Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans: from the analysis of the categorical variables, here's what I can infer:

Season: The season affect bike demand. Winter have higher demand, while spring have lower demand.

Weather: Clear weather increase bike demand, while rain and snow reduce it.

Holiday: Bike demand is slightly lower on holidays compared to regular working days.

Working Day: Bike demand is higher on working days than non working days.

Weekday: Bike demand is a bit higher on weekdays than weekends.

In summary, seasonality, weather, and work schedules all influence bike rental demand.

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

Answer: Using drop_first=True during dummy variable creation is important to avoid the dummy variable trap. This trap happens when one dummy variable is perfectly correlated with the others. This leads to multicollinearity in regression models, where the model can't tell the difference between the effects of each variable because they are perfectly correlated. Dropping the first category from the dummy variables gives the model a reference point for comparison while avoiding perfect multicollinearity. This results in more stable and easier-to-interpret coefficients.

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

Answer: From the pair-plot, it looks like temperature (temp) have the highest correlation with the target variable (cnt). Higher temperatures seems to lead to more bike rentals, likely because warmer weather encourages more outdoor activities.

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

Answer: To validate the linear regression assumptions, I did the following:

Checked linearity by looking at residual plots. The residuals were randomly scattered around zero, confirming linearity.

Tested normality of errors by plotting a histogram of the residuals. The residuals were centered around zero, supporting normal distribution.

Assessed homoscedasticity by plotting residuals against predicted values. There was no clear pattern, indicating constant variance.

Looked at Variance Inflation Factors (VIFs) to check for multicollinearity. Features with high VIFs, like atemp, were removed to improve model stability.

These steps helped ensure the linear regression model met its underlying assumptions.

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

Answer: According to the final model, the top 3 features that significantly contribute to explaining the demand for shared bikes are:

1) Year (yr): The demand in 2019 (represented as 1) was higher than in 2018, showing the service gained popularity over time.

2)Temperature (temp): Higher temperatures had a positive impact on the demand, as more people rented bikes in warmer weather.

3)Weather Situation (weathersit): Clear weather positively affected demand, while mist and light rain/snow negatively impacted it. This suggest weather conditions are a strong factor.

These key features provide important insights for predicting bike demand and adjusting business strategies accordingly.

General Subjective Questions

Question 1. Explain the linear regression algorithm in detail.

Answer: Linear regression is a statistical technique used to model the relationship between a dependent variable (target) and one or more independent variables (features). The goal is to fit a line (simple linear regression) or a hyperplane (multiple linear regression) that best represents this relationship.

The model is represented by the equation: $y = \beta0 + \beta1x + \varepsilon$ where y is the dependent variable, x is the independent variable, $\beta0$ is the intercept, $\beta1$ is the slope, and ε is the error term.

In multiple linear regression, the equation is extended to include more independent variables: $y = \beta0 + \beta1x1 + \beta2x2 + ... + \beta nxn + \varepsilon$

The objective is to estimate the coefficients $\beta0, \beta1, ..., \beta n$ such that the difference between the predicted values (y^) and the actual values of y is minimized. This is usually done using Ordinary Least Squares (OLS), which minimizes the sum of the squared residuals.

The key assumptions of linear regression are: linearity, independence, homoscedasticity, and normality of residuals.

The coefficients β represent the change in the dependent variable for a one-unit change in the corresponding independent variable.

Question 2. Explain the Anscombe's quartet in detail.

Answer: Anscombe's quartet is a set of four datasets that have nearly identical summary statistics, but look very different when plotted. Statistician Francis Anscombe created this in 1973 to show the importance of visualizing data, not just relying on numerical stats.

The four datasets have the same mean, variance, correlation and regression line. But their plots reveal very different relationships:

1. A linear relationship
2. A clear curve
3. A linear relationship with an outlier
4. A vertical relationship with one influential outlier

The key takeaway is that you should always visualize data before analyzing it. Summary stats alone can be misleading - you need to see the actual data patterns to avoid wrong conclusions.

Question 3. What is Pearson's R?

Answer: Pearson's R is a measure of the linear relationship between two continuous variables. It shows both the strength and direction of the relationship.

The formula is: $r = \Sigma(x_i - \bar{x})(y_i - \bar{y}) / \sqrt{(\Sigma(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2)}$

Where:

- r is the correlation coefficient

- x_i and y_i are the values of the two variables
- x̄ and ȳ are the means of the variables

Interpretation:

- r = 1 is a perfect positive linear relationship
- r = -1 is a perfect negative linear relationship
- r = 0 means no linear relationship

The value of r ranges from -1 to 1. The higher the absolute value, the stronger the linear relationship.

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

Answer: Scaling is the process of transforming features to be on the same scale. This is important because many machine learning algorithms are sensitive to the scale of the input data and may give more importance to variables with larger ranges.

Scaling is performed to ensure that each feature contribute equally to the model, preventing features with larger ranges from dominating. It also improve the convergence speed of optimization algorithms and can lead to better model performance.

The difference between Normalized Scaling and Standardized Scaling:

Normalization: Rescales the data to a range of [0, 1]. It transform each feature by subtracting the minimum value and dividing by the range (max - min). Use case is when you want data in a specific range, like for neural networks.

Standardization: Transform data to have a mean of 0 and a standard deviation of 1. Each feature is centered by subtracting the mean and scaled by dividing by the standard deviation. Standardization is preferred when the data follows a normal distribution or when algorithms expect data centered around 0, like linear regression and SVM.

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

Answer: The VIF becomes infinite when there is perfect multicollinearity among the independent variables. This means that one variable can be express as a perfect linear combination of other variables. In this case, the model can't separate the effects of the correlated variables, leading to the inability to calculate the unique contribution of each variable.

When perfect multicollinearity occurs, the model's matrix inversion step fails, causing the VIF calculation to approach infinity.

To fix this, you should remove or combine the collinear variables to eliminate redundancy and make the model stable.

Question 6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
Answer: A Q-Q (Quantile-Quantile) plot is a chart that compares the distribution of data to a theoretical distribution, usually the normal distribution.

How it works: The Q-Q plot shows the quantiles of the observed data against the quantiles of the theoretical distribution. If the data follows the theoretical distribution, the points will lie approximately on a straight line.

Use in Linear Regression: In linear regression, one assumption is that the residuals (errors) should be normally distributed. The Q-Q plot is used to check this assumption. If the residuals are normal, the points will fall on the diagonal line. Deviations from this line suggest the residuals are not normal.

Importance: Checking the normality of residuals is key for valid statistical inference in linear regression. The Q-Q plot helps identify iss