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

- Rentals are increasing year on year.
- If weathersit is Mist and lightsnow are negatively impacting dependent variable
- seasons summer and winter have positive impact on bike rentals. With season fall have higher numbers
- months from june to september have higher bike rentals
- Good weathersit have higher bike rentals
- holidays have lesser bike rentals
- working day have higher bike rentals marginally

Season, Mnth, weathersit, yr have high correlation with target variable. Season/Mnth have multicollinearity.

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

We create dummy variables using one hot encoding.

Example below

Here, we will be able to represent the Gender which have 3 categories using 2 columns as one of them can be derived from other two

With drop_first = False, we will have 3 columns which have multicollinearity

Assume y = B0 + B1XMale + B2XFemale + B3XTransgender

Here we must deal with 3 regression coefficients, also we can get B3 with B1 and B2 otherwise it will result in multicollinearity among these dummy variables

If we have n categories, we will use n-1 dummy variable. If drop_first = False, will result in multicollinearity

Male, Female, Transgender

Gender	One hot encoding		
Male	1	0	0
Female	0	1	0
Transgender	0	0	1

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

'temp' variable has the highest correlation with the target variable 'cnt'

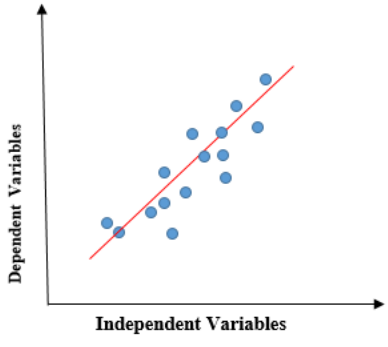
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

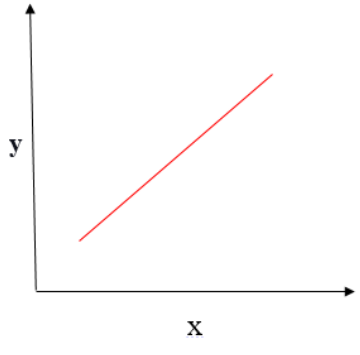
- Linear relationship is observed between the target variables 'cnt' and other independent variables 'temp','yr', 'windspeed' etc.,
- Error terms(y train – y_train_pred) is normally distributed and is shown in the analysis
- Error terms is random and no visible pattern is seen
- Error terms have constant variance, it does not have a spread systematically increasing or decreasing

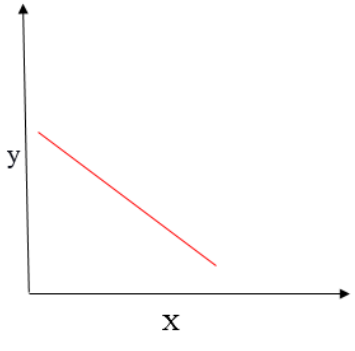
5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
'temp', 'yr', 'windspeed' are top3 features contributing significantly towards explaining the demand of the shared bikes. Model has r2 around 73% with these 3 variables alone

1.Explain the linear regression algorithm in detail.
Linear regression is a statistical regression method which shows the continuous variables relationship
It shows the linear relationship between the target variable also called as dependent variable and independent variable
If there is only feature or 1 independent variable showing the linear relationship with target variable It is called as simple linear regression. Also, if there are more than 1 independent variable showing linear relationship, then it is called multiple linear regression

Below example shows simple regression model equation, that is as independent variable is increasing, dependent variable keeps increasing. We try plotting the line which is best fit i.e., it has minimal error term (|y_pred – y_actual|).
Though we can have multiple lines, we can have one best fit line showing the linear relationship between the input and output
Y = B0 + B1X + epsilon
Here B0 intercept of the line
B1 = regression coefficient
X- independent variable
Y- dependent variable
We can interpret it as for unit increase in the independent variable, dependent variable is increased by one unit keeping rest as it is
Assume X as experience and Y as salary which usually have linear relationship. As X increases, Y increases
There are two type of linear relationship
Positive- B1 is positive
Negative – B1 is negative
Fig below shows the positive and next fig shows negative







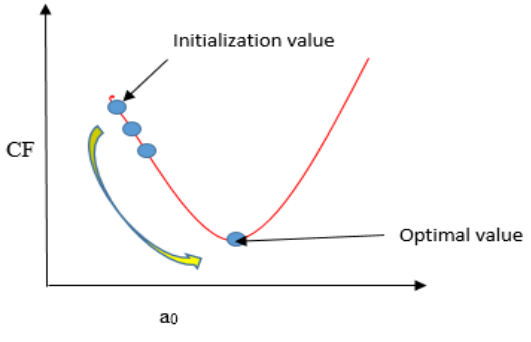
Below cost function is used for Linear regression which is mean square error. Which is average of error square of predicted and actual value. Here yi is predicted and other term is actual value

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

We try reducing the mse to find the coefficients. These can be found out using various techniques, one of them popularly used is Gradient descent

Here optimal value represents the minima which represents the minimum cost function that can be obtained varying the model parameters

The rate at which we approach the optimal value is called learning rate and it decides the algorithm time complexity. If the learning rate is low, algorithm approaches the minima in a longer time and if it is high, it overshoots the minima sometimes



$$J = \frac{1}{n}\sum_{i=1}^{n}(a_0 + a_1 \cdot x_i - y_i)^2$$

$$\frac{\partial J}{\partial a_0} = \frac{2}{n}\sum_{i=1}^{n}(a_0 + a_1 \cdot x_i - y_i)$$

$$\frac{\partial J}{\partial a_1} = \frac{2}{n}\sum_{i=1}^{n}(a_0 + a_1 \cdot x_i - y_i) \cdot x_i$$

$$\frac{\partial J}{\partial a_0} = \frac{2}{n}\sum_{i=1}^{n}(pred_i - y_i)$$

$$\frac{\partial J}{\partial a_1} = \frac{2}{n}\sum_{i=1}^{n}(pred_i - y_i) \cdot x_i$$

$$a_0 = a_0 - \alpha \cdot \frac{2}{n}\sum_{i=1}^{n}(pred_i - y_i)$$

$$a_1 = a_1 - \alpha \cdot \frac{2}{n}\sum_{i=1}^{n}(pred_i - y_i) \cdot x_i$$

The above figure shows the calculation of B0 and B1 (a0 and a1).

Multiple linear regression

Y = B0 + B1X1 + B2X2 + B3X3 ….so on + epsilon

B1, B2, B3 are regression coefficients

X1, X2, X3 are independent variables

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

Anscombe's Quartet is a group of 4 data sets which have identical descriptive statistics with difference in the data sets. If a regression model is built on this, it can fool it.

As you can see that for all the 4 data sets mean, SD and r are same but when we plot them, the scenario will become much clearer. This Anscombe's Quartet shows the importance of data visualization before applying algorithm on the data set.

This data set is created by famous scientist Francis Anscombe. It is important to plot the data to check for outliers or if relationship is linear or not. Linear regression model is good for variables with linear relationship.

Anscombe's Data										
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

Four plots for (x1,y1) (x2,y2) (x3,y3), (x4,y4) are shown below which have same regression coefficients But the scatter plots of 4 data sets are different

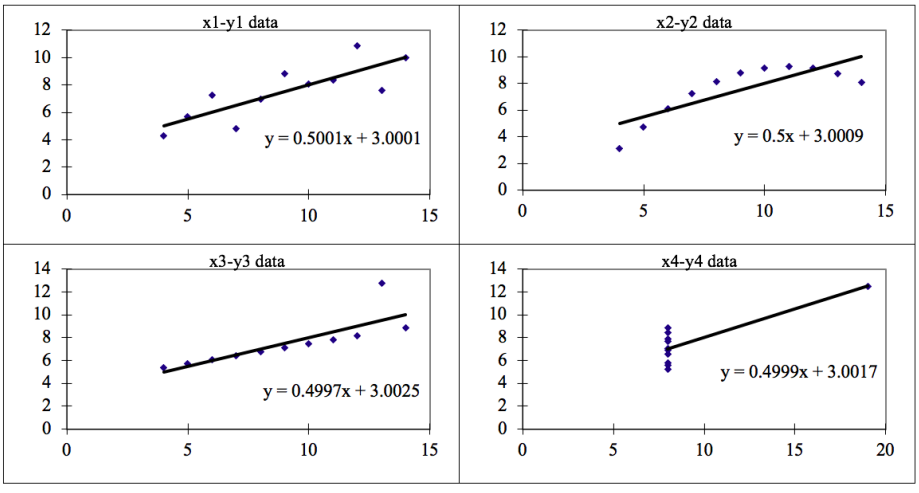
1st plot: Linear relationship exists between x1,y1 and model fits

2nd plot: The scatter plot shows non-linear relationship and regression doesn't fit well

3rd plot: Outliers are present in the data. Linear regression model could not handle the data point which is outlier

4th plot: one data point which is outlier produce high correlation coefficient and fools the regression model

We can see that it is important to look at the data visually and analyze type of relationship before implementing any ML algorithm on the dataset. Statistical properties are insufficient



3. What is Pearson's R?

Pearson correlation coefficient r shows the correlation between the two variables that is it is used to measure strength between variables. This is commonly used In Linear regression. It lies between -1 and 1. Where 1 indicates strong +ve correlation and -1 indicates strong -ve correlation between variables.

And 0 means no correlation.

Suppose we have 2 variables A and B.

+ve correlation indicate that increase in A also shows increase in B

-ve correlation indicate that increase in A show decrease in B.

No Correlation mean no indication of any pattern

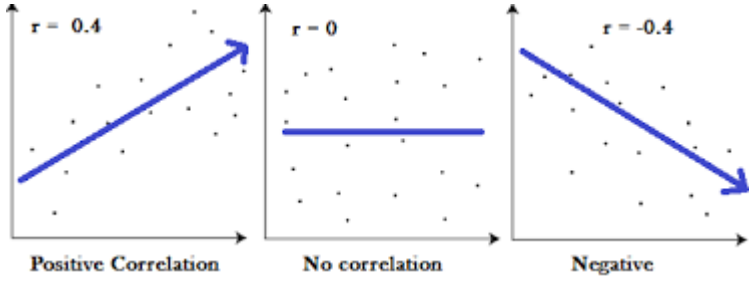
$$R = n(\sum xy) - (\sum x)(\sum y) / \sqrt{[n\sum x^2-(\sum x)^2][n\sum y^2-(\sum y)^2}$$

R pearson coefficient

x,y are variables

n sample set size

Denominators have Standard deviation of x and y



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

Scaling is a data preprocessing step applied on variables of given dataset and standardize the data into fixed range. Some Machine learning algorithms shows bias towards variables which are numerically higher. It also influences the step size of the gradient descent algo because there is a X in the formula of

step size. Scaling helps to ensure the gradient descent steps are same across all the variables which can help in converging faster

There are two types of scaling

1. Min Max Scaling or Normalization **X_new = (X - X_min)/(X_max - X_min)**
2. Standardized scaling **X_new = (X - mean)/Std**

From the above formula **for Normalization**, features are transformed into a same scale. The values will be in the range [0,1]. It can be used when there are no outliers in the given feature. This is useful when the data does not follow any distribution. Minimum/Maximum values are used in the scaling. It is initiated by MinMaxScaler

Standardization on other hand is also another scaling technique which does not limit the feature to a range bound. This can be used when the data follows Gaussian distribution. Mean and std are used for scaling the features. This scaling will make sure that the features have 0 mean and 1 SD. Outliers does not affect the scaling. It is instantiated by StandardScaler.

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

VIF = 1/1- R2

As we can see if R2 is 1, denominator becomes 0 making VIF infinity. R2 = 1 indicates perfect correlation between the features or variables. To sort this out, we can drop one variable which is causing multicollinearity between variables. We accept the VIF below 5 or 10 based on the data set.
In our dataset there is strong trend between atemp and temp

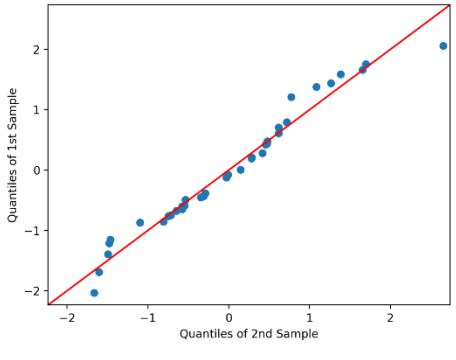
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Q-Q plot also called as quantile-quantile plot which is used to check if both distributions come from same population.
It is as plot of quantiles of one data set with the quantiles of other data set.
Quantile: 50% quantile mean that 50% of values in the dataset are below the value

In Q-Q plots, we draw the line 45% y=x and plot the quantiles of the 2 datasets to see if they lie on this line.
Y- quantiles for 1st data set
X- quantiles of 2nd data set

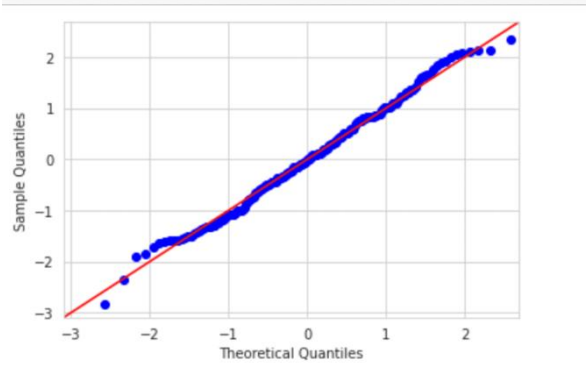
If the points lie closer to a straight line, we can say that both data sets points belong to the same population

We can also know distribution of given data set follow normal or Uniform distribution



Example below:

We Plot the quantiles of given data set against the normal distribution to see if it follows that.
In the below plot, we can see that values are normally distributed. Quantiles for Normal or Uniform distribution are Probability values



Uses of quantile-quantile plot.

- Check if the data of 2 samples are of same population set
- Skewness of data (left or right skewed)
- Check if the distributions of 2 samples are of same shape
- have common location and scale

In linear regression, Q-Q plots are used to check if the error terms or prediction errors follow normal distribution. i.e., plot training data set and test data set values on X and Y axis and check if the values are very much near to the y=x axis i.e., it is useful in visualization, checking if training and test set belong to the population with same distribution