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?**

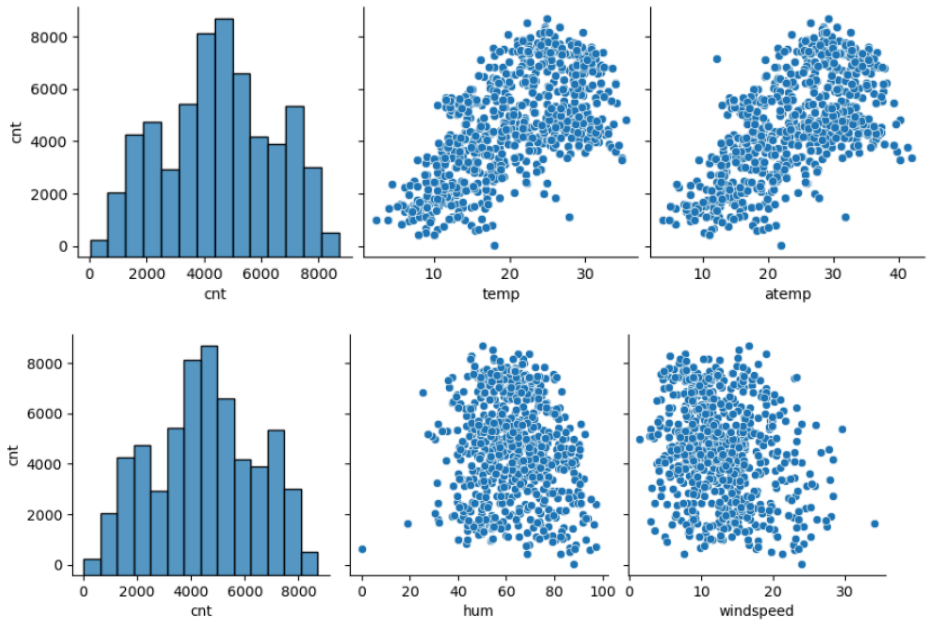
   Response : Final regression model with 80.5% R2 has 10 parameters out of which 7 are categorical variables which indicate that they are very significant in fitting the model. Without the categorical variables, the fit R2 falls down to 43.3% (refer to note book, section 6). While all categorical variables together except 'yr', the R2 fit improves to 54% which is nearly 10% improvement. Remaining 26% comes up from 'yr'.

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

   Response : It is upto us to drop either the first or any other among the dummy variables. But it is required to drop atleast 1 among all the dummy variables. This is because, any combination of "n-1" dummy variables together can explain the nth dummy variable. If we miss to drop the dummy variable here, the variable will show up in VIF with high value indicating that it is insignificant

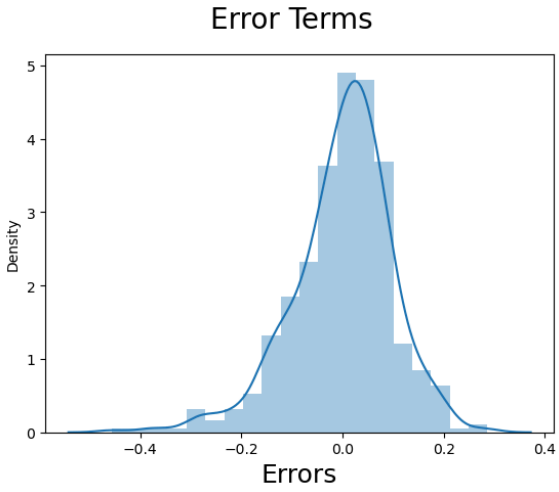
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**

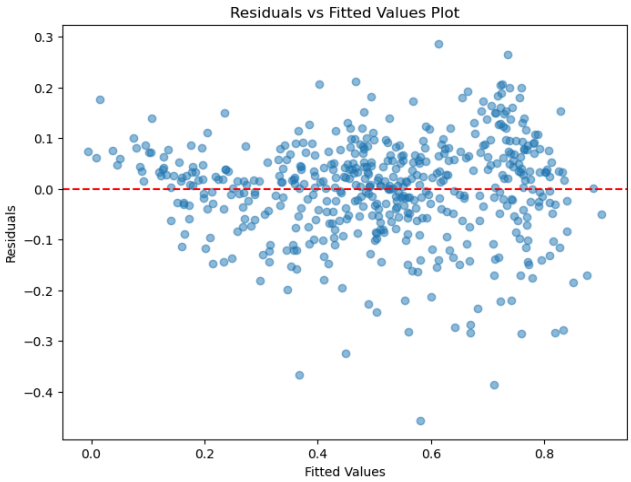
   Response: Temp



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

   Response: Errors were calculated for the train data set by taking difference between prediction and ground truth. These errors are observed to be
   1. normally distributed
   2. homoscedasticity of error terms i.e. no obvious patterns





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

   Response: ‘yr’, ‘temp’,’Light Snow’ (weather sit)

General Subjective Questions

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

   Response: Linear regression algorithm is used to fit a linear regression to a dataset if the dataset follows linear regression assumptions i.e. there exists some linear relationship b/w the x and y parameters and also the error is normally distributed. Linear regression starts with splitting the dataset to train and test sets using scikit learn packages. Then, the train dataset should be scaled to ensure that all the variables are between 0 to 1.Scaling reasons and methods are answered in previous questions. The model can then be fitted either using stats model or scikit learn packages. Stats model provides lots of probabilistic details about the fit which will be useful to understand the significance of the parameters as well as significance of the model.
   VIF factor (variance inflation factor) helps in understanding which parameters are possibly dependent on other parameters with a linear relation (R2>80% for a VIF of 5). Also RFE (Recursive feature elimination) can be used to remove the insignificant parameters in the model.
   Once the model is built, it is then validated against the test data set to ensure the R2 for test data is falling in similar range as that of train. Also, residual analysis is performed to ensure that the linear regression is a valid method for the provided data

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

   Response: Anscombe's quartet is a set of four datasets created by the statistician Francis Anscombe in 1973 to illustrate the importance of data visualization in statistical analysis. Each dataset in the quartet has the same summary statistics—mean, variance, correlation, and regression line parameters—but they differ significantly in their distribution and structure. This demonstrates that summary statistics alone can be misleading, and visualizing data is crucial for proper analysis.

3. **What is Pearson's R?**

   Response: Pearson's rrr (Pearson correlation coefficient) is a measure of the linear relationship between two continuous variables. It quantifies the strength and direction of the linear association between the variables.

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

Response: When we are building a multiple linear regression with lot of x parameters, we often see that these parameters have different ranges. Scaling is a process of bringing all the parameters to same value range.

Why is scaling performed?

Response : Parameters with larger values can dominate the loss and hence bias the training process. Scaling of all the parameters will ensure that no bias exists due to feature values. If we want any such bias to be implemented for few x parameters, we can directly add some coefficients to those parameters in the loss definition.

Gradient descent converges smoothly and quickly when features are on similar scales.

Also, when we have all the x parameters in same range, we can do comparison of their coefficients and understand significance levels of different parameters.

What is the difference between normalized scaling and standardized scaling?

Response:

Normalized Scaling alias Min-Max Scaling: Scales the features or x parameters to the range [0,1] using the formula below

X_scaled = (X-X_min)/(X_max-X_min)

However, we can also scale a parameter to the range [-1,1] using the equation below

X_scaled = 2*{(X-X_min)/(X_max-X_min)} - 1

Normalized scaling is sensitive to outliers as outliers will define the min and max values.

Standardized scaling or Z-score Normalization: This method of scaling ensures that the scaled values have a mean of 0 and a standard deviation of 1.

X_std = (X-mu)/Sigma

mu = mean of X params

Sigma = Standard deviation of X params

Standardized scaling is not impacted by outliers as the impact of outliers on the mean will become insignificant especially when the data is large.

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

Response:

It means that the parameter is fitting with a combination of other parameters with R2=1 i.e. there exists a perfect linear relation of this parameter with other parameters.

Example: Temp in F and Temp in C

T_F = 32+(1.8*T_C)

If we have both T_F and T_C in the x parameter list, then VIF for one of these will become infinite

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

Response:

Q-Q plot is a scatter plot where:

- The **x-axis** represents the quantiles of the theoretical distribution (e.g., a normal distribution).
- The **y-axis** represents the quantiles of the data being compared (e.g., the residuals of a regression model).

A Q-Q plot helps to visually assess whether the residuals of the regression model follow a normal distribution. If the points on the Q-Q plot align with the 45-degree reference line, it suggests that the residuals are normally distributed.

Deviations from the straight line in a Q-Q plot may indicate the presence of outliers, skewness, or other deviations from normality. Identifying these deviations can help diagnose and address potential issues with the model.