MATH 3330 – Fall 2024
Assignment 3 – 27 marks – Due Dec. 1 at midnight

Question 1 (13 marks): Load in the Spotify dataset into `R`. We continue our analysis from previous assignments below.

a. Carry over your regression model from Assignment 2, question 1d. Use `hatvalues` to compute h_{ii} for all $i \in [n]$. Identify the observation with the largest value of h_{ii}. What proportion of values are greater than $2p/n$? (3 marks)

b. Compute the Cooks distance of each observation. Identify the observation with the largest Cook's distance value and print it out. What proportion of values are greater than 1? (3 marks)

c. Extract the model matrix from question 1a and call this X. Add the popularity column to the matrix X. Compute the projection depth values of each row of this matrix. Identify the observation with the smallest depth value and print it out. What proportion of values are less than 0.015? (3 marks)

d. Observation 63343 has been identified as a point with low depth and high values of h_{ii}. Make the following four scatterplots:

   – 'liveness','duration_ms'
   – 'tempo', 'instrumentalness'
   – 'acousticness', 'energy'
   – 'danceability', 'duration_ms'

   with point 63343 colored red and visible. Do you think this point is a leverage point? Why or why not? (4 marks)

Question 2 (9 marks):

a. Load in the real estate data 'clean_data.csv' from eclass. Remove observations with Lotsize of 0, Sale price less than 10000 and Finished square feet smaller than 500. Add a column `ppsq` for the logarithm of the price per finished square foot. Fit the following regression model: `ppsq log(Lotsize)+Sale_date+Year_Built+District+Bdrms`. (3 marks)

b. Use VIF to assess for multicollinearity. Is there any sign of multicollinearity? Why or why not? (3 marks)

c. Perform all-subsets regression with the adjusted R^2 metric. Based on the output, if your business partner wanted you to only include a 3 category variable to summarize `District`, would you advise against this? Why or why not? (3 marks)

Question 3 (5 marks): Assume the MLR with n observations and p variables. Let $\lambda_1, \ldots, \lambda_p$ be the eigenvalues of $X^\top X$. Show that

$$E(||\beta - \hat{\beta}||^2) = \sigma^2 \sum_{i=1}^{p} \lambda_i^{-1}.$$