ASSIGNMENT BASED SUBJECTIVE QUESTION

ANSWER-1 : Bike booking in season 3 is highest having median more than 5000 followed by season 2 and season 3. Bike booking in season 1 is significantly low.

Bike booking in weathersit1 (clear, few clouds, partly cloudy) is significantly high with median of near 5000.

Bike booking in holiday is high comparing to non-holiday.

Working day showing more booking than non-working day.

ANSWER-2 : **drop_first = True** is used while converting categorical variable to numeric variable. A categorical variable with n levels can be represented by n-1 dummy variable so to avoid redundancy we drop the first dummy variable.

ANSWER-3 : From pair-plot a linear relation between temp , atemp and cnt can be seen.

ANSWER-4: By doing residual analysis.

ANSWER-5 : temp with coefficient value 0.5174 , weathersit_3 with coefficient value -0.2828 and year(yr) with coefficient value 0.2325.

GENERAL SUBJECTIVE QUESTIONS

ANSWER-1 : Linear Regression is supervised machine learning algorithm which Basically performs a regression task. Regression models predict a dependent (target) value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between the dependent and independent variables, they are considering and the number of independent variables being used.

It is of two type :

1. Simple linear regression. ($Y = \beta0 + \beta1*x$)
   Multiple linear regression ($Y = \beta0 + \beta1X1 + \beta2X2 + \ldots + \beta pXp + \varepsilon$)

   Where ,
   - Y: The response variable
   - Xi: The i th predictor variable
   - βi: The average effect on Y of a one unit increase in Xi, holding all other predictors fixed
   - ε: The error term

ANSWER-2 : Anscombe's quartet is a group of four data sets that are nearly identical in simple descriptive statistics, but the data sets have very different distributions so they look completely different from one another when you visualize the data on scatter plots.

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyse according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

ANSWER-3 : The Pearson correlation measures the strength of the linear relationship between two variables. It has a value between -1 to 1, with a value of -1 meaning a total negative linear correlation, 0 being no correlation, and + 1 meaning a total positive correlation.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Where

r = correlation coefficient

Xi = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

Yi = values of the y-variable in a sample

$\bar{y}$ =mean of the values of the y-variable

QUESTION-4 : Scaling 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.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

1. Normalization/Min-Max Scaling: x= x-min(x) / max(x)-min(x)

   It brings all of the data in the range of 0 and 1.

   sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

2. Standardization Scaling: x= x-mean(x)/sd(x)

   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).

sklearn.preprocessing.scale helps to implement standardization in python.

QUESTION-5 : If there is perfect correlation then the VIF value is infinity.

QUESTION-6 : Quantile-Quantile plot or Q-Q plot is a scatter plot created by plotting 2 different quantiles against each other. The first quantile is that of the variable you are testing the hypothesis for and the second one is the actual distribution you are testing it against. For example, if you are testing if the distribution of age of employees in your team is normally distributed, you are comparing the quantiles of your team members' age vs quantile from a normally distributed curve. If two quantiles are sampled from the same distribution, they should roughly fall in a straight line. Since this is a visual tool for comparison, results can also be quite subjective nonetheless useful in the understanding underlying distribution of a variable(s).

The Quantile-Quantile plot is used for the following purpose:

- Determine whether two samples are from the same population.
- Whether two samples have the same tail
- Whether two samples have the same distribution shape.
- Whether two samples have common location behaviour.