Assignment Questions
----

1. why drop_first = True is more important during dummy variable creation

->It helps to reduce the extra column created during dummy variable creation, and laso reduces the correlations being created among dummy variables.

2. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
> Pairwise scatterplot is helpful to vlaidate the assumptions of linear regressions. It is easy to visualize a linear relationship on a plot.
> Also check for Residual and fitted values, to check the pattern. If there is a pattern the nonlinearity data has not capture by the model.

3**. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes**

->year

General subjective questions
----
1. Explain the linear regression algorithm in detail?

Linear regression is one of the technique in which the independent variable has a linear relationship with the dependent variable,
we train the model to predict the behaviour based on the based on the given data.
The main point is to consider the datapoints and plot the line to fit the model.
liear regression suggests two variables, one is x-axis and another one is y-axis. these two should be correlated.
x = Independent variable from dataset
y = Dependent variable from dataset
Linear Regression Equation:
y = mx+c

Used in realtime scenario:
--> stock market prediction
--> risk analysis
--> sales forecasting in marketing companies

2. explain the anscombe’s quartet in detail

Anscombe's quartet is nothing but the group of four data sets which are identitical and some datasets that confuses the regression model.
it have different distributions and looks differently when it plotted on scatter plot.
mostly used for visualization purpose.Before attempting to intepret the model , the data implement in machine learning algorithm, we need to visualize the data to build a model
-> the main goal is to ananlyze the datasets is that they all share the descriptive statistics
-> Mean
-> Variance
-> SD(standard deviation)

3. What is pearson's R?

->Pearson's R , it is statistic that measures the linear correlation between two variables.
And the numerical values are between -1.0 and +1.0.
Ex : A small children height increases as age increases. It draws a line through data of two variables to show their relationship.

4. What is scaling? Whty scaling is performed? What is the difference between normalized scaling and standardized scaling?

Scaling is nothing but, data pre-processing apllied to independent variables to normalize the data within a particular range.
The collected data set contains features highly varying in magnitudes, units and range. If we don't do the scaling properly algorithm won't consider units. We need to scale and bring all the vlaues to the same level.

Normalized Scaling

-> Brings all the data in the range of 0 and 1, We can implement with the help of sklearn.preprocessing.MinMaxScaler

x = x-min(x)/max(x)-min(x)

Standardization Scaling
---
it also brings the data into a standard normal disturbution which has mean is Zero and standard deviation is 1.
x = x-mean(x)/sd(x)

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

if it is a perfect correlation between two independent variables, the VIF is infinity. In that case we get R2 is equals to 1.
where it takes to 1/(1-R2) infinity.

To overcome this problem, we need to drop one of the variables from the dataset, which is causing the collinearity.
Note: Important point is that is if VIF>10, it comes under multi-collinearity.

6. what is q-q plot? Explain the use and importance of q-q plot in linear regression?
Quantile-Quantile plot is a graphical tool to help us assess if a set of data came from normal or exponential or uniform distribution. It also helps to determine if two data sets came from populations with a common distribution.
-> It is mainly used in linear regression to assess the normality.