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?**

    a. Count of rental bikes were significantly high for Summer, fall and winter seasons than Spring.

    b. Year 2019 seen almost twice the number of rental bikes users than year 2018.

    c. Count of rental bikes were significantly high for June, July, August and September months.

    d. Count of rental bikes were lower for Jan, Feb and Dec months.

    e. Non-holiday seen higher number of rental bikes users than holiday.

    f. Clear sky or partly cloudy were the best weather for higher count of rental bikes.

    g. Count of rental bikes seen higher on holidays for casual users compared to non-holiday, it is vice-versa for registered users

    h. There is a significant difference in count of rental bikes numbers between casual and registered users during different weathersit (example – mean count of rental bikes for casual users 1-964 vs 185 and for registered users 1-3912 vs 3-1617)

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

    a. To reduce overall variables for model building, for example if we are creating dummy variable for gender categorical column that contains 'Male', 'Female', and 'Other' variables. if the person is male gender_male dummy column will have a value '1' and gender_female dummy column will have '0' and it is vice versa if the person is female. If is not either of these two then both dummy columns will have '0' value and we do not need the third dummy variable to identify 'other'. If 'n' represents number of category, consider n-1 for dummy categorical columns.

    b. It also helps to reduce the correlations created among dummy variables.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
    a. 'Registered' have the highest correlation with 'cnt' target variable
    b. Casual and temp has the 2nd highest correlation with 'cnt' target variable

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
    a. Plotted Model residual errors have a mean value of zero.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
    a. Temp – 0.43 have highest coefficient followed by year and snow and rainy weather contributing highest on negative side.

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**
    a. technique used to predict correlation between variables and how an independent variable is influenced by the dependent variable(s), is linear regression
    b. After EDA split data into train test
    c. Build a model on training data and validate it
    d. Then predict on test data
    e. Perform validation using r2 score.

2. **Explain the Anscombe's quartet in detail.**
    a. Anscombe's quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.

3. **What is Pearson's R?**
    a. Pearson's correlation coefficient, also known as Pearson's R, is a measure of the strength of correlation between two variables. It is commonly used in linear regression.
    b. The value of Pearson's R always lie between -1 and +1, the latter indicating a perfectly positive and linear correlation and the former indicating a perfectly linear negative regression. The values in between denotes the relative collinearity of two variables

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**
    a. It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range
    b. 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.
    c. Normalized - brings data in the range of 0 and 1.
    d. Standardized – brings data into std normal dist which has mean and std. dev.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**
    a. If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite. It denotes perfect correlation in variables. (1/0).
6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression?**
    a. The Q-Q plot or quantile-quantile plot is a graphical technique for determining if two data sets come from populations with a common distribution.
    b. is used to see if the points lie approximately on the line. If they don't, it means, our residuals aren't Gaussian (Normal) and thus, our errors are also not Gaussian.
    c. Importance are sample size do not need to be equal.