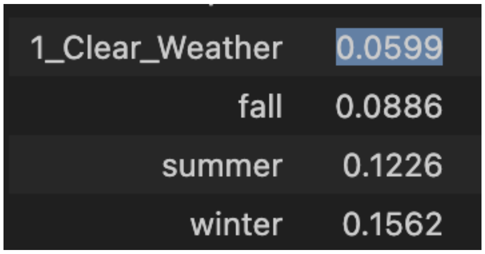
Assignment-based Subjective Questions

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

   Categorical variables have a positive impact on the target variable as their coefficient is positive.



1_Clear_Weather	0.0599
fall	0.0886
summer	0.1226
winter	0.1562

(3 marks)

2. Why is it important to use **drop_first=True** during dummy variable creation?
   Answer – Because we need to get rid of one dummy variable as for linear regression categorical variable with n levels, we need only n-1 new columns each indicating whether that level exists or not using a zero or one and drop first will drop the first one in this line (2 mark)
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

   Answer – aTemp variable has the highest correlation (1 mark)
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

   Answer – Several steps can be performed -
   1. Predict the values on test set using the same model and their R2 values should be within the 5 percent margins.
   2. Perform residuals analysis on both test set and training set
   3. Calculate mean squared errors and it should be near zero (3 marks)
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
   Answer – atemp, year and summer season variables (2 marks)

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   Answer – Linear regression is an algorithm used to decipher some kind of linear relationship (if any) between a target variable and one or more predictor variables which can be explained in terms of a mathematical equation to predict new values of target. There are generally two types of linear regression models used –

   Simple liner regression – The general form of SLR is y= mx+c where y is target variable and x is predictor variable and c is a constant.

   Multiple Linear Regression - The general form of the equations is y=B0 +B1*x1 …….B1*Xn + errors where y is the target variable and x's are the predictor variables and B's are the respective coefficients and B0 is a constant.

(4 marks)

2. Explain the Anscombe's quartet in detail.

   Answer - Anscombe's quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.

   The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

   (3 marks)

3. What is Pearson's R?

   Answer - The Pearson correlation coefficient is a descriptive statistic, meaning that it summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables.

   The Pearson correlation coefficient is also an inferential statistic, meaning that it can be used to test statistical hypotheses. Specifically, we can test whether there is a significant relationship between two variables. (3 marks)

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

   Answer- Scaling is the process to scale or measure different attributes or variables based on the same scale or metric. Scaling is performed so that different variables and attributes can be-

   1. easier to interpret and

   2. Faster convergence for gradient descent methods

   There are two ways using which we can scale the variables in linear regression algorithm-

   1. Standardizing - The variables are scaled in such a way that their mean is zero and standard deviation is one. Formula used – (x-mean(x))/sd(x)
   2. Normalized scaling aka MinMax Scaling: The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data. Formula used – (x-min(x))/(max(x)-min(x))

   (3 marks)

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

   Answer - An infinite value of the Variance Inflation Factor (VIF) for a given independent variable indicates that it can be perfectly predicted by other variables in the model. Essentially, this occurs when the coefficient of determination (R^2) approaches 1. When an independent variable is not corelated then the VIF is 1.0.

   (3 marks)

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

   Answer - The quantile-quantile(q-q plot) plot is a graphical method for determining if a dataset follows a certain probability distribution or whether two samples of data came from the same population or not. Q-Q plots are particularly useful for assessing whether a dataset is normally distributed or if it follows some other known distribution. They are commonly used in statistics, data analysis, and quality control to check assumptions and identify departures from expected distributions.

   In linear regression it can be used to compare residuals distribution to check if it follows the same Normal distribution pattern for both training and test set.

   (3 marks)