Shantanu Chitrak

• 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?

❖ I have plotted categorical variables with target variable(Dependent variable) on boxplot and infer about their effect on the dependent variable

- Clear weather demands for high rental bikes.
- Demands for rental bikes had been grown for next year.
- Season 3 has highest demand for rental bikes.
- Demand for rental bikes grows continuously till September month, September month has highest demands for rental bikes and after September demand is decreasing.
- Working day, weekday and holiday is not affecting target variable or not giving any clear information of demands on rental bikes.

2) Why is it important to use drop first=True during dummy variable creation?

❖ drop_first=True is important during variable creation because it helps to reducing extra column creation. Hence, it reduce the correlations created among the dummy variables. If we don not used drop_first=True then it become redundant with dataset.

.

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

❖ The ‘temp’ and ‘atemp’ variables has highly correlated with target variable(cnt).

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

❖ a) Linear relationship between X and Y.
b) Errors terms are normally distributed with mean 0.
c) Error terms are independent of each other.
d) Error terms have constant variance.
e) Ensuring overfitting by looking at R2 and Adjusted-R2 value.

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

❖ Features 'atemp', 'year' and month 'September' is highly correlated with target variable, so these are top features in building model.

General Subjective Questions

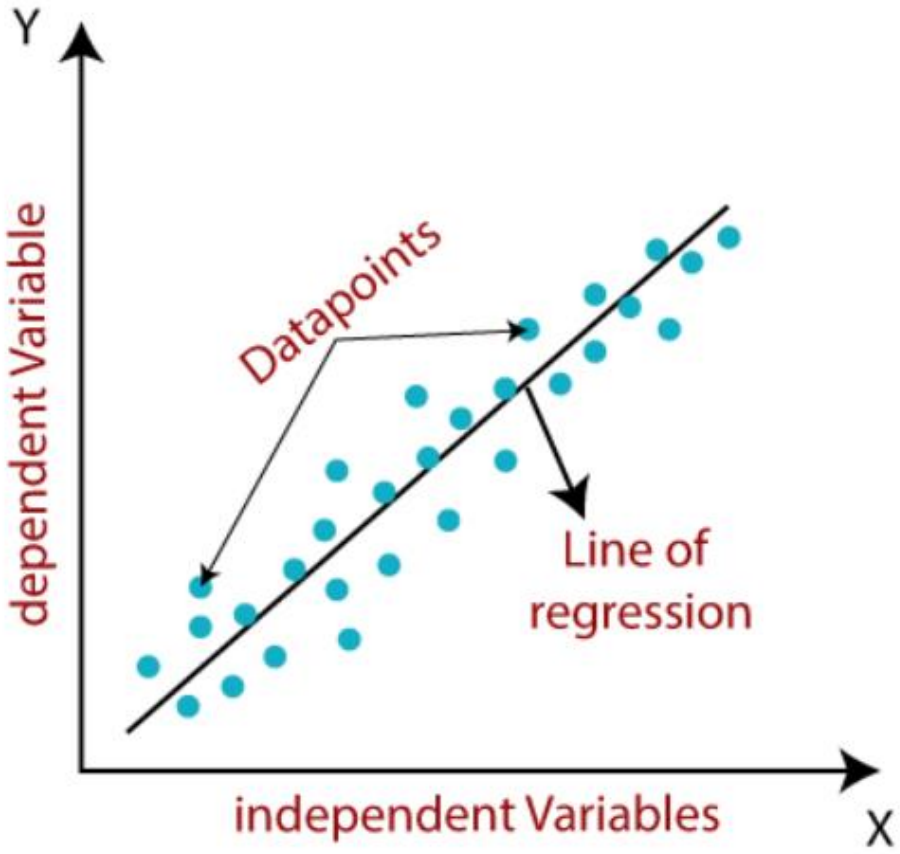
1) Explain the linear regression algorithm in detail.

❖ Linear regression algorithm is a method to analyzes the linear relationship between a dependent variable(y) and one or more independent variable (x). The linear regression model provides a sloped straight line representing the relationship between the variables.

Mathematically the relationship can be represented with the help of following equation –

- $Y = mX + c$, Where m = slope or coefficient of the regression line

  c = constant, known as Y-intercept. If X=0 then Y would be equal to c



Types of Linear Regression

Linear regression is of the following two types −

- Simple Linear Regression
- Multiple Linear Regression

Simple Linear Regression (SLR)

It is the most basic version of linear regression which predicts a response using a single feature. The assumption in SLR is that the two variables are linearly related.

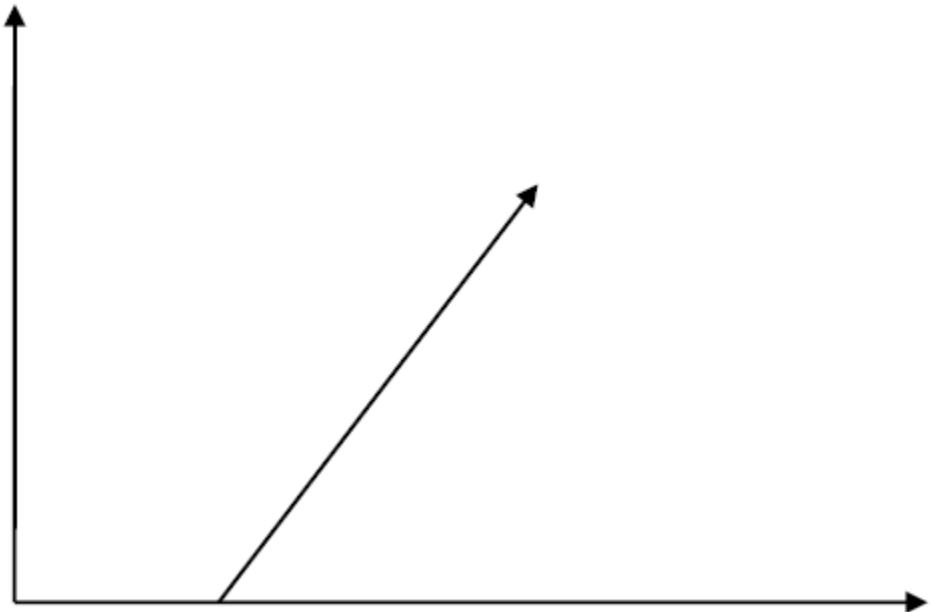
Mathematically we can explain it as follows : $Y = mX + c$

Multiple Linear Regression (MLR)

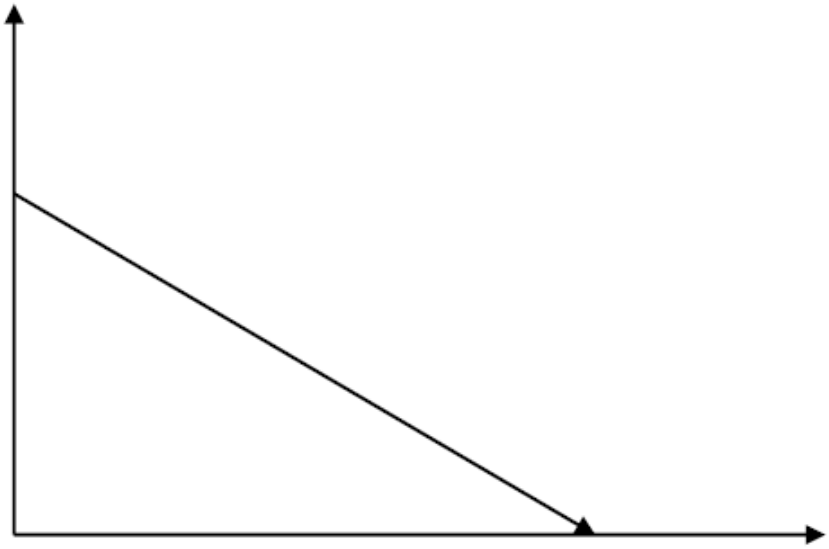
It is the extension of simple linear regression that predicts a response using two or more features. Mathematically we can explain it as follows –

$$Y = m1X1 + m2X2 + m3X3 + \ldots\ldots\ldots\ldots\ldots + c$$

A linear relationship will be called positive if both independent and dependent variable increases. A linear relationship will be called positive if independent increases and dependent variable decreases.



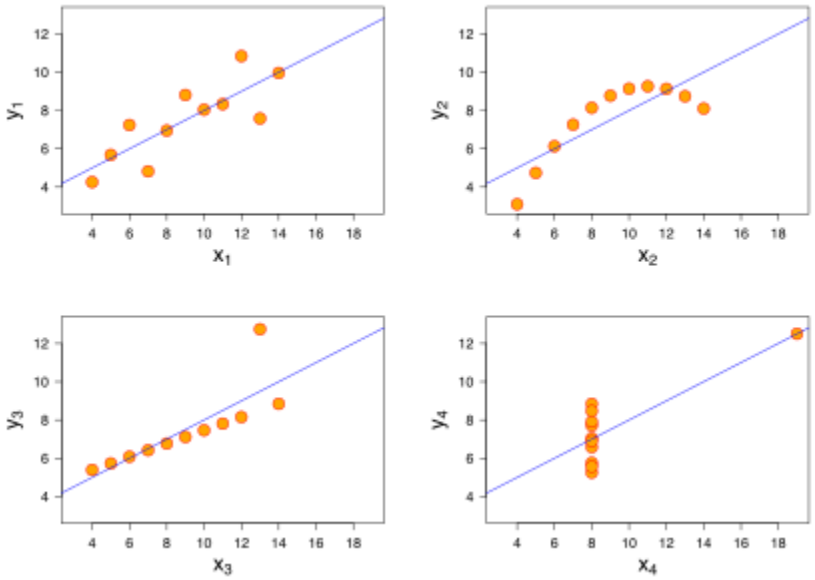
Positive Linear Relationship



Negative Linear Relationship

2) Explain the Anscombe's quartet in detail.

- ❖ Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points.



1st Figure : The relation between X and Y looks linear and we can use linear regression model to find the best fit line.

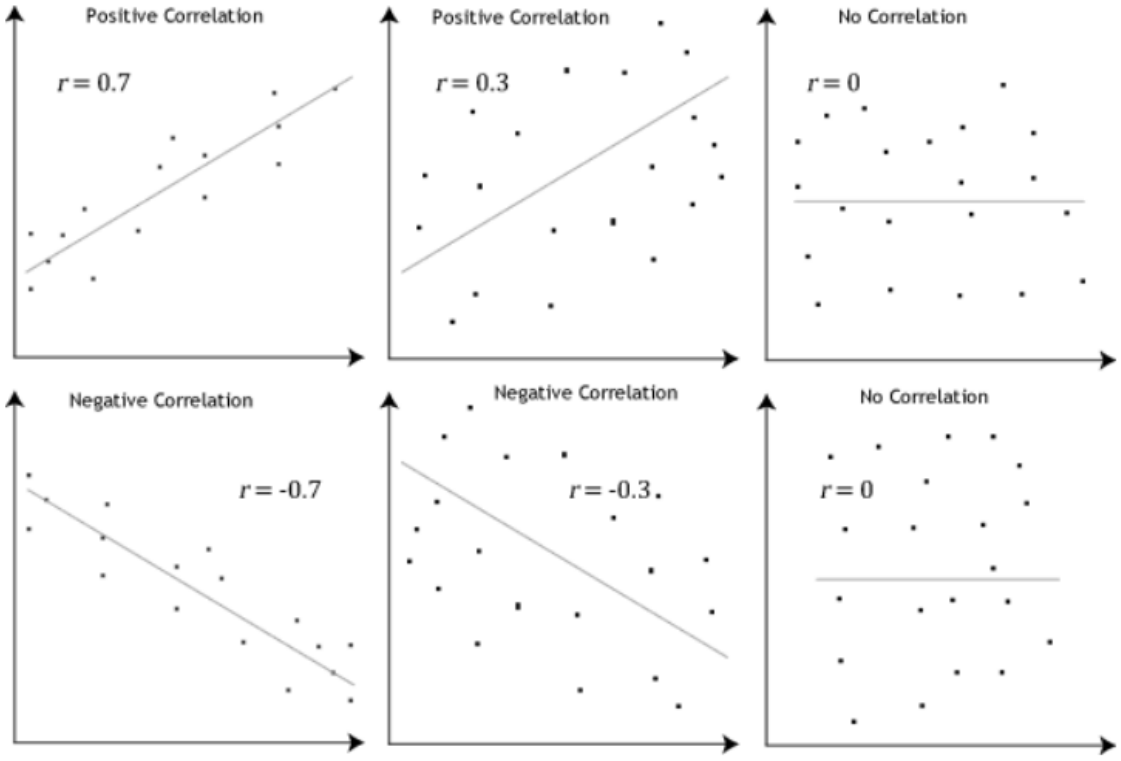
2nd Figure : From the plot we can observe that the relationship between X and Y is not linear hence linear regression is not a good option to use and we need to choose another model.

3rd Figure : The relation between X and Y looks linear and we can use linear regression model to find the best fit line but outliers throws the slope and intercept. We should do the EDA to find the reason of the outlier .

4th Figure : This dataset might also have the linear relationship between X and Y but from the plot we can conclude that we should try to acquire more data for intermediate X-values to make sure it really does the good job.

3) What is Pearson's R ?

- Correlation is a process of measuring the relationship between sets of data and how well they are related. It is the ratio between the covariance of two variables and the product of their standard deviation. It is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1.



No Correlation

R=0 indicates that there is no association between the two variables.

Positive Correlation

R value greater than 0 indicates a positive association that is, as the value of one variable increases other also increases.

Negative Correlation

R value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases.

Formula:-

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Here, r = Correlation coefficient

xi = Values of the x-variable in the sample

x bar = mean of the value of the x-variable

yi = values of the y-variable in the sample

y bar = mean of the value of the y-variable

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

- Scaling

- It is a process to bring the coefficient of the variables in a certain range by bringing all the variable in a comparable scale (interpretability)

Why is scaling performed?

- It will help us to better understand coefficient relation between then if not done then the variance in the coefficient of feature will be large.(ease of interpretation)

Faster Convergence for Gradient Descent Method.

- If we rescale the variable in the range of 0-1 then the optimization happening behind the scene becomes much faster i.e minimisation routine becomes much faster.
- When we train a network using Gradient Descent function then it becomes very fast.

MinMaxScaling is the most commonly form of scaling used in Linear Regression as it bring the variable in the range of -1 to 1 which is easy to understand.

Normalization/Min-Max Scaling:

- *It brings all of the data in the range of 0 and 1.*

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:

- *Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).*

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

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

- If there is perfect correlation between variables, then VIF = infinity. To avoid this situation we need to drop one of the variables from our dataset which leads to this multicollinearity.
- VIF value of infinity means that there is a perfect correlation which results as value 1 for R-Squared.
- VIF= 1/(1-R2)
- If R2=1 then VIF will be infinity so in this case we should drop that variable as it's information is already shown by the other set of 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 or quantile-quantile plot are plots of two quantiles against each other. Ex, the median is a quantile where 50% of the data fall below that point and 50% lie above it.
-Main purpose of Q-Q plot is to find the whether two set of data is coming from same distribution or not.
- It helps in linear regression ,when we received training and test data set separately, then we use Q-Q plot to confirm that both the data set are from populations with same distributions.
- If all points of quantile lies on or close to straight line at an angle of 45 degree from x –axis then it is called as similar distribution.
- Q-Q plot can detect outliers, shifts in scale, location, symmetry etc.