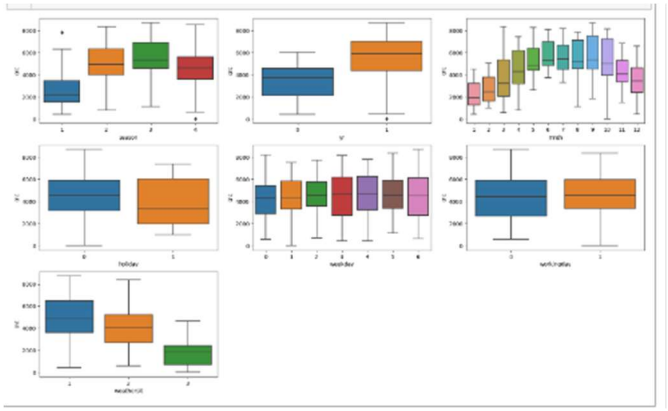
Assignment-based Subjective Questions

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

A1. In this dataset we have these categorical variables: season ,weathersit, yr, mnth, holiday, weekday, workingday,

And by seeing the boxplot we can say :

1. Bike demand is high during the fall season and low in spring.
2. Bike demand is high if weather is clear or with mist cloudy while it is low when there is light snow or heavy rain.
3. In year 2018 the demand was much lower than in 2019.
4. Month: The bike demand was lower in the beginning of the year December to March and on top since June till October, this is also comparable with weathersit.
5. Demands are lower on Holidays.
6. Bike Demands found similar during all weekdays and also no significance change based on workingday.

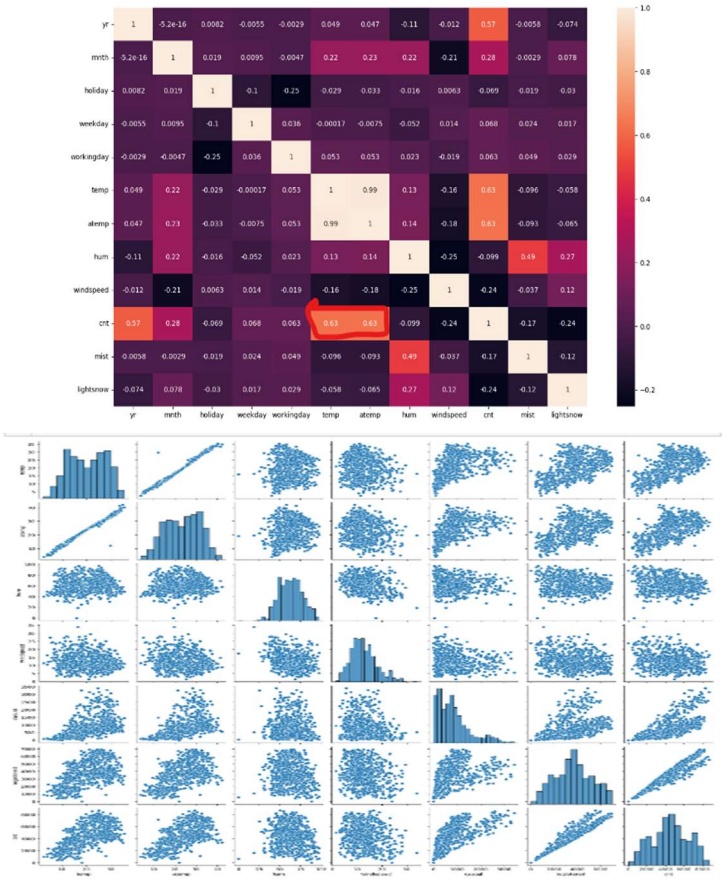


Q2. Why is it important to use drop_first=True during dummy variable creation?

A2. The key idea behind creating dummy variables is that for a categorical variable with 'n' levels, you create 'n-1' new columns each indicating whether that level exists or not using a zero or one. It is important use drop_first=True to reduce number of variables, as more number of variable may cause overfitting that is a drop in accuracy of the model.

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

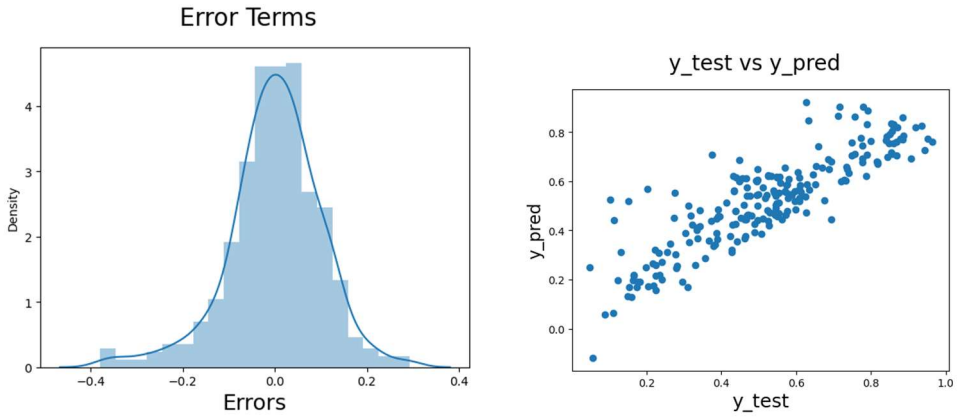
A3. Looking at the pair-plot among the numerical variables, temp and atemp has the highest correlation with the target variable with value .63, and hence dropped temp and kept atemp to build model.



Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

A4. I evaluated the model using two ways. 1. using Scatter plot, 2. Using R-squared value

ScatterPlot: by creating a scatter plot x vs y, the data points fall on a straight line in the graph, so there is a linear relationship between the dependent and the independent variables, and hence the assumption is true.



The Residual distribution is normal and mean is 0. And scatter plot shows linear relationship.

R-squared value: R2 score is used to evaluate the performance of a linear regression model, in this case I have calculated the R2 score as .745 ie, 74% accuracy.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

A5. The 3 most important features hum, atemp and mnth , seeing the VIF value and co-efficient.

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

A1. Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between a dependent variable and one or more independent features. When the number of the independent feature, is 1 then it is known as Univariate Linear regression, and in the case of more than one feature, it is known as multivariate linear regression. The goal of the algorithm is to find the best linear equation that can predict the value of the dependent variable based on the independent variables. The equation provides a straight line that represents the relationship between the dependent and independent variables. The slope of the line indicates how much the dependent variable changes for a unit change in the independent variable(s).

The equation of the best fit regression line $Y = \beta_o + \beta_1 X$

The strength of a linear regression model is mainly explained by R^2, where $R^2 = 1 - (RSS / TSS)$

RSS: Residual sums of Square, TSS: Total sum of squares

Q2: Explain the Anscombe's quartet in detail.

A2. Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

Q3. What is Pearson's R?

A3. Pearson correlation coefficient, also known as Pearson R statistical test, measures the strength between the different variables and their relationships. Therefore, whenever any statistical test is conducted between the two variables, it is always a good idea to analyse to calculate the value of the correlation coefficient to know how strong the relationship between the two variables is. Pearson's correlation coefficient can range from the value +1 to the value -1, where +1 indicates the perfect positive relationship between the variables considered, -1 indicates the perfect negative relationship between the variables considered, and 0 value indicates that no relationship exists between the variables considered.

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

A4. Scaling is a A technique to scale data is to squeeze it into a predefined interval. When you have a lot of independent variables in a model, a lot of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret. So we need to scale features because of two reasons: 1. Ease of interpretation 2. Faster convergence for gradient descent methods.

1. Standardizing: The variables are scaled in such a way that their mean is zero and standard deviation is one.

$$x = \frac{x - mean(x)}{sd(x)}$$

2. normalized 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.

$$x = \frac{x - min(x)}{max(x) - min(x)}$$

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

A5: The greater the VIF, the higher the degree of multicollinearity. In the limit, when multicollinearity is perfect (i.e., the regressor is equal to a linear combination of other regressors), the VIF tends to infinity.

The common heuristic for VIF is that while a VIF greater than 10 is definitely high, a VIF of greater than 5 should also not be ignored and inspected appropriately.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

A6. A Q-Q plot, which stands for Quantile-Quantile plot, is a graphical tool used to assess whether a given dataset follows a particular theoretical distribution, typically a normal distribution. It's also referred to as a quantile plot or a normal probability plot. The Q-Q plot compares the quantiles of the observed data to the quantiles of the theoretical distribution that you are trying to test against. This can help you visually assess how well the data matches the expected distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential.