# Homework 1

**Make a copy of this document or answer in a separate document.**
**You must submit :**

1. **.pdf of your answers to these questions**
2. **.pdf of your code / figures (File > Print > save as .pdf, *MAKE SURE THE BOTTOM ISN'T CUT OFF*)**
3. **Active link to your colab (Share > Anyone with Link, set to "Viewer") This link can be embedded in your .pdf.**

**You may have multiple submissions on Canvas.**

1. (1 point) ISLP section 2.4 problem 2.
   **Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide n and p.**
   a. *(a) We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.*
      The scenario is regression because the variables are quantitative/take on numerical values. We are most interested in inference since it is looking for an association between the factors that affect CEO salary. N = 500 since there are 500 data samples, p = 3 which are profit, number of employees, and industry.
   b. *(b) We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.*
      The scenario is a classification scenario because the outcome is qualitative (success or a failure). We are most interested in prediction because we want to know whether the product would fail or succeed. N = 20 since there are 20 similar products in the dataset, p = 13 since there are price charged, marketing budget, competition price and 10 other variables
   c. *(c) We are interested in predicting the % change in the USD/Euro exchange rate in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market.*
      The scenario is regression because the variables are numeric and the outcome is quantitative. We are most interested in prediction since we want to predict the exchange rates. N = 52 since there are 52 weeks in 2012, p = 3 since we look at the % change in the US, British, and German markets.
2. (1 points) ISLP section 2.4 problem 4.
   **You will now think of some real-life applications for statistical learning.**

a. *(a) Describe three real-life applications in which classification might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.*

   The first scenario can be to detect if a patient tests positive for a disease. The response would be “has the disease” or “does not have disease”. The predictors can be different symptoms, health history, and like blood test results. The goal of this would be predicting accurately to identify which patients have that disease

   The second scenario can be to classify which emails are spam and not spam. The response would be “spam” or “not spam”. The predictors can be the email address, contents of the email, and time of day. The goal is predicting the type of emails to show users their legitimate emails and block scams/spams.

   The third scenario is fraud detection in credit card transactions. The response would be “fraud” or “not fraud”. Predictors can include transaction amount, location, time, frequency of transaction and more. The goal of this is to predict so banks can catch fraudulent transactions to prevent loss and build customer trust.

b. *(b) Describe three real-life applications in which regression might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.*

   The first scenario can be predicting sales revenue for a store location. The response would be the sales revenue on a weekly, monthly, or yearly basis. The predictors can be similar store profits in the area, marketing budget, cost of living, need of that store, revenues of other stores in the area. The goal is to predict the revenue to see if it's worth opening up the store and gaining financial support from investors.

   The second scenario can be estimating the selling price of a house. The response would be the price in dollars. The predictors can include the square footage, number of bed/bath, land area, school zone, age of house, renovations/features, and neighborhood. The goal is to predict the estimated selling price of a home in order to list accurately.

   The third scenario is predicting movie box office revenue. The response would be the box office revenue. The predictors include marketing budget, fame of leading actors, genre, release time, social media buzz and production budget. The goal is to predict the revenue because return on investment is important.

c. *(c) Describe three real-life applications in which cluster analysis might be useful*

   The first scenario is customer segmenting in a department store in order to promote better marketing practices. The predictors can include purchase history, average spending, product category purchases, age, frequency of purchases, and location. The goal is to identify clusters of similar customers to create better marketing strategies.

   The second scenario is social media categorizing posts based on content. The predictors can be likes, shares, topics, hashtags, and captions. The goal is to group posts for recommendations and search results.

   The third scenario is city development plans. The predictors can include population density, income, crime rates, public transport, available space and green space. The goal

is to find clusters of neighborhoods with similar demographics to guide policy plans and developments for that area.

3. (1 point) ISLP section 2.4 problem 5.
    a. *What are the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression or classification? Under what circumstances might a more flexible approach be preferred to a less flexible approach? When might a less flexible approach be preferred?*
    Some advantages of a flexible approach for regression or classification are that it can fit many different possible functional forms. It can capture more complex relationships. However, a flexible model requires estimating a greater number of parameters which can lead to overfitting. A more flexible approach might be preferred when there is a large dataset with many predictors and the relationships between predictors and response complex and nonlinear. A less flexible approach may be preferred if we are mainly focused on inference because it is more interpretable.

4. (1 point) In your own words, how do we assess/quantify model accuracy? Explain whether one should use the training or the test set for this and why.
    a. We assess/quantify model accuracy by measuring how well the predictions match observed data by using measures such as the mean squared error. The smaller the MSE, the closer to the true responses. The test set should be used because it is unseen data, and the training data was used to fit the model.
5. (1 point) Python : Suppose we want to predict graduation rates. Create a histogram of graduation rates for the entire dataset.
6. (1 point) Choose 1 other variable that you think might be informative and create a visualization that explores how that variable is related to graduation rate.
7. (1 point) Create two new dataframes – one that contains the colleges in the lowest quartile of graduation rates and one that contains the colleges in the top quartile.
8. (1 point) Compute the basic statistics on both dataframes, what do you notice? Be sure to comment in particular on the predictor you observed in Q6.
9. (1 point) Create a visualization using the two dataframes in Q7 to highlight the differences in a different predictor of your choice.
10. (1 point) Create a correlation matrix using all quantitative variables, what trends do you identify?

Attach your Google Colab notebook link here :
***https://colab.research.google.com/drive/1pXeVZ8II6nVw9kGo04WHM3CTUetondRt?usp=sharing***