Assignment-based Subjective Questions

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

Actually 7 categorical variables available in the given data. By using that I have infer the few information. They are,

1. From the variable '**season**', we clear see bike demand is **low** in the '**spring'** season and **high** in the '**fall**' season.
2. From the variable '**mnth**', we clear see bike demand is **high** in '**JUL, AUG, SEP**' comparatively than other months.
3. From the variable '**weathersit'**, we clear see bike demand is **high** in '**Clear**' and **low** in '**Light Snow'**.
4. From the variable '**holiday'**, we clear see bike demand is **low** in **holidays** compare to **working day**.

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

While we create the dummy variables, for the categorical variable we are drop the first column because we can infer the same information with-out that first column also.
For an Example,
There is a field called customer rating which has the below following levels.

Customer rating
Good
Average
Bad

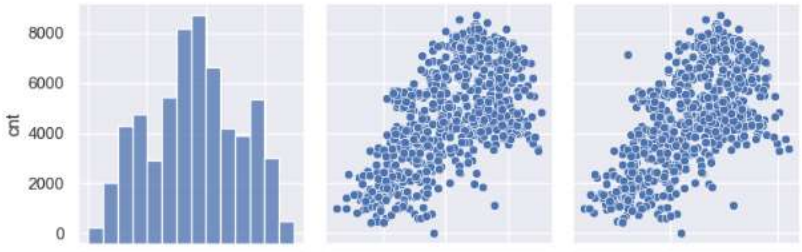
For the customer rating, we are creating dummy variable like below.

Dummy Variable	~~Good~~	Average	Bad
Good	~~1~~	0	0
Average	~~0~~	1	0
Bad	~~0~~	0	1

By seeing the table, we can clear infer the first column value, with second and third, so we are dropping the first column and it will help to reduce the number of which we add in to the model.

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

After seen the pair plot among the numerical variable, "**temp**" and "**atemp**" variable has the highest correlation with target variable "**cnt**".



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

These are the steps done for validated the assumption of linear Regression.

1. There is **a strong linear relationship** between **temp** and **cnt** variables.
2. **Residual Analysis** of the train data – check the error terms are normalized to mean zero.
3. **No visible** pattern in the error term and its **independent**.

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

Based on the final model, these are the top 3 feature contributing significantly towards explaining the demand of the shared bikes.

1. temp
2. yr
3. SEP

General Subjective Questions

1) Explain the linear regression algorithm in detail.

Linear regression is a supervised machine learning method that is used by the Train Using ML model and finds a linear equation that best describes the correlation of the explanatory variables with the dependent variable. This is achieved by fitting a line to the data using least squares. The line tries to minimize the sum of the squares of the residuals. The residual is the distance between the line and the actual value of the explanatory variable. Finding the line of best fit is an iterative process.

Linear Regression formula:-

$$Y=b0+b1*x1+b2*x2+\ldots+bn*xn+E$$

Y – target variable
x – Independent variable.
bi – co-efficient
b0 – intercept
E= error

2) Explain the Anscombe's quartet in detail.

Anscombe's quartet comprises a set of four dataset, having identical descriptive statistical properties in terms of means, variance, R-Squared, correlations, and linear regression lines but having different representations when we scatter plot on graph.
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.
Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

3) What is Pearson's R?

Pearson's r is a numerical summary of the strength of the linear association between the variables. Numerically, the Pearson coefficient is represented the same way as a correlation coefficient that is used in linear regression, ranging from -1 to +1. A value of +1 is the result of a perfect positive relationship between two or more variables. Positive correlations indicate that both variables move in the same direction. Conversely, a value of -1 represents a perfect negative relationship. Negative correlations indicate that as one variable increases, the other decreases; they are inversely related. A zero indicates no correlation.

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

What is scaling?

Scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data pre-processing step.

Why is scaling performed?

Scaling is important because, This is to ensure that no single feature dominates the distance calculations in an algorithm, and can help to improve the performance of the algorithm.

Difference between normalized scaling and standardized scaling

Normalization	Standardization
This method scales the model usi ng minimum and maximum values.	This method scales the model using the mean and s tandard deviation.
When features are on various sca les, it is functional.	When a variable's mean and standard deviation are both set to 0, it is beneficial.
Values on the scale fall between [0, 1] and [-1, 1].	Values on a scale are not constrained to a particular range.
Additionally known as scaling normalization.	This process is called Z-score normalization.
When the feature distribution is unclear, it is helpful.	When the feature distribution is consistent, it is helpful.

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

This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

I have got the VIF value is infinite, after the building the first model, then I have removed 'holiday' feature from the model and it will be fixed.

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

Quantile Quantile plots (Q-Q plots), is a graphical method for determining whether two samples of data came from the same population or not. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value.

Q-Q plots used in linear regression to compare two dataset,

1. come from populations with a common distribution
2. have common location and scale
3. have similar distributional shapes
4. Have similar tail behaviour.