Workshop #7 Exercises

1. The file heights_and_weights.csv contains data about a set of males and their heights (in cm) and weights (in kg). You want to set up a model which will predict weight from height.
* Make a scatter plot of the data. Based on the scatter plot alone, is a linear model appropriate for the data?

Build the linear model $\hat{y} = b_0 + b_1 x$.
* Make the residual plot for the data. Based on the value of R^2 and the residual plot, do you think a linear model is appropriate for the data?

It makes sense that weight = 0 would relate to height = 0.
* Build the linear regression model without an intercept: $\hat{y} = b_1 x$. To build this model, instance the model object as `LinearRegression(fit_intercept=False)`. Note: `fit_intercept` is `True` by default.
* Determine whether this, second, linear model is appropriate for the data.

---

2. The file named mutual_funds.csv contains information about 45 mutual funds that are part of the *Morningstar Funds 500* for 2008. The data set includes the following five variables:

Fund Type: The type of fund, labeled DE (Domestic Equity), IE (International Equity), and FI (Fixed Income).
Net Asset Value ($): The closing price per share on December 31, 2007.
Expense Ratio (%): The percentage of assets deducted each fiscal year for fund expenses.
Morningstar Rank: The risk adjusted star rating for each fund, from a low of 1-Star to a high of 5-Stars.
5-Year Average Return (%): The average annual return for the fund over the past five years.

The goal of this exercise is to build models to predict the 5-year average return for a domestic equity fund with a Net Asset Value of $35.53 and an expense ratio of 1.05% and a 3-Star Morningstar Rank.

* Using **only the two numerical variables** build a multivariable linear model for predicting the 5-Year Average Return. Establish if the model is appropriate for the data by considering the residual plot and calculating R^2.
* Predict the 5-Year Average Return for the fund we are interesed in (underlined above; use only the variables you need for the model)
* Next, we want to include the categorical variables (Fund Type and Morningstar Rank) in a new multilinear model. To achieve this, we must **code** them. Using `OrdinalEncoder()`, encode the *mornigstar_rank* and *fund_type*, and then build the linear model. Once you have built the model, establish if it is appropriate by considering the residual plot and R^2
* Predict the 5-Year Average Return for the fund we are interested in (underlined above; use all variables, be careful with the encoding).
* Finally, compare the two models using MAE, MSE and RMSE and decide which one performs better (use the original data to make the assessment, no need to split it into train and test).

---

3. The file weights_and_mpg.csv contains some data about cars. The goal is to build a model that can be used to predict the **mileage** of a car (i.e. the fuel efficiency) based on the car's weight.
* Build a linear model of the mileage using weight as input. Give reasons why the linear model is not appropriate for these data.
* Build a quadratic model for the mileage using weight as input by transforming the data using appropriate polynomial transformation. Provide evidence that this model is (more) appropriate for the data than the linear model.

---

4. The file `gender_classification.csv` contains data about the gender, weights and heights of 10000 people. The goal of this task is to build a logistic regression model to predict the gender of a person based on their height and weight
* Split the given data set into a **train** and **test set**. To ensure everyone gets the same results, use `random_state=1234`
* Build the logistic model using the train data
* In the next step, evaluate the model's performance on the test data. Construct the confusion matrix, and calculate the model's mean accuracy score
* Finally, make a prediction about the gender of the *median person* and the *mean person*, i.e. the persons who have median/mean height and median/mean weight.