**Final Course Project**

1. We were asked to add additional features to the data set besides the ones you selected for approval in the following file: pairings and data sets - what we added is:

* **Total In-flight Service Score**: Combining ratings for various in-flight services (such as inflight wifi service, seat comfort, inflight entertainment, etc.) into a single score.
* **Total Ground Service Score**: Combining ratings for ground services (such as ease of online booking, check-in service, baggage handling, etc.) into a single score.
* **Total Delay Score**: Combining departure delay and arrival delay into a single score to measure overall delay experience.
* **Travel Distance Category**: Categorizing flight distance into short, medium, and long-haul flights.
* **Overall Satisfaction Score**: Combining ratings for various aspects of the flight experience (in-flight service, ground service, etc.) into a single satisfaction score.
* **Time of Day**: Extracting information from departure/arrival times to categorize flights into morning, afternoon, evening, and night flights.

1. Describe your dataset & features before processing it –

* Display the first few rows of the dataset: Gives us a glimpse of the data and its structure.
* Check the data types of each column: Tells us the data type of each feature (column) in the dataset.
* Summary statistics for numerical columns: Provides summary statistics (such as mean, median, min, max, etc.) for numerical features.
* Summary of categorical columns: Provides summary statistics for categorical features.
* Check for missing values: Shows the number of missing values in each column.
* Unique values in categorical columns: Lists the unique values for each categorical feature.
* Target variable distribution: Shows the distribution of the target variable (assuming "satisfaction" is your target variable).

1. Describe the process of cleansing, fixing, prepping the dataset –

* **Handling Missing Values**: Identify and deal with missing values in the dataset. Depending on the extent of missing data, you can choose to drop rows or columns with missing values, impute missing values using statistical measures (such as mean, median, or mode), or use advanced techniques like predictive modeling to fill missing values.
* **Data Cleaning**: Clean the data to address inconsistencies, errors, or outliers. This may involve correcting typos, standardizing formats, removing duplicates, and handling outliers that could skew the analysis.
* **Feature Engineering**: Create new features or transform existing ones to extract more relevant information from the dataset. This can include creating dummy variables for categorical features, scaling numerical features, or generating new features through mathematical transformations.
* **Handling Categorical Variables**: Convert categorical variables into a suitable format for analysis. This may involve one-hot encoding, label encoding, or ordinal encoding depending on the nature of the categorical data and the requirements of the analysis.
* **Normalization and Scaling:** Normalize or scale numerical features to ensure that they have similar ranges and distributions. Common techniques include Min-Max scaling or Standard scaling.
* **Data Validation:** Validate the integrity and consistency of the dataset to ensure that it meets the requirements of the analysis. This involves checking for logical inconsistencies, verifying relationships between variables, and validating assumptions made about the data.
* **Splitting Data:** If you're planning to build predictive models, split the dataset into training and testing sets to evaluate the performance of the models on unseen data.
* **Feature Selection:** Select the most relevant features for analysis or modeling. This can be done using statistical tests, feature importance techniques, or domain knowledge.
* **Data Transformation:** Perform any necessary data transformations or normalization to prepare the dataset for modeling. This may include transforming skewed distributions, handling multicollinearity, or applying dimensionality reduction techniques.
* **Documentation:** Document the steps taken during the cleansing, fixing, and prepping process to ensure reproducibility and transparency in the analysis.

1. Describe your dataset & features after processing it –

After processing the dataset, it's essential to re-evaluate and describe the dataset and its features to understand how the preprocessing steps have affected the data. Here's how you can describe the dataset and its features after processing:

* Display the first few rows of the processed dataset: Show a glimpse of the data to ensure that the preprocessing steps haven't drastically altered its structure.
* Check the data types of each column: Verify that the data types of each feature remain consistent after preprocessing.
* Summary statistics for numerical columns: Re-calculate summary statistics for numerical features to see if there are any significant changes in their distributions or central tendencies.
* Summary of categorical columns: Re-evaluate summary statistics for categorical features to ensure that the preprocessing steps have not introduced any unexpected changes.
* Check for missing values: Confirm that missing values have been appropriately handled and that there are no remaining missing values in the dataset.
* Unique values in categorical columns: Check if the unique values in categorical features have changed after preprocessing, especially if encoding or transformation techniques were applied.
* Target variable distribution: Examine the distribution of the target variable to see if it has been affected by the preprocessing steps. Ensure that the distribution remains balanced and suitable for analysis or modeling.
* Feature engineering: Describe any new features that were created during the preprocessing steps and explain their significance in the analysis.
* Normalization and scaling: If normalization or scaling was applied to numerical features, describe the techniques used and assess their impact on the data.
* Feature selection: If feature selection techniques were applied, describe the selected features and justify their inclusion in the final dataset.

1. Describe distribution of interesting features and what can be learned about them –

For example, in the context of an airline passenger satisfaction survey dataset:

* **Flight Distance:** Visualize the distribution of flight distances to understand the typical range of distances traveled by passengers.
* **Satisfaction Ratings:** Explore the distribution of satisfaction ratings for different aspects of the flight experience (e.g., in-flight service, ground service) to identify areas of strength or improvement.
* **Delay Scores:** Examine the distribution of delay scores to understand the prevalence and impact of flight delays on passenger satisfaction.

1. Assume multiple assumptions about the data, explain why you made those assumptions based on the work you’ve done so far:

* **Assumption**: Higher levels of inflight wifi service are associated with higher overall passenger satisfaction.

**Explanation**: Based on initial exploration, you observed that passengers who rated inflight wifi service highly also tended to rate other aspects of the flight experience positively. Additionally, modern travelers often prioritize connectivity, suggesting that satisfaction with inflight wifi service may influence overall satisfaction.

* **Assumption**: Passengers traveling for business purposes are more likely to choose higher class tickets (e.g., business class) compared to those traveling for leisure.

**Explanation**: In your dataset, you noticed a higher proportion of passengers traveling for business purposes in higher class categories (e.g., business class) compared to passengers traveling for leisure. This observation aligns with common expectations that business travelers may prioritize comfort and amenities offered in higher class tickets.

* **Assumption**: Longer flight distances result in higher levels of passenger satisfaction.

**Explanation**: Your initial analysis revealed a positive correlation between flight distance and overall satisfaction ratings. This correlation may be attributed to the perception that longer flights offer more amenities, entertainment options, and comfort features, leading to higher levels of satisfaction among passengers.

* **Assumption**: Passengers who experience longer departure delays are more likely to report lower satisfaction ratings.

**Explanation**: Upon exploring the dataset, you found a negative correlation between departure delay times and passenger satisfaction ratings. This finding suggests that prolonged departure delays may contribute to passenger dissatisfaction, as delays can lead to inconvenience, missed connections, and overall negative flight experiences.

* **Assumption**: Female passengers are more likely to prioritize seat comfort compared to male passengers.

**Explanation**: Your analysis revealed differences in satisfaction ratings between male and female passengers, with female passengers consistently rating seat comfort higher. This observation may be indicative of differing preferences or comfort expectations between genders, influencing satisfaction ratings for seat comfort.

1. Prove and disprove the assumptions you have made :

* **Assumption**: Higher levels of inflight wifi service are associated with higher overall passenger satisfaction.

Approach to Prove:

Calculate the correlation coefficient between inflight wifi service ratings and overall satisfaction ratings.

Perform a hypothesis test to determine if there is a statistically significant positive relationship between the two variables.

Approach to Disprove:

Examine the distribution of overall satisfaction ratings for passengers with different levels of inflight wifi service (e.g., low, medium, high). If there is no clear trend or if satisfaction ratings are lower for higher levels of inflight wifi service, it could disprove the assumption.

* **Assumption**: Passengers traveling for business purposes are more likely to choose higher class tickets (e.g., business class) compared to those traveling for leisure.

Approach to Prove:

Calculate the proportion of passengers traveling for business purposes in each class category (e.g., economy, business, first class). If the proportion of business travelers is highest in the business class category, it would support the assumption.

Approach to Disprove:

Conduct a chi-square test of independence to determine if there is a significant association between travel purpose (business vs. leisure) and class category. If there is no significant association, it could disprove the assumption.

* **Assumption**: Longer flight distances result in higher levels of passenger satisfaction.

Approach to Prove:

Calculate the correlation coefficient between flight distances and overall satisfaction ratings. A positive correlation would support the assumption.

Approach to Disprove:

Compare the distribution of satisfaction ratings for short, medium, and long flight distances. If there is no clear trend or if satisfaction ratings are lower for longer flight distances, it could disprove the assumption.

Assumption: Passengers who experience longer departure delays are more likely to report lower satisfaction ratings.

Approach to Prove:

Calculate the correlation coefficient between departure delay times and satisfaction ratings. A negative correlation would support the assumption.

Approach to Disprove:

Compare the distribution of satisfaction ratings for different levels of departure delay times. If there is no clear trend or if satisfaction ratings are not consistently lower for longer departure delays, it could disprove the assumption.

* **Assumption**: Female passengers are more likely to prioritize seat comfort compared to male passengers.

Approach to Prove:

Compare the mean satisfaction ratings for seat comfort between male and female passengers. If female passengers consistently rate seat comfort higher than male passengers, it would support the assumption.

Approach to Disprove:

Conduct a t-test to compare the mean satisfaction ratings for seat comfort between male and female passengers. If there is no significant difference in satisfaction ratings between the two groups, it could disprove the assumption.

1. Find interesting correlations between features, explain these correlations and prove your claims using plots -