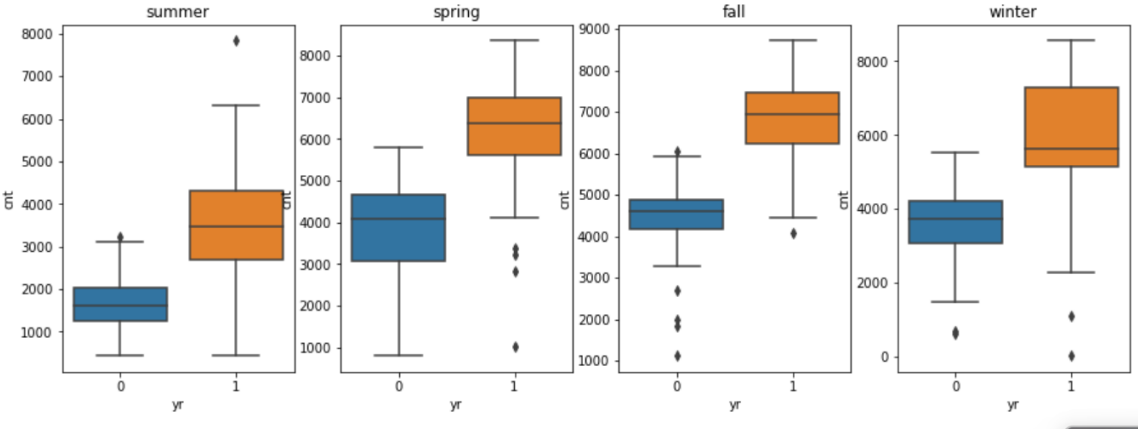
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Solution:- As per the analysis of categorical variables we can infer

List of Categorical variables “season ,mnth , weekday and weathersit”.

- The booking count has increased in 2019 as compared to 2018.
- The bookings continue most of the year but people increased could be seen for the fall and spring season and more for months of May, June, July, aug, sep and Oct.
- weathersit1 attracted more bookings which seems obvious. i.e The booking count has increased for Clear weather , on holiday and when temp is normal and humidity and wind is not high
- booking is people preferred on holiday and working day or non-working day show identical responses.



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

Solution:-

"drop_first = True"

is important to while creating dummy variables it helps in reducing the extra column creation. i.e. reduces the correlations created among dummy variables

“data_new1 = pd.get_dummies(data, drop_first= True)”

Converting categorical Variable into required format

Example:-

season : season (1:spring, 2:summer, 3:fall, 4:winter)

This is a nominal categorical variables i.e. all seasons are the same so we will convert into dummy variables for the season(n-1)

here 3 new variables could predict the 4 seasons. i.e. if summer=0 ,fall=0 and winter=0 than spring=1

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

   Solution:- "hum "has the highest correlation with the target variable and next is "temp".

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Solution:-

Normality of error terms i.e Error terms should be normally distributed

Multicollinearity check i.e There should be insignificant multicollinearity among variables.

Linear relationship validation i.e Linearity should be visible among variables

Homoscedasticity i.e. There should be no visible pattern in residual values.

Independence of residuals i.e. No auto-correlation

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks) '

Solution :- Temp, Season_4:- winter ,Month_8= aug

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.527e-16	0.018	8.71e-15	1.000	-0.034	0.034
yr	0.5072	0.018	28.354	0.000	0.472	0.542
temp	0.5758	0.022	25.757	0.000	0.532	0.620
hum	-0.1591	026	-6.198	0.000	-0.210	-0.109
windspeed	-0.1322	0.019	-6.841	0.000	-0.170	-0.094
season_2	0.1627	0.031	5.210	0.000	0.101	0.224
season_4	0.2641	0.024	10.941	0.000	0.217	0.311
mnth_3	0.0565	0.020	2.857	0.004	0.018	0.095
mnth_4	0.0392	0.025	1.547	0.122	-0.011	0.089
mnth_5	0.0568	0.026	2.219	0.027	0.007	0.107
mnth_8	0.0772	0.021	3.726	0.000	0.036	0.118
mnth_9	0.1581	0.020	8.082	0.000	0.120	0.197
mnth_10	0.0672	0.022	3.111	0.002	0.025	0.110
weathersit_2	-0.0862	0.022	-3.836	0.000	-0.130	-0.042
weathersit_3	-0.1307	0.031	-6.372	0.000	-0.171	-0.090

6.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
   Solution :

Linear regression is the statistical model that analyses the linear relationship between a dependent variable and with given set of independent variables.

Linear regression is of the following two types −
Simple Linear Regression
Multiple Linear Regression

equation − Y = mX + c

Y is the dependent variable we are trying to predict.
X is the independent variable we are using to make predictions.
m is the slope of the regression line which represents the effect X has on Y
c is a constant, known as the Y-intercept. If X = 0, Y would be equal to c.

if one or more independent variables will change the dependent variable will also change accordingly.

Furthermore, the linear relationship can be positive or negative in nature as explained below−

Positive Linear Relationship: A linear relationship will be called positive if both independent and dependent variable increases.

Negative Linear relationship: A linear relationship will be called positive if the independent increases and the dependent variable decreases.

2. Explain Anscombe's quartet in detail. (3 marks)

3. What is Pearson's R? (3 marks)

Solution:-

The Pearson correlation coefficient, r, can take a range of values from +1 to -1. The value 0 indicates that there is no association between the two variables and the value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable.

i.e. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition to the high values of the other, the correlation coefficient will be negative.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

   Solution :-

   Scaling is the process of reducing the variable size as it is important when we run bult of data it gives efficiency and increases the performance.

   a) Min-max scaling
      Xtrain-xmin/xtrain-xmax
   b) Standard Scaler
   c) xtrain=(xtrain-xtrain.mean())/xtrain.std()

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

(3 marks)

Solution:-

VIF = infinity then there is a perfect correlation.

i.e. A large value of VIF indicates that there is a correlation between the variables.

If the VIF is 5 or more, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. And we could drop this variable to increase the efficiency and correlation.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Solution:-

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

A q-q plot is used to plot the first data set against the quantiles of the second dataset.

This is important because it is used to describe if the assumption of a common distribution is justified.