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? (3 marks)

- Demand for shared bike shows a continuous rise till September. September month has highest demand. After September, demand is decreasing.
- Fall season has highest demand for rental bikes.
- Demand has decreased when there is a holiday.
- ‘weekday’ and ‘workingday’ do not show any clear trend.
- The demand is higher when there is clear weather situation (weathrsit).

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

drop_first=True is important to use, as it helps in reducing the extra columns that gets created during dummy variable creation and reduces 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? (1 mark)

After removing the ‘registered’ and ‘casual’, the numerical value columns ‘temp’ and ‘atemp’ have highest correlation with ‘cnt’

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

- Errors are normally distributed with a mean of 0. Satisfying the normality of error assumption.
- R2 value for predictions on test data (0.68) is very close to R2 value of train data (0.71).
- The prediction for test data is very close to actuals data.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Based on the final model, the top 3 features that are contributing significantly towards the demand of shared bikes are as follows:

```
         coef       std err       t         P>|t|       [0.025       0.975]
-----------------------------------------------------------------------
yr        0.2385      0.011      22.124      0.000        0.217      0.260
temp      0.5556      0.024      22.847      0.000        0.508      0.603
windspeed -0.1964     0.032      -6.109      0.000       -0.260     -0.133
```

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Linear regression is a machine learning algorithm used to model the relationship between a dependent variable and one or more independent variables.

The goal of linear regression is to find a linear relationship between the predictor variables and the response variable. In other words, it tries to fit a line (or a hyperplane in case of multiple predictors) to the data points so that the line best explains the relationship between the variables. The line is defined by the equation:

$$\boxed{y = b0 + b1x1 + b2x2 + ... + bn*xn}$$

where y is the dependent variable, x1, x2, ..., xn are the independent variables, b0 is the y-intercept (or the constant term), and b1, b2, ..., bn are the coefficients (or weights) that represent the slope of the line with respect to each predictor variable.

The linear regression algorithm works by minimizing the difference between the predicted values and the actual values of the response variable. This difference is called the residual or the error. The algorithm tries to find the values of the coefficients that minimize the sum of the squared errors (SSE) between the predicted values and the actual values. The SSE is defined by the equation:

$$\boxed{SSE = \Sigma(yi - \hat{y}i)\hat{}2}$$

where yi is the actual value of the response variable for the ith observation, and ŷi is the predicted value of the response variable for the ith observation.

The process of finding the optimal values of the coefficients is called training the model. The algorithm uses a training dataset, which consists of input (predictor) variables and corresponding output (response) variable values, to find the optimal values of the coefficients.

There are different methods to train the linear regression model, including the ordinary least squares (OLS) method, gradient descent, and stochastic gradient descent. The OLS method is the most common method used in linear regression. It involves finding the values of the coefficients that minimize the sum of the squared errors using a closed-form solution.

Once the model is trained, it can be used to make predictions on new data by simply plugging in the values of the predictor variables into the equation and calculating the predicted value of the response variable**.**

2. Explain the Anscombe’s quartet in detail. (3 marks)

Anscombe’s quartet is a set of four datasets that have identical statistical properties but appear quite different when plotted.

i. The quartet consists of four sets of data, each with eleven (x, y) points:
   a. Set I: (10, 8.04), (8, 6.95), (13, 7.58), (9, 8.81), (11, 8.33), (14, 9.96), (6, 7.24), (4, 4.26), (12, 10.84), (7, 4.82), (5, 5.68)

   b. Set II: (10, 9.14), (8, 8.14), (13, 8.74), (9, 8.77), (11, 9.26), (14, 8.10), (6, 6.13), (4, 3.10), (12, 9.13), (7, 7.26), (5, 4.74)

   c. Set III: (10, 7.46), (8, 6.77), (13, 12.74), (9, 7.11), (11, 7.81), (14, 8.84), (6, 6.08), (4, 5.39), (12, 8.15), (7, 6.42), (5, 5.73)

    d. Set IV: (8, 6.58), (8, 5.76), (8, 7.71), (8, 8.84), (8, 8.47), (8, 7.04), (8, 5.25), (19, 12.50), (8, 5.56), (8, 7.91), (8, 6.89)

ii. Each of the four datasets have the same summary statistics:
    a. The mean of x is 9
    b. The mean of y is approximately 7.5
    c. The variance of x is approximately 11
    d. The variance of y is approximately 4.1
    e. The correlation between x and y is approximately 0.82

iii. Despite having the same summary statistics, the datasets have very different patterns when plotted. For example:
    a. Set I appear to have a linear relationship between x and y.
    b. Set II appears to have a non-linear relationship between x and y, possibly a logarithmic relationship.
    c. Set III appears to have an outlier that is affecting the correlation coefficient and the regression line.
    d. Set IV appears to have a vertical relationship between x and y, with all x values being the same.

iv. The quartet was created by the statistician Francis Anscombe in 1973 to demonstrate the importance of graphing data before analyzing it. The quartet shows that summary statistics can be misleading and that it is important to examine the data visually to identify any patterns or outliers that may affect the analysis.

3. What is Pearson's R? (3 marks)

Pearson's R is a statistical measure that quantifies the strength and direction of the linear relationship between two variables. It is denoted by the symbol "r" and takes values between -1 and +1.

A value of +1 indicates a perfect positive correlation between two variables, meaning that as one variable increases, the other variable also increases. A value of -1 indicates a perfect negative correlation, meaning that as one variable increases, the other variable decreases. A value of 0 indicates no correlation between the variables.

Pearson's R is calculated as the covariance between two variables divided by the product of their standard deviations. The formula for Pearson's R is as follows:

r = (Σ[(x - mean(x)) * (y - mean(y))]) / [sqrt(Σ(x - mean(x))^2) * sqrt(Σ(y - mean(y))^2)]

where x and y are the two variables being correlated, mean(x) and mean(y) are their respective means.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Scaling refers to the process of transforming a variable or set of variables to a similar scale or range. In data analysis, scaling is performed to standardize or normalize the data to ensure that different variables are on a common scale, which makes it easier to compare them and to perform certain analyses.

Scaling is performed for various reasons, including:

- To improve the performance of certain algorithms.
- To facilitate the interpretation of data.

- To improve the convergence of optimization algorithms.

There are two main types of scaling: normalized scaling and standardized scaling.

- Normalized scaling involves transforming the data so that it falls within a specific range, typically between 0 and 1. This is done by subtracting the minimum value from each data point and dividing by the range of the data (i.e., the difference between the maximum and minimum values). Normalized scaling is appropriate when the range of the data is known and fixed.
- Standardized scaling involves transforming the data so that it has a mean of 0 and a standard deviation of 1. This is done by subtracting the mean from each data point and dividing by the standard deviation. Standardized scaling is appropriate when the range of the data is not known or when it is not fixed.

The main difference between normalized scaling and standardized scaling is that normalized scaling preserves the shape of the distribution of the data, whereas standardized scaling transforms the data so that it has a specific mean and standard deviation. Additionally, standardized scaling is more appropriate when the range of the data is not known or when it is not fixed, while normalized scaling is more appropriate when the range of the data is known and fixed.

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

(3 marks)

The VIF value may be calculated as infinite, when one of the independent variables in the regression model is perfectly correlated with a linear combination of the other independent variables. In other words, the variable is redundant and can be expressed as a linear combination of the other variables in the model. When this occurs, the denominator of the formula for VIF becomes zero, leading to an infinite value. A high VIF value indicates that there is a high degree of multicollinearity among the independent variables, which can affect the reliability and interpretability of the regression coefficients.

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

(3 marks)

A Q-Q (quantile-quantile) plot is a graphical technique used to compare the distribution of a sample to a specified theoretical distribution. In linear regression, Q-Q plots are often used to assess the normality assumption of the residuals, which is an important assumption for valid inference in regression analysis.

Here's how a Q-Q plot works:

- The observed data are sorted in ascending order.
- The corresponding quantiles of the theoretical distribution are calculated.
- The observed quantiles are plotted against the expected quantiles.

If the observed data follow the specified theoretical distribution, the points on the Q-Q plot should form a roughly straight line. Deviations from a straight line indicate departures from the specified distribution.

If the Q-Q plot of the residuals shows deviations from a straight line, this indicates departures from normality. This could be due to various reasons such as outliers, heavy-tailed distributions, or skewness in the data. In such cases, it may be necessary to consider alternative modeling techniques.