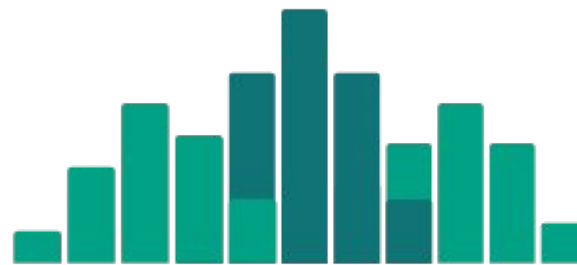# Machine Learning Workshop

Predicting Housing Prices with Multiple Linear Regression Techniques
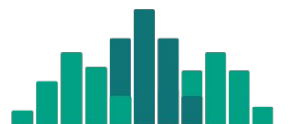By: Vibhuti Gandhi and Jason Dang

SFU DATA SCIENCE

Topics

- Data Pre-Processing
- Exploratory Data Analysis
- Data Cleaning and Feature Engineering
  - Missing Data
  - Categorical Data & Dummy Variables
  - Dimensionality Reduction
  - Feature Transformation
- Machine Learning
  - Multiple Linear Regression
  - Data Modelling
  - Overfitting & Underfitting
  - Regularization

SFU DATA SCIENCE

Aim of this Workshop

- Introduce you to Multiple Linear Regression

- Build upon the existing statistical and python knowledge that you have

- See how data science workflows occurs

- Less Coding, More discussion on the How, Why and What

- Most importantly, you are encouraged to take these topics, expand upon them and build your own data science project.

Data

- Ames Housing dataset compiled by Dean De Cock
- Data contains 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa.

Variables of Interest:
- Response Variable:
  - SalePrice - the property's sale price in dollars
- Explanatory Variables:
  - LotArea - Lot size in square feet
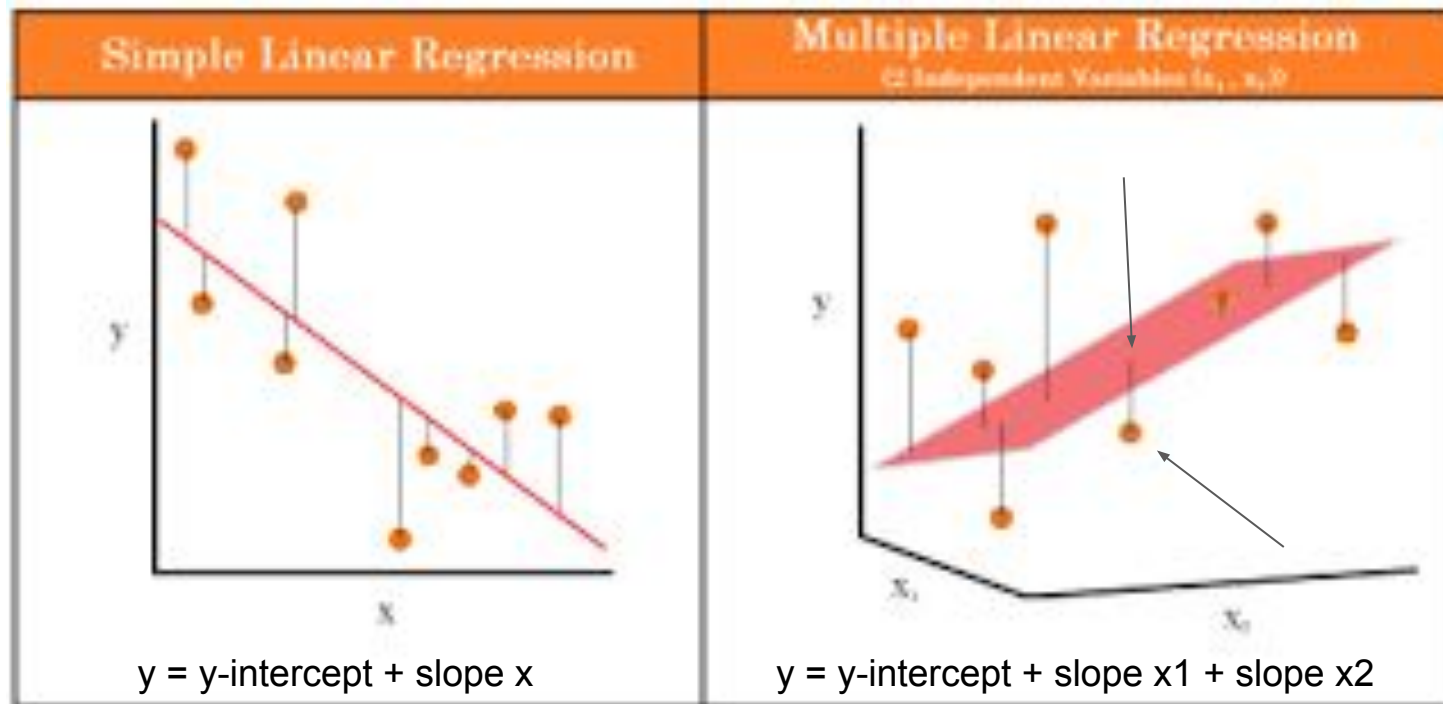 - Foundation - Type of foundation

Analysis

- We have a **continuous response variable**
- We have **one numeric** (LotArea) and **one categorical** (Foundation) **explanatory variable**
- We want to describe the relationship between the response and two explanatory variables simultaneously
- We can use a **multiple linear regression model** for this purpose!!!
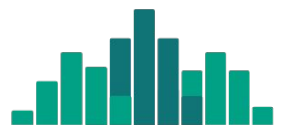
What is a Multiple Linear Regression

The **multiple linear regression model** is a model for describing the relationship between a dependent variable and two or more independent variables simultaneously

Simple Linear Regression	Multiple Linear Regression (2 Independent Variables (x₁, x₂))
y = y-intercept + slope x	y = y-intercept + slope x1 + slope x2

Notation

Let Y_i be the sale price of the ith house

Let x_1i and x_2i be the LotArea and Foundation respectively of the ith house
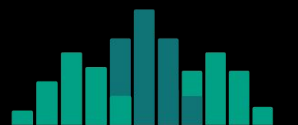
For our categorical predictor (Foundation):
- Need k-1 = 5 indicator variables, one for each of levels 2-6
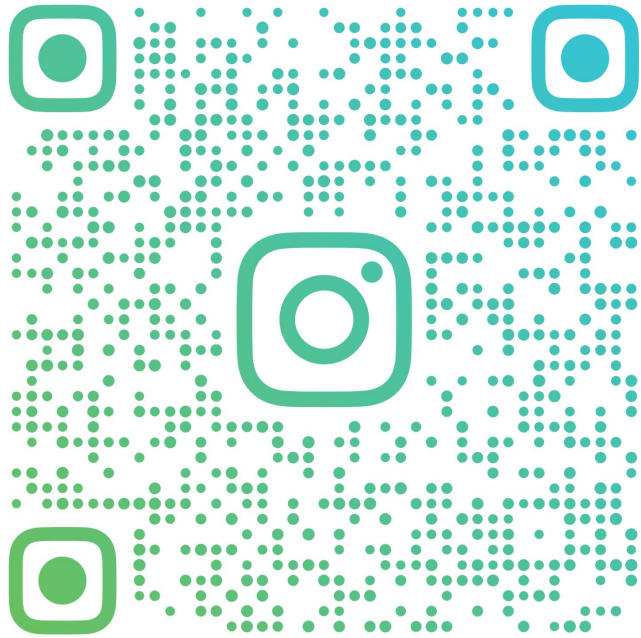- Let x_3i = 1 if the ith house has foundation Brick & Tile
- and x_3i = 0 otherwise

Model Specification

$$Y_i = \beta_0 + \sum_{i=1}^{7} \beta_j x_{ji} + \varepsilon_i$$

$$Y_i = \beta_0 + \sum_{i=1}^{7} \beta_j x_{ji} + \varepsilon_i$$

SFU DATA SCIENCE

Join our Social Media Platforms

SFUDSSS

SFU DATA SCIENCE DISCORD

SFU DATA SCIENCE