Assignment-based Subjective Answers:

Answer 1: Categorical Variables affect the target variable as follows:

- Fall Season has seen highest sale of bikes.
- Sales have increased considerably from 2018 to 2019 indicating an increase in demand.
- Month of July has seen the highest sales of bikes.
- Median sales of bikes is more when there is no holiday.
- Clear skies or few clouds weather has affected most sales.

Answer 2: It is important to use Drop_first = True during dummy variable creation because what it's essentially doing is dropping one of the variables to make it n-1 categories. As it is understood that if it is 0 for all other variables, it will be 1 for the dropped variable. Hence we actually need only n-1 variables to learn the model.

Answer 3: Looking at the pairplot of the numeric variables, we can see that temp and atemp variables seem to have a high correlation with the target variable. However we have to keep in mind that both the variables may be highly correlated to each other as well leading to multicollinearity.

Answer 4: To validate the assumptions of the linear regression model:

- Firstly, I have done a residual analysis where I have plotted a histogram of the error terms(y_train – y_train_pred). The distplot showed the error terms were normally distributed which is one the major assumptions of the LR model.

- Secondly, I plotted a scatter plot with the residuals on the y-axis. This plot showed the residuals are scattered around the mean zero and no visible pattern was seen among the residuals. This also proved the homoscedasticity of the error terms and negated heteroscedasticity. This is also one of the major assumptions of the model which was proved right.

Answer 5: Based on the final model the top 3 features that explained the demand for bikes are:

1. Temperature
2. Year
3. Weather – light snow + light rain.

General Subjective Answers:

Answer 1: Linear Regression Algorithm is a regression technique where a continuous variable is predicted with the help of single or multiple independent variables.

Machine Learning Algorithms are of 3 types : Regression, Classification and Clustering. In regression techniques one of the techniques where a continuous variable is to be predicted and there is some kind of linear relation is established with the target variable, linear regression is used.

When the independent variable is singular, it is called simple linear regression and when the independent variables are more than one, it is called multiple linear regression.

Linear regression fundamentally works on the equation y = mx + c, where m is the slope and c is the y intercept. Hence the equation of a straight line is used. The equation can also be interpreted as y = B0 + B1*x. B1 is the coefficient of the independent variable and B0 is the intercept.

It uses a technique called gradient descent to produce the Best Fit Line that best explains the spread of the data-points. The line represents the predicted values by the model. The difference between the actual values and the predicted values is taken, which is called the error terms. Ordinary Least Squares Method is used to minimize the error terms and R-Square value is generated which explains the variance of the dependant/target variable.

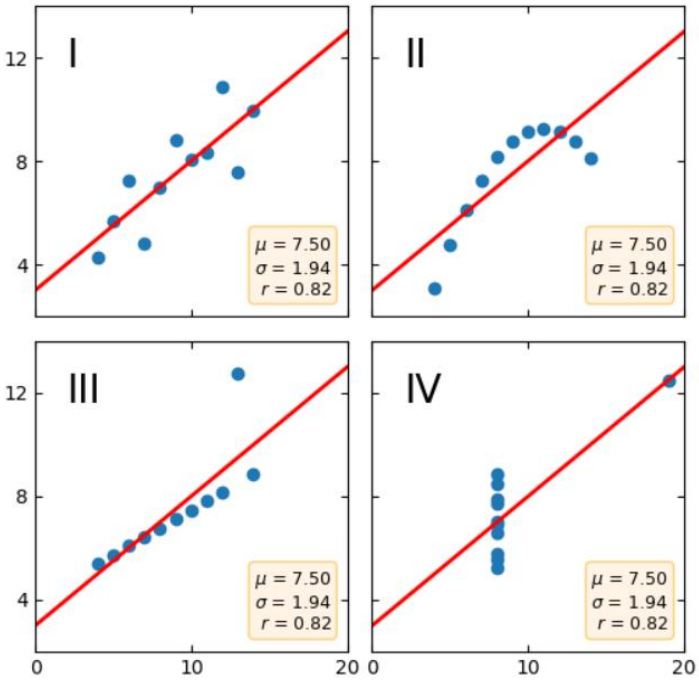
Few assumptions are made while formulating the Linear Regression Algorithm which are:

- The error terms should be normally distributed and should not show heteroscedasticity.
- The independent variables should not show multicollinearity. This means the independent variables are correlated to each other.
- The mean of the error terms should be zero and the error terms should show constant variance.
- There is a linear relation between the target and the independent variables.

Answer 2: Anscombe's Quartet is a group of datasets (x,y) that have the same mean, standard deviation and regression line but which are qualitatively different.

It is basically used to emphasize the importance of looking at the dataset graphically and not only the statistic properties.

The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.



We can see in this plot that all four plots show unique relations between x and y variables with unique variability and correlation strengths. But all the datasets have the same statistic summary like mean, variance, correlation coefficient and regression line.

Answer 3: Pearson's r is also called Pearson's Correlation Coefficient is a measure of the linear correlation strength between 2 sets of data. It is a normalized measure of covariance, hence it is always between -1 and 1. It measure the direction of the relationship between 2 variables.

If r is between 0 to 1, it means there is a positive linear relationship and if r is between -1 to 0, it means there is a negative or inverse relationship between the variables.

Answer 4: Scaling is basically redefining the values of a variable within a particular given range. Scaling is done to avoid overestimation or underestimation of coefficients or predicting power. It also avoids misrepresentation or mismatch of scale and brings all the variables to a comparable scale.

Normalised scaling scales the data to a specified user defined range like 0 to 1 or -1 to 1. Whereas Standardized scaling scales the data around mean 0 and a standard deviation of 1. Most times Normalised scaling is preferred because it completely negates the chances of getting outliers by defining a specific fixed range, whereas Standardized scaling does not necessary do so.

Answer 5: An infinite value of VIF means that the independent variables are perfectly collinear. This case of perfect multicollinearity means that some variables can perform perfect multiple regression on the

other variables with R-square = 1. When R2 = 1, VIF will be 1 / 1-1, which is infinity. In such a case we need to identify and drop one of the variables that is showing perfect multicollinearity.

Answer 6: Q-Q plot is a technique of plotting probability distributions of 2 samples of datasets. It is useful in checking if the data is normally distributed between quantiles of one dataset against quantile of another dataset.

In Linear Regression Q-Q plot is useful in plotting actual values vs predicted values and check if they are forming a roughly straight line which means they are normally distributed. For e.g we can plot y_test (target variable values of test dataset) against y_pred (predicted values of test dataset) and check if they are aligning on a straightish line. If they align, then the model is working fine.