From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Cnt is the dependent variable and we have categorical as 'weathersit' which is negatively correlated with the Cnt by 0.3, 'season' which is positively correlated with 0.4 correlation, 'mnth' this is also positively correlated with 0.28 correlation and 'weekday' this positively correlated with 0.068 correlation.

Why is it important to use drop_first=True during dummy variable creation?

Drop original variable for which the dummy was created and drop first dummy variable for each set of dummies created.

Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

There is a LINEAR RELATION between 'temp', 'atemp' and 'cnt', the correlation for 'temp' and 'atemp' is 0.63 with 'cnt'

How did you validate the assumptions of Linear Regression after building the model on the training set?

We perform the hypothesis testing on the model where we consider H0:B1=B2=...=Bn=0 and H1: Bi!=0, hence we can see that all our coefficients are not equal to zero which means we can reject the null hypothesis.

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

As per our final Model, the top 3 predictor variables that influences the bike booking are:

- Temperature (temp)
- Weather Situation 3 (weathersit_3)
- Year (yr)

Explain the linear regression algorithm in detail.

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.

Explain the Anscombe's quartet in detail

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points.

What is Pearson's R?

Pearson's r is a bivariate statistical model that analyses two variables. Pearson's correlation may ALWAYS be used to test an associative research hypothesis as long as the variables being analysed are both quantitative.

What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling.

It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (**μ)** zero and standard deviation one (**σ**).

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

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables

What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution.