Assignment-based Subjective Questions

1. The model which I developed takes 18 independent variables to predict the dependent variable cnt. All these 18 independent variables together explains around 80% of the variation in the target variable.

2. If we don't use drop_first=True while creating dummy variables. Then the get_dummies method (or equivalently OneHotEncoder in sklearn) will new feature for every single level in the categorical variable under consideration. Surely the sum of these encoded features in every row will be 1 which ensures multicollinearity which will lead to unusually high or infinite VIF.

3. I did not use pairplot, instead I used heatmap of the correlation matrix which gives us the same information. From this heatmap we can see that atemp, temp both have highest correlation with target variable and yr has the second highest correlation with target variable.

4. After building the model on the training set, we did residual analysis using 2 residual plots. One is kdeplot to check the normality of the residuals and the other is scatterplot of residuals vs actuals which is used to check the homoscedasticity and independence of the residuals.

5. The top 3 variables are atemp, hum, yr. Out of which atemp & yr have the highest +ve impact on the target variable cnt and the hum has the highest -ve impact on the target variable cnt.

General Subjective Questions

1. Linear regression is the one of the most basic form of regression algorithms in machine learning. The model consists of a single parameter and a dependent variable has a linear relationship. When the number of independent variables increases, it is called the multiple linear regression models. We denote simple linear regression by the following equation given below. y = mx + c + e where m is the slope of the line, c is an intercept, and e represents the error in the model. The best-fit line is determined by varying the values of m and c for different combinations. The difference between the observed values and the predicted value is called a predictor error. The values of m and c get selected to minimum predictor error. Of course we'll extend the same concept to multiple linear regression. That apart, to make inference from this model we need to check a set of assumptions.

2. Anscombe's quartet consists of 4 data sets which have almost equal statistical summary, still they are entirely different when plotted as a graph. So Anscombe's quartet is used to demonstrate the significance of analyzing the given data set in great detail (if dimension permits, visually) before jumping to conclusion based on simple statistical summary.

3. Pearson's correlation coefficient R determines the strength of the linear relation between given 2 numerical variables. It always ranges from -1 to 1. R=1 implies there's perfect +ve linear correlation between the 2 variables. R=-1 implies there's perfect -ve linear correlation between the 2 variables. R=0 implies there's absolutely no linear correlation between the 2 variables.

4. Feature scaling is done for 2 reasons, it makes the coefficients of the scaled features much more easily interpretable. Gradient descent will converge faster on scaled features. There are 2 commonly used scaling techniques, one is normalization where we scale the feature down to 0 to 1 range and the other is standardization where we scale and convert the feature into a distribution with mean=0 and st.dev=1

5. VIF is a measure how well a feature under consideration is linearly explained by other features. If VIF='inf', it implies that there's perfect multicollinearity among predictor variables. In the sense, one of the predictor variables can be exactly determined by the remaining predictor variables. This usually occurs when we don't use drop='first' while applying OneHotEncoding on nominal categorical variables.

6. Q-Q plots stands for Quantile-Quantile plots. This is plot of quantiles of a sample distribution against quantiles of a theoretical distribution. In the context of linear regression, one of the assumptions of linear regression (for inference purposes) is that the residuals are normally distributed around 0. We also try to verify this assumption by drawing the kdeplot of the residuals, which will look like the normal curve if our assumption is correct. But strictly technically speaking, we cannot visually determine if a distribution is normal or not. Even the t-distribution is visually indistinguishable from a normal curve. So we have to use this Q-Q plot and if our assumption is correct, then the quantiles of our sample distribution will highly linearly correlate with the line y=x