AI1010 Group Project Report

1 Data Exploration & Preprocessing

1.1 Dataset Overview

The dataset contains **35,000 records** and **79 features**: **31 numerical** and **48 categorical**. The target, `OfficeCategory`, defines a **five-class classification task** with roughly balanced categories (each about 20%), providing a fair basis for model training.

1.2 Missing Values

Several features contain severe missingness (Table 1). For instance, `RecreationQuality` and `MiscellaneousFeature` exceed 99% missing, and `ConferenceRoomQuality` likely indicates "no conference room" rather than random loss.

Feature	**Missing Rate**
RecreationQuality	100.0%
MiscellaneousFeature	99.9%
AlleyAccess	99.5%
ExteriorFinishType	84.3%
ConferenceRoomQuality	73.8%

Table 1: Highly missing features.

Missing values in numerical columns were imputed with **median** or **constant** values depending on data skewness, while categorical variables used **most frequent** or a constant placeholder (`__MISSING__`). Business-relevant missingness (e.g., `ConferenceRoomQuality`) was retained as an indicator variable.

To Be Added

Strategies TBD

1.3 Categorical Features

Categorical features were encoded numerically before modeling. Low-cardinality variables applied one-hot encoding, while higher-cardinality ones adopted target encoding to avoid dimensionality explosion.

To Be Added

Strategies TBD

1.4 Outlier Analysis

Outliers were identified using the interquartile range (IQR) method across numerical features. Most variables exhibited less than 5% of potential outliers, suggesting generally well-behaved distributions. A few features, such as `FinishedBasementArea2` (11.8%) and `EnclosedBalconyArea` (15.5%), showed heavy right-tailed distributions due to a large proportion of zero values and a few extremely high observations.

Upon visual inspection with boxplots, these extreme points were found to correspond to valid but rare property characteristics (e.g., unusually large basements or balconies) rather than data-entry errors. Since tree-based models such as Random Forest and Gradient Boosting are robust to moderate outliers, no records were removed. Instead, potential transformations (e.g., $\log(1 + x)$) and indicator variables were considered to mitigate skewness and capture the presence of extreme cases.

Overall, outlier analysis confirmed that most numerical features are clean and that a few highly skewed variables will be handled through feature engineering rather than exclusion.

1.5 Class Imbalance

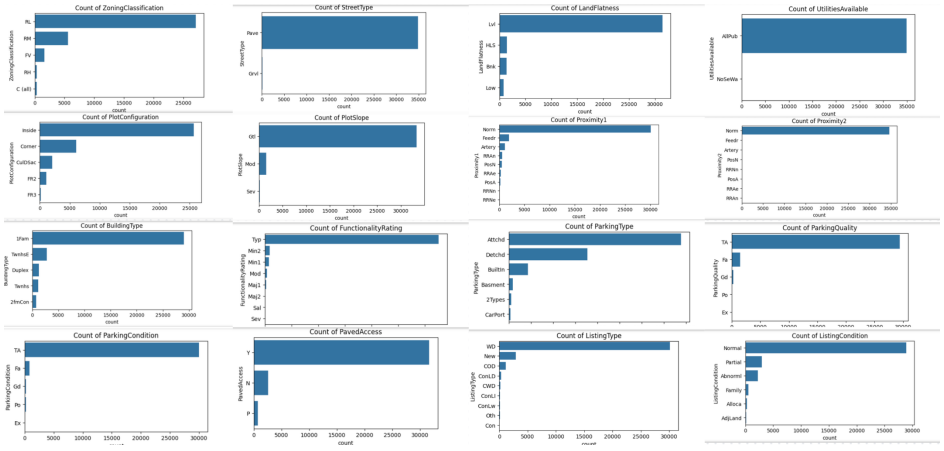

Figure 1: Feature imbalance visualization.

Some categorical variables (`ZoningClassification`, `StreetType`, `UtilitiesAvailable`) are extremely imbalanced, providing limited predictive value. Such columns may be dropped or merged.

1.6 Feature Correlation

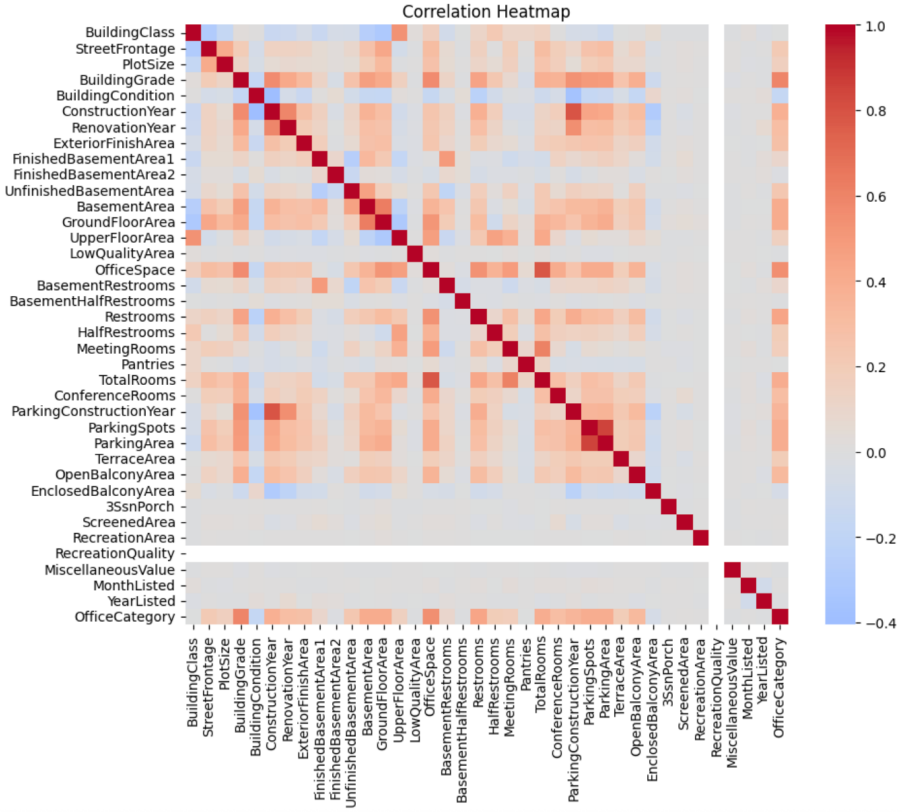

Figure 2: Correlation heatmap among numerical features.

High correlations were observed between `ParkingArea`–`TerraceArea` and `ConstructionYear`–`RenovationY` A cluster of amenities (`TotalRooms`, `ParkingSpots`, `OpenBalconyArea`) reflects property size; redundant variables may be combined or reduced (e.g., via PCA).

1.7 Insights from Data Exploration

The data exploration reveals that:

- The target variable is balanced, reducing bias risk.
- Several features suffer from heavy missingness requiring imputation or indicator treatment.

- Certain categorical variables are highly imbalanced and low-value.
- Multiple numerical features are strongly correlated, suggesting dimensionality reduction.
- No critical outliers were detected beyond legitimate extreme properties.

2 Feature Engineering Summary

2.1 New Features and Motivation

Feature / Transformation	Description	Motivation / Intuition
Log1p(PlotSize)	Apply $\log(1 + x)$ transform to `PlotSize`.	Reduce right-skew and stabilize tree splits.
Business Missing Indicator (`ConferenceRoomQuality`)	Replace empty strings with NaN and add missing flag.	Missing value often implies "no conference room," a potential signal.
Frequency Encoding (`RoofType`, `ExteriorCovering1`)	Encode categories by their empirical frequency.	Preserve rarity/commonness information while avoiding high-dimensional one-hot.
Target Encoding (`ZoningClassification`, `BuildingType`)	Use smoothed conditional probabilities $P(y\|x)$.	Capture class propensity patterns unavailable to one-hot encoding.
Wide Feature Builder	$\sim$30 domain-driven ratios and interactions (e.g., `BuildingAge`, `OverallQuality`, `QualityAreaProximity`).	Inject expert priors on quality, utilization, and location efficiency.
Statistical Aggregation (Z-scores)	Group-wise z-scores and relative shifts by zoning/building type.	Contextualize each property relative to its peers.

Table 2: Newly engineered features and their motivations.

2.2 Feature Selection Method

Feature selection was conducted through a **sequential ablation study** using a RandomForest backend. Each feature block was toggled on/off and the pipeline retrained to measure marginal impact on validation accuracy and weighted F1. Features with positive or negligible performance change were retained, while negative contributors were pruned. All runs were logged and versioned for reproducibility.

2.3 Impact on Model Performance

Configuration	Active Features	Accuracy	Weighted F1
Baseline	Core preprocessing only	0.6957	0.6911
+ Log1p + useful encoders	Log1p, `RoofType`, `ExteriorCovering1`, `ZoningClassification`, `BuildingType`	0.7081	0.7048
+ High-value wide features	Add top 7 domain features (`BuildingAge`, `OverallQuality`, etc.)	**0.7543**	**0.7529**
+ All statistical aggregations	Add all z-scores/shifts	0.7444	0.7432

Table 3: Performance impact of feature engineering configurations.

Best performance: Accuracy = 0.7543 (an improvement of +0.0586 over baseline). Main contributors were the domain-driven wide features and selective encoders, while excessive statistical aggregations slightly reduced accuracy.

2.4 Summary

Feature engineering improved the model substantially by addressing skew, categorical complexity, and domain structure. Sequential ablation confirmed that selective, interpretable features boosted accuracy from **0.6957** to **0.7543**, demonstrating that targeted transformations and domain insights significantly enhance predictive performance.

3 Model Selection & Tuning

Initial modeling will compare baseline algorithms (e.g., Logistic Regression, Random Forest, Gradient Boosting). Hyperparameter optimization will follow grid or Bayesian search, evaluated via stratified cross-validation to preserve class balance.

To Be Added

Insert final model selection results.