1. A website has trained a linear regression model to predict the amount of minutes that a user will spend on the site. The formula they have obtained is t = 0.8d + 0.5m + 0.5y + 0.2a + 1.5 where t is the predicted time in minutes, and d, m, y, and a are indicator variables (namely, they take only the values 0 or 1) defined as follows:

   - d is a variable that indicates if the user is on a desktop.
   - m is a variable that indicates if the user is on mobile device.
   - y is a variable that indicates if the user is young (under 21 years old).
   - a is a variable that indicates if the user is an adult (21 years old or older)

   If a user is 30 years old and on a desktop, then d = 1, m = 0, y = 0, and a = 1. If a 45-year-old user looks at the website from their phone, what is the expected time they will spend on the site?

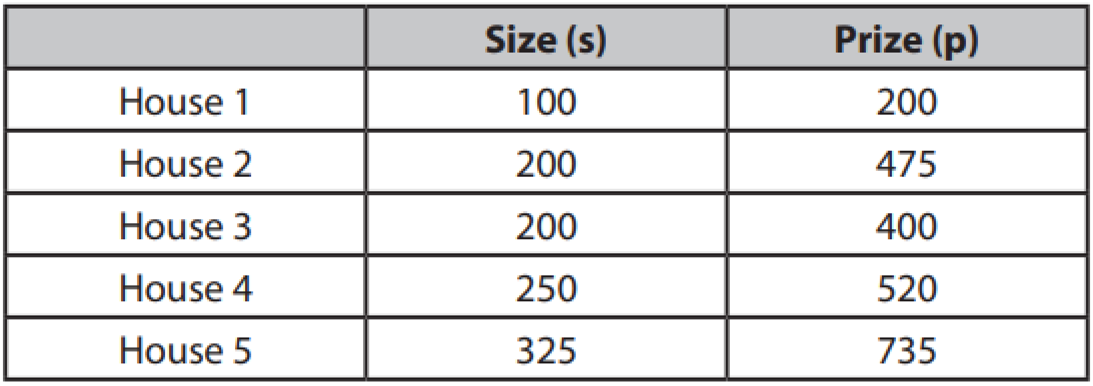
2.2 Minutes

2. Imagine that we trained a linear regression model in a medical dataset. The model predicts the expected lifespan of a patient. To each of the features in our dataset, the model would assign a weight. For the following quantities, state if you believe the weight attached to this quantity is a positive number, a negative number, or zero.

   Note: if you believe that the weight is a very small number, whether positive or negative, you can say zero.

   i. Number of hours of exercise the patient gets per week **positive**
   ii. Number of cigarettes the patient smokes per week **positive**
   iii. Number of family members with heart problems **zero**
   iv. Number of siblings of the patient **negative**
   v. Whether or not the patient has been hospitalized **negative**

2. The following is a dataset of houses with sizes (in square feet) and prices (in dollars).



	Size (s)	Prize (p)
House 1	100	200
House 2	200	475
House 3	200	400
House 4	250	520
House 5	325	735

Suppose we have trained the model where the prediction for the price of the house based on size is the following: p = 2s + 50

i. a. Calculate the predictions that this model makes on the dataset.

House 1 – 250

House 2 – 450

House 3 – 450

House 4 – 550

House 5 - 700

ii. b. Calculate the mean absolute error of this model.

Mean absolute error = (|250-200|+|475-400|+|450-400|+|550-520|+|735-700|)/5 = 240/5 = 48

iii. c. Calculate the root mean square error of this model.

Root Mean Square error = root(12750/5) = 2550^1/2 = 50.49

1. What is linear regression and for which kind of data do we use this model?

Linear regression is a statistical method used for modeling the relationship between a dependent variable (also called the target or response variable) and one or more independent variables (also called predictor variables or features). The goal of linear regression is to find the best-fitting linear relationship between the variables, allowing us to make predictions or understand the strength and direction of their association.

2. What is the equation of the linear regression line?

The equation of a simple linear regression line is given by:

$y=mx+b$

Where:

y is the dependent variable (target)

x is the independent variable (feature)

m is the slope of the line

b is the y-intercept

3. Explain the slope and intercept in the linear regression equation.

The slope (m) represents the change in the dependent variable for a unit change in the independent variable. It indicates the direction and steepness of the relationship. A positive slope indicates a positive correlation, while a negative slope indicates a negative correlation.

The intercept (b) is the value of the dependent variable when the independent variable is zero. It provides the starting point of the line on the y-axis.

4. If we increase the slope of a line in which direction the line moves (clockwise or anti-clockwise).

Increasing the slope of a line will cause it to become steeper. If you increase the slope in a positive direction, the line will move in a counterclockwise direction, and if you increase it in a negative direction, the line will move in a clockwise direction.

5. Explain the different evaluation metrics and their formula.

Mean Squared Error (MSE): Measures the average squared difference between actual and predicted values.

Root Mean Squared Error (RMSE): The square root of MSE, providing a more interpretable error metric in the same units as the target variable.

Mean Absolute Error (MAE): Measures the average absolute difference between actual and predicted values.

R-squared (R^2) Score: Represents the proportion of the variance in the dependent variable that is explained by the independent variable(s).

6. What is the difference between simple and multiple linear regression models?

Simple Linear Regression: Involves only one independent variable. The relationship between the dependent and independent variables is represented by a straight line.

Multiple Linear Regression: Involves two or more independent variables. The relationship is modeled as a hyperplane in a higher-dimensional space. The equation becomes:

$Y = b0+b1x1+b2x2+..+bpxp$

Where p is the number of independent variables
bi are coefficients
xi are the corresponding independent variables