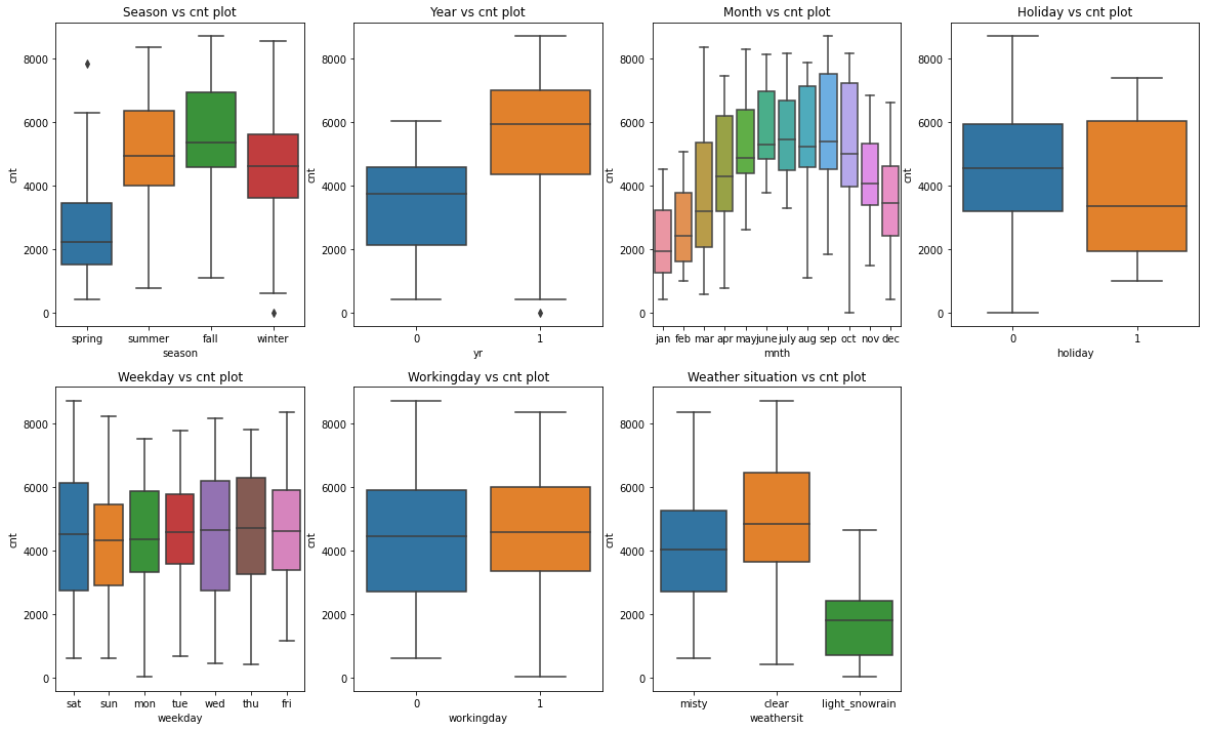
Assignment-based Subjective Questions

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



From the above box plots, the following points can be inferred -

1. The highest number of bookings are made in the fall season.
2. Bookings in 2019 are greater than that in 2018.
3. In a year, bookings increase till mid of the year and then decrease towards the year end.
4. Bookings decrease on holidays.
5. Bookings are almost similar on working and non working days.
6. Clear weather attracts more bookings.

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

drop_first=True during dummy variable creation helps in removing the first dummy column created from the categorical variable.

For example, if a categorical variable can take ‘n’ distinct values, then during dummy variable creation, ‘n’ variables will be created. But if we do drop_first=True, then ‘n-1’ variables will be created and the first one will be dropped.
This helps in reducing the correlation among the dummy variables.

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

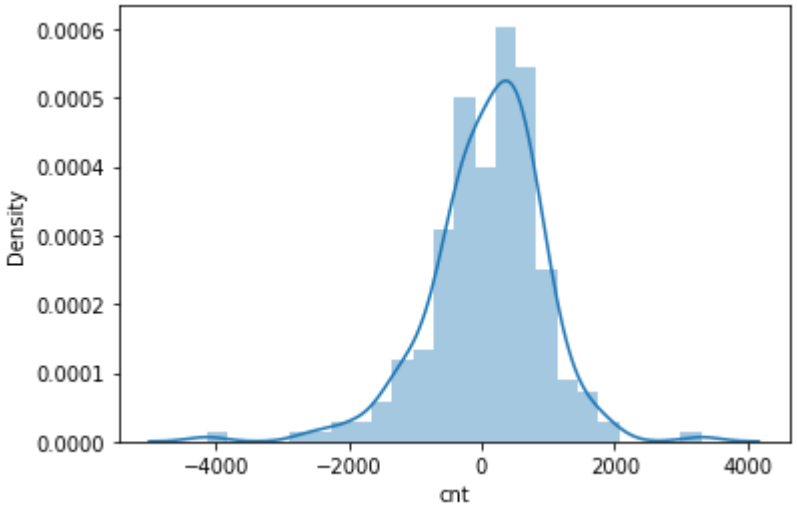
The variables ‘temp’ and ‘atemp’ have the highest correlation with the target variable ‘cnt’.

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

The following assumptions of linear regression have been validated after model building -

1. Linear relationship between independent and dependent variables
    - Pair plot was built before the model building to confirm the linear relationship between the variables.

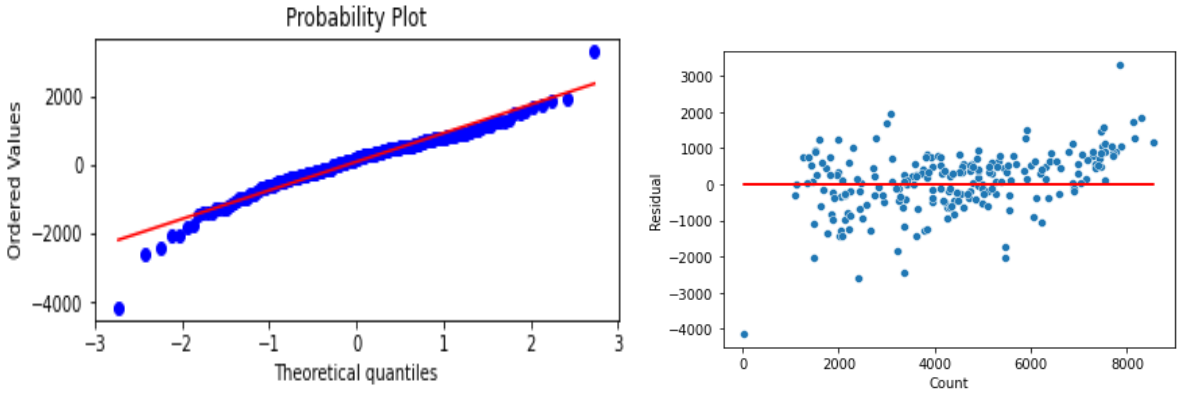
2. Normality of error terms
    - Distribution plot formed to check normality



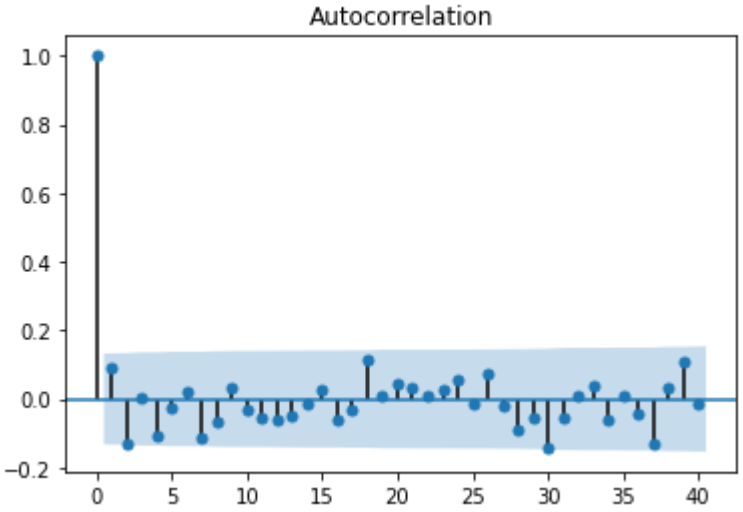
3. Multicollinearity check

- VIF values of all the columns used in the model is < 5.

4. Homoscedasticity of residuals
    - Q-q plot and scatter plot built to check if residuals have constant variance



5. Residuals should be independent
    - Plot built to check the autocorrelation between residuals.



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 final model, the top 3 features which contribute significantly to the demand of shared bikes are -

1. Temp -> coeff. 3576.6992
2. Year -> coeff. 1998.9114
3. Weathersit_light_snowrain -> coeff. -1658.8462

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear Regression is a type of supervised machine learning algorithm in which we can analyse the linear relationship between the independent variable(s) and the dependent variable and then use it to predict the values of the target dependent variable. It can be applied only on numeric values. The target variable needs to be a continuous variable.

Linear regression algorithm finds the line of best fit “Y = m * X + c” by means of ordinary least squares method where m = slope and c = intercept.

Linear Regression can be further divided into -

1. Simple Linear Regression - when the value of a dependent variable is predicted using a single independent variable. Here, equation of line of best fit is Y = m * x + c

2. Multiple Linear Regression - when the value of a dependent variable is predicted using two or more independent variables. Here, equation of line of best fit is Y = m1 * x1 + m2 * x2 + … + mN * xN + c

Assumptions made in linear regression algorithm -

1. Linear relationship between dependent and independent variables
    - The linear relationship between the independent and dependent variables can either be -
    - Positive - increase in independent variables increases the dependent variable.
    - Negative - increase in independent variables decreases the dependent variable.

2. Multi - coliinearity
    - In case of multiple linear regression, the independent variables should have low or dependency on each other.

3. Normality of error terms
    - The error terms or residuals should be normally distributed

4. Homoscedasticity
    - The residuals should have constant variance i.e. no visible pattern

5. Autocorrelation
    - The error terms should have negligible correlation i.e. there is no dependency in them

The linear regression model is mostly evaluated using the R-squared value, which is derived by

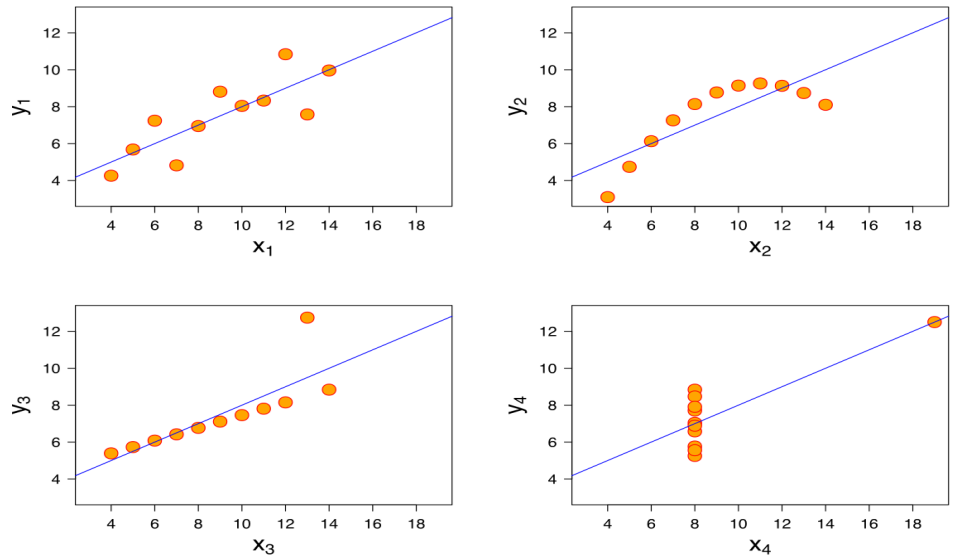
$$R^2 = 1 - \frac{RSS}{TSS}$$

2. Explain the Anscombe's quartet in detail.

Anscombe's Quartet is a combination of four data sets each having 11 data rows (x,y). It was developed by statistician Francis Anscombe. It is special in the way that all the four datasets are similar in terms of descriptive statistics but on plotting the data values, the picture is entirely different.

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

It is evident that the datasets have the same sum, average and s.d. But on plotting the scatter plot, we get the following insights.



1. The 1st dataset seems to have a simple linear relationship.
2. The 2nd dataset is not normally distributed.
3. The 3rd dataset has a linear relationship but the regression line is not able to predict the only outlier.
4. The 4th dataset shows that an outlier is sufficient to have a high correlation even though other data points are not at all correlated.

The main motive of Anscombe's Quartet is to explain the importance of data visualization in the field of data science and machine learning as the data sets which had equal descriptive statistics are so much dissimilar to each other in actual and this could only be inferred by the plotting the data values.

3. What is Pearson's R?

Pearson's R is a statistic which is used to measure the strength of linear relationship between two variables. Its value ranges between -1 to +1. If the variables have a positive linear relationship, the Pearson's R is also positive and if the variables have a negative linear relationship, the Pearson's R is also negative.

The value of Pearson's R can be derived by

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

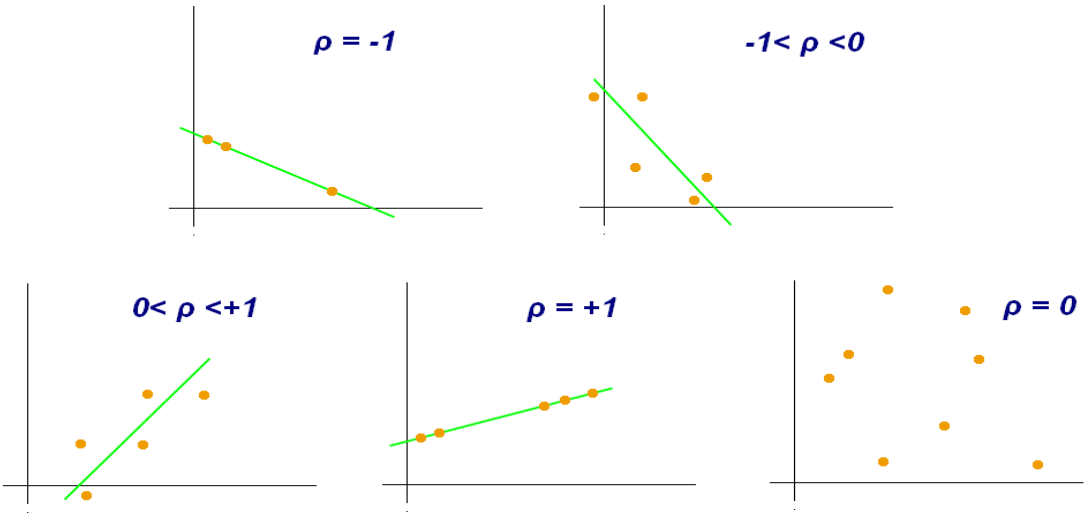
x_i = x - values in the sample

$\bar{x}$ = mean of x - values

y_i = y - values in the sample

$\bar{y}$ = mean of y - values

The different values of Person's R can be explained with the help of the following graphs.



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

Scaling can be defined as the process of transforming different numerical variables in order to bring them in a certain range for easy modelling. It is done during the preprocessing step to prepare data for model building so that highly varying values in the dataset are easily handled. If scaling is not done before model building, then the model tends to give higher weightage to variables having high values and less weightage to variables having lower values which will result in an incorrect and non reliable model.

For example, if we don't do feature scaling, then a model will find 100 kgs greater than 2 tons, which is exactly not the case. Such a model will yield wrong results and hence, it is important to scale the features.

The differences between normalized and standardized scaling include -

S.No.	Normalized Scaling	Standardized Scaling
1.	Minimum and maximum values are used for scaling.	Mean and standard dev. values are used for scaling.
2.	The scaled values lie between [-1,+1].	There is no definite range of scaled values.
3.	The scaled values are affected by outliers.	The scaled values are not affected by outliers.
4.	It is used when there is no idea about the distribution of values.	It is used when it is known that the values are normally distributed.
5.	It is also known as scaling normalization.	It is also known as Z - score normalization.
6.	Done using MinMaxScaler provided by sklearn library.	Done using StandardScaler provided by sklearn library.
7.	Formula applied - $x = \frac{x - min(x)}{max(x) - min(x)}$	Formula applied - $x = \frac{x - \mu_x}{\sigma_x}$

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

VIF (Variance Inflation Factor) is a statistic which is used to determine the collinearity of a variable with other independent variables. It is calculated as $VIF_i = \frac{1}{1 - R_i^2}$.

A higher value of VIF denotes high correlation between the variables. If VIF is infinite, 1 - R^2 = 0, then R^2 = 1, which means that the two columns are perfectly correlated with each other. In such a case, it is important to remove either of the correlated variables to remove perfect multicollinearity, so that the built model does not give any ambiguous predictions.

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

A q-q (quantile-quantile) plot is a probability plot used to check if two datasets come from populations with the same or common distribution.

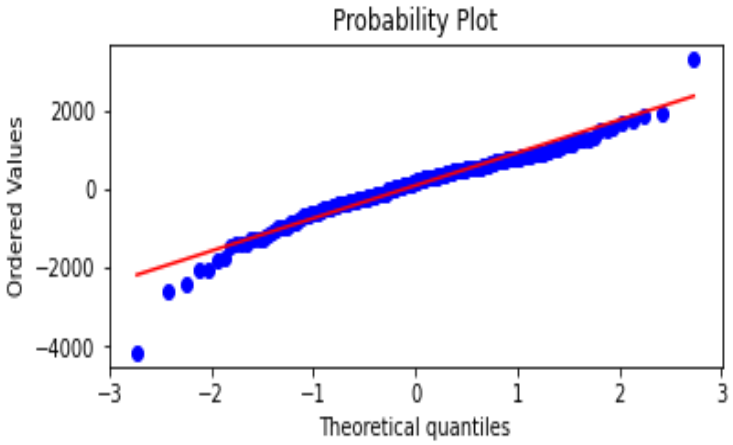
Use of q-q plot :-
In a q-q plot, quantiles of two data sets are plotted against each other along with a line at an angle of 45 degrees. In case the two datasets are from the populations with common distribution, the points lie on the 45 degrees reference line. The greater is the distance of points from the reference line, higher are the chances that the data sets belong to populations with different distributions.

A q-q plot helps in knowing the following details about the two data sets plotted -

1. Whether the data sets belong to populations with same distribution
2. Whether the data sets have a common location and scale
3. Whether the data sets have similar distribution shapes

Apart from these, q-q plots can also be used to check following in Linear Regression -

1. Skewness of distribution
2. Normality and homoscedasticity of the residual terms



Importance of q-q plot :-

When we have two data sets, it is necessary to test the assumption that they belong to populations with common distribution. If true, location and scale estimators can pool both datasets to find estimates of the common location and scale. If false, then q-q plot can help in finding the nature of difference between the two data sets in a better way than the analytical methods.