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?**

Here are a few conclusions I came to after studying categorical data from the dataset that related to the dependent variable (Count).

1. Fall has the highest median, which is expected as weather conditions are most optimal to ride bike followed by summer.
2. Median bike rents are increasing year on as year 2019 has a higher median then 2018, it might be due the fact that bike rentals are getting popular and people are becoming more aware about environment.
3. Overall spread in the month plot is reflection of season plot as fall months have higher median.
4. People rent more on non-holidays compared to holidays, so reason might be they prefer to spend time with family and use personal vehicle instead of bike rentals.
5. Overall median across all days is same but spread for Saturday and Wednesday is bigger may be evident that those who have plans for Saturday might not rent bikes as it a non-working day.
6. Working and non-working days have almost the same median although spread is bigger for non-working days as people might have plans and do not want to rent bikes because of that
7. Clear weather is most optimal for bike renting, as temperate is optimal, humidity is less, and temperature is less.

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

If you don't remove the first column (redundant), dummy variables will be correlated. Some models may be adversely affected by this, and the effect is increased when the cardinality is low. Iterative models, for example, may have difficulty convergent, and lists of variable importance may be distorted. Another argument is that having all dummy variables results in multi collinearity between them. We lose one column to keep everything under control.

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

Ans: “temp” and “atemp” has the highest correlation (0.63) with the target variable.

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

The residuals distribution should be normal and centered at 0. (The average is 0). Making a distplot of the residuals to determine whether they follow a normal distribution allows us to test this residuals assumption.
The residuals are scattered around mean = 0

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

Following are the top 3 features contributing significantly towards explaining the demands of the shared bikes:

1. temp (0.3821)
2. weathersit_drizzle (-0.3200)
3. year (0.2407)

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear Regression is a type of supervised Machine Learning algorithm that is used for the prediction of numeric values. Linear Regression is the most basic form of regression analysis. Regression is the most commonly used predictive analysis model. Linear regression is based on the equation,

“y=ax+b".
It presumes that the predictor(s)/independent variable(x) and the dependent variable(y) have a linear relationship. When performing a regression analysis, we determine the line that best represents the relationship between the independent and dependent variables. When the dependent variable is a continuous data type and the predictor(s) or independent variable(s) can be of any continuous, nominal, categorical, etc. data type. The best fit line that demonstrates the link between the dependent variable and predictors with the least amount of error is what the regression method seeks to find. The output/dependent variable in a regression is a function of the independent variable, the coefficient, and the error term. Regression is broadly divided into simple linear regression and multiple linear regression.

1. **Simple Linear Regression:** SLR is used when the dependent variable is predicted using only one independent variable.
2. **Multiple Linear Regression:**MLR is used when the dependent variable is predicted using multiple independent variables.

The equation for Multiple Linear Regression will be:

$$y_i=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+...+\beta_px_{ip}+\in$$

where, for i=n observations:
yi=dependent variable
xi=explanatory variables
*ϐ*0=y-intercept (constant term)
ϐp=slope coefficients for each explanatory variable
ϵ=the model’s error term (also known as the residuals)

2. Explain the Anscombe’s quartet in detail.

Anscombe's quartet consists of four data sets with virtually similar simple descriptive statistics, but when represented graphically, the distributions are very different. The mean, sample variation of x and y, correlation coefficient, linear regression line, and R-Square value make up the simple statistics. Anscombe's Quartet demonstrates how graphing can nevertheless reveal significant differences between numerous data sets with many comparable statistical features. The graphs are shown below:

Statistical Properties:

- A straightforward linear relationship appears to exist in the first scatter plot (top left).
- While there is a relationship between them, it is not linear, as seen by the second graph (top right), which is not normally distributed.
- Although the distribution in the third graph's bottom left corner is lincar, a different regression line should be used. One outlier cancels out the estimated regression, having enough of an impact to reduce the correlation coefficient from 1 to 0.816.
- The fourth graph, shown in the bottom right corner, illustrates a case where a single high-leverage point might result in a high correlation coefficient even when the other data points do not support a relationship between the variables.

3. What is Pearson’s R?

The strength of a relationship between two variables is measured by Pearson's R. It is calculated by dividing the covariance of two variables by the sum of their standard deviations. Its range of values is +1 to -1.

- A value of 1 denotes a complete linear positive correlation. It implies that if one variable rises, the others will follow suit.
- Zero indicates there is no association.
- A result of -1 indicates a completely negative connection. It implies that if one variable rises, another will fall. 4. What is scaling? Why is scaling performed?

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

To keep a variable within a specified range, scaling is used. In a linear regression study, scaling is a pre-processing step. To speed up the computation of gradient descent, we scale a variable. The gradient descent process will take a very long time if the data contains both small variables (values in the range of 0-1) and big variables (values in the range of 0-1000). The step size of gradient descent is typically low for accuracy.

Normalized Scaling	**Standardized scaling**
Called min max scaling, scales the variable. such that the range is 0-1	Values are centered around mean with a unit standard deviation
Good for non- gaussian distribution	Good for gaussian distribution
Value id bounded between 0 and 1	Value is not bounded
Outliers are also scaled	Does not affect outliers