Assignment-based Subjective Questions

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

Ans- a. The demand of bikes for rent are high in fall season and low in spring season.

b. The demand of bikes for rent is very low in month of January compared to other months.

c. The demand is high in clear weather and low in light_snow_rain weather

Maybe company would not be operating properly where the demand is low.i.e. in spring season, month of January etc.

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

Ans:- drop_first = True is important to use because it helps in reducing the extra column created during dummy variable creation and reduces the correlations created among dummy variables which helps in reducing the time and memory used.

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

Ans:- Temperature variable has the highest correlation with the target variable.

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

Ans:- 1. By checking the VIF values of the fetures for no multicolinearity.

2. By checking normality of residual by plotting a histogram.

3. By checking the homoscedasticity (or constant variance).

4. By checking out No auto correlation of residual.

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

Ans:- - 1. Light_Snow_Rain: As the coefficient of 'Light_Snow_Rain' is the highest and since it is negative, change in 'Light_Snow_Rain' feature will inversely affect the demand for bikes.

2. Year: As the coefficient of 'Year' is also high and is positive, the change in Year feature will directly affect the demand for bikes.
3. Season_Spring has negative effect with 3rd highest coefficient.

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

Ans:- Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.

In other words, Linear regression is a type of regression analysis used for predicting the unknown value of the dependent variable based upon the known value of the independent variable**.**

Linear regression establishes the relationship between two variables by fitting a linear equation to observed data. For example, If we increase the marketing budget it will eventually increase product sales. This is a kind of positive relationship between the dependent and independent variables.

Simple Linear regression:

It is a type of regression analysis wherein the values of a dependent and independent variable vary linearly that is why this model is called a linear regression model. For example, the price of the house increases with an increase in its area or size.

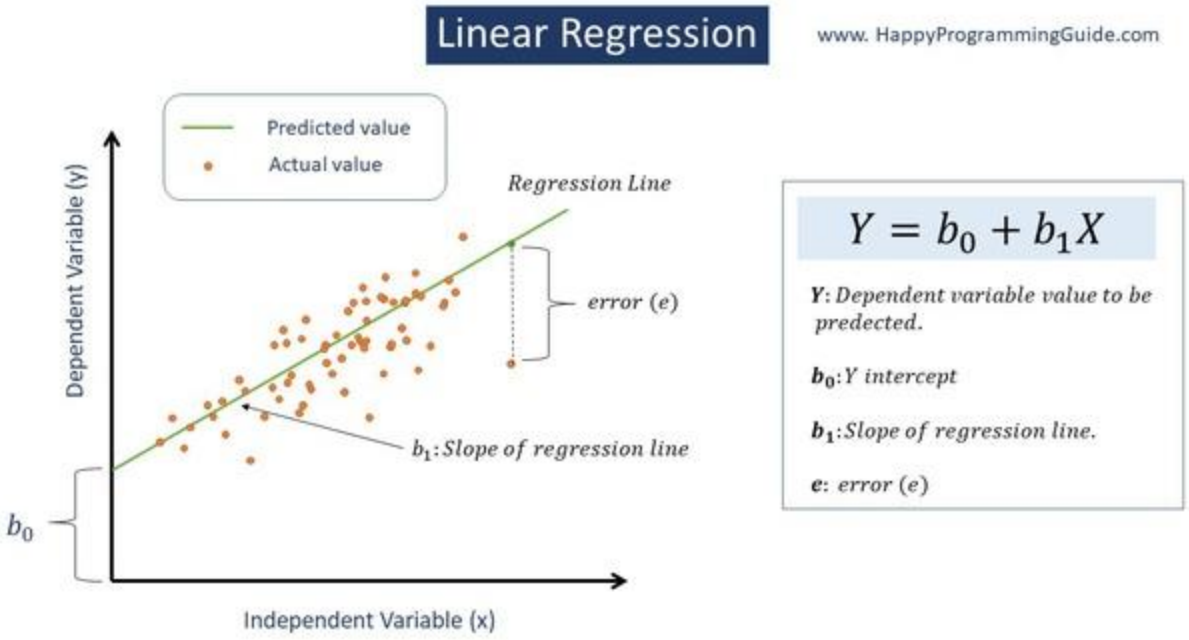
If we try to plot the graph of dependent and independent variables it is mostly a straight line and not the curve.

In simple linear regression, the value of the dependent variable is calculated based upon the single independent variable.

The equation for the simple linear regression is as shown in the image below.

- Here "Y" is the dependent variable plotted on Y-axis whose value we are trying to predict. It is a continuous quantity i.e it can be any integer value.
- "X" is the independent variable whose value we already know.
- "B0" is a "Y" intercept.
- "B1" is the slope of the regression line.
- Greenline shown here is the regression line.

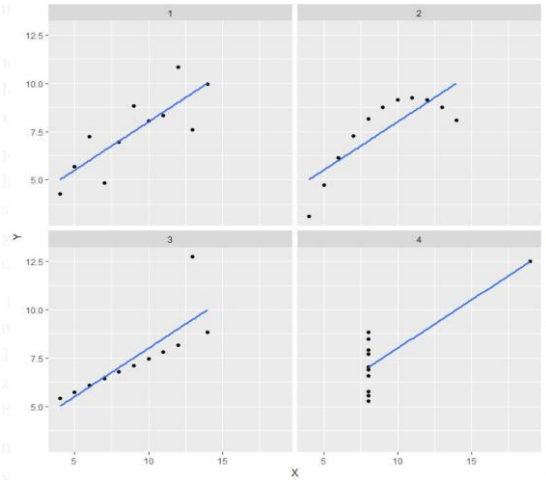
- The distance between the predicted value and the actual value is known as the error in prediction. So our aim while solving the linear regression problem is to fit the regression line in such a way that the error should be least.



Q.2 - Explain the Anscombe's quartet in detail.

Ans:- Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

For Eg:-



- 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.

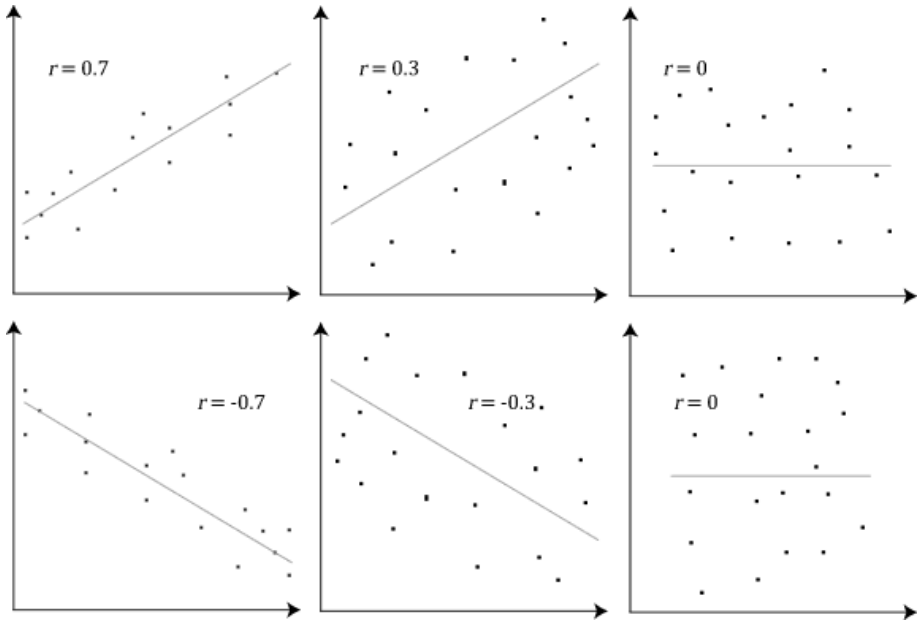
Application:

The quartet is still often 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.

Q-3. What is Pearson's R?

Ans:- Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations.

The stronger the association of the two variables, the closer the Pearson correlation coefficient, *r*, will be to either +1 or -1 depending on whether the relationship is positive or negative, respectively. Achieving a value of +1 or -1 means that all your data points are included on the line of best fit – there are no data points that show any variation away from this line. Values for *r* between +1 and -1 (for example, *r* = 0.8 or -0.4) indicate that there is variation around the line of best fit. The closer the value of *r* to 0 the greater the variation around the line of best fit. Different relationships and their correlation coefficients are shown in the diagram below:



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

Ans:- Scaling is the procedure of measuring and assigning the objects to the numbers according to the specified rules. In other words, the process of locating the measured objects on the continuum, a continuous sequence of numbers to which the objects are assigned is called as scaling.

Scaling is performed to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step.

The most common techniques of feature scaling are Normalization and Standardization. Normalization is used when we want to bound our values between two numbers, typically, between [0,1] or [-1,1].
While Standardization transforms the data to have zero mean and a variance of 1, they make our data unitless.

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

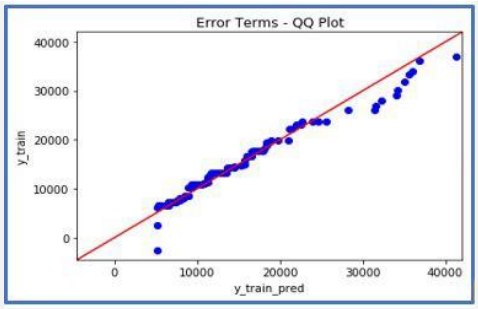
Ans:- Sometime the value of VIF is Infinite due to perfect correlation. i.e it indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

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

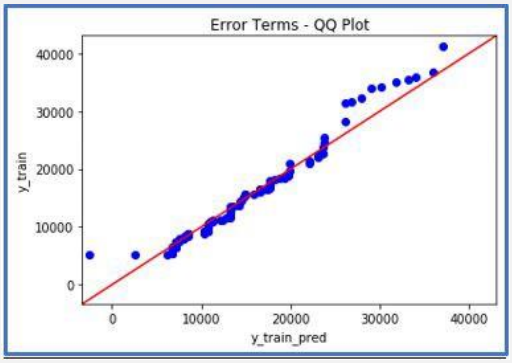
Ans:- Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

Few advantages:
a) It can be used with sample sizes also
b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.
It is used to check following scenarios:
If two data sets —
i. come from populations with a common distribution
ii. have common location and scale
iii. have similar distributional shapes
iv. have similar tail behavior
Interpretation:
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.
Below are the possible interpretations for two data sets.
a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x - axis
b) Y-values < X-values: If y-quantiles are lower than the x-quantiles.



c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.



d) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis