Assignment-based Subjective Questions

1. Most of the categorical variables have negligible effect on the dependent variable, as seen from the correlation matrix and the regression run, which showed the most important features to be ‘registered’, ‘casual’ and ‘temp’, while the others had less to no impact.

2. The **drop_first=True** parameter is important when creating dummy variables to avoid the dummy variable trap, wherein the columns created (dummy) are not independent and contain redundant information. This happens due to the nature of one-hot encoding, where each and every variable is encoded, wherein the same information can be encoded without the creation of any one column.

3. The variable with the highest correlation with the target variable is ‘registered’.

4. By constructing the correlation matrix and calculating the variational inflation factor and checking for violation of such assumptions as multicollinearity.

5. As mentioned before, as per the model, the factors that contribute the most to the calculation of the dependent variable are: ‘registered’, ‘casual’ and the average of ‘atemp’ and ‘temp’.

General Subjective Questions

1. The Linear regression model seeks to establish a linear relationship between a dependent variable and one or multiple independent variables. The relationship is given by the equation: **Y = B0 + B1(X1) + B2(X2) + . . . + Bn(Xn),** where the intercept is given by the B0 value, while the coefficients are a measure of the slope or the relationship between that independent variable and the dependent variable. Some important steps are:
    a. This is a supervised learning algorithm, where the data is labeled and the regression line is fitted initially with random coefficients.
    b. The residual sum of squares is then calculated, which is also known as the cost function or how bad the model is.
    c. The purpose of ordinary least square regression is to minimize the cost function by tweaking the coefficients to find the optimal coefficients.

d. While an analytic solution mostly does not exist, an approximation does seem to generalize well on the data, which gives accurate predictions.

2. Anscombe's quartet is a famous example in statistics consisting of four sets of data. The interesting thing about these datasets is that even though they have very similar summary statistics, like mean, variance, and correlation, they look completely different when visualized. In other words, if you just looked at the calculated numbers for these datasets, they would seem almost identical. But if you plotted them on a graph, you'd see they have very distinct shapes and distributions.
    a. **Dataset 1:** This data appears to follow a roughly linear relationship with some scatter around the line. It reflects a moderately strong positive correlation between x and y.
    b. **Dataset 2:** This dataset shows a clear non-linear pattern, possibly resembling a curve. The x and y values here don't exhibit a linear relationship.
    c. **Dataset 3:** This one appears to have a very tight, almost perfect linear relationship between x and y, except for one point that acts as a significant outlier. This outlier can significantly skew the interpretation if you rely solely on summary statistics.
    d. **Dataset 4:** This dataset visually appears like the x values have a constant relationship with y, except for one outlier. Despite the outlier, the remaining points show almost no variation in y for different x values.

3. Pearson's R is a technique for calculating the correlation coefficient between any two variables. This establishes a linear relationship between the two variables.
    a. The value ranges from -1 to 1.
    b. The interpretation is as follows:
        i. -1 : The variables are perfectly negatively correlated, or have an inverse relationship
        ii. 0 : the variables do not have any correlation, or are independent of each other
        iii. 1 : the variables are perfectly positively correlated, or are dependent on each other.
    c. The formula is : $\sum (x_i - \overline{x})(y_i - \overline{y}) / \sqrt{[\sum (x_i - \overline{x})^2 \sum (y_i - \overline{y})^2]}$

4. Scaling is simply transforming the values of the dataset to be between 0 and 1 or within some specified range. Scaling is done to better compare values or for efficient working of certain machine learning algorithms. The constraint of keeping the values within a certain range keeps the values within the algorithm bounded, and keeps them from vanishing (going to 0) or blowing to infinity.
    a. **Normalization**: This is done when we want all the features to have equal weightage in the final algorithm, where each value is transformed to be between 0 and 1. One such technique is mini-max scaling
    b. **Standardization**: This is done when we want to preserve the underlying distribution of the dataset, hence we subtract by the mean and divide by the standard deviation, to convert the overall mean to 0 and the overall SD to be 1. It is easier to work with, while also preserving the underlying distribution.

5. VIF (Variance Inflation Factor) tends to infinity in the scenario of **perfect multicollinearity** among the independent variables in a regression model.

    Multicollinearity occurs when there's a high degree of correlation between two or more independent variables. This correlation can cause problems in regression analysis because it becomes difficult to isolate the independent effect of each variable on the dependent variable.

    When this correlation becomes so strong that one independent variable can be perfectly predicted from a linear combination of the others, then VIF approaches infinity. This signifies that the variance of the estimated coefficient for that particular variable is infinitely inflated, making its interpretation unreliable.

6. A Q-Q plot, also known as a Quantile-Quantile plot, is a graphical tool used to assess how well the distribution of a dataset aligns with a theoretical distribution, like a normal distribution. In linear regression, it specifically helps us check if the **errors** (residuals) of the model follow a normal distribution.

   Working:

   - A Q-Q plot plots the quantiles of your data (errors in linear regression) against the quantiles of the theoretical distribution (usually normal).
   - Quantiles divide the data into equal-sized portions. So, the 0.1 quantile represents the value below which 10% of the data lies, and the 0.9 quantile represents the value below which 90% of the data lies.
   - If the errors follow the theoretical distribution (often normal), the points in the Q-Q plot should fall approximately along a straight diagonal line.

   Importance:

   - Many statistical tests used in linear regression, like hypothesis testing for coefficients, rely on the assumption of normally distributed errors.
   - A Q-Q plot helps you visually verify this assumption. Deviations from the straight line indicate potential problems with normality.
   - If the errors are not normal, it can affect the reliability of your p-values and confidence intervals in the regression analysis.

   Interpretation:

   - **Straight diagonal line:** This indicates good agreement with the theoretical distribution (often normal).
   - **Curvature:** This suggests the errors might not be normally distributed. The direction of the curvature can provide clues about the type of non-normality (e.g., heavier tails, skewed distribution).

   Overall, the Q-Q plot is a simple but valuable tool for checking normality of residuals in linear regression, helping to ensure the reliability of your analysis.