Data Preparation Report (Stage 2-1)

1. Handling Missing Data

- **Numeric columns**: Missing values imputed using the **median** strategy (robust to outliers).
- **Categorical columns**: Missing values imputed using the **most frequent** (mode) value.

Output File:
- `01_missing_data_handled.csv` → Cleaned dataset with all missing values handled.

---

2. Encoding Categorical Variables

- Applied **One-Hot Encoding** to convert categorical variables into numeric format.
- Used `drop="first"` to avoid the dummy variable trap (reduces multicollinearity).
- Used `sparse_output=False` so encoded results are stored as a dense array (compatible with pandas DataFrame).

Output File:
- `02_categorical_encoded.csv` → Dataset with categorical features encoded into numeric columns.

---

3. Deliverables Summary

1. **01_missing_data_handled.csv** → Preprocessed dataset after handling missing values.

2. **02_categorical_encoded.csv** → Preprocessed dataset after encoding categorical variables.

These outputs complete the Stage 2-1 requirements: - Preprocessed dataset with missing data handled.
- Encoded dataset ready for further steps (scaling, splitting, clustering).

Notes

- This script is designed for execution in Google Colab with Google Drive mounted.

- File paths may be adjusted for local or other environments.
- Target column (e.g., *Churn*) should be excluded from encoding if required by later modeling tasks.