIGADO
MERCI
TACK
Grazie
bedankt
grazie
obrigado

Appendix

Codes and Data Source

Codes and Data Source

Pearson Correlation Matrix Code:

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
# Load the dataset
df = pd.read_csv('/mnt/data/Invistico_Airline.csv')
# Convert satisfaction column to numeric: Satisfied = 1, Neutral or
Dissatisfied = 0
df['satisfaction_numeric'] = df['satisfaction'].map({'Satisfied': 1,'Neutral or
Dissatisfied': 0})
# Select relevant numerical service columns + satisfaction
service_cols = ['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', 'Cleanliness',
'satisfaction_numeric']
# Create correlation matrix
corr = df[service_cols].corr()
# Plot heatmap
plt.figure(figsize=(12, 10))
sns.heatmap(corr, annot=True, cmap='coolwarm', fmt=".2f", linewidths=0.5)
plt.title('Correlation Heatmap: Service Ratings and Satisfaction')
plt.tight_layout()
plt.show()
```

- **OLS and Cluster codes attached as Python files in Canvas submission**
- https://www.kaggle.com/code/firuzjuraev/airlines-customer-satisfaction-classification/input - Kaggle data source on Invistico customer satisfaction
- https://www.kaggle.com/datasets/sjleshrac/airlines-customer-satisfaction
    - CSV File attached in canvas Submission