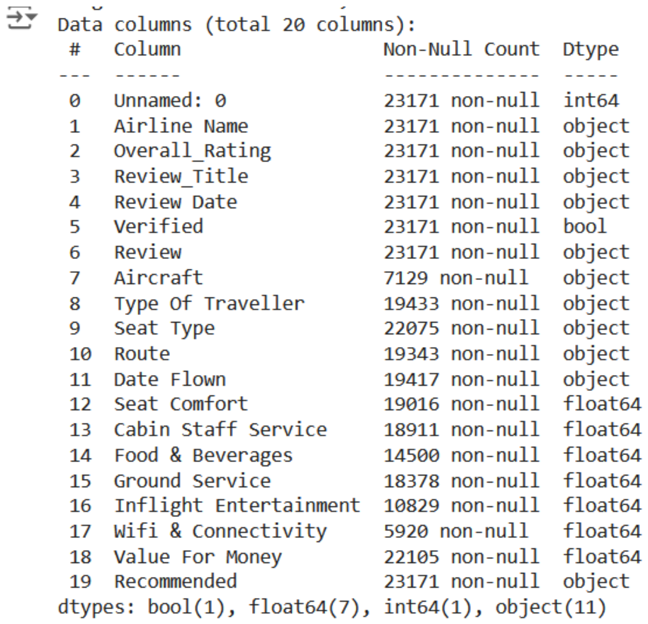
Data Collection and Preprocessing Phase

Date	15 July 2024
Team ID	739653
Project Title	Airline Review Classification
Maximum Marks	6 Marks

Data Exploration and Preprocessing Template

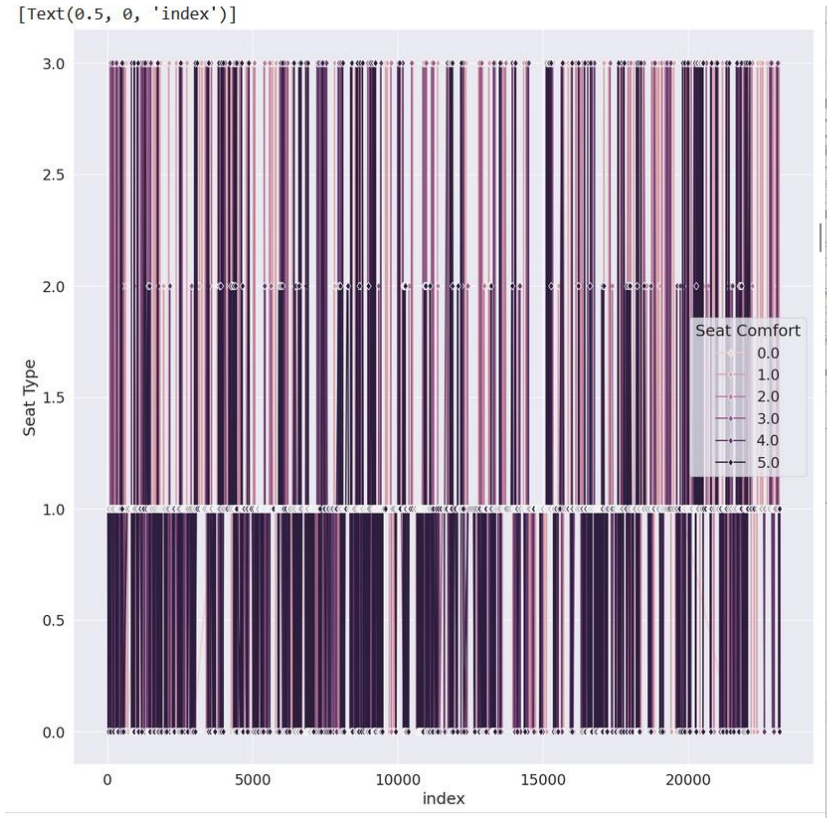
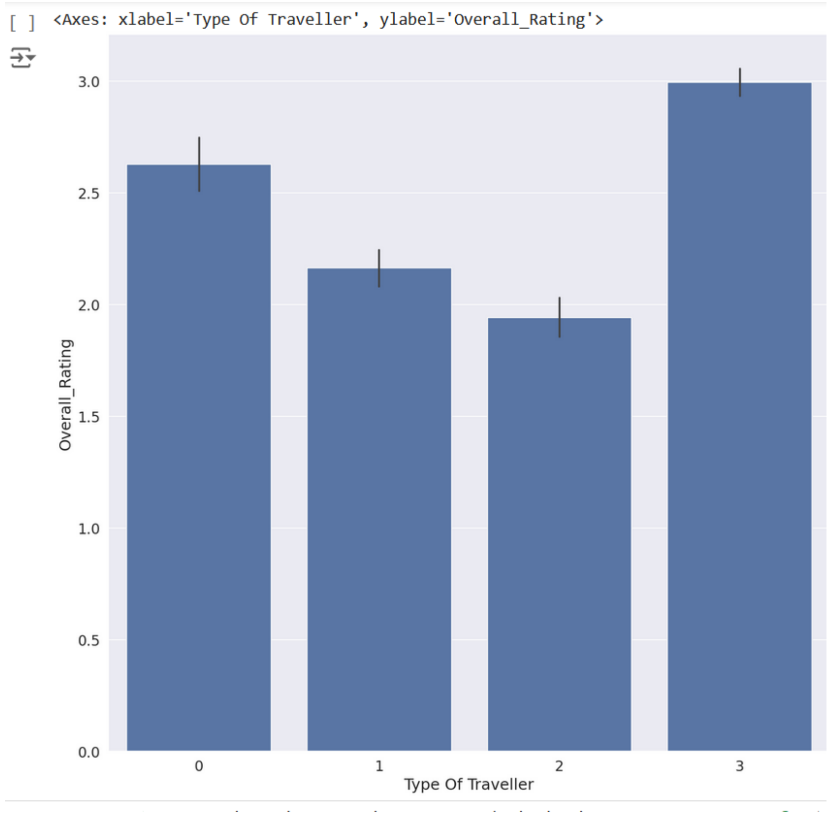
Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

<table>
<tr><th>Section</th><th>Description</th></tr>
<tr><td>Data Overview</td><td>

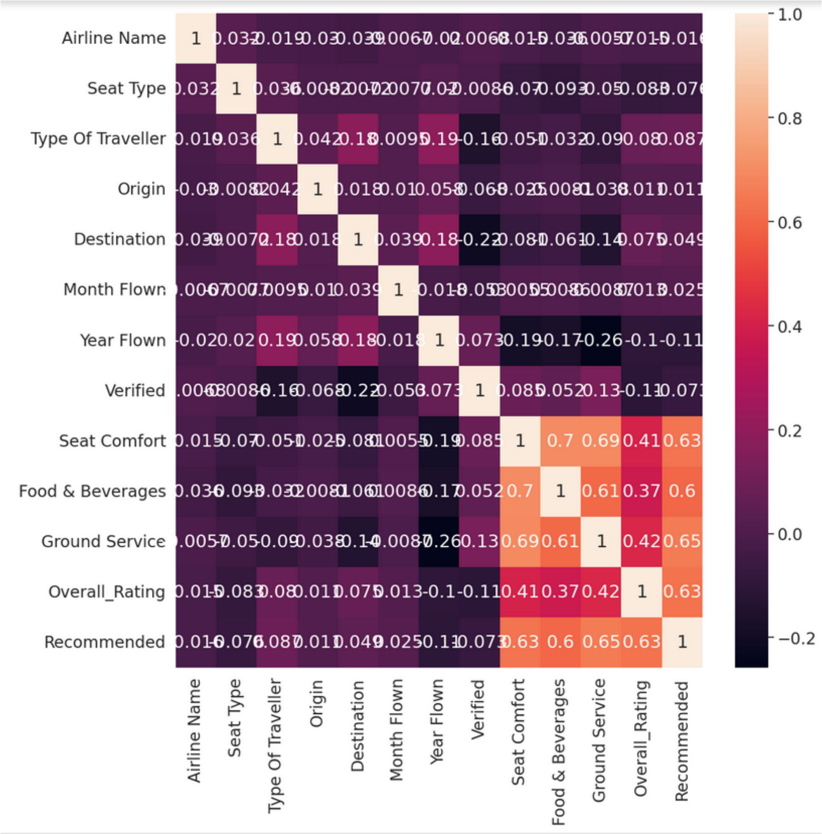


```
Data columns (total 20 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Unnamed: 0              23171 non-null  int64
 1   Airline Name            23171 non-null  object
 2   Overall_Rating          23171 non-null  object
 3   Review_Title            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
dtypes: bool(1), float64(7), int64(1), object(11)
```

</td></tr>
</table>

Univariate Analysis	 [Text(0.5, 0, 'index')] Seat Type 3.0 2.5 2.0 1.5 1.0 0.5 0.0 0 5000 10000 15000 20000 index Seat Comfort 0.0 1.0 2.0 3.0 4.0 5.0
Bivariate Analysis	 [ ] <Axes: xlabel='Type Of Traveller', ylabel='Overall_Rating'> Overall_Rating 3.0 2.5 2.0 1.5 1.0 0.5 0.0 0 1 2 3 Type Of Traveller

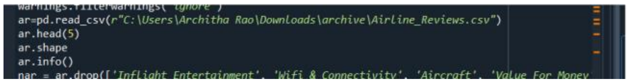
<table>
<tr>
<td>Multivariate Analysis</td>
<td>



</td>
</tr>
<tr>
<td>Outliers and Anomalies</td>
<td>-</td>
</tr>
<tr>
<td colspan="2">

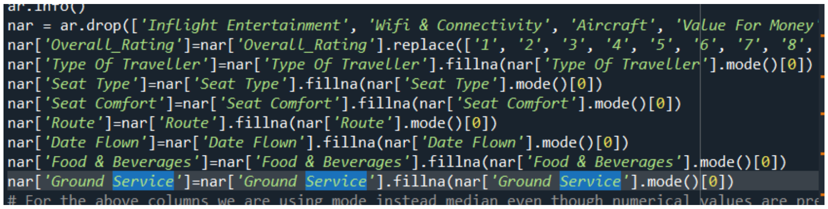
Data Preprocessing Code Screenshots

</td>
</tr>
<tr>
<td>Loading Data</td>
<td>



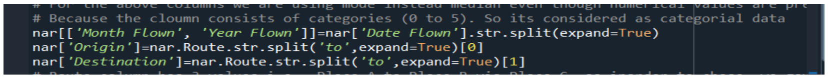
```
warnings.filterwarnings('ignore')
ar=pd.read_csv(r"C:\Users\Architha Rao\Downloads\archive\Airline_Reviews.csv")
ar.head(5)
ar.shape
ar.info()
nar = ar.drop(['Inflight Entertainment', 'Wifi & Connectivity', 'Aircraft', 'Value For Money
```

</td>
</tr>
<tr>
<td>Handling Missing Data</td>
<td>



```
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
```

</td>
</tr>
</table>

Data Transformation	 # For the above columns we are using mode instead median even though numerical values are pre # 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]
Feature Engineering	Attached code in final submission.
Save Processed Data	-