# what is the function of P-values in high dimensional linear regression?

In high-dimensional linear regression, **p-values** serve the following functions:

1. **Hypothesis Testing**: P-values help determine whether the relationship between each independent variable and the dependent variable is statistically significant. A low p-value (typically less than 0.05) indicates that you can reject the null hypothesis, suggesting that the variable contributes meaningfully to the model.
2. **Feature Selection**: In high-dimensional datasets where the number of predictors is greater than the number of observations, p-values can help identify which variables are significant predictors. This aids in selecting a more parsimonious model by retaining only those variables with significant p-values.
3. **Model Interpretation**: P-values provide insight into the strength of the evidence against the null hypothesis for each coefficient in the model. This helps in understanding the importance of different predictors in relation to the outcome variable.

# Let's say you have a categorical variable with thousands of distinct values. how would you encode it?

For encoding a categorical variable with thousands of distinct values, consider these methods:

1. **Target Encoding:** Replace categories with the mean of the target variable.
2. **Frequency Encoding:** Use the count of each category.
3. **Leave-One-Out Encoding:** Mean of the target, excluding the current observation.
4. **Clustering**: Group similar categories and encode with cluster labels.
5. **Hashing Trick:** Use a hash function to map categories to a fixed number of dimensions.
6. **Dimensionality Reduction:** Apply PCA or similar techniques if needed.

Avoid one-hot encoding due to high dimensionality unless you can reduce the number of categories

.

# Describe how PCA works?

Principal Component Analysis (PCA) is a technique used to reduce the dimensionality of data while retaining as much variance as possible.

It works by first standardizing the data, ***[rescale the values into a range [0,1] .]***

then calculating the covariance matrix to understand how variables relate.

Next, it computes eigenvalues and eigenvectors from this matrix, sorts them, and selects the top eigenvectors that capture the most variance.

Finally, PCA projects the original data onto these principal components, resulting in a simpler representation with fewer dimensions while preserving essential information.