Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer:

As per the given dataset, if we take the categorical variable 'weathersit' and consider its effect on target variable 'cnt'. Upon doing EDA on between both of these, it can be derived that during the weathersit_3 (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered), there is a decrease in bike hiring stats by 0.333164 units.

After building the model on inclusion of categorical features such as yr and season we will be able to see a change in the value of R-squared and adjusted R-squared. Hence, it can be concluded that categorical features were helpful in explaining a greater proportion of variances in the data sets

---

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: When creating dummy variables (dummy encoding), it's recommended to set drop_first=True to avoid introducing redundant features. Without this, the resulting dummy variables could be highly correlated, as the first column is typically used as the reference group in the encoding process.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

Total Marks: 1 mark (Do not edit)

Answer: The numerical feature 'registered' (with a correlation of 0.95) initially shows the highest correlation with the target variable 'cnt' when all features are included. However, after data preparation and dropping 'registered' due to multicollinearity, the numerical variable 'atemp' (with a correlation of 0.63) emerges as having the strongest correlation with the target variable 'cnt'.

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: After developing the model on the training set, I conducted the following analysis:

Assumptions of Linear Regression:

1. There is a linear relationship between the independent variables (X) and the dependent variable (Y).
2. The error terms are normally distributed with a mean of zero (note: this refers to the errors, not the variables X or Y).
3. Residual analysis of the training data indicates that the residuals are normally distributed.
4. Based on this, the assumptions for linear regression are considered valid.
5. Independent variables were included or excluded from the models based on Variance Inflation Factor (VIF) and p-values to mitigate multicollinearity.

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)

Answer: According to our final model, the top three predictor variables influencing bike bookings are:

1. Temperature (temp): With a coefficient of 0.375922, a one-unit increase in temperature leads to an increase of 0.375922 units in bike hire numbers.
2. Weather Situation 3 (weathersit_3): Representing conditions like Light Snow, Light Rain + Thunderstorm, Scattered Clouds, and Light Rain + Scattered Clouds, the coefficient value of -0.333164 suggests that, with respect to weathersit_3, a one-unit increase in this variable results in a decrease of 0.333164 units in bike hire numbers.
3. Year (yr): The coefficient of 0.232965 indicates that a one-unit increase in the year (i.e., moving from one year to the next) results in an increase of 0.232965 units in bike hires.

Given the influence of these variables, it is recommended to prioritize them when planning for maximum bike bookings.

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Linear regression is a statistical method used to model the relationship between a dependent variable (YYY) and one or more independent variables (XXX). The goal is to find the best-fitting line (in simple linear regression) or hyperplane (in multiple linear regression) that minimizes the difference between observed and predicted values.

Linear regression makes several assumptions: there must be a linear relationship between XXX and YYY, errors should be normally distributed with constant variance (homoscedasticity), and predictors should not be highly correlated (no multicollinearity).

The model is usually trained using the Normal Equation or Gradient Descent. The Normal Equation provides a closed-form solution, while Gradient Descent is an iterative optimization method used for larger datasets.

The performance of the model is evaluated using metrics like R-squared, which measures how well the model explains the variability of the target variable.

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Anscombe's Quartet is a set of four datasets that have nearly identical simple descriptive statistics, yet very different distributions and relationships between variables. It was created by statistician

Francis Anscombe in 1973 to demonstrate the importance of visualizing data before drawing conclusions based solely on statistical measures.

The four datasets in Anscombe's Quartet all have the same:

- Mean of xxx (the independent variable),
- Mean of yyy (the dependent variable),
- Variance of xxx,
- Variance of yyy,
- Correlation between xxx and yyy, and
- The least-squares regression line.

However, when plotted, the datasets reveal distinct patterns:

1. Dataset 1 shows a linear relationship, where a straight line fits the data well.
2. Dataset 2 shows a strong linear relationship but includes one extreme outlier that significantly affects the regression line.
3. Dataset 3 shows a nonlinear relationship, where the data points follow a curve rather than a straight line.
4. Dataset 4 shows a case where most data points are clustered in a horizontal line with one outlier that has a large influence on the regression line.

Anscombe's Quartet emphasizes the importance of using graphical methods, such as scatter plots, to understand the underlying patterns in data, rather than relying solely on summary statistics, which can be misleading.

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Pearson's rrr, also known as the Pearson correlation coefficient, measures the strength and direction of the linear relationship between two continuous variables. It provides a value between -1 and +1:

- r=+1r = +1r=+1 indicates a perfect positive linear relationship, meaning that as one variable increases, the other increases in a perfectly proportional way.
- r=−1r = -1r=−1 indicates a perfect negative linear relationship, meaning that as one variable increases, the other decreases in a perfectly proportional way.
- r=0r = 0r=0 indicates no linear relationship between the variables.

Values closer to +1 or -1 suggest a stronger linear relationship, while values near 0 indicate a weaker or no linear relationship. Pearson's rrr assumes that the relationship between the variables is linear and that the data are normally distributed. It is sensitive to outliers, which can significantly affect the correlation value.

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. It is extremely important to rescale the variables so that they have a comparable scale. If

we don't have comparable scales, then some of the coefficients as obtained by fitting the regression model might be very large or very small as compared to the other coefficients.
Normalized scaling means to scale a variable to have values between 0 and 1, while standardized scaling refers to transform data to have a mean of zero and a standard deviation of 1

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

The Variance Inflation Factor (VIF) is a measure of colinearity among predictor variables within a multiple regression. It is calculated by taking the the ratio of the variance of all a given model's betas divide by the variane of a single beta if it were fit alone.
If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to compare the distribution of a dataset to a theoretical distribution, such as a normal distribution. It plots the quantiles of the data against the quantiles of the chosen theoretical distribution. If the data follows the theoretical distribution, the points in the Q-Q plot will lie approximately along a straight line.

<u>Use of a Q-Q Plot</u>

- Check for normality: A Q-Q plot is primarily used to assess if a dataset follows a specific distribution, commonly the normal distribution. In the context of linear regression, we often use it to check if the residuals (errors) are normally distributed.
- Detect skewness and kurtosis: The plot helps identify if the data has skewness (asymmetry) or kurtosis (heavy tails) compared to the normal distribution. Deviations from the straight line indicate departures from normality.

<u>Importance of a Q-Q Plot in Linear Regression</u>

In linear regression, one of the key assumptions is that the residuals (the differences between observed and predicted values) should be normally distributed. This assumption is important because it impacts the validity of hypothesis tests, confidence intervals, and significance tests of the model coefficients.

- Assessing normality of residuals: A Q-Q plot of the residuals helps visually assess whether the residuals are normally distributed. If the residuals follow a normal distribution, the points will closely align with the reference line in the plot. If there are large deviations from the line, it suggests that the residuals are not normally distributed, which could invalidate certain statistical inferences (like p-values and confidence intervals).

- Detecting model issues: A Q-Q plot can also help identify issues such as outliers or non-linear relationships in the data. For example, if the residuals show a pattern or if the points deviate significantly from the line, it may indicate that the model is not capturing some aspect of the data (e.g., non-linearity or heteroscedasticity).