Assignment Based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable ?

'season' and 'weathersit' are two categorical columns where I applied dummy encoding and they are quite significant according to the p-values for the prediction of 'cnt'. I did not apply dummy encoding to 'mnth', 'weekday', instead I normalized them because they are ordered categorical variables and label encoding was already applied.

Q2. Why is it important to use drop_first=True during dummy variable creation ?

Dummy encoding is used on categorical variables so that we can feed them into a linear regression model. Basically, we create 'n' new columns if 'n' is the number of classes in the categorical column. But, It turns out that we can choose to use only 'n – 1' columns instead of 'n' to feed the same number of classes to the model. It works because we can encode one class when all the other classes are negative. So, to use this concept in practice, we use **drop_first=True** argument in **pd.get_dummies()** method of pandas which gives us n – 1 new columns corresponding to the classes of the column we applied the method on.

Q3. Looking at the pair plot among the numerical variables, which one has the highest correlation with the target variable ?

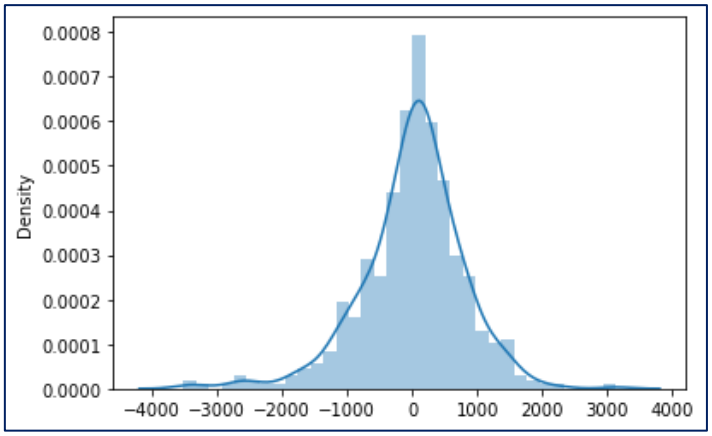
I used pair plot on the data initially, but the output had many too many small plots which made the interpretation difficult. According to correlation matrix, among the numerical variables, **'temp'** has the highest correlation with the target variable with the value of 0.63 followed by 'yr' with value of 0.57.

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set ?

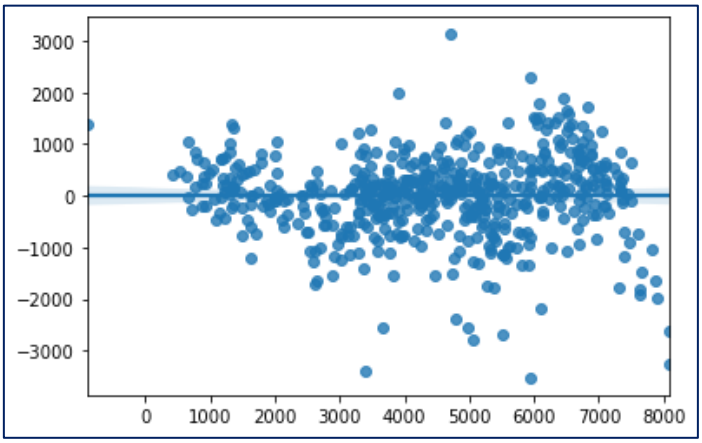
Linear Regression comes with 4 assumptions which are:-

- Linear relation between dependent and independent variables
- Normal distribution of error terms, centered around zero.
- Constant variance of the error terms
- Independent variables have no relation among them

The assumption of linear relation was validated by building the linear regression model with an Adjusted R2 score of 0.836.
Assumption of normal distribution of error terms was validation using the distribution plot of residuals as shown below:-



The assumption of constant variance of error terms with respect to predicted value was validated using the scatter plot as shown below:-



The last assumption of 'homoscedasticity' was validated by removing columns having high VIF values. The final VIF values of the variables look like this, highest value being 2.92:-

	Features	VIF
0	const	74.16
9	season_winter	2.92
2	mnth	2.78
5	temp	2.77
8	season_spring	2.75
6	hum	1.94
11	weathersit_mist	1.56
10	weathersit_light_rain	1.36
7	windspeed	1.23
1	yr	1.05
3	weekday	1.01
4	workingday	1.01

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes ?

According to absolute values of coefficients, three most significant variables for prediction of 'cnt' are:-

S.No.	Column	Coefficient
1.	temp	4264.98
2.	yr	1986.37
3.	weathersit_light_rain	-1845.75

All the variables were normalized between 0 and 1 and multicollinearity was taken care of so the coefficient values are interpretable.

General Subjective Questions

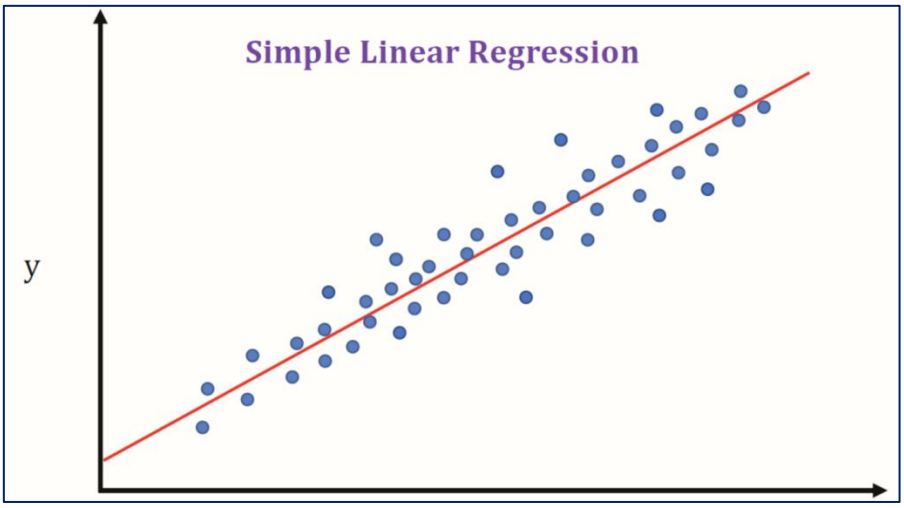
Q1. Explain the linear regression algorithm in detail.

Linear Regression is an algorithm which we use to find the best fit line between one or more independent variables and one dependent variable.

The use of Linear Regression is that we are able to predict the value of dependent variable for the given values of independent variable(s) on which the model may or may not have been trained upon.

Once trained on data, it is able to predict dependent variable on never before seen values of independent variables. Model having only one independent variable is called simple linear regression and more than one independent variables version is called multiple linear reg.

The data distribution and the best fit line for simple linear regression looks like this:-



By finding the best fit line, we are mathematically discovering the linear relationship between independent and dependent variable which is already present in the data. Because we are finding the linear relationship, Linear Regression comes with few assumptions about the data.

Assumption of Linear Regression:-

1. Linearity: There exists a linear relationship between input and target variables
2. Normality: The error terms are normally distributed having mean of 0
3. Homoscedasticity: The distribution of error terms have constant variance
4. Independence: The independent variables are no correlation between them

Working of Linear Regression:-

As you know that we are finding the best fit line so we use the formula for a line to find it.

$$y = m * X + c$$

where, y = Dependent variable
X = Independent variable
m = Slope of line
c = Intercept of line

Essentially, we are finding the best value of 'm' and 'c' for the data which is able to capture most of the variance in the dataset.

For multiple linear regression, the concept remains the same with few changes:-

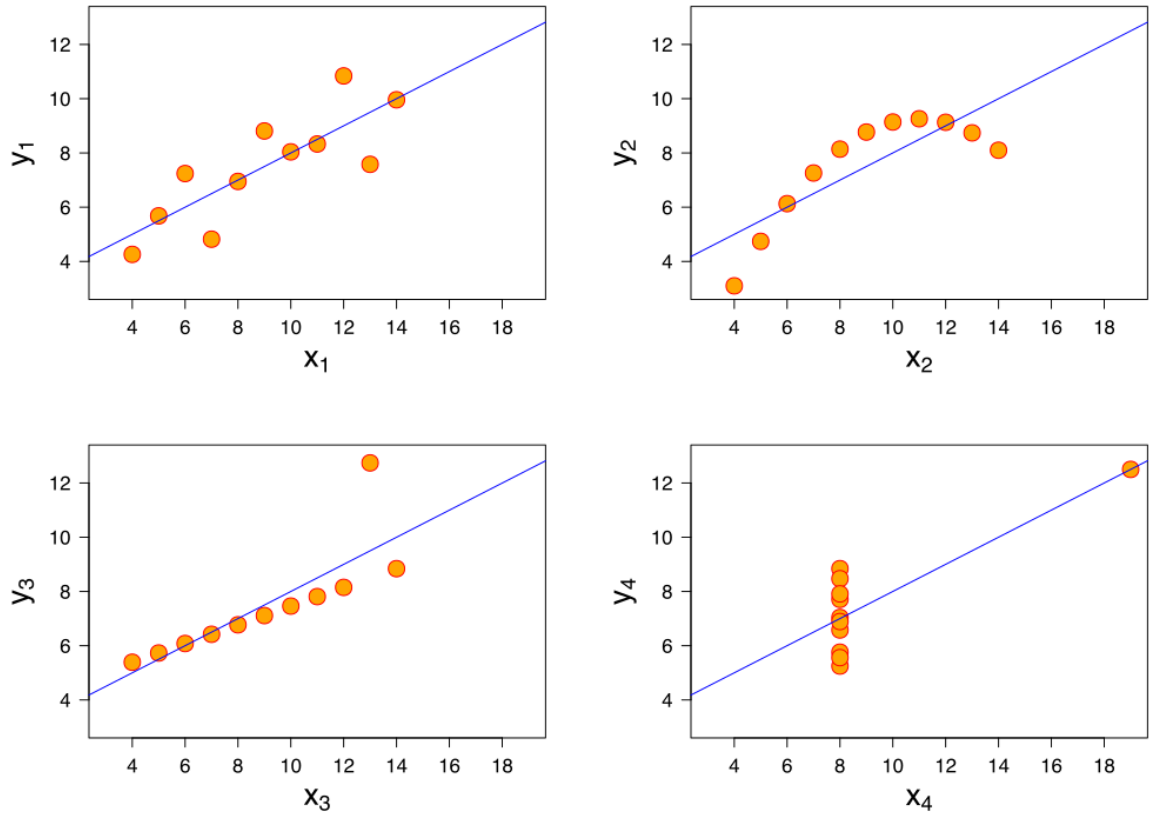
1. Dimensionality of input data increases
2. There are as many values of 'm' as there are number of columns

Let us see how the model learns the best fit line:-

- Initially, the model starts with random values of 'm' and 'c' and the target is predicted.
- Naturally, the predicted value is not good and a loss is calculated to measure how bad the prediction was
- The model's parameters i.e. 'm' and 'c' are updated via Gradient Descent to make better predictions next time the data is passed into model
- By passing the data through the model many times and updating the values of 'm' and 'c' each time to make better predictions, we are able to find the optimal values of the parameters which is able to explain maximum possible variance by linear regression.

So, that was the basic idea of Linear Regression.

Q2. Explain the Anscombe's Quartet in detail.



Anscombe's Quartet is a set of four datasets, where each dataset has same summary statistics but very different plots.

All the four datasets have same mean, variance, regression line, R2 score, correlation between x and y but the distribution of the data is totally different.

It depicts the importance of:-

1. Visualizing the dataset
2. Treating outlier values

All the four graphs have same best fit line which tells us few things:-

1. Simple Linear Regression is not always the best choice for modelling
2. Presence of outliers distorts the linear relationship between input and target variables

Q3. What is Pearson's R ?

Before, Pearson's R, let's me explain what covariance is. Covariance is joint variability of two variables. A positive covariance means that both variables grow together and a negative covariance means when one grows, the other falls in value. The magnitude of covariance is not easy to interpret because it depends on the values of the two variables.

It is given by,

$$cov(X, Y) = E[(X - E[X]) * (Y - E[Y])]$$

where, E is expected value

Correlation or Pearson's R: It is a measure of linear correlation between two variables. It can have values between -1 to 1. It is basically a normalized version of covariance so that the value of covariance also makes sense, along with the sign.

A correlation of 1 means that input and target variable are perfectly positively correlated.

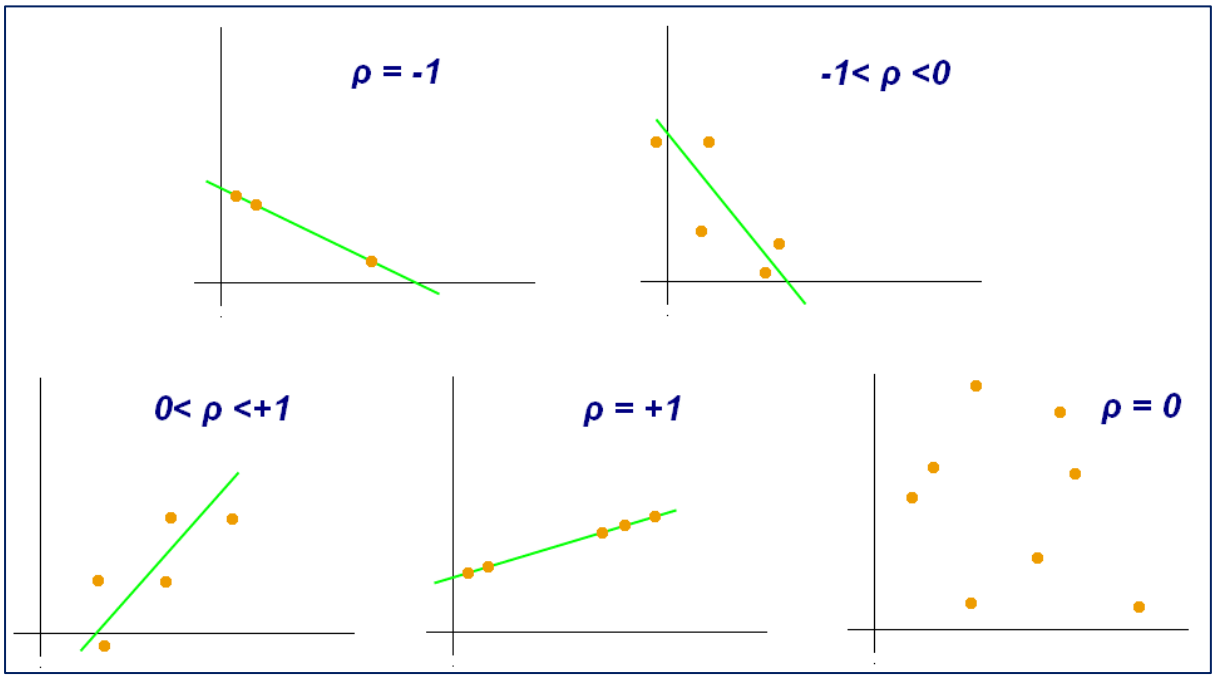


Fig: Pearson's R value for different datasets

It is given by,

$$corr(X, Y) = \frac{cov(X, Y)}{\sigma(X) * \sigma(Y)}$$

where, σ is Standard Deviation

R Squared = Square of R for Simple Linear Regression and Square of coefficient of multiple correlation which is correlation between predicted and actual values of a variable.

It is this R Squared value that we use to determine how much variance of the dependent variable is explained using the independent variables.

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling is the process of bringing all the variables in a dataset in more or less the same range of values. There are two types of scaling, which are:-

1. Min Max Scaling (Normalization) – values between 0 and 1

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

2. Z – score Scaling (Standardization) – values centered around mean with unit std. dev.

$$x_{scaled} = \frac{x - mean}{sd}$$

Advantages of Scaling:-

- Having all the column in similar range makes the process of gradient descent faster
- Interpretability of the model increases because the absolute value of coefficients will be comparable
- The model does not put more weightage to variables having large values

Difference between Normalization and Standardization:-

- In normalization, the values are compressed to take values between 0 and 1. The minimum value will take zero value and the maximum value will be 1 and the intermediate values will take decimal values between zero and 1.
- In standardization, we subtract the mean which tells us whether it is below the average or above the average value and then we divide it with standard deviation which finally tells us how many standard deviations away it is from mean.

Both normalization and standardization have their own advantages and disadvantages. Former takes care of outliers well, latter does not destroy the inherent structure of the data. Which one to use in our model totally depends on the data and domain knowledge.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

VIF stands for variance inflation factor which is given by,

$$VIF = \frac{1}{1 - R^2}$$

VIF becomes infinite when 1 - R^2 = 0 or R^2 = 1 which means that the variable having VIF = 1 can be perfectly explained using other independent variables.

When we have this kind of a situation, we should drop the variable having infinite VIF and then see how the values of VIF changes for the rest of the variables.

Q6. 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 plot the quantiles of a sample distribution against quantiles of a theoretical distribution. A quantile determines how many values are above or below a certail limit.

If the sample and theoretical distribution is same, the plot is a straight line otherwise there is some pattern. To check the distribution of our sample, we test it against different theoretical distribution eg. Normal, Uniform, Exponential, etc. If, with anyone of these theoretical distribution, our sample distribution follows a straight line, then our sample also follows the same distribution.

Using Q-Q plots, we can determine-

- If two populations are of the same distribution
- If residuals follow a normal distribution
- Skewness of distribution

A Q-Q plot for normal theoretical and sample normal distribution looks like this:-