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? (3 marks)
A.The final Multiple Linear Regression model contains many predictor variables that are categorical in nature and some of them have been encoded to dummy variables. spring, winter falls under season category and have been dummy encoded. weathersit_2 and weathersit_3 falls under weathersit category and have been dummy encoded. Similarly, month variablesfall under mnth category and have been dummy encoded. We can infer from above image that these variables

Q2.Why is it important to use drop_first=True during dummy variable creation? (2marks)
of linear regression. So, if we have k levels where k >=3, we only use k-1 levels while dummy variable encoding. The dropped level gets handled by intercept as a base case.

Q3.Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

and registered in our pre-processed training data for model training. casual + registered= cnt. This might leak out the crucial information and model might get overfit.So, excluding these two variables atemp is having highest correlation with target
variable cnt which is followed by temp.As per the correlation heatmap, correlation coefficient between atemp and cnt is
0.65. And correlation coefficient between temp and cnt is 0.64

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
Residual Analysis:
We need to check if the error terms are also normally distributed (which is infact, one of the major assumptions of linear regression).The residuals are following the normally distribution with a mean 0.

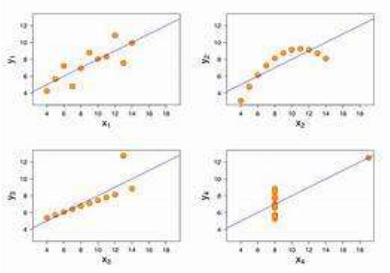
Linear relationship between predictor variables and target variable: This is happening because all the predictor variables are statistically significant (p-values
Error terms are independent of each other: Handled properly in the model. The predictor variables are independent of each other. Multicollinearity issue is not there because VIF (Variance Inflation Factor) for all predictor variables are below 5.

Q5.Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
A5.Top 3 featuressignificantly contributing towards demand ofshared bikes are:
1.temp
2.yr
3.Weathrsit

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
A1.Linear Regression finds the best linear relationship between the independent and dependent variables.It is a method of finding the best straight-line fitting to the given data.In technical terms, linear regression is a machine learning algorithm that findsthe best linear-fit relationship on any given data, between independent and dependentvariables. It is mostly done by the Sum of Squared Residuals Method.The assumptions of linear regression are: a. The assumption aboutthe form of the model: It is assumed that there is a linear
b. Assumptions about the residuals: 1) Normality assumption: It is assumed that the error terms, ε(i), are normally distributed. 2) Zero mean assumption: It is assumed that the residuals have a mean value of zero, i.e., the error terms are normally distributed around zero. 3) Constant variance assumption: It is assumed that the residual terms have the same (but unknown) variance,sigma square. This assumption is also known as the assumption of homogeneity or
c. Assumptions about the estimators: 1) The independent variables are measured without error. 2) The independent variables are linearly independent of each

2. Explain the Anscombe’s quartet in detail. (3 marks)
A, Anscombe's quartet comprisesfour data sets that have nearly identicalsimple descriptive statistics, yet have very different distributions and appear very different when graphed.



Here are the key points about Anscombe’s quartet:
1. Identical Statistics: All four datasets have the same mean, variance, correlation, and linear regression line. This means that if you only look at these summary statistics, the datasets appear to be very similar.
2. Different Graphical Representations: When plotted, each dataset looks very different:

Dataset 1: Appears to fit a linear regression model well.
Dataset 2: Shows a clear non-linear relationship.
Dataset 3: Contains an outlier that influences the regression line.
Dataset 4: Also has an outlier, but in a different pattern than Dataset 3.

Q What is Pearson’s R? (3 marks)

Pearson’s R, also known as the **Pearson correlation coefficient**, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. Here are the key points:
1. Range and Interpretation: The value of Pearson’s R ranges from -1 to 1:

+1 indicates a perfect positive linear relationship.
-1 indicates a perfect negative linear relationship.
0 indicates no linear relationship1.

Q3.What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
A.caling in data science refers to adjusting the range of your data so that it fits within a specific scale. This is particularly important when dealing with variables that have different units or vastly different ranges.
Why is Scaling Performed?
Scaling is performed for several reasons:

Improves Model Performance: Many machine learning algorithms, especially those that use distance measurements (like k-nearest neighbors or support vector machines), perform better when the data is scaled.
Speeds Up Convergence: Algorithms like gradient descent converge faster when the data is scaled.
Ensures Fairness: Scaling ensures that no single feature dominates others due to its scale, leading to more balanced and fair models1.

Difference Between Normalized Scaling and Standardized Scaling

Normalized Scaling (Min-Max Scaling):

Definition: Transforms the data to fit within a specific range, usually [0, 1] or [-1, 1].
Formula: Xnew=Xmax−XminX−Xmin

Use Case: Useful when you want to ensure that all features are on the same scale, especially when there are no outliers2.

Standardized Scaling (Z-Score Normalization):

Definition: Transforms the data to have a mean of 0 and a standard deviation of 1.
Formula: Xnew=stdX−mean

Q5.You might have observed that sometimes the value of VIF is infinite. Why does this happen?
A5.The Variance Inflation Factor (VIF) measures how much the variance of a regression coefficient is inflated due to multicollinearity in the model. When the value of VIF is infinite, it indicates perfect multicollinearity. Here's why this happens:

Perfect Multicollinearity: This occurs when one predictor variable in a regression model is an exact linear combination of other predictor variables. In such cases, the denominator in the VIF formula (which involves the R-squared value of the regression of one predictor on the others) becomes zero, leading to an infinite VIF.

Mathematical Explanation: The VIF for a predictor variable (X_i) is calculated as:
$VIF(Xi)=1-Ri21$
where (R_i^2) is the R-squared value obtained by regressing (X_i) on all other predictor variables. If (R_i^2 = 1), which means perfect multicollinearity, the denominator becomes zero, making the VIF infinite.

Q6.What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
A.A Q-Q plot plots the quantiles of the data against the quantiles of a specified theoretical distribution. If the data follows the theoretical distribution, the points on the Q-Q plot will lie approximately along a straight line.

Use and Importance in Linear Regression
Assessing Normality of Residuals: In linear regression, one of the key assumptions is that the residuals (errors) are normally distributed. A Q-Q plot helps to visually assess this assumption. If the residuals are normally distributed, the points on the Q-Q plot will fall along a straight diagonal line1.
Identifying Deviations: The Q-Q plot can reveal deviations from normality, such as skewness or kurtosis. For example:
Left-Skewed Data: Points will deviate from the line, arching upwards on the left.
Right-Skewed Data: Points will deviate downwards on the right.
Heavy Tails: Points will deviate from the line at the ends, indicating more extreme values than expected under normality2.