Report for mult-class classification dataset

Travel Behavior Insights

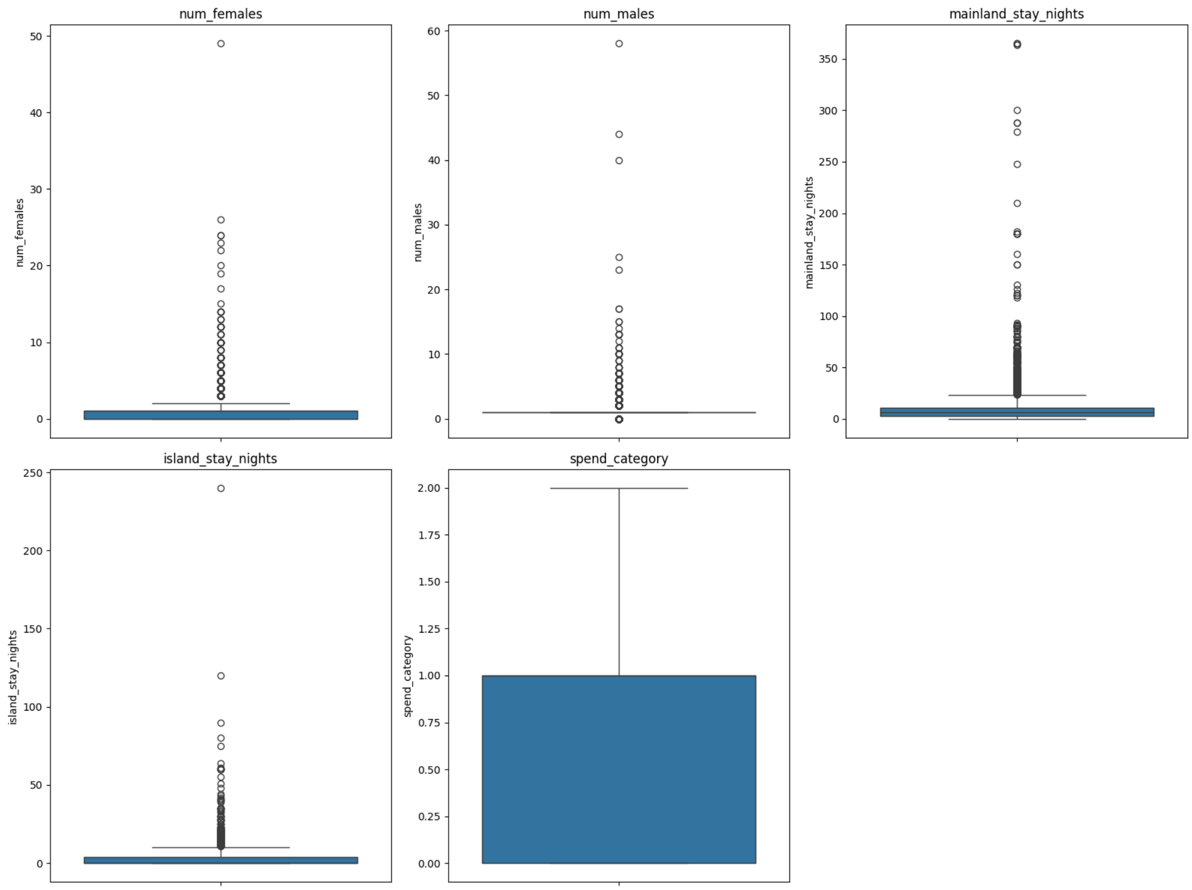
1. Objective

The goal of this project is to predict the spend category of travelers using trip-related information such as travel companions, stay duration, activities, and accommodations. This is a **multi-class classification problem**, and the aim is to identify models that generalize well on unseen test data.

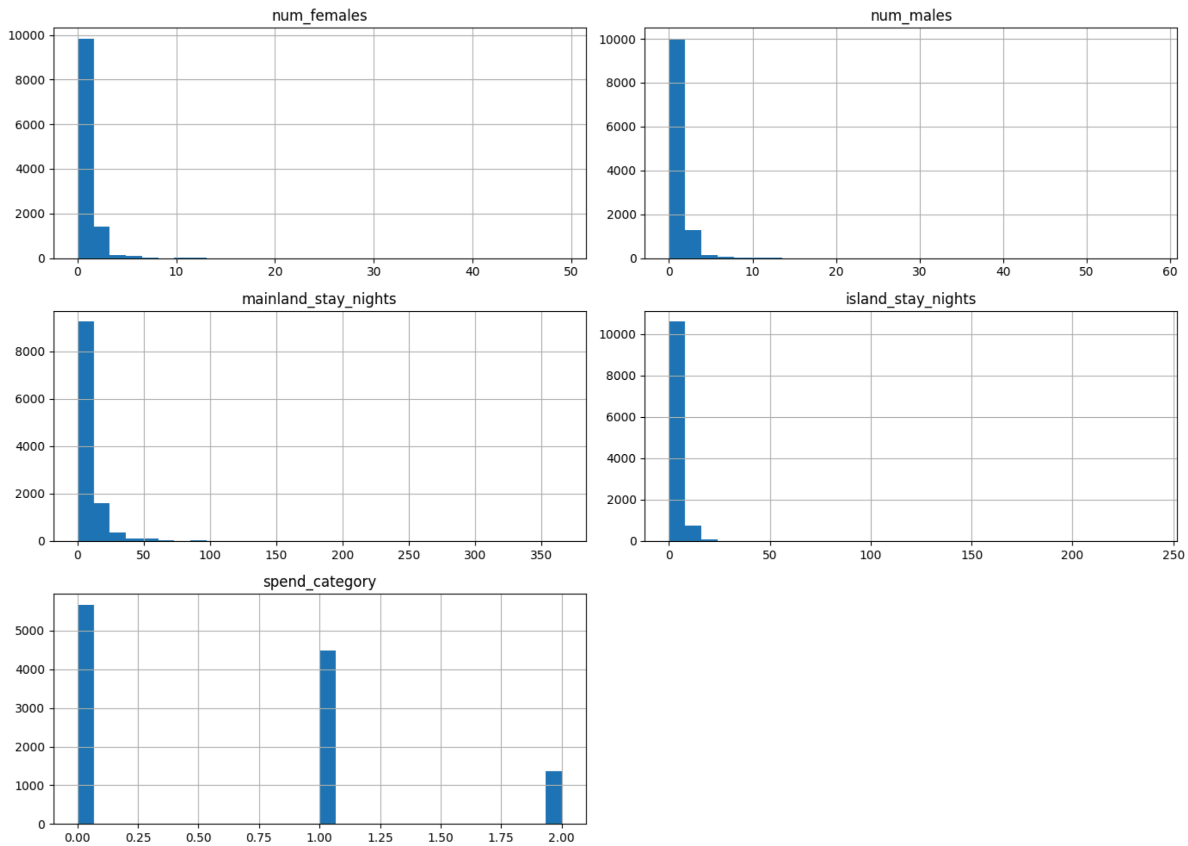
---

2. Exploratory Data Analysis (EDA)

- Dataset contains ~12k rows with numeric and categorical features.
- Target variable: `spend_category`.
- Key observations:
  - `num_males` and `num_females` are small integers with occasional outliers.
  - Stay durations (`mainland_stay_nights`, `island_stay_nights`) are skewed.
  - Many categorical features have missing values (<2%) or redundant information.
- Visualizations used:
  - **Boxplots**: detect outliers and determine clipping thresholds.



- ◦ **Histograms**: examine distribution and skewness.



- **Summary statistics**: verify ranges and detect anomalies.

	num_females	num_males	mainland_stay_nights	island_stay_nights
count	11505.000000	11505.000000	11505.000000	11505.000000
mean	0.949066	1.012169	9.206780	2.522555
std	1.295324	1.273400	14.868802	5.170178
min	0.000000	0.000000	0.000000	0.000000
25%	0.000000	1.000000	3.000000	0.000000
50%	1.000000	1.000000	6.000000	0.000000
75%	1.000000	1.000000	11.000000	4.000000
max	49.000000	58.000000	365.000000	240.000000

	spend_category
count	11505.000000
mean	0.625120
std	0.686026
min	0.000000
25%	0.000000
50%	1.000000
75%	1.000000
max	2.000000

3. Handling Missing Values

- Rows with missing target (`spend_category`) were dropped.
- Columns with >40% missing values were dropped.
- Rows with <2% missing in a column were dropped.

- Manual imputations for categorical features:
  - `travel_companions`: “Alone”
  - `days_booked_before_trip`: “61-90”
- Columns dropped for irrelevance:
  - `arrival_weather`
- After handling missing values, about ~11k rows were left.

4. Feature Engineering

- **Outlier clipping (winsorization)**:
  - `num_males`, `num_females`: clipped at [0, 5]
  - `mainland_stay_nights`: clipped at [0, 30]
  - `island_stay_nights`: clipped at [0, 21]
- **Derived features**:
  - `total_people = num_males + num_females`
- Dropped `num_males` and `num_females` (summarized in `total_people`).
- **Encoding & scaling**:
  - Categorical: One-hot encoding
  - Numeric: StandardScaler + PolynomialFeatures (degree=2) for logistic regression; RobustScaler for SVM, NN, Naive Bayes

5. Models, Hyperparameters, and Test Scores

Model	Preprocessing	Best Parameters	Test Dataset Accuracy	Notes
Logistic Regression	StandardScaler + PolyFeatures	`max_iter=1000`, `multi_class='multinomial'`, `class_weight='balanced'`, `solver='lbfgs'`	0.704	Polynomial features helped capture non-linear relationships.

Model	Preprocessing	Best Parameters	Test Dataset Accuracy	Notes
SVM (RBF)	RobustScaler	`kernel='rbf'`, `C=3`, `gamma='scale'`, `class_weight='balanced'`	0.706	Performed best; handles small dataset + high-dimensional one-hot features well.
Neural Network (MLPClassifier)	RobustScaler	`hidden_layer_sizes=(64,32)`, `activation='relu'`, `solver='adam'`, `learning_rate='adaptive'`, `batch_size=32`, `alpha=0.0005`, `max_iter=500`, `early_stopping=True`	0.687	Slight overfitting observed; early stopping helped stabilize training.
Naive Bayes (GaussianNB)	RobustScaler	Default	0.247	Poor performance due to strong independence assumptions on mixed-type features.

Observations:

- **SVM achieved the highest test accuracy (0.706)**, likely due to its ability to handle a mix of categorical and numeric features after robust scaling.
- Neural network underperformed due to **limited dataset size (~11k rows)**.
- Logistic regression improved with polynomial features but could not surpass SVM.
- Naive Bayes assumptions do not hold, leading to very poor performance.
