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?

By analyzing each categorical variable with the target variable i.e., 'Count' here the following inferences I have made:

- Year- The median of Boombikes rental in year 2019 is 6000 while the median in 2018 is 3000 .Thus it shows the demand for Boombikes rental is more in 2019 as compared to in year 2018.
  This indicates, Year can be a good predictor for the target variable.
- Holiday- From the median we can see that People seem to rent more in holidays compared to non-holidays. So, reason might be they like to spend more time with their families like Family Cycling Holidays.
- Workingday- Working and Non-working days have almost the same median .Thus, there is no significant change in bike demand .This indicates, workingday is not quite a good predictor of Target variable.
- Month- Month June to Oct has high bike demands .Mainly the fall season has high bike demands and January is the lowest bike demand month.
- Weathersit- The bike demand is high when weather is clear as temperature is optimal and humidity is less.
- Season- The demand of bike is less in the month of spring when compared to Fall season which has the highest median as the weather condition is optimal to ride bikes followed by summer.
- Weekday- Overall median across all days are similar i.e. the demand of bike is almost similar throughout the weekdays.

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

drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations

created among dummy variables. I.e. if we have a categorical variable with 'n' levels, then we need 'n-1' columns representing the dummy variables.

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

Temperature variable has the highest correlation coefficient of 0.64 with the target variable i.e. "count", which means if the temperature increases by one unit the number of bike rentals increases by 0.64 units.

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

The assumptions of Linear Regression can be validate after building the model on the training set:

1. Linear regression states only linear relationship between dependent and independent variables. This can be validated by plotting a scatter plot between the features and the target.
2. Multicollinearity is a state of very high inter-correlations among the independent variables and it can be validate by using Pair-plots and Heatmaps for identifying highly correlated features.
3. Homoscedasiticity can be validate by using scatter plot of residual values vs. predicted values.
4. The fourth assumption is that the error (residuals) follows a normal distribution and it can be validated by plotting a q-q plot.
5. Autocorrelation occurs when the residual errors are dependent on each other. Autocorrelation can be tested with the help of Durbin-Watson test.

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

- The **temperature** variable is having the **highest coefficient 0.64**, which means if the temperature increases by one unit the number of bike rentals increases by 0.64 units.

- **Spring (-0.55), Mist cloudy (-0.18), light snow (-0.23)** variables have **negative coefficient**. The coefficient value signifies how much of the dependent variable changes given a one-unit shift in the independent variable while holding other variables in the model constant.
- **Year** Variable also has a **positive coefficient of 0.59** with the Target variable.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

- Regression analysis is a technique of predictive modeling that helps to find out the relationship between scalar response and one or more explanatory variables (also known as dependent and independent variables).
- For example, you might guess that there's a connection between how much you eat and how much you weigh.
- Linear Regression Algorithm is a machine learning algorithm based on supervised learning.

Regression analysis is used for three types of applications:

1. Finding out the effect of Input variables on Target variable.
2. Finding out the change in Target variable with respect to one or more input variable.
3. To find out upcoming trends.

The types of regressions:

1. Simple Linear Regression- Changing only one variable at a time .
2. Multiple Linear Regression- changing more than one variable at a time
3. Logistic Regression- used to model the probability of a certain class or event existing such as pass/fail, win/lose.

- Mathematically, linear regression equation is:

$$Y = mX + b$$

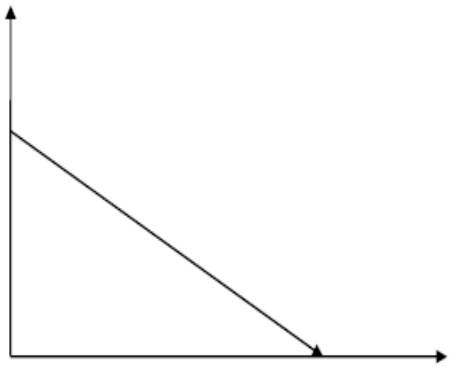
- where Y is Dependent Variable, X is Independent Variable, m is the slop of the regression line which represents the effect X has on Y and b is a constant or intercept.

Positive Linear Relationship

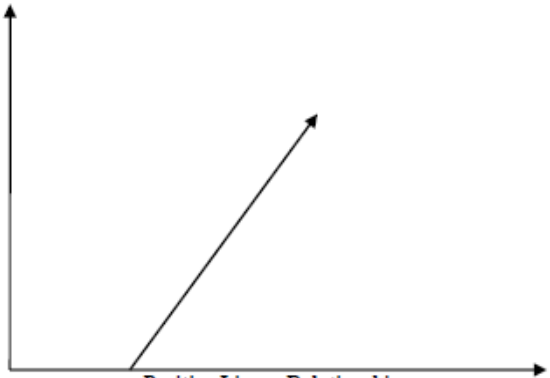
A linear relationship will be called positive if both independent and dependent variable increases. It can be understood with the help of following graph.

Negative Linear relationship

A linear relationship will be called positive if independent increases and dependent variable decreases. It can be understood with the help of following graph.



Negative Linear Relationship



Positive Linear Relationship

Assumptions of Simple Linear Regression

1. The target variable and the input variables are linearly dependent i.e. 'x' and 'y' should display some sorts of a linear relationship otherwise, there is no use of fitting a linear model between them.

2. Error terms are normally distributed with mean zero (not x, y).

- If the error terms are not normally distributed then the P-value obtained during hypothesis test to determine the significance of the coefficients become unreliable.

3. Error are independent of each other i.e. the error terms should not be dependent on one another (like in a time series data wherein the next value is dependent on the previous one).

4. Error terms have constant variance i.e. Homoscedasticity.

- The variance should not increase as the error value changes.

- If the error terms are Heteroscedasticity then we cannot make inferences about the model.

2. Explain the Anscombe's quartet in detail.

- Anscombe's Quartet defined as a group of four data sets which are nearly identical in simple descriptive statistics, which provides same statistical information that involves variance, and mean of all x, y points in all four datasets.

- But there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

- It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties.

- It states about the importance of visualizing the data before applying various algorithms to build models which shows that the data features must be plotted in order to see the distribution of the samples that can help to identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc.

- The Linear Regression can be only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets.

3. What is Pearson's R?

Pearson's R also known as the Pearson correlation coefficient (PCC) is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1.

For example, like expecting the age and height of a sample of teenagers from a high school to have a Pearson correlation coefficient significantly greater than 0, but less than 1 (as 1 would represent an unrealistically perfect correlation).

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

Scaling: Scaling is the procedure of measuring and assigning the objects to the numbers according to the specified rules i.e. the process of locating the measured objects on the continuum, a continuous sequence of numbers to which the objects are assigned.

Why is scaling performed?

In regression, it is important to scale the features so that the predictors have a mean of 0. This makes it easier to interpret the intercept term as the expected value of Y when the predictor values are set to their means.

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 it leads to incorrect modeling. Thus scaling is done to bring all the variables to the same level of magnitude.

What is the difference between normalized scaling and standardized scaling?

Normalization/Min-Max Scaling:

- It brings all of the data in the range of 0 and1.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:

- 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).

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

- One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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

If there is perfect correlation, then VIF = infinity which 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.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

6. 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 which is used to assess if a set of data plausibly came from some theoretical distribution such as a Normal,

exponential or uniform distribution. It helps to determine if two data sets come from populations with a common distribution.

In linear regression we use Q-Q plot to confirm whether the training and test data sets that we got separately are from populations with same distributions.