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?
   Ans: Categorical variables has the same importance as of numerical variables. But we need to adjust the categorical variable to proper firmat to analyse them
   Two such methods are One hot encoding and n-1 encoding.
   In one -hot encoding , for n values of categorical variable there will be n variables generated. But n-1 form encoding , we will have only n-1 variables generated. One entry will be eliminated and all 0's will be considered as the eliminated entry. When we adjust categorical variable, we will be using 0's and 1's.
   In the case study , we have categorical variables such as seasons. Eventhough the values are already given in numbers (like 1, 2,3, 4 etc), if we are converting them using the above methods and use the given numbers for calculation , then the largest number will get more priority and the model will not be correct.
   Pandas library will help to convert the catogorical variables using pd.get_dummies.
   From the conclusions we can already see that the converted categorical variables play a good role in predicting the model for usage of shared bike.

2. Why is it important to use drop_first=True during dummy variable creation?
   Ans: drop_first=True is a property used in n-1 from encoding to convert categorical variables in to numerical forms in 0's and 1's. the syntax for the same is
   column = pd.get_dummies(df['Colmnname'],drop_first=True).
   If we are not using drop_first=True then all the values of categorical variable will be a new column. Say a categorical variable has 3 values , while converting 3 new fields will be added to the dataset. By using drop_first=True , only 2 new variables will be added. And for the 3rd variable value will be hypothetically 1 when other 2 variables are 0.

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

   Ans: The highest correlation is with registerd users and then Casual users . But since these two cannot be used for data prediction as this will cause data leakage. Next variable which has more correlation is temperature in celciius.

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

   Ans: Assumptions are :

   1.Dependent and independent variables should have a linear relationship.

   2. Error terms should be normally sistributed.

   3. Error terms should not be dependednt on each other.

   R2_Scope : Calculated using : r2_score(y_true=y_test_m,y_pred= y_test_m_predict) using actual and predicted variable. And this should be almost equat to r2.

   Residual analysis : Residual – which is difference btween actual and predicted will be ploted in a distplot. This should have a normal distribution with mean almost equal to 0.

   Plot actual and predicted values in a scatter plot. These values should be similar.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
   Ans: Temperature-> it has a possitive effect.
   Climate -> if climate is bad like snow, mist etc . it has a negative effect .
   Seasons also has a considerable effect on bike demand.
   Also Demand will be more on working day.

General Subjective Questions

1. Explain the linear regression algorithm in detail.
   Ans: Linear Algorithm is a machine learning algorithm which works based on the relation ship between dependent and independent variables. This is one of the regressionalgorithm technique. Supervised learning is used in this process. This is used for creating the model and prediction of values. The main assumptions are :

   1.Dependent and independent variables should have a linear relationship.

   2. Error terms should be normally sistributed.

   3. Error terms should not be dependednt on each other and distribution has with constant variance.

   Steps involved are:

   1. Identify dependent and independent variables.
   2. Create training and test set.
   3. Train the model and learn the coefficients.
   4. Evaluate the model using training and test set.

   This is mainly used for predictive analysis.

2. Explain the Anscombe's quartet in detail.
   Ans: Anscombe's Quardrent is mainky 4 dataset which are almost identical. But thsese will have some peculiarities in regression model .
   Franscis Anscombe brought this model forward to state that we shouldn't only depend on statistical measures while analysisng data.
   Anscombe's Quadrent shows the importance of plotting the model after the analysis is done to validate the fit of the model. This illustrate the importance of analysing the data graphically before analysis the relationship statically.

3. . What is Pearson's R ?
   Ans: Pearsons's R is Pearson Correlation coefficient which helps to mesure the degree of correlation. If the value lies between ± 0.50 and ± 1, it is said to have a high correlation. This is a way of measuring linear correlation. The value lies between -1 to +1.
   r>0 indicates a possitive correlation. And r<0 indicates a negative correlation.
   This is used when we analyse the P value. We P value is smaaller than .05 , we reject null hypothesis, which states a linear relation between two variables.

   Limitation of Pearson's Coefficient is that it cannot distinguish

   Betweenindependent and dependent variables.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
   Ans:Scaling a process which is done while data preparation. When we create a model ,model will not be proper when the variables are of different scales. So we use scale to minimise the number as well as to bring the data to same scale.Doing scaling will nly have a difference in coefficient values only.
   This is done basically Standardization and Normalization.
   Standardization : (X-Xmean)/Std Deviation of X.
   This brings the data into standard normal distribution format. With mean 0 and std Deviation as 1.
   Normalization : Also known as min max scaling . This basicxally bring data range between 0 and 1. Formula is
   (X-Xmin)/(Xmax - Xmin).

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
   Ans: VIF value infinity indicates a perfect correlation between the variables. VIF = 1/(1-(R2)) = infinity . This can be an indication of multicollinearity. So we will have to eliminate one of the variables.

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

   Ans:Q-Q plot is Quantile -Quantile plot.

   This is a graphical technique which help to identify whether two datasets are part of population with a common distribution.

   Advantages:

   1. The sample size do not need to be equal.
   2. Many distributional aspects can be simultaneouly tested.

   A Q-Q plot is a plot of quantiles where quantiles of 1^{st} data set plot against quantiles of 2^{nd} dataset. On a Q-Q plot , a normally distributed data appears roughly as a straight line.

   Q-Q plot compares how close two distributions are. This is used to assess normality of linear regression.