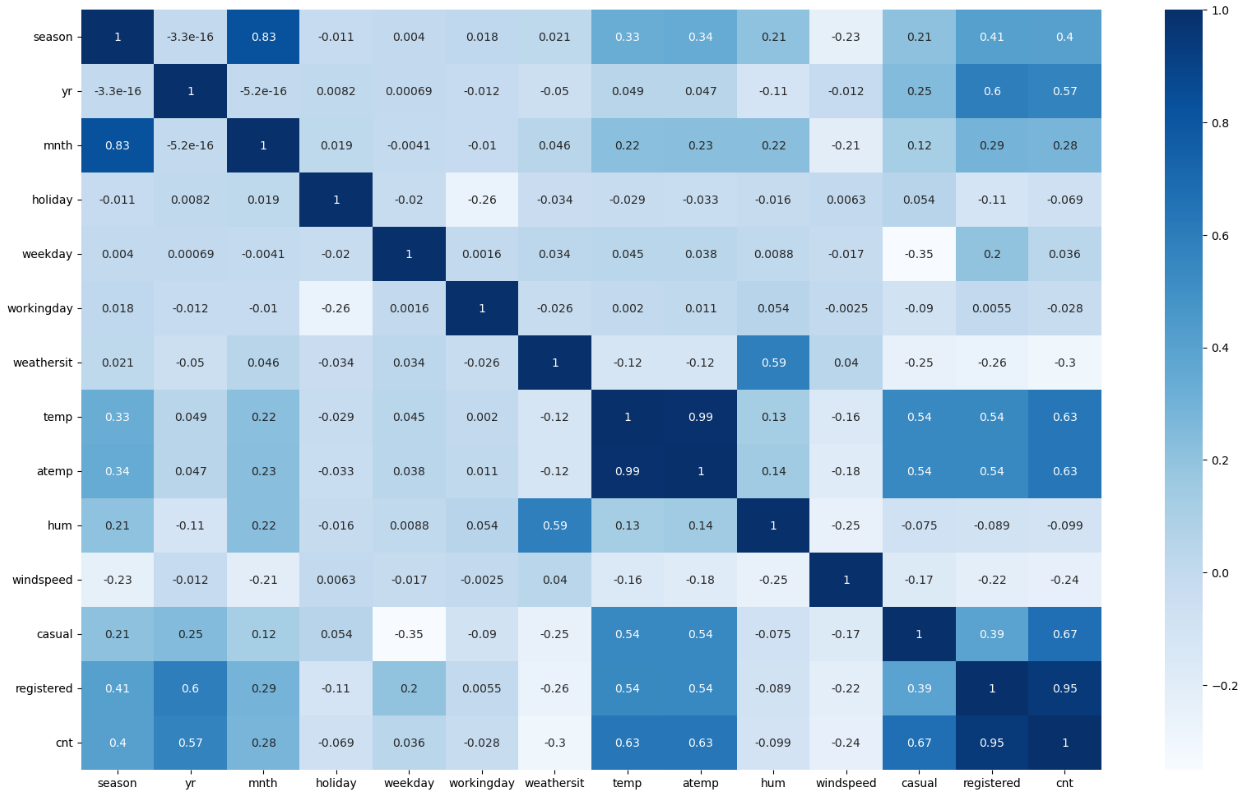
Linear Regression Subjective Questions

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?**

   More bikes are rented during fall season, when the sky is clear and when there is no holiday. These variables have a good relationship with the target variable ‘cnt’. There is also a good correlation of these categorical variables with ‘cnt’.
   Correlation for season with ‘cnt’ is 0.4,
   Correlation for weathersit with ‘cnt’ is -0.3(negative correlation : when one value increases other decreases),
   Correlation for working day with ‘cnt’ is -0.028(negative correlation),
   Correlation for holiday with ‘cnt’ is -0.069 (negative correlation)



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

   If we don’t use drop_first=True, it will be an extra variable
   For example: weathersit has 3 values: clear, mist, snow. Two variables will be enough to represent these values as below

   Let’s assume clear is dropped using drop_first

	Mist	snow
Clear	0	0
Mist	1	0
Snow	0	1

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

   From a pair plot a linear relationship is clearly visible for the variables registered, casual with ‘cnt’, But there might exist a collinearity, Then temp, atemp has a clear visible pattern with target variable.

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

   I will validate my assumptions of the model using the following criteria:
   1. R-squared value and Adj R squared value should increase when a new variable is added to model, Otherwise there will be no use adding the extra attribute.
   2. Checking the P value. It should be less than 0.05, This shows collinearity and checks if the model is significant.
   3. VIF <5 while adding the variables and removing them one after other.
   4. There should not much difference of r squared between train and test datasets. High difference indicates overfitting.

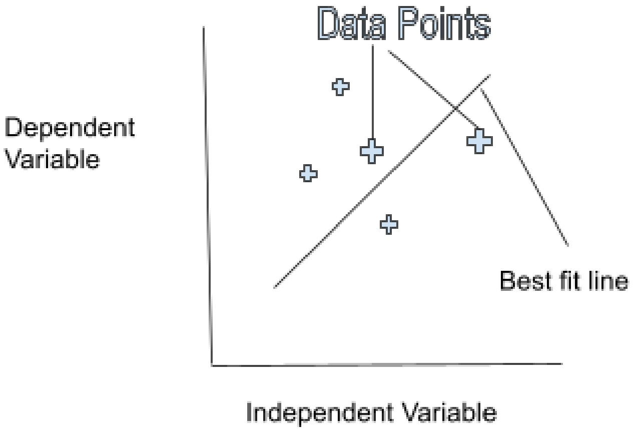
5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**

   Top 3 features contributing significantly towards explaining the demand of bikes in my model are casual, snow, spring. Then yr and weekday.

General Subjective Questions

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

   Linear Regression is a supervised ML model, which shows the linear relationship of independent variables with dependent variable. This algorithm focuses to reduce error but finding out best possible line to fit the data points to make error less.



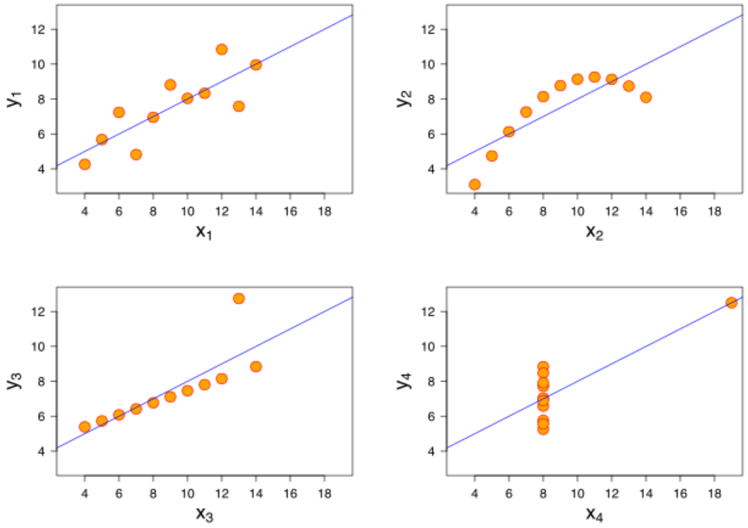
The steps involved in the algorithm are:

1. Exploratory Data Analysis and visualization to identify patterns.
2. Dummy variable creation for categorical variables.
3. Splitting into test and train datasets.
4. Then build the model starting with one variable and add other variables one by one while checking R- squared and p value.
5. R squared should increase and p value<0.05
6. Adding all the variables and build the model.
7. Calculate VIF, It should be less than 5.
8. Remove one variable with high VIF and again build the model and calculate VIF.
9. Continue the process
10. Finally we will have variables with VIF<5 and p<0.05. We can stop here.
11. Residual Analysis, They should be a normal distribution
12. Then calculate r squared for train and test datasets, There should not be much difference then the model is significant. High difference indicates overfitting.

2. Explain the Anscombe’s quartet in detail.

Anscombe’s quartet states that EDA and visualization of data are important rather than completely dependent on statistical analysis. There might be the same results for statistical analysis but the distribution of the data might be different.

For eg : The regression line for all the datasets is same, but the distribution is different.



3. What is Pearson’s R?

Pearson’s R also called Pearson correlation coefficient is a correlation coefficient that measures linear correlation between two sets of data. R calculates the error(distance) of the data point from the regression line. The formula for R is below.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Where x : x values of data points
y : y values of data points
$\bar{X}$: mean of x values
$\bar{Y}$: mean of y values

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

   There would be some variables with high magnitude. In this case only values are considered while building a model but not the units, So scaling helps to map the values in the same magnitude so that high magnitude values will not be able to overpower the coefficients of the model.

   a. Normal Scaling(Min Max Scaling) brings all the values in between 0 and 1. Formula is x=(x-min(x))/(max(x)-min(x))
   b. Standard scaling replaces the values with z score, This brings the data to normal distribution where mean = 0 , Standard Deviation = 1.

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

   High VIF or Infinite VIF indicates high collinearity of variables, When VIF is infinite, the variables are almost the same as each other. From the Assignment temp, atemp has high VIF, both the variables are collinear.

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

   Q-Q plots are also known as Quantile-Quantile plots, they plot the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential. These plots help to identify if two populations are of the same distribution, check if residuals have normal distribution and to check skewness of the data.