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)

Ans: The following conclusions concerning their impacts on the dependent variable—in this case, the count of bike rentals (cnt)—would be possible to draw from the linear regression analysis considering categorical variables:

Seasonality: There's a good chance that the season variable, which divides the data into seasons like spring, summer, fall, and winter, has a significant impact on bike rentals. Due to variations in weather, different seasons can have a substantial impact on rental trends; generally speaking, warmer seasons see greater rentals and colder ones, lower.

Month (mnth) of the Year:
Just like with seasons, rental counts can also be impacted by the month.
For instance, rental activity typically rises in months with better weather.
Month's inclusion as a categorical variable allows the model to account for variations in demand monthly.

Day of the Week (weekday): This variable can provide light on trends that occur every week, including the distinctions between weekdays and weekends.
Weekend rental trends may vary based on events and leisure activities, while weekday rental patterns may be constant and may be tied to commuting.

Weather Situation (weathersit): It is anticipated that this variable, which represents the daily weather (clear, hazy, light rain, etc.), will significantly affect the number of rentals. While bad weather may discourage individuals from renting bikes, clear and moderate weather is probably linked to higher rentals.

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

Ans: drop_first=True is used during dummy variable creation to avoid the issues of multicollinearity, which can negatively affect the performance of linear regression model. While creating a dummy variable from a categorical features with N possible values, we transforming those values into feature corresponding to one of the original categorical values. When all the dummy variables are included, we are essentially introducing multicollinearity because the sum of all the dummy variable will always equal. To avoid this from happening we use drop_frst=True.

3. Looking at the pair-plot among the numerical variables, which

one has the highest correlation with the target variable? (1 mark)

Ans: In the pari plot, ‘temp’ and ‘atemp’ show the highest correlation with the target vairbale ‘cnt’, as indicated by the updwared trend in scatter plot against bike rental counts.

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

Ans: To confirm that the linear regression model was appropriate for the training data, I evaluated its key assumptions. First, I scrutinized the scatter plot of residuals versus predicted values, ensuring no apparent patterns emerged that would contradict the assumption of a linear relationship.

Next, I checked for constant variance (homoscedasticity) in the residuals across all predicted values; a lack of a pattern, such as a funnel shape, would affirm this condition.

I then turned to the distribution of the residuals. Using a histogram and Q-Q plot, I verified that the residuals approximated a normal distribution centered around zero, which is crucial for the validity of the model's statistical tests.

Finally, to tackle the issue of multicollinearity, I examined the correlation between independent variables and considered the variance inflation factors (VIFs). Adjustments were made to the model as necessary to reduce multicollinearity, such as excluding or combining highly correlated variables.

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

Ans: The final linear regression model identified temperature ('temp'), feeling temperature ('atemp'), and year ('yr') as the three most influential features in predicting bike-sharing demand. Temperature and its perceived counterpart directly influence rental behavior due to their impact on outdoor comfort. The 'yr' variable captured the increase in bike-sharing popularity over time, reflecting a growing trend or greater awareness and usage of the service from one year to the next.

General Subjective Questions

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

Linear regression is a statistical method used for predictive analysis. It establishes a relationship between a dependent variable (Y) and one or more independent variables (X) using a linear equation. The core idea is to find the best-fitting straight line through the data points.

Ans: This line is represented by the equation Y = a + bX + e, where:

- Y is the dependent variable you're trying to predict,

- X is the independent variable you're using for prediction,

- a is the y-intercept, which represents the value of Y when X is 0,

- b is the slope of the line, which indicates how much Y changes for a one-unit change in X,

- e is the error term, accounting for the variation in Y not explained by the X variables.

In simple linear regression, we use one independent variable to predict the dependent variable. In multiple linear regression, several independent variables are used.

The process involves fitting the model by calculating the coefficients a (intercept) and b (slope) that result in the smallest possible difference between the predicted values and the actual values. This difference is known as the residual for each data point, and the best-fitting line is found by minimizing the sum of the squares of the residuals, a method known as Ordinary Least Squares (OLS).

The model's effectiveness is evaluated using metrics like R-squared, which measures the proportion of variance in the dependent variable that can be explained by the independent variables in the model. The goal is a model that accurately predicts Y and provides insights into the relationships between the variables.

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

Anscombe's quartet comprises four distinct datasets created by the statistician Francis Anscombe in 1973. Each dataset contains eleven points and is designed to have nearly identical statistical properties, specifically the same mean, variance, correlation, and linear regression line (to two decimal places) when applied to X and Y variables.

Ans: The purpose of Anscombe's quartet is to demonstrate the importance of visual data analysis before statistical analysis. Despite the four datasets having nearly identical statistical characteristics, they have vastly different distributions and appear very different when graphed. Each dataset tells a unique story:

1. The first set forms a typical pattern that one would expect when assuming a linear relationship, fitting the assumptions of linear regression well.
2. The second is clearly curved, indicating a non-linear relationship that linear regression would fail to capture adequately.
3. The third dataset shows a linear relationship with an outlier influencing the regression line significantly.
4. The fourth has a low variance in X with one outlier driving the correlation.

Anscombe's work highlights the risks of relying solely on summary statistics and underscores the value of graphical summaries in data analysis. It serves as a powerful reminder that identical statistical properties can lead to misleading conclusions without graphical representation.

3. What is Pearson’s R? (3 marks)

Ans: Pearson's R, also known as Pearson's correlation coefficient, is a measure of the linear correlation between two variables, X and Y. It is a value between -1 and 1, where 1 means that there is a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 means no linear correlation. The coefficient is obtained by dividing the covariance of the two variables by the product of their standard deviations. It quantifies the degree to which a relationship between two variables can be described by a line. In practical terms, Pearson’s R assesses how well two variables move together: if Pearson’s R is positive, as one variable increases, the other tends to increase as well; if it's negative, as one increases, the other tends to decrease. The magnitude of the coefficient indicates the strength of the correlation; the closer to 1 or -1, the stronger the correlation. Pearson's R is a widely used statistical tool because of its simplicity and the intuitive interpretation of its value.

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

Ans: Scaling is a method used in data preprocessing to adjust the range of variable values. The process ensures that each feature contributes equally to the analysis, particularly important in algorithms that compute distances or assume normality.

Scaling is performed to bring all numerical variables into a specific range, preventing variables with larger scales from dominating the model's learning process. It helps improve the convergence speed in gradient descent, maintains numerical stability, and is essential for techniques like PCA that are sensitive to variances.

Normalization and standardization are two common scaling methods. Normalization, often referred to as Min-Max scaling, rescales data to a fixed range, typically [0, 1], without distorting differences in the ranges of values. It subtracts the minimum value and divides by the range.

Standardization, on the other hand, rescales data to have a mean of 0 and a standard deviation of 1, transforming it into a distribution with a z-score. This doesn't bind values to a specific range, which can be useful for algorithms that require data to have a Gaussian distribution.

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

(3 marks)

Ans: The Variance Inflation Factor (VIF) quantifies how much the variance of an estimated regression coefficient increases due to multicollinearity. A VIF becomes infinite when the independent variable is perfectly collinear with other independent variables in the model, meaning there's an exact linear relationship between them. This situation implies that the independent variable in question can be perfectly predicted from the others with no error, leading to a division by zero in the VIF formula, hence an infinite value. It signals the need to remove or combine features to alleviate multicollinearity.

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

(3 marks)

Ans: A Q-Q (quantile-quantile) plot is a graphical tool to assess if a set of data potentially came from some theoretical distribution such as a Normal, Uniform, or Exponential distribution. In the context of linear regression, a Q-Q plot is used to verify the assumption that the residuals are normally distributed. It plots the quantiles of the residuals against the expected quantiles of a normal distribution. If the points on the plot fall approximately along a straight line, it suggests that the residuals exhibit normality. The use and importance of a Q-Q plot in linear regression lie in its ability to visually check for deviations from normality, which is critical since many inferential techniques in linear models assume normally distributed error terms. Deviations from this assumption can affect the validity of statistical tests and confidence intervals calculated from the model.