CS-349, Fall 2024
Final Project: Proposal
Group 18: Jianwei Lyu, Ziyu Li, Junyang Ding, Xiaoqin Bai, Zexi Zhang

Kaggle Competitions
https://canvas.northwestern.edu/courses/217335/files?preview=20075840

(0.5 points) What task will you address, and why is it interesting? This should be as simple as a couple of sentences.

The project aims to create a predictive model for determining the values of sneakers on platforms such as StockX using features such as brand, model, and market trends in area. By analyzing historical data and trends, the model helps guide sneaker enthusiasts and investors to make smart and more cost-effective decisions. As a sneaker fan, I want this model to help people navigate the market efficiently, saving both time and money. This project combines my love for sneakers with data analysis to create something of value for the sneaker community.

(1.0 points) How will you acquire your data? This element is intended to serve as a check that your project is doable -- so if you plan to collect a new data set (which I discourage), be as specific as possible.

We eventually decided to get the database from a random sample of all Off-White x Nike and Yeezy 350 sales from between 9/1/2017 on StockX. There are 99,956 total sales in the data set; 27,794 Off-White sales, and 72,162 Yeezy sales.
(https://stockx.com/news/the-2019-data-contest/)

We might use data collected through API if necessary.
(https://github.com/druv5319/Sneaks-API)

(1.0 points) Which features/attributes will you use for your task?

There are a total of 6 features of training data: Order Date, Brand, Sneaker Name, Release Date, Shoe Size and Buyer Region, with Sale Price and Retail Price as targets for prediction. All of these features should be considered.

However, with the fact that irrelevant data may form noises and affect model performance, feature selection or data dimensionality reduction may be performed, including cross-feature correlation detection through the calculation of metrics such as Pearson Correlation Coefficient, and pre-processing techniques such as PCA. After the feature selection step, we will further determine which of these features are most representative and use them for model training.

(2.5 points) What will your initial approach be? What data pre-processing will you do, which machine learning techniques (decision trees, KNN, K-Means, Gaussian mixture models, etc.) will you use, and how will you evaluate your success (Note: you must use a quantitative metric)?

Generally, you will likely use mean-squared error for regression tasks and precision-recall for classification tasks. Think about how you will organize your model outputs to calculate these metrics.

Data preprocessing
Handling Missing Data: For missing values we will use statistical methods (e.g., most common value, or means for numerical values) or drop records with insufficient information.
Feature Engineering: Convert order dates and release dates into useful features (e.g., day of the week, month, time since release). We will encode categorical data (Brand, Sneaker Name, Buyer Region) using one-hot encoding or label encoding.
Outlier Detection: We will identify and remove any outliers in sale prices using z-scores or IQR to prevent skewing the model.
Feature Selection/Dimensionality Reduction: By applying Pearson Correlation, we hope to remove highly correlated features.

Machine learning techniques:
We will start with linear regression, and for neural networks, we would possibly consider using Recurrent Neural Networks. For the outcomes, we will apply mean squared error to check the accuracy since we focus on the price prediction, and to align with the metrics, the workflow of the neural network should be given a specific shoe and the desired time, the network should output the predicted price.