1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
   - The demand for bikes is least in spring.
   - The demand bike increased in the year 2019 when compared with year 2018.
   - Bike Rentals are more in partly cloudy weather

2. Why is it important to use drop_first=True during dummy variable creation ?
   - If you don't drop the first column then your dummy variables will be correlated. This may affect some models adversely and the effect is stronger when the cardinality is smaller

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
   - FeelsLike('atemp') (and/or) 'temp' has the highest co-relation.

4. How did you validate the assumptions of Linear Regression after building the model on the training set ?
   - By plotting the predicted values vs actual**.**
     - **There is a** *linear relationship* **between FeelsLike (atemp) and count.**
   - Error terms are normally distributed with mean zero(not X, Y):

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
   - Temperature plays significant role in bike demand.
   - Climate also plays important role Spring season the demand is least and 'Fall' has maximum.
   - Whether also has impact in the demand, people prefer to use the rental when the sky is clear.

Assignment-based Subjective Question

1. Explain the linear regression algorithm in detail

Linear regression is simple statistical regression method used for predictive analysis and shows the relationship between the continuous variables. Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), consequently called linear regression. If there is a single input variable (x), such linear regression is called simple linear regression. And if there is more than one input variable, such linear regression is called multiple linear regression. The linear regression model gives a sloped straight line describing the relationship within the variables.

The above graph presents the linear relationship between the dependent variable and independent variables. When the value of x (independent variable) increases, the value of y (dependent variable) is likewise increasing. The red line is referred to as the best fit straight line. Based on the given data points, we try to plot a line that models the points the best.

To calculate best-fit line linear regression uses a traditional slope-intercept form.

$$Y= mx+c$$

y= Dependent Variable.
x= Independent Variable.
c= intercept of the line.
m = Linear regression coefficient.

Need of a Linear regression

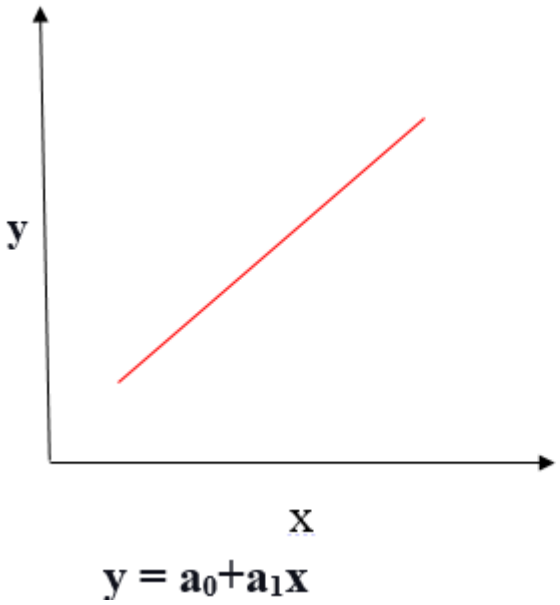
As mentioned above, Linear regression estimates the relationship between a dependent variable and an independent variable. Let's understand this with an easy example:

Let's say we want to estimate the salary of an employee based on year of experience. You have the recent company data, which indicates that the relationship between experience and salary. Here year of experience is an independent variable, and the salary of an employee is a dependent variable, as the salary of an employee is dependent on the experience of an employee. Using this insight, we can predict the future salary of the employee based on current & past information.

A regression line can be a Positive Linear Relationship or a Negative Linear Relationship.

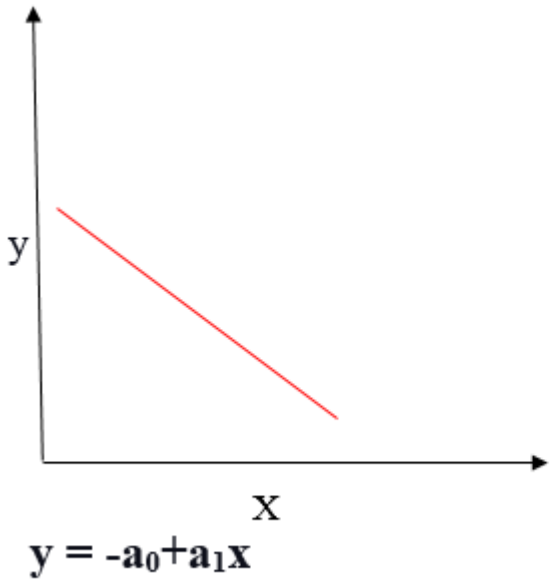
Positive Linear Relationship

If the dependent variable expands on the Y-axis and the independent variable progress on X-axis, then such a relationship is termed a Positive linear relationship.



Negative Linear Relationship

If the dependent variable decreases on the Y-axis and the independent variable increases on the X-axis, such a relationship is called a negative linear relationship.



The goal of the linear regression algorithm is to get the best values for a0 and a1 to find the best fit line. The best fit line should have the least error means the error between predicted values and actual values should be minimized.

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. 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.

3. What is Pearson's R?

   It is a statistic that measures the linear correlation between two variables. It is the covariance of two variables, divided by the product of their standard deviations. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.

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

   It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

   Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

   Normalization/Min-Max Scaling: It brings all of the data in the range of 0 and 1.

   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).

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. This 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. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multi collinearity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
   The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.
   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. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.