LINEAR REGRESSION APPLICATION:

* One of the main and basic application of linear regression is prediction of quality of wine.
* The quality of wine is tested by experts by tasting it.
* Orley Ashenfelter, a economic professor in Princeton, claims that he can predict the wine quality without tasting it.
  + He used linear regression that uses
    - Input: a set of independent variables
    - Output: a dependent variable
  + For example, in case of wine quality,
    - Input: Age, season temp, harvest Rain, Winter rain, etc.
    - Output: price in 1990-1991 wine auctions.
* One variable Linear Regression:
  + It uses one independent variable as input and produces an output.
  + I.e.; the model is
    - yi = b0 + (b1)(xi) + ei
    - yi ⇒ dependent variable for ith observation
    - xi ⇒ independent variable for ith observation
    - ei ⇒ error term for ith observation
    - b0 ⇒ intercept co-efficient
    - b1 ⇒ regression coefficient for independent variable. (slope of the line)
  + It finds the best observation in the model with low error term.
  + But SSE can be hard to interpret bcoz:
    - Depends on N
    - Units are hard to understand(can be square units)
  + So, we use Root-Mean Square Error(RMSE) = sqrt(SSE/N)
  + Another error measure for Linear Regression is R2(R square)
    - This compares best model to baseline model
    - Baseline model doesn’t use any variables.
    - It predicts same outcome regardless of the independent variable.
    - Then the error term is
      * r^2 = 1 - (SSE(best model) / SSE(baseline))
    - If R^2 == 0 ---> no improvements over baselinr
    - If R^2 == 1 ---> perfect predictive model
* Multi linear regression :
  + Allows us to use all of these variables to improve our predictive ability.
  + Model is:
    - yi = b0 + b1(x1i) + b2 (x2i) + ………. + bk (xki) + ei
    - yi ⇒ dependent variable for ith observation
    - xji ⇒ jth independent variable for ith observation
    - ei ⇒ error term for ith observation
    - b0 ⇒ intercept co-efficient
    - bji ⇒ regression coefficient for jth independent variable. (slope of the line)
  + Not all the available variables should be used.
* To use linear regression model:
  + Model1 = lm(price ~ HarvestRain, data= wine)
  + Price ⇒ dependent variable
  + HarvestRain ⇒ independent variable
  + Data ⇒ the dataset you rae going to use
  + Lm ⇒ linear model
* For multivariable linear regression:
  + Model1 = lm(price ~ HarvestRain + WinterRain, data= wine)
* We can det the error temrs of the models from residuals variable in the dataframe.
* The coefficients column information in the summary output helps to find out what independent variables can be included in the linear model.
* CORRELATION:
  + A measure of linear relationship between variables
  + +1 ⇒ perfect +ve linear relationship
  + 0 ⇒ no linear relationship
  + -1 ⇒ perfect -ve linear relationship
  + cor(wine$winterRain, wine$Price)
  + cor(wine) ⇒ returns for all variables.
* The data that we used to build the model is called training data
* The data on which we check how well does the model perform is called test data
* To test the model:
  + predict(model\_name, newdata = data\_set\_name)
* We can check the difference between the R^2 values of train model and test model.
* R^2 values cannot decrease when new variables are added to the predicting model.