- **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?**

Answer –

Categorical variables affected the dependant variable to a large extent, way more than the numerical variables.

At the first glance of pairplot between numerical variables and target variable, correlation seemed to be highest with temp/atemp. However, after VIF analysis, both temp/atemp and hum got very high VIF values i.e they could be explained through other variables and therefore were dropped from the final model. Except windspeed, the final model has all dummy encoded categorical variables.

Following are the inferences about categorical variables:

1. Out of all weather situations, rain has the highest negative influence on demand
2. Year - positive correlation i.e. demand is likely to increase with each year
3. Out of all seasons, 'fall' has the highest positive influence on demand
4. Out of all days, 'monday' has a positive influence on demand
5. Out of all months, 'september' has the highest positive influence on demand
6. Working day has a positive influence on demand

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

Answer-

It is important to use drop_first=True during dummy variable creation for the following reasons:

1. Number of predictor variables increase, which is not actually required, as rest of the variables can explain the dropped variable
2. More importantly, the VIF of all dummy variables comes as infinite if the first dummy variable is not dropped. This makes selection of features for elimination ambiguous and unreliable.

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

Answer –

‘temp’ variable shows the highest correlation with the target variable, as shown by the pair plot.

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

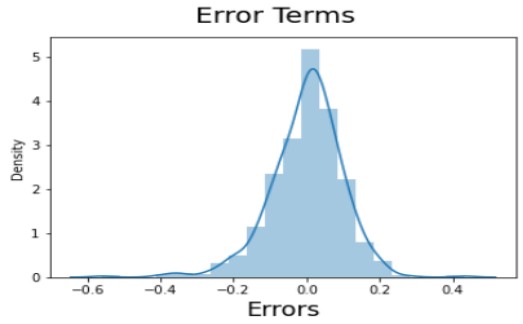
   Answer-

   Assumptions of Linear regression were validated after building the model on the training set by doing Residual analysis. Following steps were performed:

   1. Verifying the distribution of error terms to be normal by creating the distplot.

```
fig = plt.figure()
sns.distplot(res, bins = 20)
fig.suptitle('Error Terms', fontsize = 20)
plt.xlabel('Errors', fontsize = 18)
```

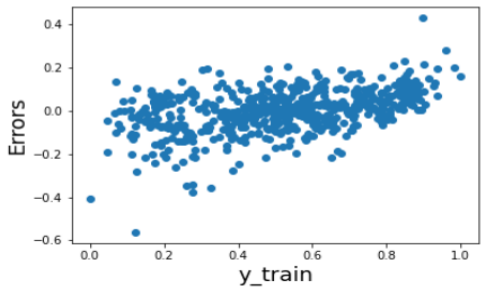
```
Text(0.5, 0, 'Errors')
```



   2. Verifying that the spread/variance of error terms has no pattern when measured against target variable. The spread was found to be almost uniform, neither increasing nor decreasing with y_train.

```
plt.scatter(x= y_train, y = res)
fig.suptitle('Spread of Error Terms', fontsize = 20)
plt.ylabel('Errors', fontsize = 18)
plt.xlabel('y_train', fontsize = 18)
```

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



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

   Answer-

   Following 3 features were found to be contributing most significantly towards explaining the demand of the shared bikes:

   1. Weather situation – Rain (negative influence)
   2. Season – Fall (positive influence)
   3. Year (positive influence)

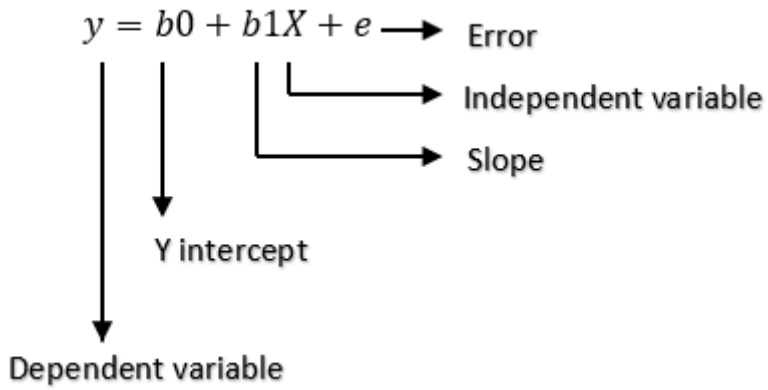
- ## General Subjective Questions

1. Explain the linear regression algorithm in detail

Answer-

Linear regression is a statistical technique to understand the relationship between one dependent variable and one or more independent variables (explanatory variables). The objective of linear regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X. It predicts the target variable by means of fitting a straight line through the given data points. Using the training data, a regression line is obtained which will give the minimum error. This linear equation is then used to apply for new data. That is, if we give X as an input, our model should be able to predict Y with minimum error.

The linear regression model is represented by the following equation:



Metrics for model evaluation:

1. Linear regression most often uses mean-square error (MSE) to calculate the error of the model.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

test set; predicted vaue; actual value

2. A better metric for the linear regression model is R squared. R-squared (R^2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. Whereas correlation explains the strength of the relationship between an independent and

dependent variable, R-squared explains to what extent the variance of one variable explains the variance of the second variable. So, if the R^2 of a model is 0.50, then approximately half of the observed variation can be explained by the model's inputs.

R2 = 1−(Unexplained Variation/Total variation)

The value of R2 always lies between 0 and 1. The higher , the better.

Assumptions of Linear Regression:

1. There is a linear relationship between X and Y
2. Residuals or Error terms are normally distributed with mean zero
3. Error terms are independent of each other
4. Error terms have constant variance

2. **Explain the Anscombe's quartet in detail.**

Answer-

Anscombe's quartet comprises of four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

These were the 4 data sets used:

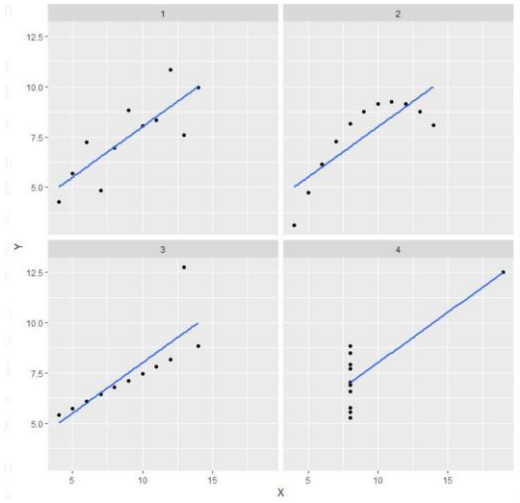
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	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

Statistical Properties were found to be exactly same as shown below:

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

Scatter Plots looked as follows:



In the first one(top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.

In the second one(top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.

In the third one(bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.

Finally, the fourth one(bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

Application:
The quartet is used to illustrate the **importance of looking at a set of data graphically** before starting to analyze according to a particular type of relationship, and **the inadequacy of basic statistic properties** for describing realistic datasets.

3. What is Pearson's R?

Answer-

A measure of linear correlation between two sets of data.

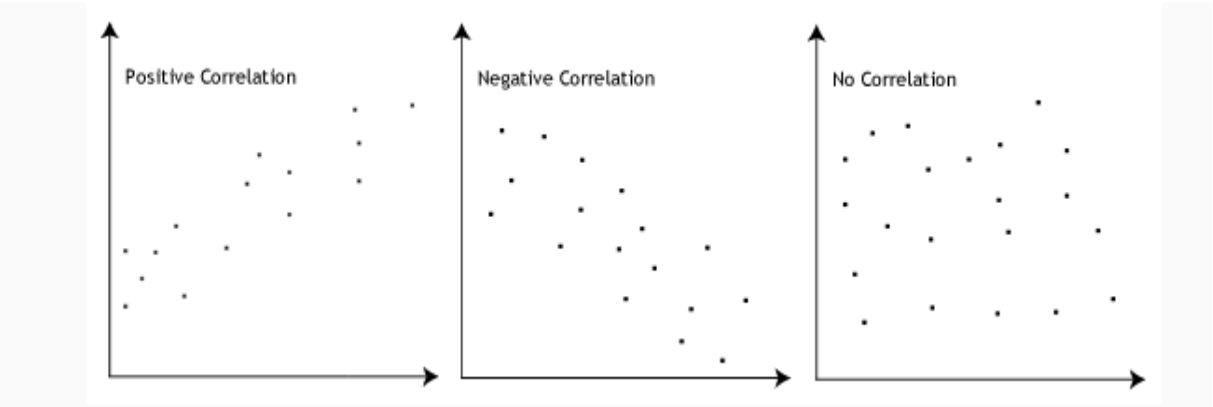
r = slope * (std deviation of x/std deviation of y)

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



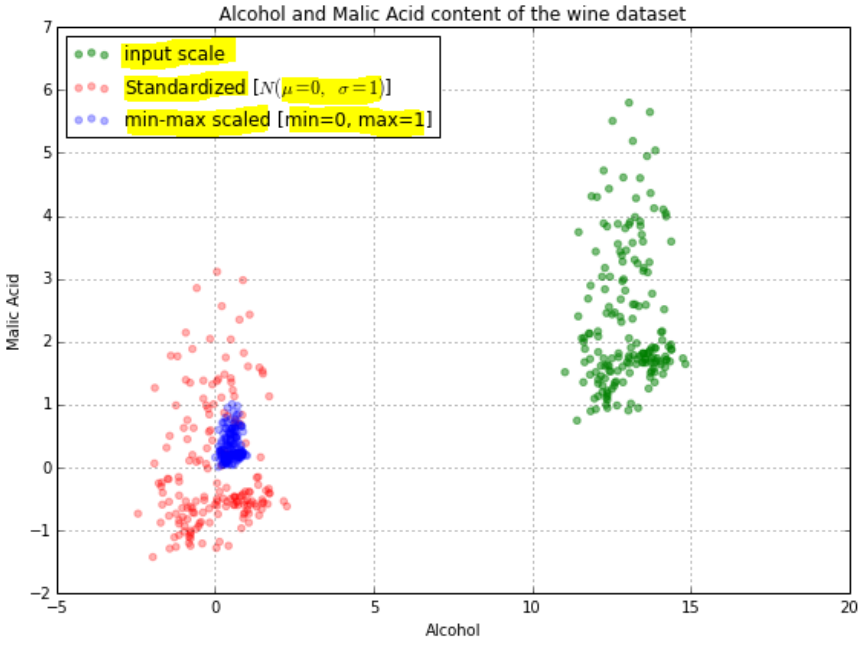
4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
Answer –

Scaling refers to the process of standardizing the values of independent numerical variables so that they can be compared on the same scale of measurement.

Scaling is helpful in Distance-based algorithms like Linear regression, K-Nearest neighbours, Neural networks and also in faster convergence of optimization algorithms like Gradient descent.

Scaling affects only the scale and does not change the underlying distribution.

Following image shows the effect of two popular types of scaling, namely normalization (min-max scaling) and standardization:



Normalized scaling rescales the values of variables to fall within [0,1].

$$x(new) = (x - min(x))/(max(x) - min(x))$$

- During normalisation, we lose outliers which is a slight disadvantage
- It does not affect the values of dummy variables

Standardization rescales the values such that the distribution has a mean of 0 and std deviation of 1.

$$x(new) = (x - mean(x))/std\ dev(x)$$

- Affects the values of dummy variables

Both techniques find their use in different situations. Like in PCA, where we do dimension reduction by plotting our 3D data into 1D(say), we make use of standardisation. In Image processing, it is required to normalise pixels before processing.

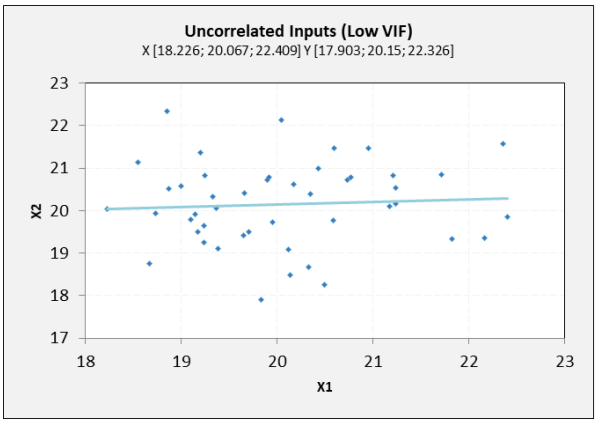
5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**

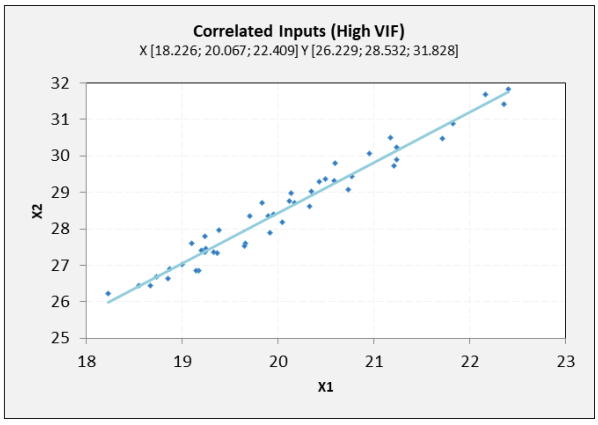
   Answer-

   There are generally two possibilities where VIF tends to be infinite:

   1. The value of VIF as infinite shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity.

      The second image shows perfect correlation between the input variables x1 and x2. The first image is what is actually desired.





      To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

   2. The value of VIF may tend to be infinite if the no. of predictor variables is way more than the no. of observations in the sample i.e. the sample size. To solve this, either increase the sample size or use a stepwise feature selection for model building.

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

The quantile-quantile (Q-Q) plot is a graphical method for determining whether two samples of data came from the same population or not. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value.

If we run a statistical analysis that assumes our residuals are normally distributed, we can use a Normal Q-Q plot to check that assumption. It's just a visual check but it allows us to see at-a-glance if our assumption is plausible, and if not, how the assumption is violated and what data points contribute to the violation. A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. For the reference purpose, a 45% line is also plotted, if the samples are from the same population then the points are along this line.

The Quantile-Quantile plot is used for the following purpose:

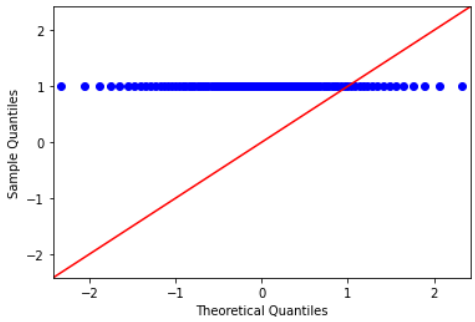
1. Determine whether two samples are from the same population.
2. Whether two samples have the same tail
3. Whether two samples have the same distribution shape.
4. Whether two samples have common location behavior.

How to Draw Q-Q plot

1. Collect the data for plotting the quantile-quantile plot.
2. Sort the data in ascending or descending order.
3. Draw a normal distribution curve.
4. Find the z-value (cut-off point) for each segment.
5. Plot the dataset values against the normalizing cut-off points.

Advantages of Q-Q plot

1. Since Q-Q plot is like probability plot. So, while comparing two datasets the sample size need not to be equal.
2. Since we need to normalize the dataset, so we don't need to care about the dimensions of values.



// Q-Q Plot For a Uniform Distribution //