**Figure 6**

## Population Variance

**Feature Engineering**

Clustering, on the other hand, is an unsupervised learning task that involves grouping similar data points together based on their inherent patterns or similarities. It aims to discover hidden structures or clusters in the dataset without any prior knowledge or labels.

A real-world example of clustering is customer segmentation for a retail business.

By analyzing customer data such as purchasing behaviour, demographics, and preferences, clustering algorithms can group customers into distinct segments.

This segmentation can help retailers personalize marketing campaigns, target specific customer groups, and optimize product recommendations.

**Clustering**, on the other hand, is used for market segmentation, anomaly detection, document clustering, or recommendation systems, to uncover patterns and structure in the data.

Both techniques have practical significance and applications in a wide range of fields, enabling us to gain insights and make informed decisions from complex datasets.

## Modelling

### Normalization

Normalization is an important topic as it leads to better predictions. Features are on the same scale.

Three features selection techniques that will be used to categorical variables will use the sklearn’s library HotOnesEncoder will be used to transform categorical values into their own feature as either a 0 or 1.

Regression can be classified as linear and non-linear. In order to use linear regression, there must be a linear relationship between the independent and dependent variables.

## Validation

To use multiple linear regression, there must be an assumption of linearity between the response variable and the independent variables. In our scenario, an anova test can a hypothesis test can be used to check whether the slope of each independent variable is equal to zero or not. If it is equal to zero, then there is no relationship with hourly wage. The equation of the line is:

In addition, four assumptions must be held with random errors . These errors are independent and normally distributed, and have a mean of zero and common variance. This can be checked by

Multiple Linear Regression (MLR) uses the method of least squares to achieve the best fitting line in a hyperplane.

The estimated regression coefficients will have large standard errors, causing imprecision in confidence and prediction intervals. Adding or deleting a predictor variable may cause significant changes in the values of the other regression coefficients.

We want to minimize the error when we build a predictive model to make the prediction as accurate as possible.

This is the main goal. Linear regression is used when the variable is numerical discrete or continuous. Least squares to minimize. Relies on assumptions of linearity.

Normalization makes the chance of better preictions. Outliers will reduce our accuracy. Highlyover correlated features will increase the risk of overfitting.

And also, please also check the distribution of the errors in a linear regression.

The distribution of the errors should be normal.

The errors should be normally distributed

Talk about minimizing least squares is a common method in deermining the line of best fit.

How can you tell whether a regression analysis exhibits multicollinearity? Look for these clues:

The value of R2 is large, indicating a good fit, but the individual t-tests are nonsignificant.

The signs of the regression coefficients are contrary to what you would intuitively expect the contributions of those variables to be.

A matrix of correlations, generated by computer, shows you which predictor variables are highly correlated with each other and with the response y.

We use least square to minimize the error.

Supervised learning is the machine-learning task of inferring a function from labelled training data.

So, please check the word "labelled" here. Okay? So, we are inferring a function from the label training data.

Unsupervised learning is the machine-learning task of inferring a function to describe a hidden structure from unlabeled data.

Please here also check the "unlabeled" part, okay?

So, we are trying to understand the, we are trying to describe the hidden structure of the unlabeled data, when we go to unsupervised learning.

Clustering is a descriptive method. We want to use this to explore our data and see patterns in our data. We also mentioned that cluster analysis or clustering is the task of grouping a set of objects in such a way that the objects inside the same group are more similar to each other than to those in other groups.

We will use centroid-based clustering, hard part is selecting k. Looking for hiding structures in the dataset.

## Approach

* When comparing regression you can do an anova test with reduced model-full model.