特征工程

将数据转变成适合机器学习模型的格式

Transforms：

- Normalization (Min-Max Scaling)：$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$，通常用于KNN和神经网络，对于数据没有任何特定分布的要求。能够不显著改变特征的分布，且确保没有任何一个单独特征能影响整个统计。能够生成较小的标准差，但受离群值影响很大。
- Z-score Normalization：$z = \frac{x_i - \mu}{\sigma}$，期望数据是遵从高斯分布的，由于受离群值影响更小，更常用
- Log Transform：$x' = \log x$，能够处理有偏数据，并将数据近似于正态，使其能够用分析工具（t-test, Anova）

独热编码 One hot Encoding：对于标签数据，其大小没有意义，因此可以使用独热编码方式将其变成二元向量