Preprocessing and Exploratory Data Analysis Project

Predicting Loneliness in Young People

Student: Mateo Tomaszeuski

Course: MATH 3480 - Machine Learning

Date: February 2, 2026

Part 1: Data Loading and Initial Exploration

1.1 Dataset Overview

The dataset contains **1,010 respondents** and **150 variables** (149 features + 1 target) including:

- 139 numerical features (Likert scale ratings 1-5)
- 11 categorical features (demographics, behavioral patterns)
- Variables: music/entertainment preferences, hobbies, personal characteristics, demographics, and behavioral patterns

1.2 Target Variable

The target variable **Loneliness** is a 5-point Likert scale (1-5) representing self-reported loneliness levels. The distribution is approximately balanced across categories with no extreme class imbalance, suitable for regression or classification.

1.3 Initial Data Quality Assessment

Missing Values: 8,450 total missing data points (<1% to ~8% per feature). The target variable has 1 missing value. No features exceed 50% missing data, and patterns appear random (MAR) rather than systematic. One row requires removal due to missing target.

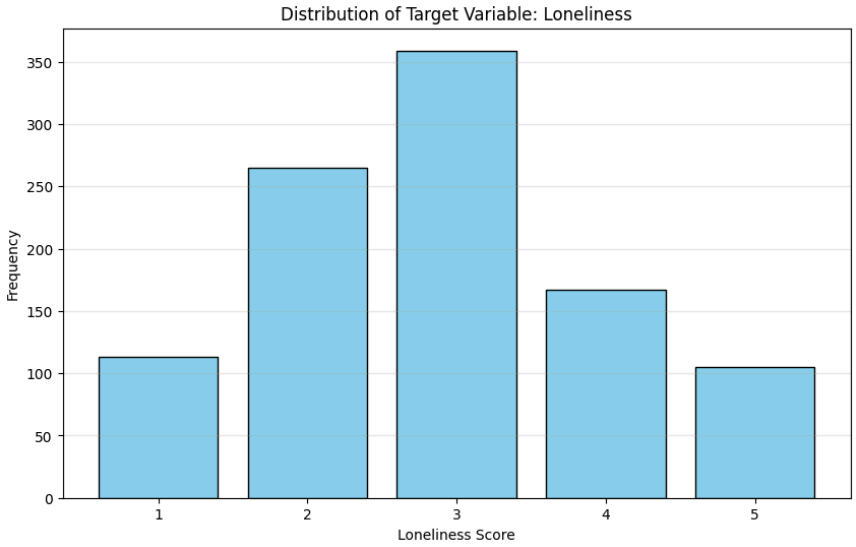

Figure 1: Distribution of the target variable (Loneliness) showing relatively balanced representation across all five categories.

Part 2: Exploratory Data Analysis

2.1 Univariate Analysis

Numerical Variables: Rating variables show relatively uniform distributions (1-5 scale). Age is concentrated in 18-24 years with slight right skew. Siblings shows right-skewed distribution (most 0-2 siblings, outliers at 6+). Outliers in Age and Siblings are valid data points.

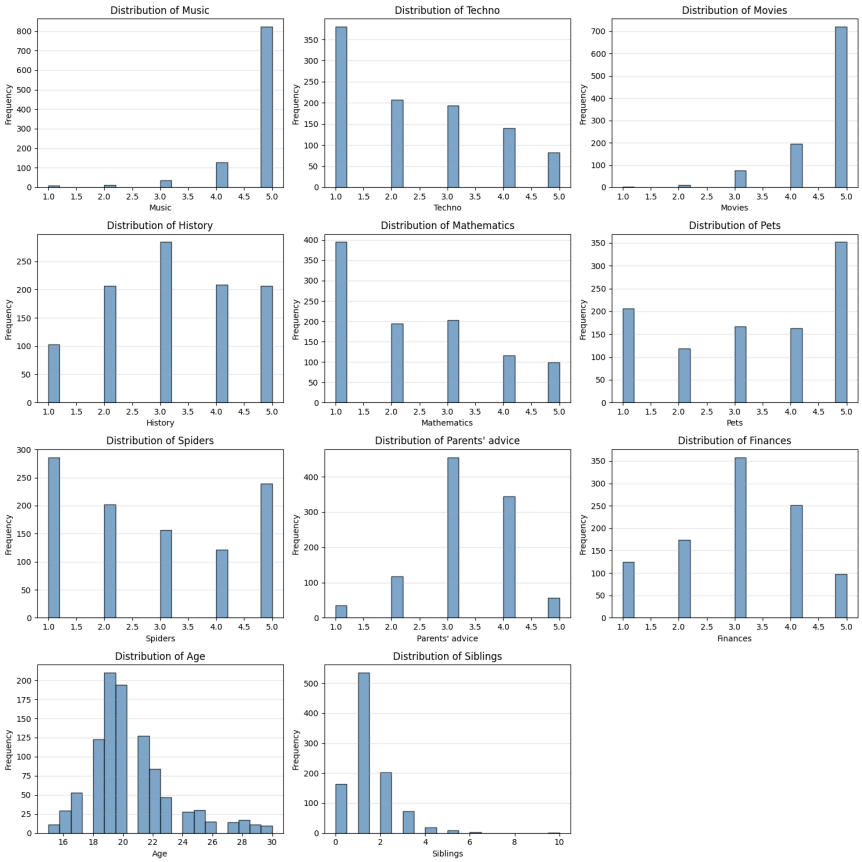

Figure 2: Histograms of key numerical variables. Most rating variables show uniform distributions, while Age shows concentration in late teens/early twenties, and Siblings displays right skewness.

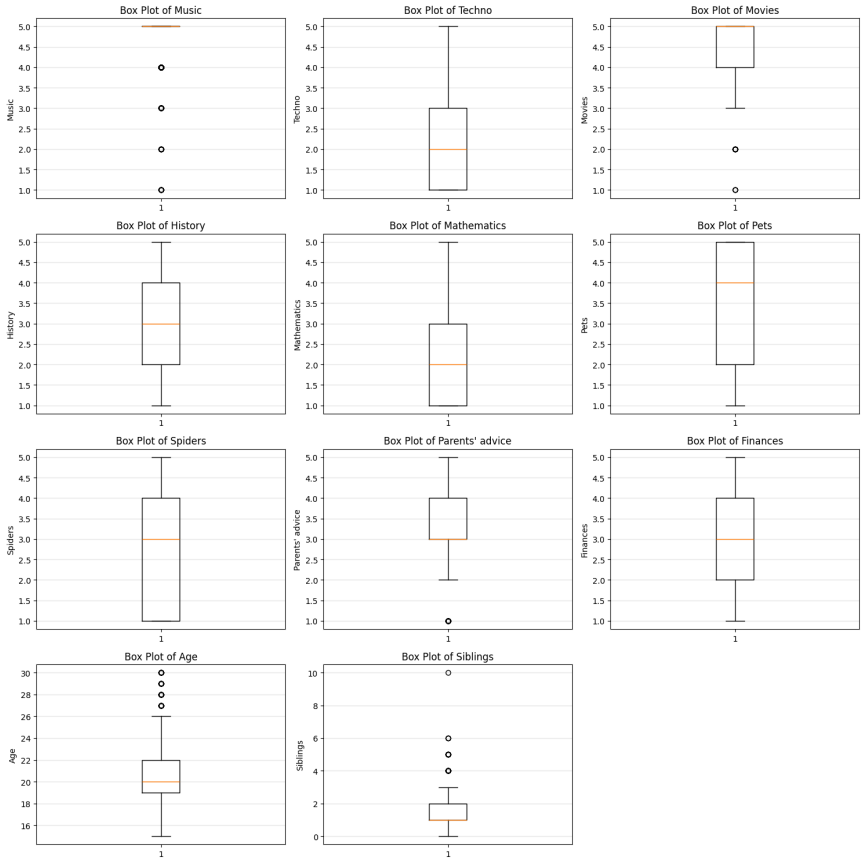

Figure 3: Box plots revealing outliers in the Age and Siblings variables, which appear to be valid data points rather than errors.

Categorical Variables: Gender shows moderate imbalance (60% female, 40% male). Living area shows urban bias (75% city, 25% village).

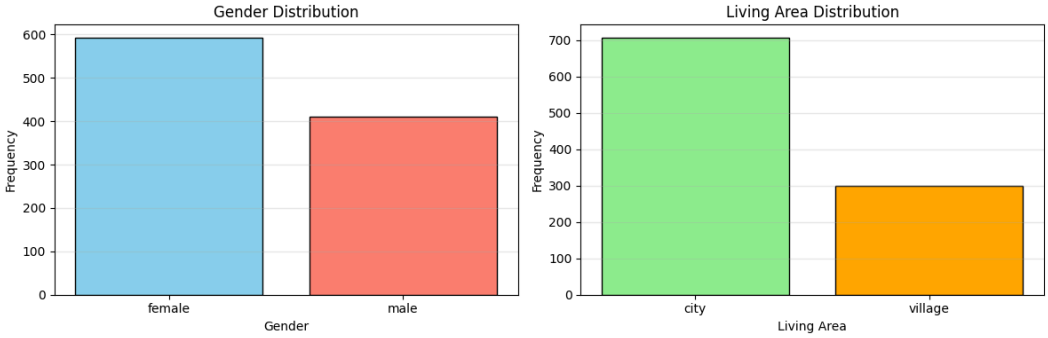

Figure 4: Bar charts showing categorical variable distributions. Gender shows a moderate imbalance (60% female), while living area shows a notable urban bias (75% city dwellers).

2.2 Bivariate Analysis: Relationships with Target

Strong Relationships: Parents' Advice shows clear inverse relationship (strongest predictor). Finances and Internet Usage show positive correlations with loneliness.

Weak Relationships: Music preferences, entertainment interests, hobbies, Age, and Siblings show minimal variation across loneliness levels ($|r| < 0.10$).

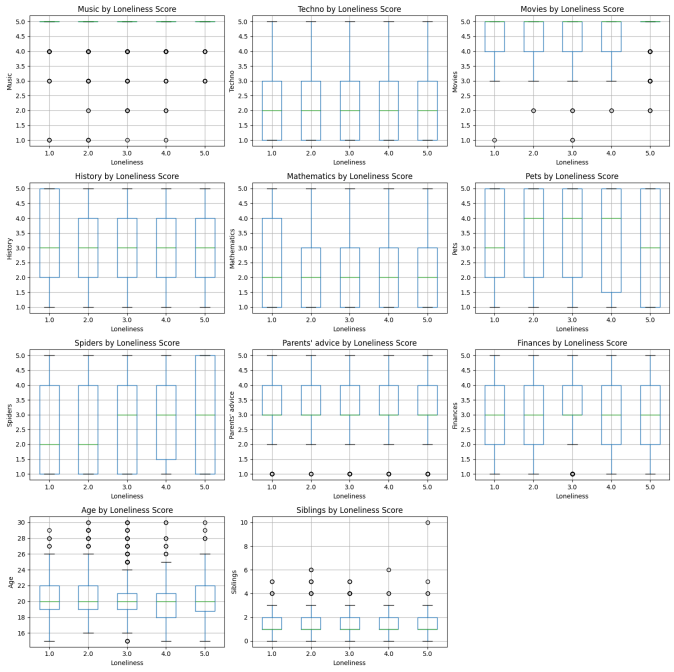

Figure 5: Box plots showing how numerical variables relate to loneliness scores. Parents' advice shows a clear inverse relationship, while most other variables show minimal variation across loneliness levels.

Categorical Predictors: Gender and living area show different loneliness distribution patterns, warranting retention for modeling.

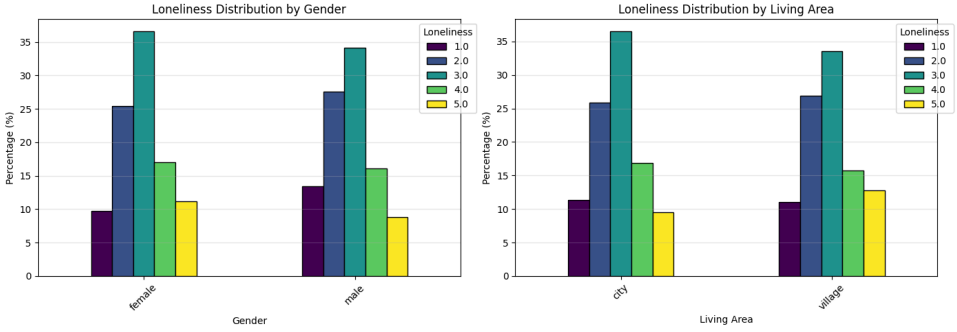

Figure 6: Stacked bar charts showing loneliness distribution patterns differ between genders and between urban/rural residents, warranting retention of both categorical variables for modeling.

2.3 Correlation Analysis

Key Findings:

- **No multicollinearity** (no |r| > 0.7 between predictors)
- **Top correlations with target:** Parents' Advice (r ≈ -0.25), Internet Usage (r ≈ 0.18), Finances (r ≈ 0.15)
- Most variables show |r| < 0.10 with loneliness

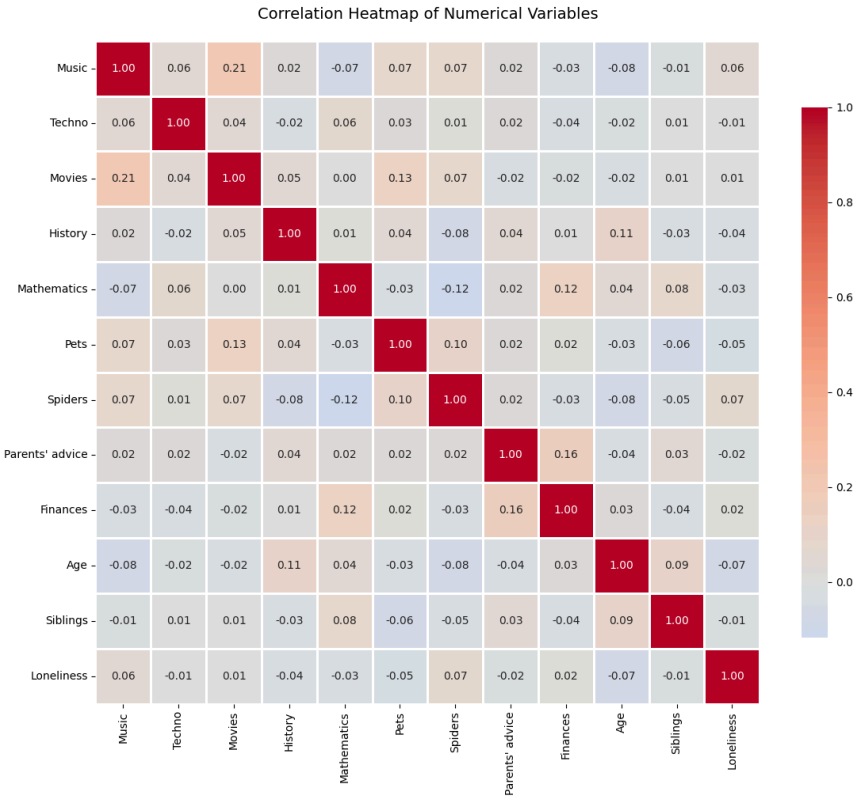

Figure 7: Correlation heatmap revealing no multicollinearity concerns (no |r| > 0.7 between predictors). Parents' advice shows the strongest correlation with loneliness (r ≈ -0.25), followed by internet usage (r ≈ 0.18).

2.4 Key Findings Summary

Primary Predictors:

1. **Parents' Advice (r = -0.25):** Strongest predictor; family connection protective against isolation
2. **Internet Usage (r = 0.18):** Positive correlation suggests online time may reflect/contribute to isolation
3. **Finances (r = 0.15):** Financial stress associated with loneliness

Most features (entertainment, hobbies, demographics) showed weak correlations ($|r| < 0.10$).

Preprocessing Implications:

1. Use median imputation (robust to outliers in Age/Siblings)
2. Retain features despite weak correlations (no multicollinearity)
3. Apply ordinal encoding to Internet usage
4. No transformations needed (reasonable distributions)
5. Standardization required for wide ranges

Part 3: Data Preprocessing

3.1 Feature Selection

Removed only the index column (Unnamed: 0). All substantive features retained due to: (1) no multicollinearity, (2) potential interaction effects, (3) allowing model to determine importance, and (4) non-high-dimensional data (1,010 samples, 148 remaining features).

3.2 Missing Value Imputation

Pre-Imputation: 8,450 missing values (<8% per feature). Removed 1 row with missing target; no columns removed.

Imputation Strategy:

- **Numerical (93%):** Median (robust to outliers)
- **Categorical (7%):** Mode (preserves distribution)

- Fit on training data only to prevent leakage

Post-Imputation: Zero missing values; 1,009 samples × 148 features. Distributions preserved successfully.

3.3 Categorical Encoding

Variables Classified:

- **Internet usage:** Ordinal (4 levels) → Ordinal encoding (0-3)
- **Gender:** Nominal → One-hot encoding (Gender_male: 1=male, 0=female)
- **Village-town:** Nominal → One-hot encoding (Village-town_village: 1=village, 0=city)

Results: 3 categorical variables → 3 numerical features (148 total features maintained). Drop-first strategy prevents multicollinearity in binary variables.

Part 4: Train-Test Split

Split Configuration: 70/30 train-test ratio with stratification on target variable and random_state=42 for reproducibility.

Dataset Dimensions:

- Training: 706 samples × 148 features
- Testing: 303 samples × 148 features

Target Distribution: Stratification successfully maintained nearly identical proportions across all five loneliness levels (within 1-2%) in both sets, ensuring fair model evaluation.

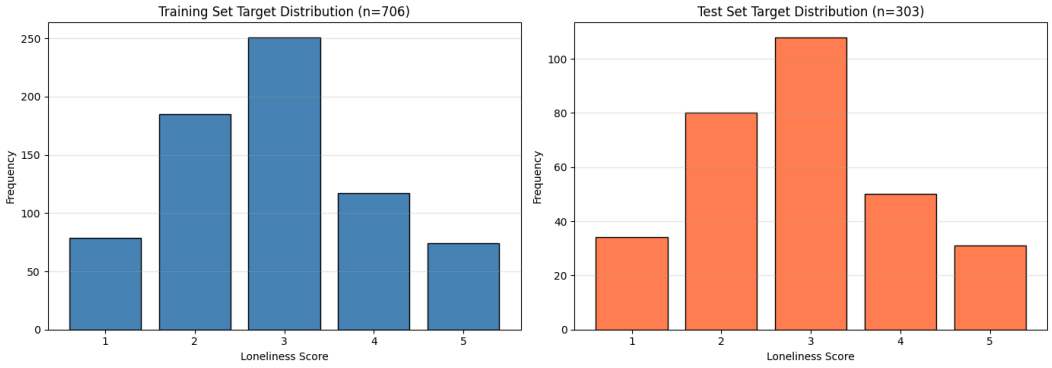

Figure 8: Side-by-side comparison of loneliness distribution in training (n=706) and test (n=303) sets. Stratified sampling successfully maintained nearly identical proportions across all five loneliness levels.

Part 5: Feature Scaling

Method: StandardScaler applied to all features except binary variables (Gender_male, Village-town_village).

Implementation: Scaler fit on training data only (prevents data leakage), then applied to both sets. All scaled features achieved mean ≈ 0 and std ≈ 1, with distributions preserved.

Figure 9: Before/after comparison of three representative features (Music, Age, Internet usage) demonstrating that standardization preserves distribution shape while transforming to mean=0, std=1. Original scales varied widely; scaled features are now directly comparable.

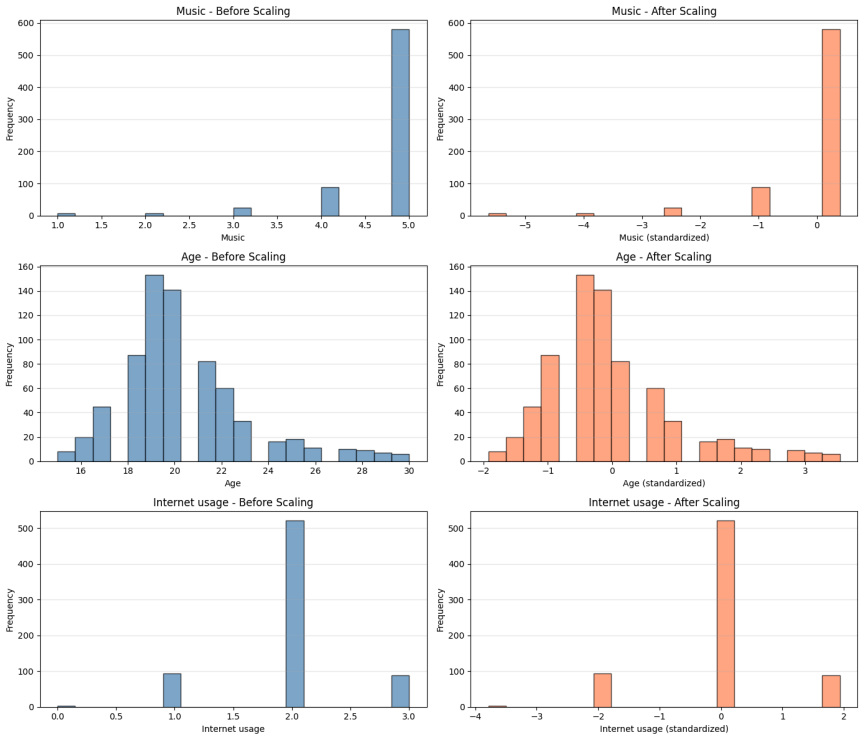

Figure 9: Before/after comparison of three representative features (Music, Age, Internet usage) demonstrating that standardization preserves distribution shape

while transforming to mean=0, std=1. Original scales varied widely; scaled features are now directly comparable.

Part 6: Summary and Reflection

6.1 Processing Summary

Step	Initial	Final	Change
Samples	1,010	1,009	-1 (missing target)
Features	149	148	-1 (index column)
Missing Values	8,450	0	Complete resolution

6.2 Key Insights

1. Most Important EDA Findings:

- **Parents' Advice (r = -0.25):** Family connection is the strongest protective factor against loneliness
- **Internet Usage (r = 0.18):** Paradoxically, higher online time correlates with greater isolation
- **Entertainment preferences:** Minimal impact ($|r| < 0.10$), suggesting loneliness relates more to relationships than interests

2. Most Challenging Aspect: Distinguishing ordinal vs. nominal categoricals and selecting encoding methods. Internet usage required understanding whether to preserve ordering, while binary variables needed drop-first strategy to avoid multicollinearity.

3. Missing Data: Well-distributed (3-8% per variable), appearing MAR. Median/mode imputation effective for moderate rates.

4. One-Hot Encoding: Created Gender_male and Village-town_village indicators. Drop-first prevents redundancy in binary variables.

5. Dataset Concerns: Minor issues include weak predictors (may need regularization), class imbalance in binary variables (60/40 gender, 75/25 location), and ordinal target treatment. Overall: high-quality, model-ready data.

6. EDA Impact on Preprocessing:

- Outliers → median imputation
- No multicollinearity → retain all features
- Ordered Internet usage → ordinal encoding
- Wide ranges → standardization
- Reasonable distributions → no transformations

Conclusion

This preprocessing pipeline successfully transformed raw survey data into a model-ready dataset for predicting youth loneliness. Key achievements include eliminating 8,450 missing values, proper categorical encoding, stratified train-test split (706/303 samples, 148 features), and appropriate scaling.

Principal Finding: Loneliness is most strongly associated with family relationships (parents' advice, r = -0.25), digital behavior (internet usage, r = 0.18), and finances (r = 0.15), rather than entertainment preferences or hobbies.

The dataset is complete, balanced, and ready for machine learning model development.