Assignment-based Subjective Questions

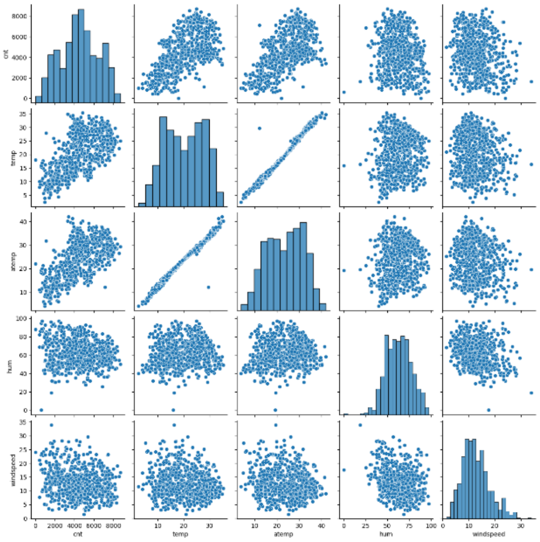
1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

   The categorical variables season, month, year, weekday, working day, and weather situation significantly impact the dependent variable 'cnt'.

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

   Using drop_first = True is crucial as it removes the extra column created during dummy variable creation, thereby reducing multicollinearity among the dummy variables. This simplification helps in minimizing the correlation between these variables

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



   The 'temp' & 'atemp' variables have highest correlation when compared to the rest with target variable as 'cnt'

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

   The approach by which I have validated the Linear Regression Model are based on the following five assumptions:

   - *Normality of error terms*: The error terms should be normally distributed.

- *Multicollinearity check:* There should be insignificant multicollinearity among variables.
- *Linear relationship validation*: A linear relationship should be visible among variables.
- *Homoscedasticity*: There should be no visible pattern in residual values.
- *Independence of residuals*: There should be no autocorrelation in the residuals.

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

   Based on the final mode, the Top 3 features that has been contributing a significant impact towards explaining the demand of the shared bikes are temperature, year and season

General Subjective Questions

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

Linear regression is a simple algorithm that models the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data. The main goal is to predict the dependent variable based on the independent variables. It calculates the best-fit line by minimizing the sum of the squared differences between observed and predicted values (least squares method). The equation of the line is y=β0+β1x1+β2x2+...+βnxny = \beta_0 + \beta_1x_1 + \beta_2x_2 + \ldots + \beta_nx_ny=β0+β1 x1+β2x2+...+βnxn.

Key assumptions include linearity, independence, homoscedasticity, and normality of residuals.

2. **Explain the Anscombe's quartet in detail.**

Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics, such as mean, variance, and correlation, yet they exhibit very different distributions and relationships when graphed. The quartet illustrates the importance of graphing data before analyzing it. Each dataset includes pairs of x and y values. Despite their statistical similarities, the first dataset shows a linear relationship, the second a curvilinear pattern, the third includes an outlier influencing the relationship, and the fourth has a vertical outlier. The quartet emphasizes that summary statistics alone can be misleading without visual inspection.

3. What is Pearson’s R?

Pearson's R, or Pearson's correlation coefficient, measures the strength and direction of the linear relationship between two variables. It ranges from -1 to 1, where 1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear relationship. It is commonly used to understand how two variables are related in a dataset.

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

Scaling is adjusting data features to a common scale. It's performed to ensure that features contribute equally in machine learning algorithms, improving model performance and convergence speed.

Normalized Scaling scales data to a fixed range, typically [0, 1], making it suitable for algorithms sensitive to data magnitude differences.

Standardized Scaling transforms data to have a mean of 0 and a standard deviation of 1, preserving the distribution shape. It's useful when the data has outliers or different distributions.

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

The value of VIF (Variance Inflation Factor) can become infinite when there is perfect multicollinearity among the predictor variables. This means one predictor variable is a perfect linear combination of other predictor variables, leading to an undefined or infinite value in the VIF calculation. Essentially, the regression model cannot differentiate between the perfectly correlated variables, causing the VIF to spike to infinity

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

A Q-Q plot (Quantile-Quantile plot) is a graphical tool that compares the distribution of your data to a theoretical distribution, usually the normal distribution.

Use and Importance in Linear Regression:

1. **Assess Normality:** It helps check if the residuals (errors) of the model are normally distributed, which is an assumption in linear regression.
2. **Identify Deviations:** Points that deviate from the line in a Q-Q plot indicate that the residuals are not normally distributed.
3. **Detect Outliers:** It can highlight outliers or unusual data points that could affect the regression model.