What Is Regression?

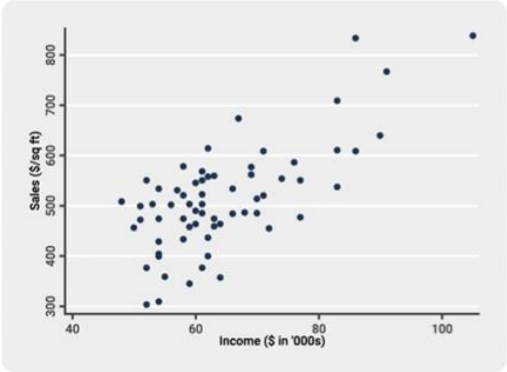
Supervised learning is a type of machine learning in which the model is trained on a labeled dataset, meaning each training example is paired with an output label. The goal is for the model to learn the mapping from inputs to outputs and make predictions for new, unseen data. There are various types of supervised learning algorithms, which can be broadly categorized into regression and classification algorithms. The goal of regression algorithms is to predict a continuous numerical value. Examples include predicting house prices, stock prices, or temperatures. In contrast, the goal of classification algorithms is to predict a discrete label or category. Examples include classifying emails as spam or not spam, identifying the species of an iris flower, or determining if a patient has a certain disease. The output of a regression algorithm is a continuous value, while the output of a classification algorithm is a categorical value.

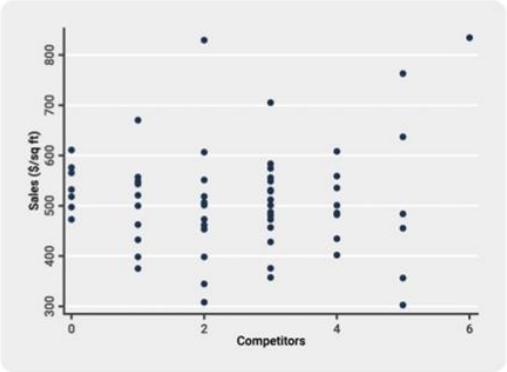
Regression, at its core, is simply comparing one variable with another variable and finding the relationship between the two. Regression also determines how strongly one variable influences another. Linear regression assumes that the relationship between variables can be plotted using a straight line (hence the term 'linear').

Linear regression strives to model a relationship between two variables by fitting the observed data to a linear equation. Of the two variables, one variable is evaluated as an explanatory (or independent) variable, while the other is considered a dependent variable. The linear regression equation can be expressed as $Yi=B0+BiXi$ where Xi is the explanatory variable and Yi is the dependent variable. Additionally, the slope of the line is defined as Bi, and B0 is the intercept.

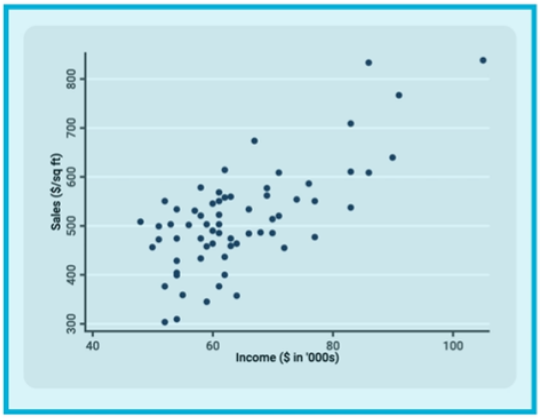
X stores the years of experience as the explanatory variable, and Y stores the salary as the dependent variable. As years of experience rise, so does salary. Thus, these variables are positively correlated, and the data 'fits' around the sloped red line.

TWO-WAY SCATTER PLOTS





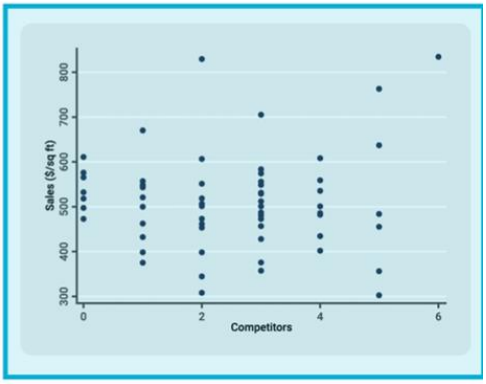
TWO-WAY SCATTER PLOTS

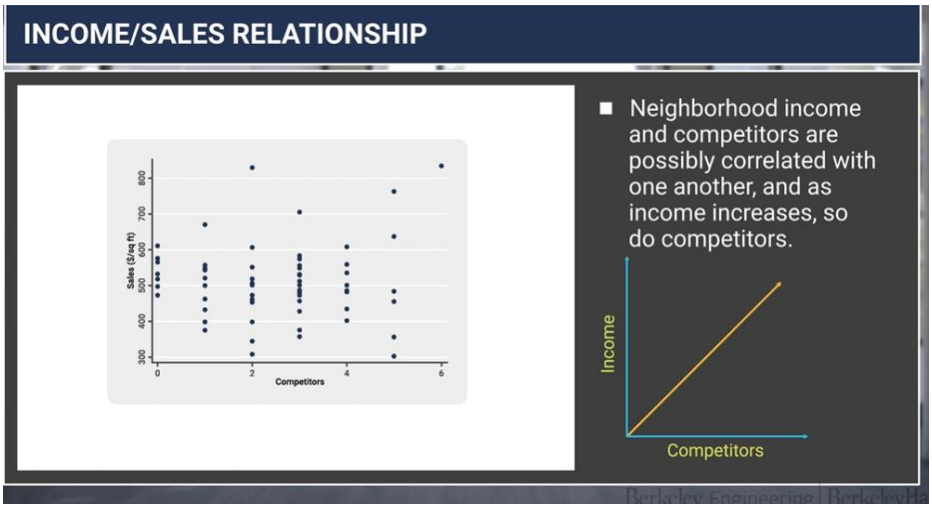


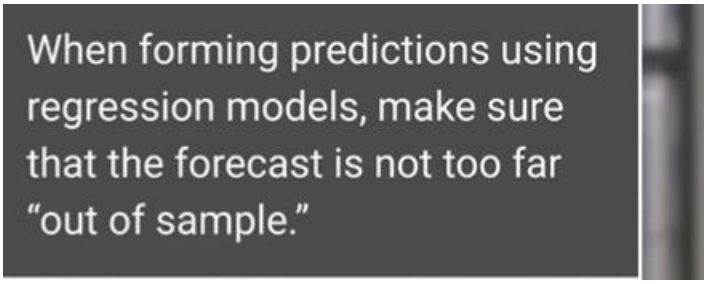
Shows a positive correlation between the neighborhood income and sales

TWO-WAY SCATTER PLOTS

Shows a weak relationship between the number of competitors and sales



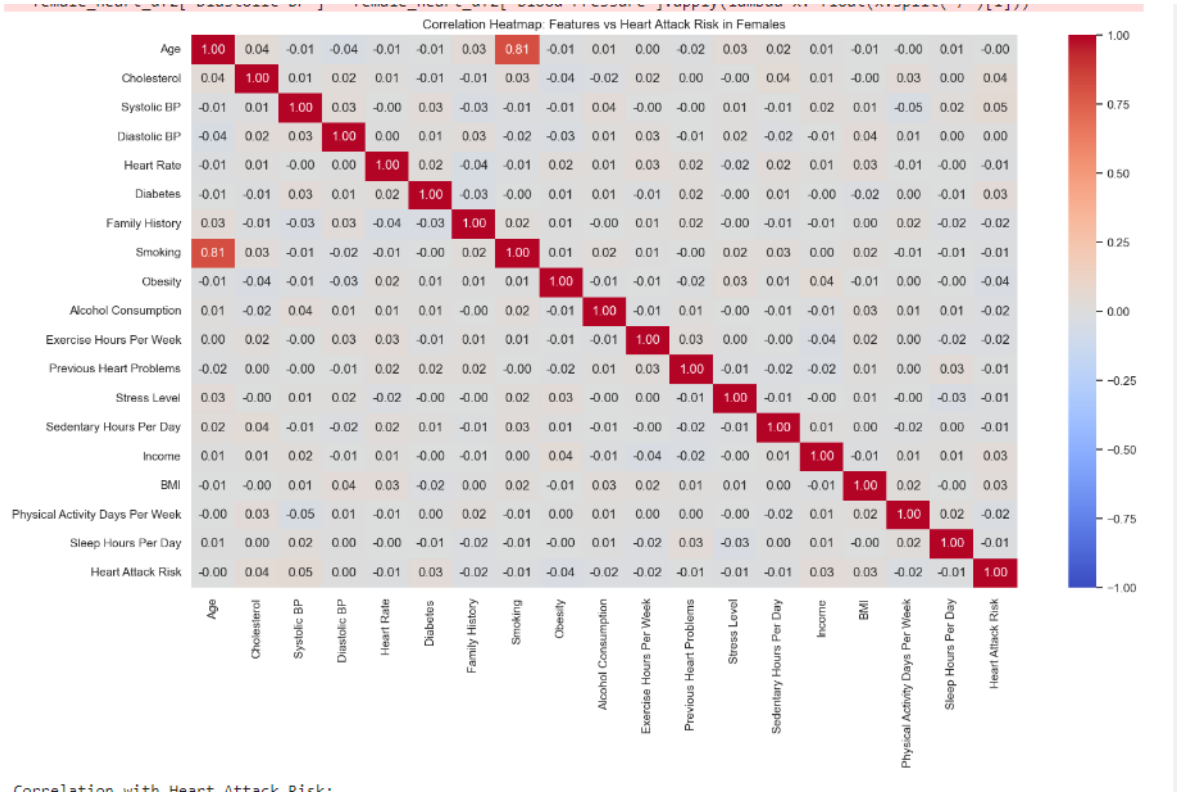




Questions

1. What is the difference btw Linear Regression and Logistic Regression?

2. I am working in this project [Heart Attack Risk Prediction for Female Populations].



I have the following questions. (See Table for reference)

Scenario 1

If you want to analyze four variables that are a mix between numerical and categorical values.

e.g., X Variables are: Smoking (Yes/No) and Cholesterol Level (235)

Y Variable is: Heart Attack Risk (Yes/No)

1. Is the study considered Classification?
2. How do you name the study if is a mix of Linear Regression and Classification?

Scenario 2

If you want to analyze two variables that are Categorical

e.g., X Variable-Independent is: Smoking (Yes/No)

Y Variable-Dependent is: Heart Attack Risk (Yes/No)

1. Is the study consider Classification?

2. How can I make the study Linear Regression?
3. Do I insert another variable but Numerical?

Supervised Learning	Number of Variables	Variable Type	Example
Regression	1 variable	Continuous, Numerical	Cholesterol Level 235 mg/d
	>1 variable	Continuous, Numerical	Cholesterol Level 235 mg/d Stress Level 9
Classification	1 variable	Discrete, Categorical	Smoke Yes/No
????? Scenario 1	>1 variable	Discrete, Categorical	Smoke Yes/No Heart Attack Risk (Yes/No)
????? Scenario 2	>1 variable	Continuous, Numerical Discrete, Categorical	Smoke Yes/No Cholesterol Level 235 mg/d Heart Attack Risk (Yes/No)