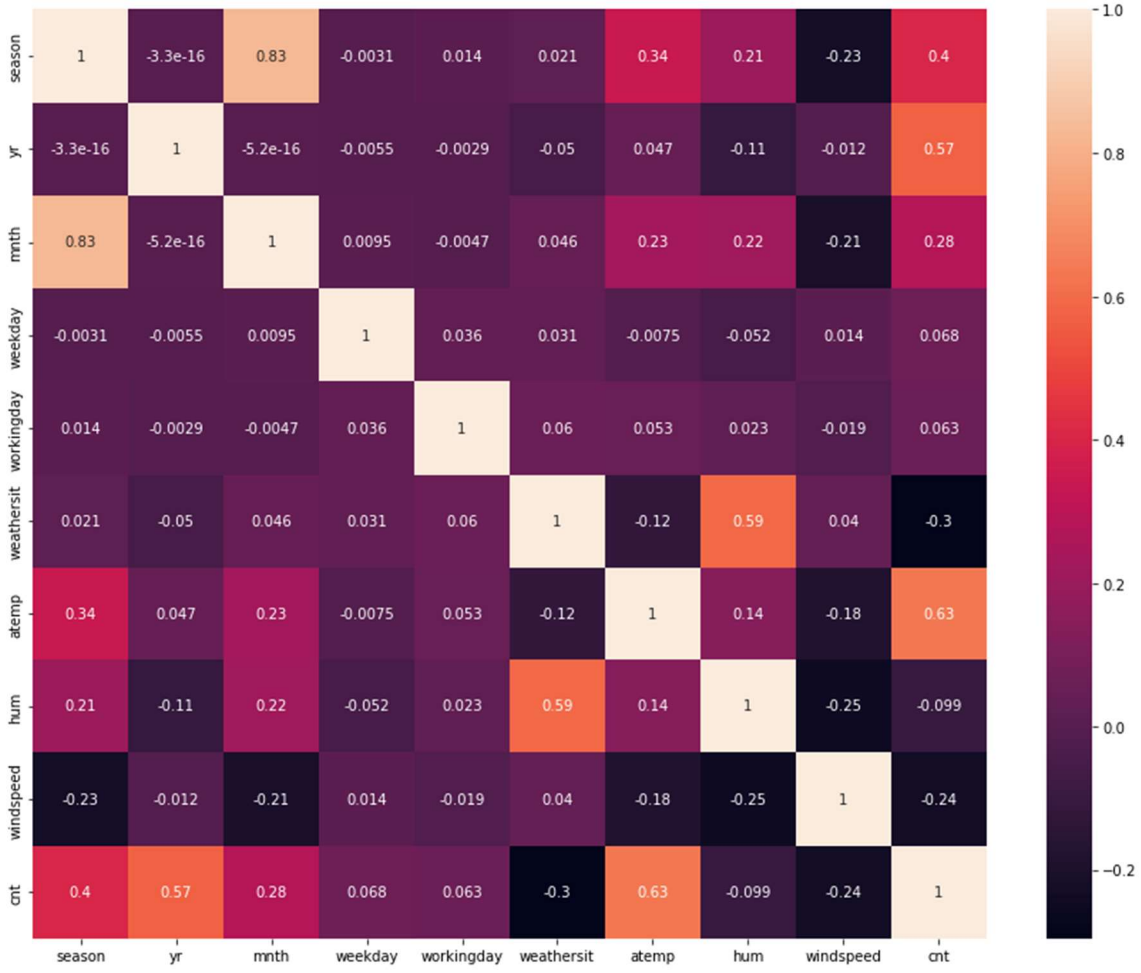
Assignment-based Subjective Questions

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

   Categorical variable like

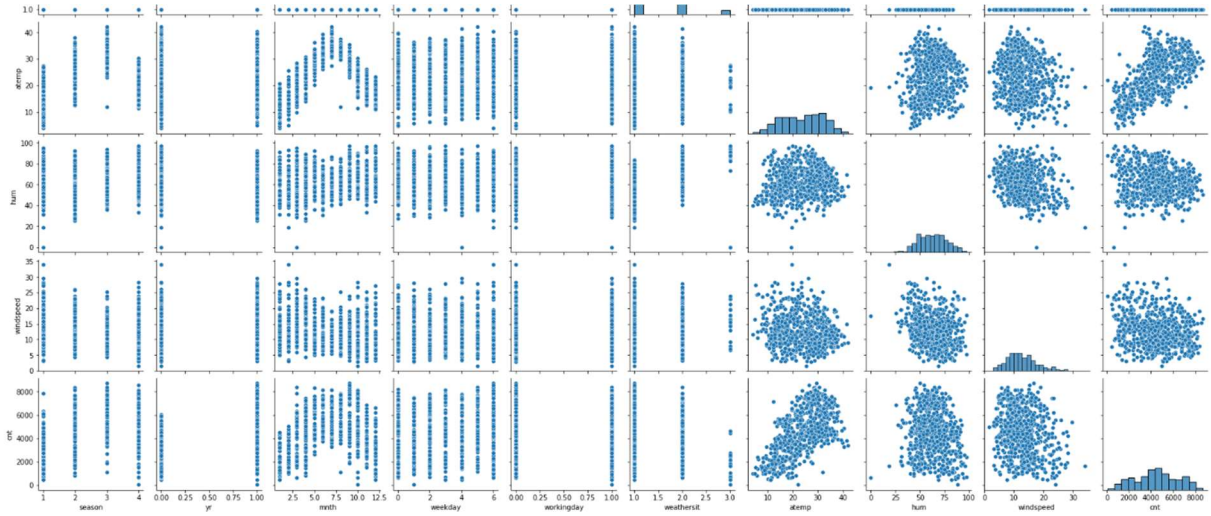
   a. Season, year ad month has high correlation with Count which is 0.4,0.57, 0.28 respectively and this says the relation is linear.
   b. Season and month has high correlation with Atemp (0.34,0.23), Humidity (0.21, 0.22), and wind speed(-0.23,-0.21) so there is a positive correlation between (season, month) with humidity and Atemp.



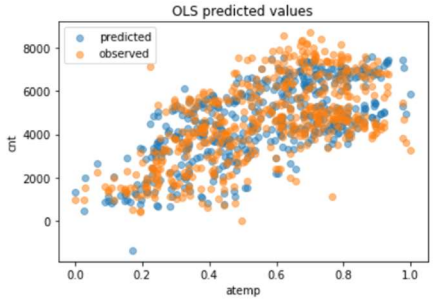
2. **Why is it important to use drop_first=True during dummy variable?**

   During a creation of Dummy variable, the number of columns required to represent a categorical column is n-1 where n is the number of categorical variables. And the reason for this is the model becomes complex and the computation speed required will also increase depending on thenumber of dependent columns. Hence it is advised to keep the number of columns to n-1 in dummy encoding and to do that its important we use drop_first=True.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
   a. From the pair plot it is evident that the highest amount of corr is between variable ‘cnt’ and all other numeric variables (windspeed, a temp, humidity)
   b. A temp and hum has a linear correlation.
   c. Wind speed and hum has a negative correlation.
   d. Wind speed and a temp has a positive linear correlation.



4. How did you validate the assumptions of Linear Regression after building the model on the training set?
   After the training set, I verified the variables against the predicted and actual using a scatter plot to validate if the model is linear or not.
   Also the value of R-squared is approximately 0.83 and it gives me a sense of understanding that my assumption of model is linear is True.



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 model, the top 3 features contributing to the model are
   1. Year
   2. ATemp
   3. Weathersit

General Subjective Questions

1. Explain the linear regression algorithm in detail.

- Reading and Understanding the Data use pandas library to read CSV files.
- EDA
  - Missing Value Treatment-
    - use drop na method to drop rows with more na values. Also use mean and median to fill na cells if possible
  - Outlier Treatment
    - Use inter quartile method to treat outliers
  - Univariate Analysis
    - Descriptive Statistics
      - Mean/Median/Mode/Skewness
      - Transformations
  - Multivariate Analysis
    - Correlations
    - Box Plots, pair plots, heat maps
    - Chi-square Test to understand the relationship between two categorical variables
- Encoding
  - Target Variable Encoding - for Y variable
  - Dummy variable Encoding - for X variables(categorical variables)
- Scaling - for numerical variables use min max scaler for scaling
- Feature Selection
  - RFE
- Base Model Building - First Model building in python
- Regression Model Assumptions – Validation(assuming linear, the variables are taken and dropped based on VIF and P value)
- Final Model Building (final model with est R-squared and least P value(<5%) and VIF less than 5 to be maintained)
- Results Interpretation- based on the prediction of train set, validate the model to interpret data on test set.
- Submission

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is a phenomenon where the it consist of 4 data set with a similar or identical descriptive statistics results but deviate when they are graphed. It essentially helps understand the effect of outliers on the summary or predictability of the model.

3. What is Pearson's R?

Pearson's R or Pearson's covariation coefficient describes the linear relationship between 2 data sets. And it is given by the equation as below.

X & Y are 2 data sets

Std X and std Y are respective standard deviations

P=cov(X,Y)/std X. std Y

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

Scaling is a process of converting the numbers with large variance in a data set to value near the mean which is set at zero or distributed between the range 0 to 1

Scaling is performed so that the variables in the data set are of same or identical scale values , which helps us with our analysis.

Standardized scaling is a type of scaling where the data point are distributed around mean which is centered around zero and it is given by the equation as below

Si =be standard scaling of ith term

X-mean

X-i

SD-Standard deviation

Si=(X-i- X-mean)/(SD)

Normalized or min max scaling is a type of scaling where the data point are distributed between 0 and 1 and it is given by equation as below

NI= (X-i -Xmin)/(X-max – X-min)

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

The value of VIF turns infinity when the model Rsquared turns to 1.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression
   a. Q-Q give a comparison plot between 2 probability distribution by plotting their quantiles against each other.

   b. The Q-Q- plot is a non decreasing plot when viewed from left to right and if the 2 distributions are equally distributed it follows a 45deg st line plot.

c. Q-Q plot is used to find out if 2 data sets come from same distribution or not. A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.