


Data Collection and Preprocessing Phase

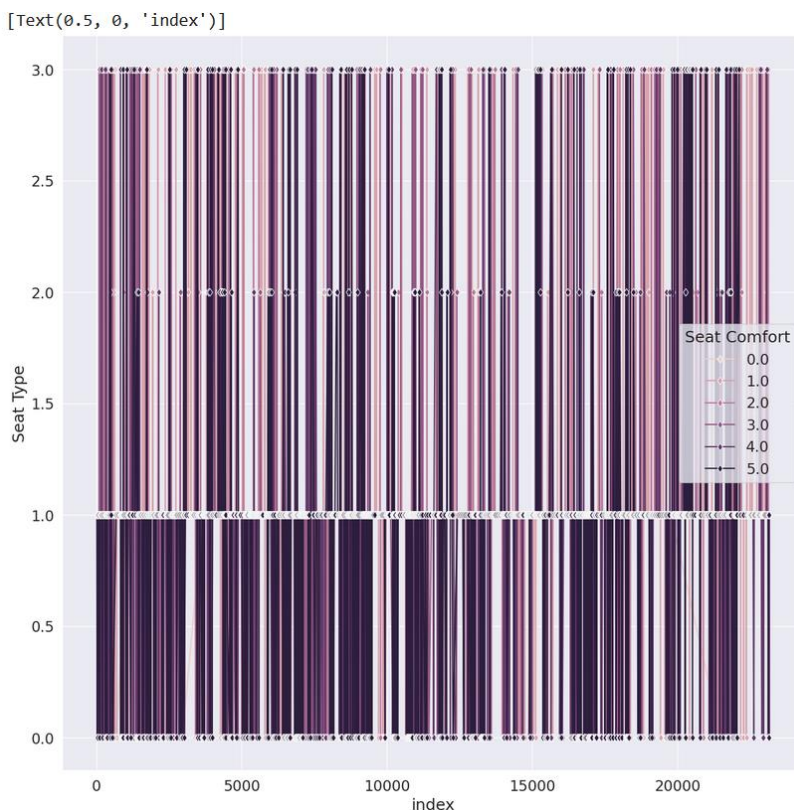
Date	15 July 2024
Team ID	740057
Project Title	Airline Review Classification
Maximum Marks	6 Marks

Data Exploration and Preprocessing Template

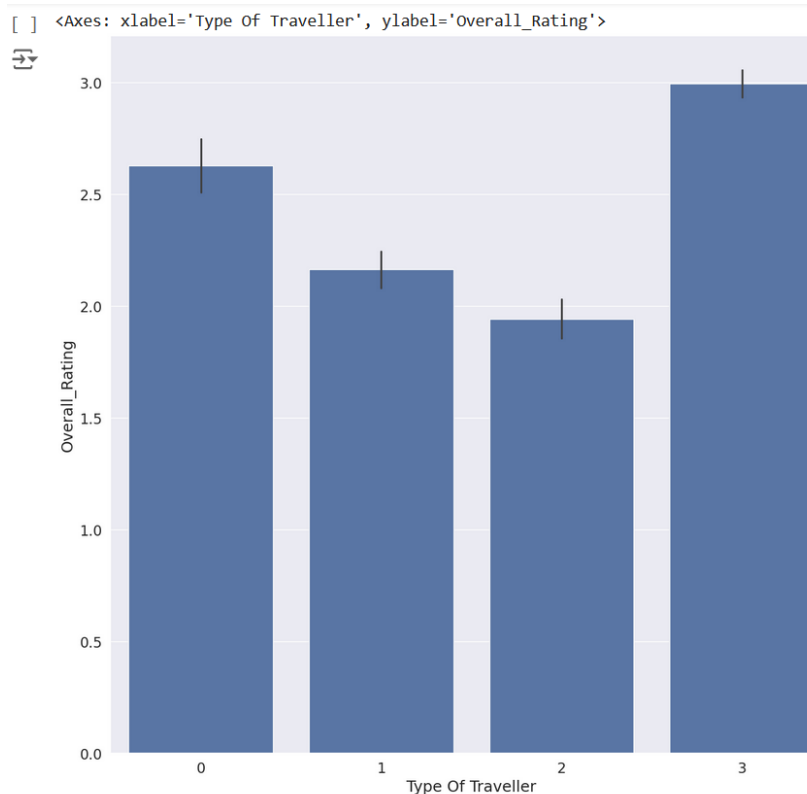
Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

Section	Description																																																																																				
Data Overview	<div><div></div><div>Data columns (total 20 columns):</div><table><thead><tr><th>#</th><th>Column</th><th>Non-Null Count</th><th>Dtype</th></tr></thead><tbody><tr><td>0</td><td>Unnamed: 0</td><td>23171 non-null</td><td>int64</td></tr><tr><td>1</td><td>Airline Name</td><td>23171 non-null</td><td>object</td></tr><tr><td>2</td><td>Overall_Rating</td><td>23171 non-null</td><td>object</td></tr><tr><td>3</td><td>Review_Title</td><td>23171 non-null</td><td>object</td></tr><tr><td>4</td><td>Review Date</td><td>23171 non-null</td><td>object</td></tr><tr><td>5</td><td>Verified</td><td>23171 non-null</td><td>bool</td></tr><tr><td>6</td><td>Review</td><td>23171 non-null</td><td>object</td></tr><tr><td>7</td><td>Aircraft</td><td>7129 non-null</td><td>object</td></tr><tr><td>8</td><td>Type Of Traveller</td><td>19433 non-null</td><td>object</td></tr><tr><td>9</td><td>Seat Type</td><td>22075 non-null</td><td>object</td></tr><tr><td>10</td><td>Route</td><td>19343 non-null</td><td>object</td></tr><tr><td>11</td><td>Date Flown</td><td>19417 non-null</td><td>object</td></tr><tr><td>12</td><td>Seat Comfort</td><td>19016 non-null</td><td>float64</td></tr><tr><td>13</td><td>Cabin Staff Service</td><td>18911 non-null</td><td>float64</td></tr><tr><td>14</td><td>Food & Beverages</td><td>14500 non-null</td><td>float64</td></tr><tr><td>15</td><td>Ground Service</td><td>18378 non-null</td><td>float64</td></tr><tr><td>16</td><td>Inflight Entertainment</td><td>10829 non-null</td><td>float64</td></tr><tr><td>17</td><td>Wifi & Connectivity</td><td>5920 non-null</td><td>float64</td></tr><tr><td>18</td><td>Value For Money</td><td>22105 non-null</td><td>float64</td></tr><tr><td>19</td><td>Recommended</td><td>23171 non-null</td><td>object</td></tr></tbody></table><div>dtypes: bool(1), float64(7), int64(1), object(11)</div></div>	#	Column	Non-Null Count	Dtype	0	Unnamed: 0	23171 non-null	int64	1	Airline Name	23171 non-null	object	2	Overall_Rating	23171 non-null	object	3	Review_Title	23171 non-null	object	4	Review Date	23171 non-null	object	5	Verified	23171 non-null	bool	6	Review	23171 non-null	object	7	Aircraft	7129 non-null	object	8	Type Of Traveller	19433 non-null	object	9	Seat Type	22075 non-null	object	10	Route	19343 non-null	object	11	Date Flown	19417 non-null	object	12	Seat Comfort	19016 non-null	float64	13	Cabin Staff Service	18911 non-null	float64	14	Food & Beverages	14500 non-null	float64	15	Ground Service	18378 non-null	float64	16	Inflight Entertainment	10829 non-null	float64	17	Wifi & Connectivity	5920 non-null	float64	18	Value For Money	22105 non-null	float64	19	Recommended	23171 non-null	object
#	Column	Non-Null Count	Dtype																																																																																		
0	Unnamed: 0	23171 non-null	int64																																																																																		
1	Airline Name	23171 non-null	object																																																																																		
2	Overall_Rating	23171 non-null	object																																																																																		
3	Review_Title	23171 non-null	object																																																																																		
4	Review Date	23171 non-null	object																																																																																		
5	Verified	23171 non-null	bool																																																																																		
6	Review	23171 non-null	object																																																																																		
7	Aircraft	7129 non-null	object																																																																																		
8	Type Of Traveller	19433 non-null	object																																																																																		
9	Seat Type	22075 non-null	object																																																																																		
10	Route	19343 non-null	object																																																																																		
11	Date Flown	19417 non-null	object																																																																																		
12	Seat Comfort	19016 non-null	float64																																																																																		
13	Cabin Staff Service	18911 non-null	float64																																																																																		
14	Food & Beverages	14500 non-null	float64																																																																																		
15	Ground Service	18378 non-null	float64																																																																																		
16	Inflight Entertainment	10829 non-null	float64																																																																																		
17	Wifi & Connectivity	5920 non-null	float64																																																																																		
18	Value For Money	22105 non-null	float64																																																																																		
19	Recommended	23171 non-null	object																																																																																		

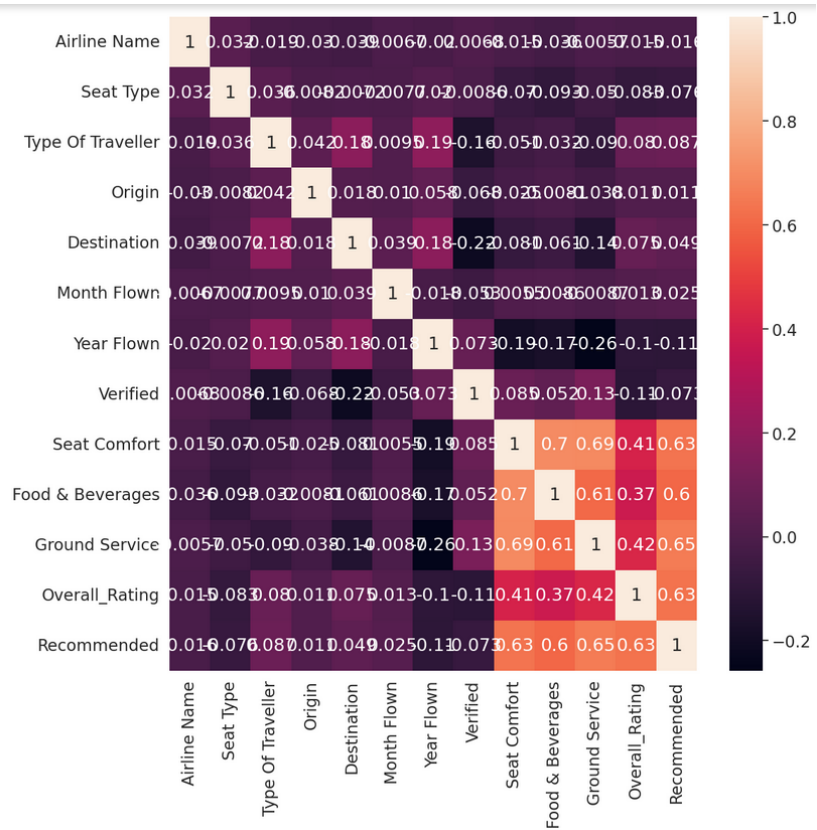
Univariate Analysis



Bivariate Analysis



Multivariate Analysis



Outliers and Anomalies

-

Data Preprocessing Code Screenshots

Loading Data

```
warnings.filterwarnings('ignore')
ar=pd.read_csv(r"C:\Users\Architha Rao\Downloads\archive\Airline_Reviews.csv")
ar.head(5)
ar.shape
ar.info()
nar = ar.drop(['Inflight Entertainment', 'Wifi & Connectivity', 'Aircraft', 'Value For Money
```

Handling Missing Data

```
ar.info()
nar = ar.drop(['Inflight Entertainment', 'Wifi & Connectivity', 'Aircraft', 'Value For Money
nar['Overall_Rating']=nar['Overall_Rating'].replace(['1', '2', '3', '4', '5', '6', '7', '8',
nar['Type Of Traveller']=nar['Type Of Traveller'].fillna(nar['Type Of Traveller'].mode()[0])
nar['Seat Type']=nar['Seat Type'].fillna(nar['Seat Type'].mode()[0])
nar['Seat Comfort']=nar['Seat Comfort'].fillna(nar['Seat Comfort'].mode()[0])
nar['Route']=nar['Route'].fillna(nar['Route'].mode()[0])
nar['Date Flown']=nar['Date Flown'].fillna(nar['Date Flown'].mode()[0])
nar['Food & Beverages']=nar['Food & Beverages'].fillna(nar['Food & Beverages'].mode()[0])
nar['Ground Service']=nar['Ground Service'].fillna(nar['Ground Service'].mode()[0])
# For the above columns we are using mode instead median even though numerical values are pre
```

Data Transformation	<pre># For the above columns we are using mode instead median even though numerical values are pr # Because the cloumn consists of categories (0 to 5). So its considered as categorial data nar[['Month Flown', 'Year Flown']]=nar['Date Flown'].str.split(expand=True) nar['Origin']=nar.Route.str.split('to',expand=True)[0] nar['Destination']=nar.Route.str.split('to',expand=True)[1]</pre>
Feature Engineering	Attached code in final submission.
Save Processed Data	-