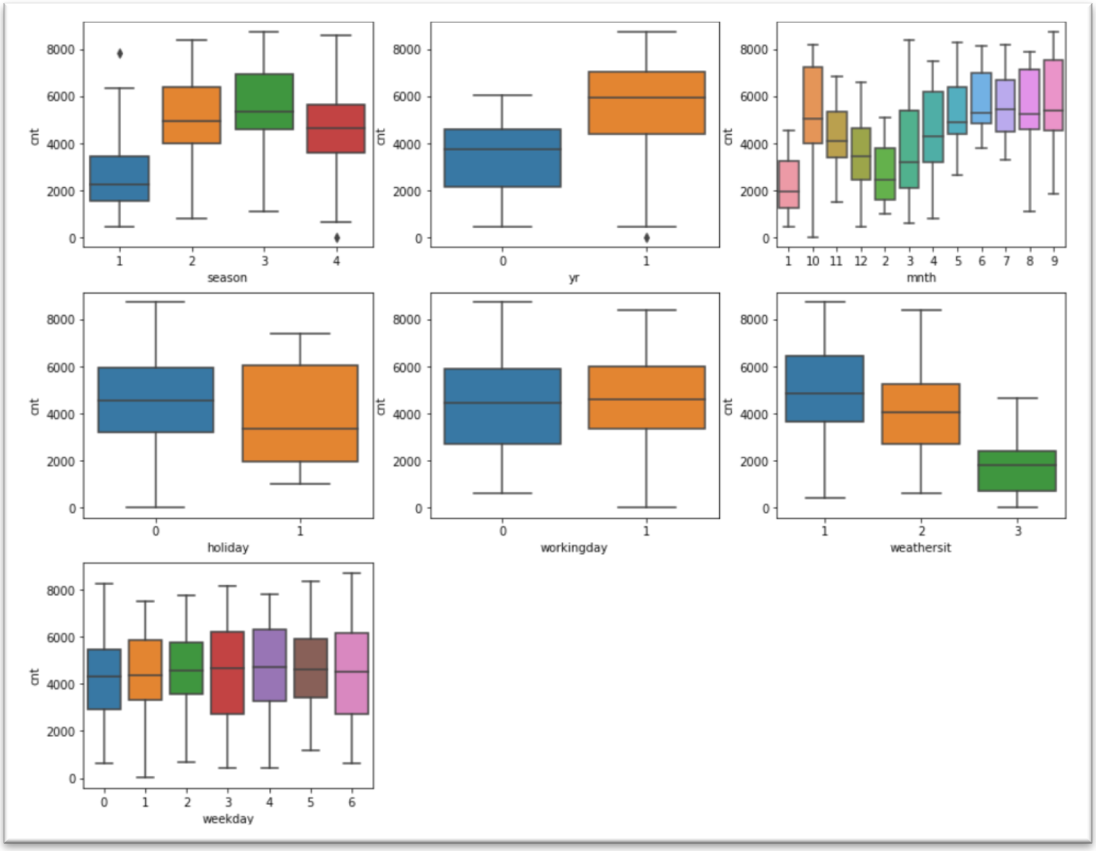
Assignment-based Subjective Questions

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

Refer to the image below, it is clear that except season and month, all the other categorical variables are having linear relationship with the Target variable. While Year has clear positive correlation, weathersit has clear negative correlation.



The slopes of the mentioned categorical variables can be found below as well.

	coef	std err	t	P>|t|
const	0.3106	0.026	12.172	0.000
yr	0.2305	0.008	27.848	0.000
holiday	-0.0937	0.026	-3.581	0.000
weathersit	-0.0743	0.010	-7.753	0.000
temp	0.4746	0.134	3.551	0.000
atemp	0.0590	0.141	0.418	0.676
hum	-0.1676	0.038	-4.382	0.000
windspeed	-0.1913	0.027	-7.195	0.000
mnth_8	0.0615	0.017	3.667	0.000
mnth_9	0.1238	0.017	7.480	0.000
season_2	0.1069	0.011	9.596	0.000
season_4	0.1429	0.011	13.083	0.000

Dummy variables

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

Simply put, if we don't drop the first column then dummy variables will be correlated
Also, the first column can be inferred or calculated using the binary method.

Let's take the example of gender. 3 Genders: Male, Female, Others

Gender	Male	Female	Others
Suraaj	1	0	0
Aastha	0	1	0
Rohit	0	0	1

The above can also be inferred using the below column

Gender	Male	Female
Suraaj	1	0
Aastha	0	1
Rohit	0	0

However, there's any category with hundreds of values, it's not recommended to drop the first column. That will make it easier for the model to "see" all the categories quickly during learning (and the adverse effects are negligible).

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

Target variable ('cnt') has the highest correlation with 'atemp' (~0.63).

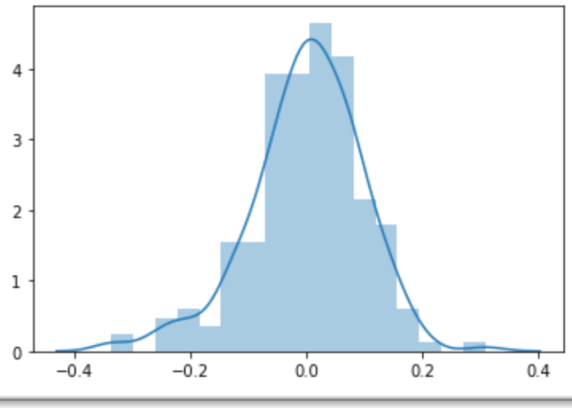
cnt	atemp	0.629747
cnt	temp	0.626290
cnt	yr	0.569728
cnt	season	0.404584
cnt	mnth	0.278191
cnt	weekday	0.067534
cnt	workingday	0.062542
cnt	holiday	-0.068764
cnt	hum	-0.098060
cnt	windspeed	-0.233517
cnt	weathersit	-0.295929

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

1. Target variable (cnt) and input variables are linearly dependent.

2. Checked the distribution of error: Normal distribution (Mean = 0)

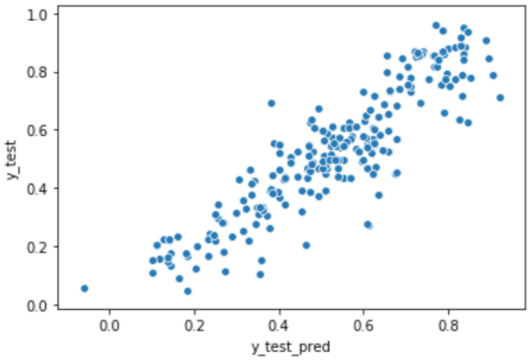
```
#Residual
residual_test=y_test-y_test_pred
sbn.distplot(residual_test)
plt.show()
```



3. Checking the relationship between predicted values and actual values. Error terms have constant variance (homoscedascity)

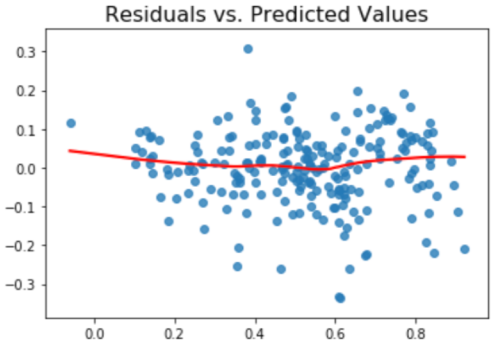
```
#checking the relationship between predicted values and actual values
sbn.scatterplot(x=y_test_pred,y=y_test)
plt.ylabel('y_test')
plt.xlabel('y_test_pred')
```

```
Text(0.5, 0, 'y_test_pred')
```



4. Checking the variance between errors and predicted values. Error terms are independent.

```
#checking the variance of errors
sbn.regplot(x=y_test_pred, y=residual_test, lowess=True, line_kws={'color': 'red'})
plt.title('Residuals vs. Predicted Values', fontsize=16)
plt.xlabel='Predicted'
plt.ylabel='Residuals'
```



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

   The below image shows the top features contributing significantly towards explaining the demand of the shared bikes. These are arranged on the basis of their coefficients (slopes)

1. Temp
2. year
3. season_4 (dummy of season)

temp	0.529628
const	0.311485
yr	0.230523
season_4	0.143449
mnth_9	0.123619
season_2	0.107364
mnth_8	0.060539
weathersit	-0.074528
holiday	-0.094144
hum	-0.166530
windspeed	-0.193410

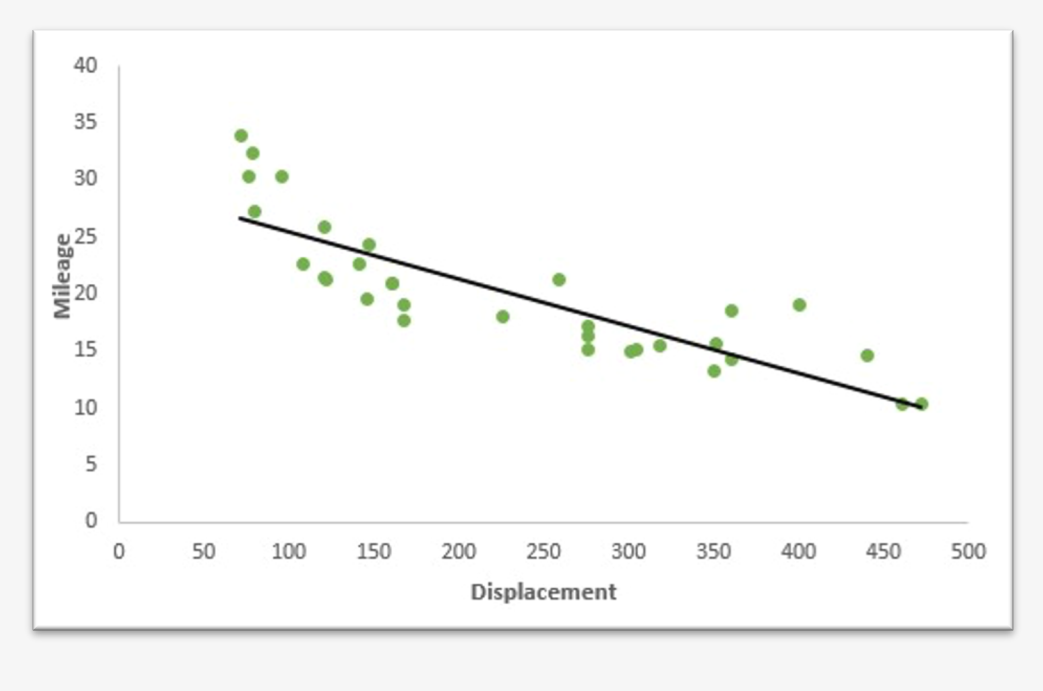
General Subjective Questions

1. Explain the linear regression algorithm in detail.

Regression analysis is used to model the relationship between a dependent variable and one or more independent variables.
Linear Regression is the simplest form of regression.

Technique:

1. The relationship between the dependent variable and independent variables is assumed to be linear in nature.
2. We can observe that the given plot represents a somehow linear relationship between the mileage and displacement of cars. The green points are the actual observations while the black line fitted is the line of regression



When we have only 1 independent variable and 1 dependent variable, it is called simple linear regression.
When we have more than 1 independent variable and 1 dependent variable, it is called Multiple linear regression.

The equation of multiple linear regression is listed below -

$$y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + \varepsilon$$

Here, 'y' is the dependent variable to be estimated
X are the independent variables
ε is the error term
βi's are the regression coefficients.

Assumptions of linear regression:

1. There must be a linear relation between independent and dependent variables.
2. Homoscedasticity
3. Error terms should be independent.
4. Error terms should be normally distributed with mean 0 and constant variance.

Estimating the parameters

To estimate the regression coefficients βi's we use principle of least squares which is to minimize the sum of squares due to the error terms i.e.

$$\text{Min} \sum \varepsilon^2 = \text{Min} \sum (y - (\beta_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k))^2$$

2. Explain the Anscombe's quartet in detail.

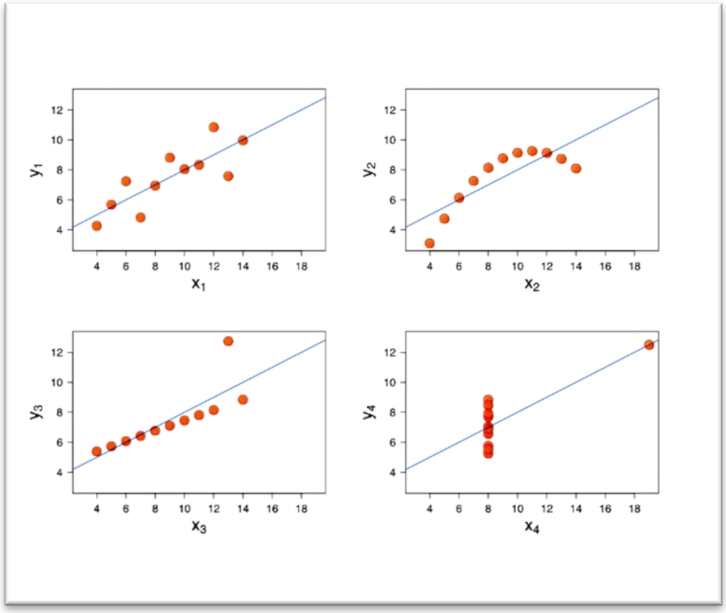
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

When we use .describe() on the above data, we will get same summary (mean, standard deviation & correlation)

Summary

Set	mean(X)	sd(X)	mean(Y)	sd(Y)	cor(X,Y)
1	9	3.32	7.5	2.03	0.816
2	9	3.32	7.5	2.03	0.816
3	9	3.32	7.5	2.03	0.816
4	9	3.32	7.5	2.03	0.817

But when we plot the datasets on the graph, we will get



Now, graphs were completely different even though the summary was exactly similar. This dataset came to be known as **Anscombe's quartet**.

Therefore, it's recommended to visualize before trusting a dataset

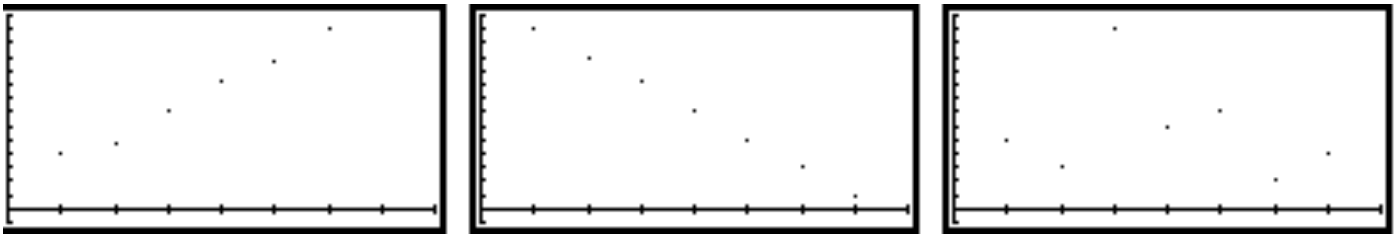
3. What is Pearson's R?

Pearson's correlation coefficient (r) is a measure of the strength of the association between the two variables.
Pearson's r is a numerical summary of the strength of the linear association between the variables. 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 with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

The Pearson's correlation coefficient varies between -1 and +1 where:

- **r = 1** means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- **r = -1** means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- **r = 0** means there is no linear association

The figure below shows some data sets and their correlation coefficients.



The first data set has an r=0.996, the second has an r = -0.999 and the third has an r= -0.233

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

Scaling is the method used to standardize the range of features of data. Since, the range of values of data may vary widely, it becomes a necessary step in data preprocessing while using machine learning algorithms.

In ***min-max scaling or normalized scaling***, you transform the data such that the features are within a specific range e.g. [0, 1].

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

where x' is the normalized value.

Standardization *(also called **z-score normalization**)* transforms your data such that the resulting distribution has a mean of 0 and a standard deviation of 1. It's the definition that we read in the last paragraph.

$$x' = \frac{x - x_{mean}}{\sigma}$$

where x is the original feature vector, xmean is the mean of that feature vector, and σ is its standard deviation.

- Normalization is good to use when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data.

- Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if we have outliers in your data, they will not be affected by standardization.

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

If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables.

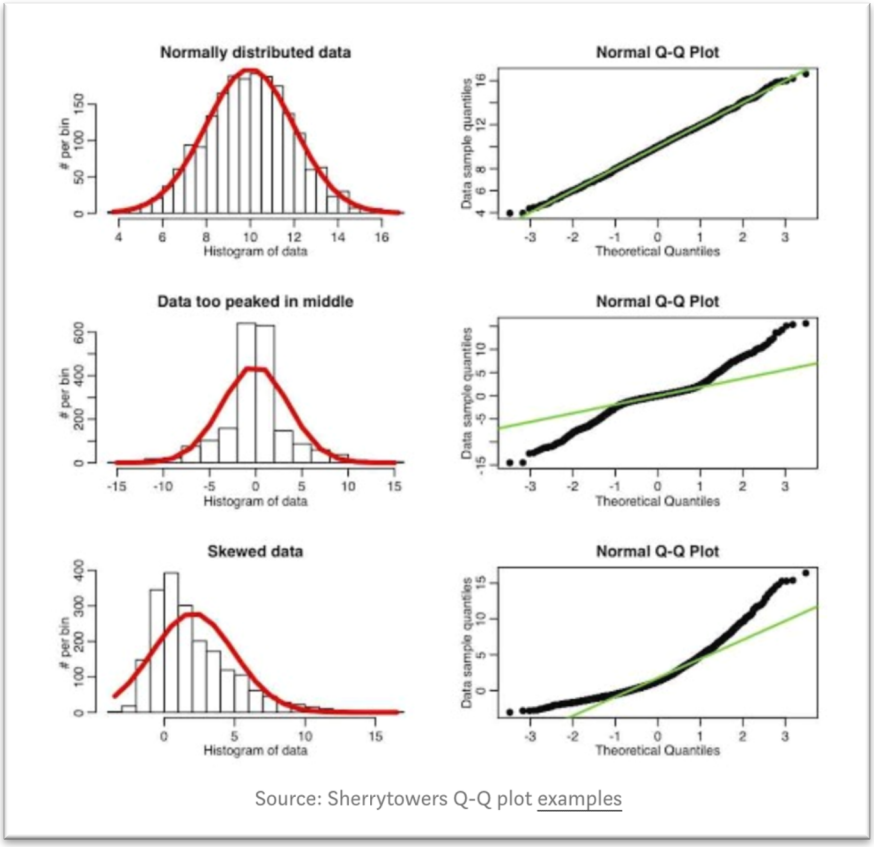
In VIF, each feature is regression against all other features. If R2 is more which means this feature is correlated with other features.

- VIF = 1 / (1 – R2)
- When R2 reaches 1, VIF reaches infinity

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

Q-Q plots are used to find the type of distribution for a random variable whether it be a Gaussian Distribution, Uniform Distribution, Exponential Distribution or even Pareto Distribution, etc. You can tell the type of distribution using the power of the Q-Q plot just by looking at the plot.



Source: Sherrytowers Q-Q plot examples

Q-Q plot importance in 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.

It is used to check following scenarios:
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