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?

A) The fitted regression equation is:

cnt = 0.4028 + 0.5272*atemp -0.1545*spring -0.0974*holiday-0.2195*snow -0.2285 * hum

Based on the analysis and the final fitted equation, the categorical variables have an inverse relations hip with demand.

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

A) It is useful in whether to drop or not drop the first category of categorical variable encoded.

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

A) The casual and registered user count has very strong relationship with demand count. But, this is obvio us. In terms of variables having significant predictive power, 'temp' and 'atemp' have highest correlati on with demand at close to 0.63.

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

A) Validated Autocorrelation, and normaility based on DW and JB test statistics from lr.summary() and m ulticolinearity through VIF.

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

A) Top 3 features: 'atemp', 'snow'('weathersit') and 'hum'

General Subjective Questions

1. Explain the linear regression algorithm in detail.

A) Linear regression is based on assumption of linear relationship between target and independent variable s. The algorithm finds the best fit line to explain the relationship between the dependent and independe nt variables. The best fit line is identified by minimizing the error (the distance between the points on li ne and actual values), i,e. Actual – Predicted. These errors are termed as residuals. There are several as sumptions related to the residuals for linear regression which includes constant variance of residuals, n ormal distribution of residuals, absence of serial correlation among residuals.

2. Explain the Anscombe's quartet in detail.

A) The quartet refers to 4 data points which look very similar (identical) based on summary statistics such as mean, median, etc, but very look very different when plotted. This demonstrates importance data vis ualization.

3. What is Pearson's R?

A) Pearson's R or Pearson'r correlation tells the strength of linear relationship between two variables. It is the ratio of product of variance of the two variables to the covariane of the two variables.

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

A) Scaling is done to increase interpretability of the model coefficients as scaled variables lead to compara ble coefficients which are same scale wither standard or nomal. Standardized scaling scales the variable using the standard normal distribution scale, i,e, it uses Z-value with 0 as mean 1 as standard deviation of the data.

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

A) When pearson's r is 1, VIF becomes infinity as it is calculated as 1/(1-r2)

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

A) Q-Q plot is visual tool to test normaility of distribution of a variable as it plots the theoretical (normal) vs actual distribution of the data points.