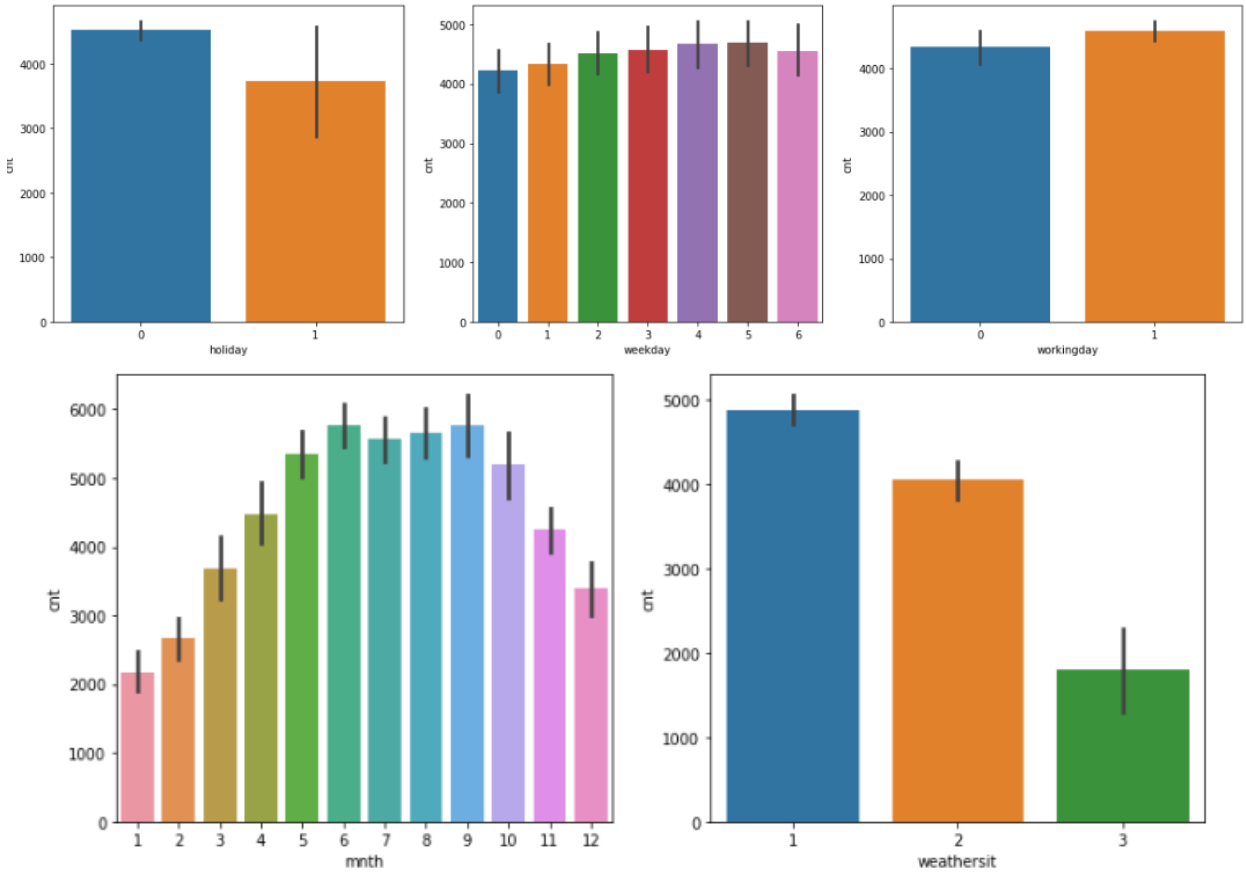
Assignment-based Subjective

Questions 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
Answer 1: From the original data set, we can determine the presence of the following categorical variables:

Season,yr,month,holiday,workingday,weekday and weathersit

Effects on dependent variable:

- Season clarifies that the demand was higher in summer and fall season
- Weekday does not show a high difference as the distribution across all weekdays is almost same
- Whereas, the month's distribution is same as the season's for summer and fall months it is at peaks
- Working day and holiday mark the similarity here where the demand seemed high on a working day as compared to a non working day and same for holiday- high demand when there is no holiday
- With the above inference we can consider that the demand was there from the population commuting to maybe workplaces mostly
- The weathersit(weather situation) is another important factor as it clearly predicts that the demand was high for a clearer weather of type:1



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

Answer: We use dummy variables to handle the categorical variables with k levels and so for the level after creating the dummy we get k individual variables with binary values as 0 or 1.
Consider, and example of season variable here from our dataset:
Season ={1 :'spring',2:'summer',3:'fall',4:'winter'}
After creating the dummy we get 4 variables as 1,2,3,4 for each of the season with values 0 or 1

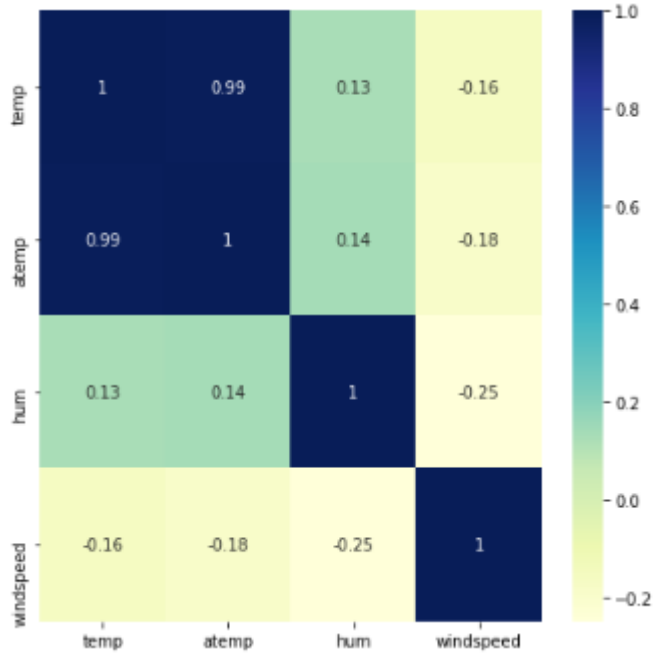
1	2	3	4	Season
0	0	0	1	4= winter
1	0	0	0	1=spring
0	1	0	0	2=summer
0	0	1	0	3=fall

The same result can be obtained by using k-1 variables so why make the model heave and not drop the first variable.
The result of k-1 variables is shown below:

2	3	4	Season
0	0	0	all 0s = spring
0	1	0	3= fall
1	0	0	2=summer
0	0	1	4=winter

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

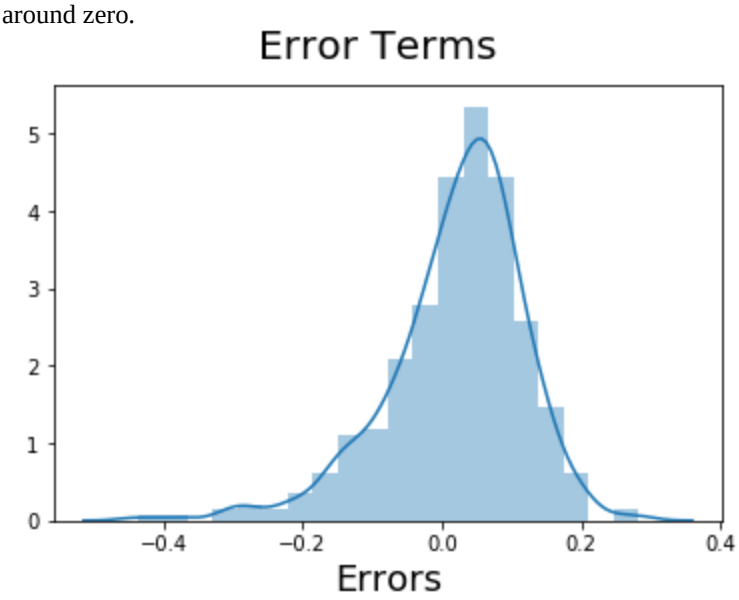
Answer:
From the pair-plot is evident that the temp variable and atemp has the highest correlation with the target variable (cnt) followed by hum and windspeed.
But we are considering only temp here and not atemp as the temp and atemp are both highly correlated together so we have dropped the them



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

Answer:

- It is assumed that there is a linear relationship between the dependent and independent variables which is known as ‘linearity assumption’, which was plotted after training the dataset
- Error terms are approximately normally distributed
- It is assumed that the residuals have a mean value of zero, i.e., the error terms are normally distributed around zero.



- There is no multicollinearity as well because the independent variables are linearly dependent on each other
- The pair-wise covariance was also not an obstacle
- The VIF was also not very high and in the permissible range of 0-5

Question: 5

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

Answer:

The top 3 variables based on the model are **temp,hum and windspeed** which can be proven with the RFE ranking while building the model and their correlation form the pair-plot as well

SUBJECTIVE Questions

Question1:

Explain the linear regression algorithm in detail.

Answer:

- **Linear Regression** is a machine learning algorithm based on **supervised learning**.

- It performs regression task
- Such regression algorithms predict a target value based on independent variables.
- Liner Regression can be of types:

  Single Linear Regression

  Multiple Linear Regressions

These are linear or multiple based on the independent variables we are passing in this model

- For the linear regression, we split the data into train and test data set and pass the independent variables to predict the target variables and then further make predictions on the test set
- The Function for linear regression would be same as the equation of a straight line $y = mx + c$ (where m = slope and c = constant)

Where x will be the input predictor in terms of Machine learning and we will get the intercept and the coefficients and y will be the target variable

Some example or use cases of linear regressions:

- Risk Management- Using regression to the analysis of Risk Management in the financial and insurance sector.
- Using regression to predict the change in price of stock or product.

Question 2:

Explain the Anscombe's quartet in detail.

Answer:

- Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics
- These 4 datasets have very different distributions and appear very different when graphed. Each dataset consists of eleven (x,y) points.
- They were constructed in 1973 by the statistician Francis Anscombe to demonstrate the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties
- Anscombe's Quartet reminds us that graphing data prior to analysis is good practice, outliers should be removed when analyzing data, and statistics about a data set do not fully depict the data set in its entirety

Question 3:

What is Pearson's R?

Answer**:**

- The Pearson coefficient is a type of correlation coefficient that represents the relationship between two variables that are measured on the same interval or ratio scale.
- It has a value between +1 and −1. A value of +1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation
- It's the covariance of the two variables divided by the product of their standard deviations.
- It is represented by a greek letter rho ρ
- It has the formula;

- $$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} \quad \text{(Eq.1)}$$

  Where cov is the covariance
  Sigma(X),sigma(y) = standard deviation of x and y

Question 4:

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

Answer:

- Scaling is a technique to standardize the independent features present in the data in a fixed range.
- It is performed during the data pre-processing to handle highly varying magnitudes or values or units
- If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.
- The feature scaling also prevents the danger of outliers
- Normalization usually means to scale a variable to have a values between 0 and
- Standardization transforms data to have a mean of zero and a standard deviation of 1.

Question 5:

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

Answer:

- VIF gives how much the variance of the coefficient estimate is being inflated by collinearity.
- The VIF can be conceived as related to the R-squared of a particular predictor variable regressed on all other includes predictor variables.:
- VIF of X1 = 1/(1 - R-squared of X1 on all other Xs).If you only have 1 X or that X is orthogonal with all the other Xs; then VIF = 1/(1-0) = 1 - so no variance inflation
- If two Xs are perfectly correlated
- VIF = 1/(1-1)= 1/0 = infinity
  Which marks that when two predictors who are perfectly correlated with each other in that case the VIF will be infinite
  Question 6:
  What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

  Answer:
  - Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution
  - Also, it helps to determine if two data sets come from populations with a common distribution.
  - This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.
  - A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.
  - In linear regression if we have following scenarios, then Q-Q plot is helpful:
    If two data sets —
    1. Come from populations with a common distribution
    2. Have common location and scale
    3. Have similar distributional shapes
    4. Have similar tail behavior