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

- Booking has increased from year 2018 to 2019.
- Fall season has more no of bookings and its increasing from year 2018 to 2019.
- We see in reference to the month (may, June, July, august, september)the count is more than other months.
- When the weather is clear booking count is increased which is obvious as compare to rainy, snow or misty days.
- Thursday, Friday, Saturday and Sunday the count is almost same in the year of 2019 as compare to other days in the year 2018.
- There is slight difference in Working-day and non-working , though working day is more than non-Working-day because might have plans or just want to stay at home.
- On holidays and on weekends , more customers go for bike bookings.

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

A variable with n levels presented by n-1 dummy variables. drop_first=True is used as it helps in minimising the extra columns created during the creation of dummy variables. Thus, it reduces the correlation created among dummy variables.

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

By looking at the pair-plot among the numerical values, temp and atemp has the highest correlation with the target variable(count).

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

- By doing the residual analysis, it shows the normal distribution.
- There is linear relationship between target variables and predictors.

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

Looking at the final model, I can see top three features are 'temp', 'year' and 'summer_season' with VIF's 2.71, 1.93, 1.49 resp. which explains the demand of the shared bikes.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is the machine learning algorithm which is used for supervised learning. Linear regression performs the task to predict a dependent variable based on the independent variable. So, this technique of regression finds out a linear relationship between a dependent variable and the other given independent variables. This means, it creates a best straight fit line to check the linear relationships between independent and target variables.

Mathematically, we can write a linear regression equation as:

$$y = a + bx$$

Where a and b given by the formulas:

$$b(slobe) = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$a(intercept) = \frac{n\sum y - b(\sum x)}{n}$$

Here, x and y are two variables on the regression line.

b = Slope of the line

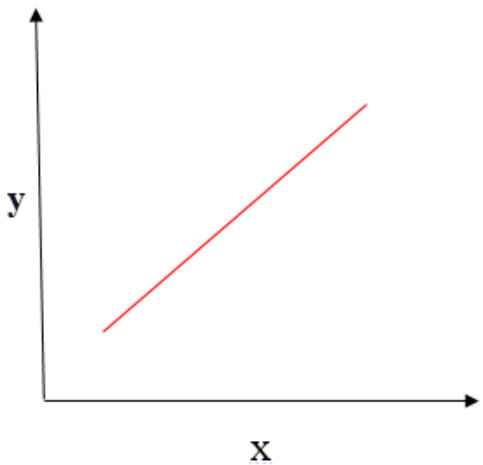
a = y-intercept of the line

x = Independent variable from dataset

y = Dependent variable from dataset

Positive Linear Relationship:

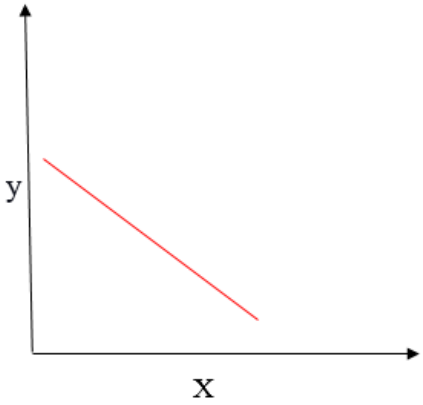
If the dependent variable expands on the Y-axis and the independent variable progress on X-axis, then such a relationship is termed a Positive linear relationship.



$y = a_0 + a_1 x$

Negative Linear Relationship:

If the dependent variable decreases on the Y-axis and the independent variable increases on the X-axis, such a relationship is called a negative linear relationship.



$y = -a_0 + a_1 x$

The goal of the linear regression algorithm is to get the best values for a0 and a1 to find the best fit line. The best fit line should have the least error means the error between predicted values and actual values should be minimized.

There are two types of liner regression:

- Simple linear regression
- Multiple linear regression

Simple Linear Regression:

A simple linear regression model attempts to explain the relationship between a dependent variable and one independent variable using a straight line.

The independent variable is also known as the predictor variable, and the dependent variables are also known as the output variables.

The equation of the best fit regression line :

$$Y = \beta_0 + \beta_1 X$$

Multiple Linear Regression:

This is used when the number of independent variables is more than 1.Multiple linear regression is needed when one variable is not sufficient to create a good model and make accurate predictions.

The formulation for predicting the response variable now becomes this

$$Y=\beta 0+\beta 1X1+\beta 2X2+\ldots+\beta pXp+\epsilon$$

Cost function

The cost function helps in figuring out the best possible values for a0 and a1, which will provide the best fit line for the data points. Cost function optimizes the regression coefficients or weights and measures how a linear regression model is performing. The cost function is used to find the accuracy of the **mapping function** that maps the input variable to the output variable. This mapping function is also known as **the Hypothesis function**.

Mean Squared Error (MSE) cost function is used in linear regression, which is the average of squared error that occurred between the predicted values and actual values.

By simple linear equation y=mx+b we can calculate MSE as:

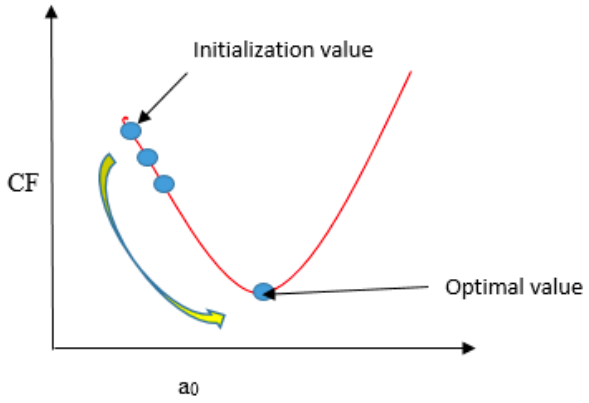
Let's y = actual values, yi = predicted values

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(y_i - (mx_i + b))^2$$

The goal of the linear regression algorithm is to get the best values for a0 and a1 to find the best fit line. The best fit line should have the least error means the error between predicted values and actual values should be minimized.

Gradient descent

Gradient descent is a method of updating a0 and a1 to minimize the cost function (MSE). A regression model uses gradient descent to update the coefficients of the line (a0, a1 => xi, b) by reducing the cost function by a random selection of coefficient values and then iteratively update the values to reach the minimum cost function.



2. Explain the Anscombe's quartet in detail.

Anscombe's quartet were constructed in 1973, by famous statistician Francis Anscombe to give the glimpse of graphing data before analysing

It includes four datasets and 11 (x,y) points which have almost same statistical properties. Though they appear quite different when graphed.

Those 4 datasets and 11 points are given below to get the understanding:

	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

3. What is Pearson's R?

Correlation coefficients are used to measure how strong a relationship is between two variables. Pearson's is most famous correlation coefficient and largely used in linear regression which is denoted by 'r'. It varies from -1 to +1.

Where -1 is the negative linear correlation and +1 is positive correlation of the data and 0 means there is no correlation.

The formula of Pearson's r is :

$$r = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

Where:

N	=	number of pairs of scores
Σxy	=	sum of the products of paired scores
Σx	=	sum of x scores
Σy	=	sum of y scores
Σx^2	=	sum of squared x scores
Σy^2	=	sum of squared y scores

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

It is extremely important to rescale the variables so that they have a comparable scale. If we don't have comparable scales, then some of the coefficients as obtained by fitting the regression model might be very large or very small as compared to the other coefficients. This might become very annoying at the time of model evaluation. As you know, there are two common ways of rescaling:

Standardisation brings all the data into a standard normal distribution with mean 0 and standard deviation 1.

MinMax scaling, on the other hand, brings all the data in the range of 0-1. The formulae used in the background for each of these methods are as given below:

Formula of Normalized scaling:

$$x = \frac{x - min(x)}{max(x) - min(x)}$$

Formula of Standardized scaling:

$$x = \frac{x - mean(x)}{sd(x)}$$

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

Below is the formula used to calculate the VIF:

$$VIF_i = \frac{1}{1 - R_i^2}$$

Where, 'i' is the i-th variable.

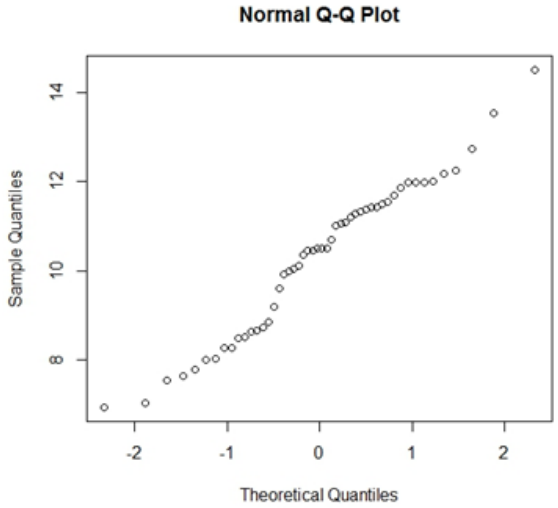
So, if the value of R-squared becomes 1 then the denominator of the formula becomes 0 and the overall value of VIF becomes infinite which shows the perfect correlation in variables.

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

The Q-Q plot is also known as quantile-quantile plot which is a graphical technique for determining if two data sets come from populations with a common distribution.

It is a scatterplot only which can be plotted using two sets of quantiles against each other. In case both sets of quantiles came from the same distribution, then points

forming a line should be almost straight. As we can see from the example of a Q-Q plot where both sets of quantiles come from Normal distributions.



Q-Q plot in Linear Regression: It is used to see if the points lie approximately on the line. If they don't, it means, our residuals aren't Gaussian (Normal) and thus, our errors are also not Gaussian.

Importance of Q-Q plot:

I. The sample size does not needs to be equal.

II. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers.

III. The q-q plot can provide more insight into the nature of the difference than analytical methods.