Ben Li, MSAI, 1st year Graduate Student
Rimsha Kayastha, MSDS, 2nd year Graduate Student

Description of the problem:
This project focuses on developing predictive and analytical models using the 2019 "Apartment for Rent Classified" dataset. The primary goal is to predict apartment rental prices based on available features. This can assist real estate platforms in offering smarter search features and assist renters in making informed decisions. The rising demand for rental housing and the variety of available listings make it challenging for tenants to efficiently identify suitable apartments. The specific questions this project aims to address are: Can we accurately predict rental prices based on the apartment's features, and can we identify patterns or clusters in the rental market that can aid in recommendation systems? Another goal is identifying abnormal listings through unsupervised methods.

Summary of the data:
The dataset contains 10,000 apartment rental listings collected from different cities in the United States. Each row represents a listing with 22 features such as price, square_feet, etc. Primary variables to predict rent mainly include square footage, city, and number of bedrooms and bathrooms. Grouping analysis can be based on rent and square footage. Identifying anomalies can use features such as price and square footage. Initial data exploration reveals potential issues: missing values; a large range of prices with a high standard deviation; a strong correlation between bathrooms and square feet as well as bedrooms and square feet. The correlations seem to violate some initial assumptions about price and size of houses, but the data includes houses from all over the United States, so pricing may vary significantly.

Methods:
Considering the growing rental market pressures, we have chosen to leverage this dataset to analyze housing attributes, and find underlying patterns in the current renting market. We are aiming to support accurate rent price prediction in two-folds: employing a linear regression model for easier interpretability, and a MLP (Multi-Layer Perceptron) model for capturing complex patterns. Similarly, we intend to implement an unsupervised learning framework using K-means clustering with an autoencoder-based dimensionality reduction to uncover market trends, such as identifying overpriced areas or undervalued properties, and detect pricing anomalies. Through these methods, we aim to provide budget-conscious renters with the tools to make informed decisions and gain deeper insight into market dynamics.

Preliminary results:
This dataset consists of 10,000 entries with mixed data types (numeric, categorical, and temporal), requiring careful preprocessing for each data type before data modeling. Data quality checks reveal that, while a few other features seem to have missing values, the 'amenities', 'pets_allowed', and 'address' columns seem to have a significant number of missing values. Correlation analysis done on the numerical features shows moderate correlation values, indicating that while size and room quantities are meaningful price predictors, approximately 60-85% of price variation stems from other factors, possibly categorical and temporal.

References:
Dataset: https://archive.ics.uci.edu/dataset/555/apartment+for+rent+classified