Assignment-Based Subjective Questions

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

Ans. The given assignment, I first identified Categorical Variable "Season, yr, mnth, holiday, weekday, workingday weathersit against the target variable cnt, and EDA Visualization was made. As from the Stats this was very clear that weather Situation ~~area~~ median where around 50,000 approximately and similar we could see in 'season' and yr. And the final model building shows a significant growth of R^2 and adjusted R^2 for yr, season etc.

Q.2. Why is it important to use drop_list = True during dummy variable creation?

Answer:- Yes this is highly advisable to use drop_list = True, this will avoid redundant feature appear while we create correlation and the reference group. This will avoid to create more dummy variable during dummy creation.

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

Ans → Initially I saw numerical variable "registered" having highest correlation with the target variable 'cnt'. During data preparation we drop 'register' variable due to its multi-collinearity the numerical variables 'atemp' got the highest correlation with the target variable 'cnt'.

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

Ans. Below are the steps performed to carry out the analysis.

1. Test for normal distribution of errors terms (we call it as residuals) by visualizing plot of the error terms.
2. Add and drop the variables based on each model VIF and P-values to avoid multicollinearity

Q.5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans. Below are the top 3 features which significantly explain the demand.

1. atemp - Feeling temperatures in Celcius
2. yr - year (0: 2018, 1: 2019)
3. Winter - It's subcategory of Season (4: Winter)

General Subjective Questions

Q. 1. Explain the linear regression algorithm in detail.

Ans Its machine learning algorithm which is based on supervised learning. This performs regression tasks, and models based on independent variables. This is basically used to find out relationship between different variable and its predictive (forcasting) value. In others or simple word we can say this is against dependent variable and independent variables. If we talk about the linear formula its $y = mx + c$ where c is intercept and m is slope, however for the multivariable this comes with more generic formula.

y (dependent variable)

x (independent variable)

lets explain the linear regression formula as below.

$$y = \beta_0 + \beta_1 x + \beta_2 x_2 + \dots \beta_n x_n$$

where

y is predictive value

β_0 is the constant term

$\beta, \beta_1, \beta_2, \beta_3 \dots \beta_n$ are the model parameters

x_1, x_2, x_3 are the feadures values.

Q.2. Explain the Anscombe' Quartet in detail.

Ans. To be frank, this was unknown term for me however I google and try to learn the concept. This is basically a group of 4 data set which looks similar in terms of average value, Average value of y. The variance of X and y and the correlation between them is also similar, hence forth linear regression model looks the same however on the plot they tell different story.

Q.3. what is Pearson's R?

Ans. This is basically the strength of the linear regression between the variables. And it varries between -1 and 1, And its feature is if variables goes up and down together then its co-efficient will always be Positive and vice a versa. Below I tried to explain the theory

r = -1 (perfectly linear with negative slope)

r = 0 (means no linear association)

r > 0 < 5 (means weak association)

r > 5 < 8 (mean moderat association)

r > 8 (means very strong association)

* You can find the formula as below

$$r = \frac{N\Sigma_{xy} - (\Sigma_x)(\Sigma_y)}{\sqrt{[N\Sigma x^2 - (\Sigma_x)^2][N\Sigma_y^2 - (\Sigma_y)^2]}}$$

where

N = number of pairs of scores

Σ_{xy} = Sum of the product of pairs scores

Σ_x = Sum of x scores

Σ_y = Sum of y scores

Σ_{x^2} = Sum of squared x scores

Σ_{y^2} = Sum of squared y scores

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

Ans:- This is the technique to standardize the independent features present in the data in a fixed range, and this is performed during the data processing that handle highly varying magnitude.

We do to rescale the variables so that this can be comparable. Imagine if we dont have comparable then coefficient obtained by fitting regression model will

be very large or small compared to the other co-efficient. And normalized scale means between 0 to 1.

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

Ans Very simple answer would be that particular variable have severe collinearity. And this variable can be denoted as linear combination of other variable, which mean square of multiple correlation of any predictor variable with the other predictors approaches unity.

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

Ans As per my understanding Q-Q plot means Quantile-Quantile plot, which help us to do assessment on set of data come from some theoritical distribution such as normal or exponential.

Q-Q plot is scatterplot created by two set of Quantiles against each other. If both set of variables come from same set of distribution, then we would see them forming a line, which may be straight line.