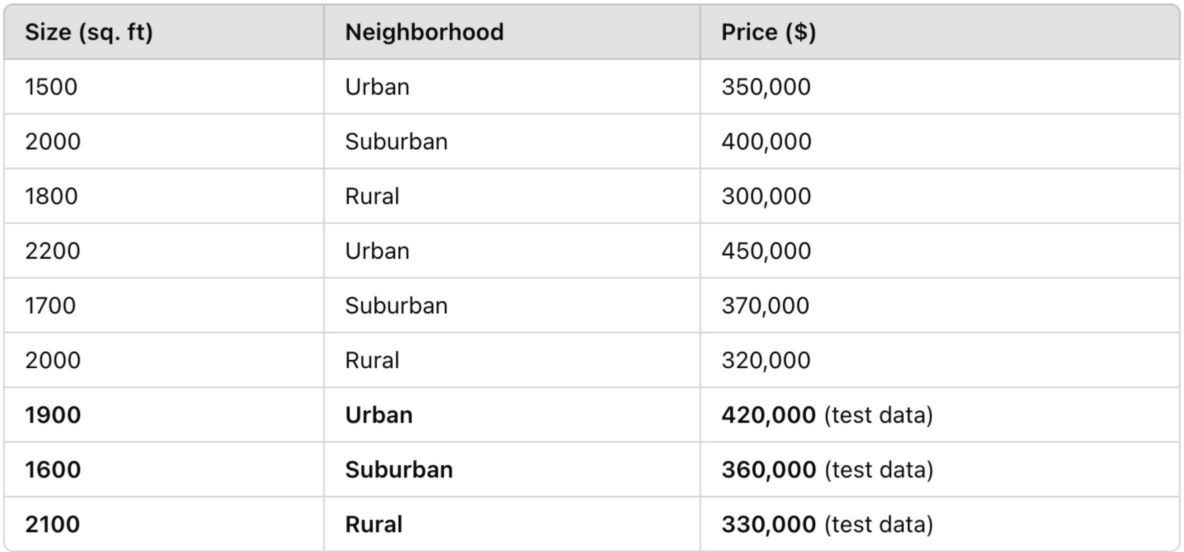
CS 5805: Machine Learning I
Spring 2025
Homework 1
Date Assigned: 1/27/2025
Date Due: 2/5/2025

1. (5 points) Prove that the Hat matrix is symmetric.
2. (5 points) Prove that the Hat matrix is idempotent.
3. (10 points) Which of the following types of curves can be estimated accurately from a given dataset (assuming it satisfies the functional form) using linear regression methods? Here y is the dependent variable, and x is the independent variable. "Estimate accurately" means if we substitute values for the coefficients in the curves below, generate data, and learn a regression model from that data, the model will recover the values we used to generate the data in the first place.
   a. $$y = \sum_{i=0}^{n} a_i x^i$$
   b. $$y = ax + b \cdot sin(x) + c \cdot log(x) + d$$
   c. $$y = ae^{(bx)}$$
   d. $$y = ax + bx + c$$
   e. $$y = (ax + b)/(cx + d)$$
4. (25 points) You are provided the following housing dataset where the goal is to predict price from size and neighborhood. It has 6 rows of training data and 3 rows of test data.



Size (sq. ft)	Neighborhood	Price ($)
1500	Urban	350,000
2000	Suburban	400,000
1800	Rural	300,000
2200	Urban	450,000
1700	Suburban	370,000
2000	Rural	320,000
1900	**Urban**	**420,000** (test data)
1600	**Suburban**	**360,000** (test data)
2100	**Rural**	**330,000** (test data)

Because neighborhood is a qualitative variable, we will be creative and encode it using three binary features called “is_urban”, “is_suburban”, and “is_rural”. Thus, the dataset now has 4 features and one dependent variable (price). The first row, for instance, will have a 1 for “is_urban” and 0 for both “is_suburban” and “is_rural”. The second row, because it is suburban, will have a 1 for “is_suburban” and 0 for the other two features.

Compute the hat matrix for this dataset and apply it to make predictions for new homes. What do you observe? If you encounter any issues, how can you address them?

Also conduct a regression using scikit-learn’s regression model and make predictions for the same test data. Do your answers match the scikit-learn model’s answers? Do both sets of answers match the ground truth from the test data? Why/why not?

5. (20 points) Consider the food delivery time prediction problem and dataset described in https://www.kaggle.com/datasets/denkuznetz/food-delivery-time-prediction/data. This dataset has been subsetted into training (700 instances) and test datasets (300 instances) for your convenience and supplied with this assignment. First, fit a regular linear model to this data. Next, compute the hat matrix from the training data and inspect its diagonal to identify the influence or leverage of individual data points. For each data point, also look at the residual. Explain what influence and residual together tell you about the role of individual data points in the overall regression model. Give specific examples to validate your conclusions.

6. (35 points) For the same training and test dataset as the problem above, implement the following regression approaches, and compare and contrast the results.
    a. **Regression by Successive Orthogonalization**
    - Perform regression by successive orthogonalization to predict the delivery time using the provided features.
    - Report the coefficients for each predictor at each step.
    b. **Lasso Regression**
    - Perform Lasso regression to predict the delivery time. Use cross-validation to select the best regularization parameter.
    - Identify which features are eliminated or retained in the final model.
    c. **Ridge Regression**
    - Perform Ridge regression to predict the delivery time. Use cross-validation to select the best regularization parameter.
    - Compare the Ridge regression coefficients with those from Lasso and the successive orthogonalization model.
    d. **Model Comparison**
    - Use appropriate metrics (e.g., Mean Squared Error, R-squared, RSS) to evaluate the performance of all three models (successive orthogonalization, Lasso, and Ridge).
    - Based on the results, discuss which method performed best and why. Consider model interpretability, regularization effects, and predictive accuracy.

Include sufficient plots, graphs, and commentary to support your observations. Provide a link to your notebooks that we can evaluate as necessary.