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? (3 marks)**

Answer :-

In the bike sharing dataset, lets consider the effect of the categorical variable 'weathersit' on the target variable 'cnt'. While performing EDA, I visualized the relationship between the categorical variables and the target variable. It was seen that during the weather situation 1 (Clear, few clouds, partly cloudy, a high number of bike rentals were made, with the median being 50,000 approximately. Similarly, certain inferences could be made 'season' and 'yr' as well.

Also, during model building on inclusion of categorical features such as yr,season etc we saw a significant growth in the value of R-squared and adjusted R-squared. This implies that the categorical features were helpful in explaining a greater proportion of variance in the dataset.

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

Answer:-

During dummy value creation (dummy encoding) it is advisable to use drop_first=True, otherwise we will get a redundant feature i.e. dummy variables might be correlated because the first column becomes a reference group during dummy encoding.

drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not furnished and semi_furnished, then It is obvious unfurnished. So we do not need 3rd variable to identify the unfurnished. Example

Hence if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

3. **3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**

Answer:-

The numerical variable 'registered' has the highest correlation with the target variable 'cnt', if we consider all the features. But after data preparation, when we drop registered due to multicollinearity the numerical variable 'atemp' 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? (3 marks)**

Answer:-

We Performed multiple operation on dataset to achieve accuracy, we taste whole dataset analysis it, and then we removed unnecessary variable from data sets. We also perform EDA and hypothesis testing on dataset, we also used F-Staitsics.

F-Statistics is used for testing the overall significance of the Model. The higher the F-Statistics, the more significant the Model is.
Validating the assumption of Linear Regression Model :

Linear Relationship

Homoscedasticity

Absence of Multicollinearity

Independence of residuals

Normality of Errors

All the predictor variables have VIF value less than 5. So we can consider that there is insignificant multicolinearity among the predictor variables.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Answer:-

'''As per the final model, the top 5 predictor variables that influences bike booking are:

Temperature (Temp)

A coefficient value of '0.564438' indicated that a temperature has significant impact on bike rentals

Light Rain & Snow (weathersit =3)

A coefficient value of '-0.307082' indicated that the light snow and rain deters people from renting out bikes

Year (yr)

A coefficient value of '0.230252' indicated that a year wise the rental numbers are increasing

It is recommended to give utmost importance to these three variables while planning to achieve maximum bike rental booking.As high temperature and good weather positively impacts bike rentals, it is recommended that bike availability and promotions to be increased during summer months to further increase bike rentals.

General Subjective Questions

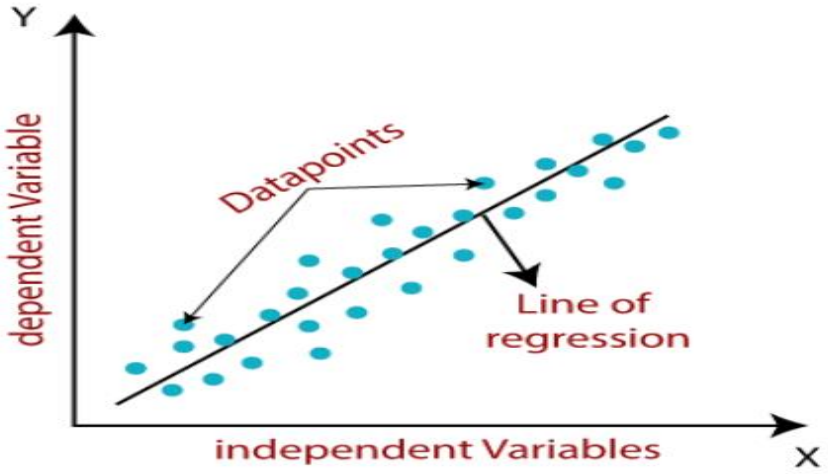
1. Explain the linear regression algorithm in detail. (4 marks)

Answer:-

Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as **sales, salary, age, product price,** etc.

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

The linear regression model provides a sloped straight line representing the relationship between the variables. Consider the below image:



Mathematically, we can represent a linear regression as:

$y= a0+a1x+ \varepsilon$

Here,

Y= Dependent Variable (Target Variable)

X= Independent Variable (predictor Variable)

a0= intercept of the line (Gives an additional degree of freedom)

a1 = Linear regression coefficient (scale factor to each input value).

ε = random error

The values for x and y variables are training datasets for Linear Regression model representation.

Types of Linear Regression

Linear regression can be further divided into two types of the algorithm:

Simple Linear Regression:

If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.

Multiple Linear regression:
If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

Finding the best fit line:
When working with linear regression, our main goal is to find the best fit line that means the error between predicted values and actual values should be minimized. The best fit line will have the least error.
The different values for weights or the coefficient of lines (a0, a1) gives a different line of regression, so we need to calculate the best values for a0 and a1 to find the best fit line, so to calculate this we use cost function.

2. Explain the Anscombe's quartet in detail. (3 marks)

Answer:-

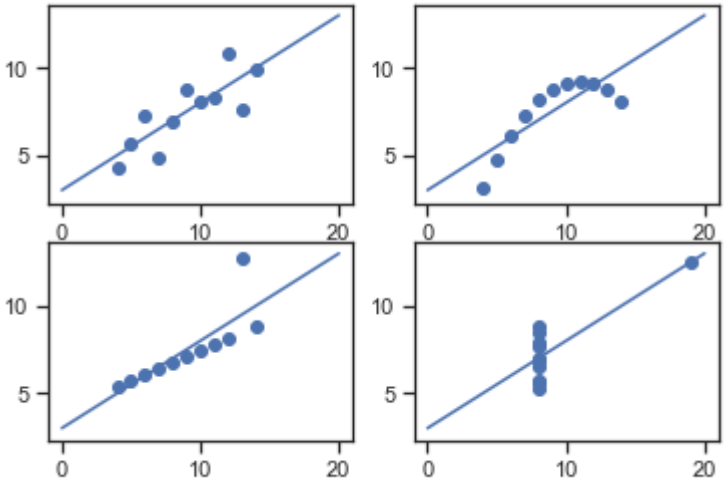
- Anscombe's Quartet is the modal example to demonstrate the importance of data visualization which was developed by the statistician Francis Anscombe in 1973 to signify both the importance of plotting data before analyzing it with statistical properties.
- It comprises of four data-set and each data-set consists of eleven (x,y) points. The basic thing to analyze about these data-sets is that they all share the same descriptive statistics(mean, variance, standard deviation etc) but different graphical representation. Each graph plot shows the different behavior irrespective of statistical analysis.

x1	y1	x2	y2	x3	y3	x4	y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Apply the statistical formula on the above data-set,
Average Value of x = 9
Average Value of y = 7.50
Variance of x = 11

Variance of y =4.12
Correlation Coefficient = 0.816
Linear Regression Equation : y = 0.5 x + 3

However, the statistical analysis of these four data-sets are pretty much similar. But when we plot these four data-sets across the x & y coordinate plane, we get the following results & each pictorial view represent the different behavior.



- Data-set I — consists of a set of (x,y) points that represent a linear relationship with some variance.
- Data-set II — shows a curve shape but doesn't show a linear relationship (might be quadratic?).
- Data-set III — looks like a tight linear relationship between x and y, except for one large outlier.
- Data-set IV — looks like the value of x remains constant, except for one outlier as well.

3. What is Pearson's R? (3 marks)

Answer:-

- When we try to infer something from what we have heard or read, the first step we do is relate a few of the parameters or scenes, etc. with each other and then proceed.
- Correlation means to find out the association between the two variables and Correlation coefficients are used to find out how strong the is relationship between the two variables.
- The most popular correlation coefficient is Pearson's Correlation Coefficient. It is very commonly used in linear regression.
- Pearson's Correlation coefficient is represented as 'r', it measures how strong is the linear association between two continuous variables using the formula:

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

Where,

r = Pearson Correlation Coefficient

x_i = x variable samples

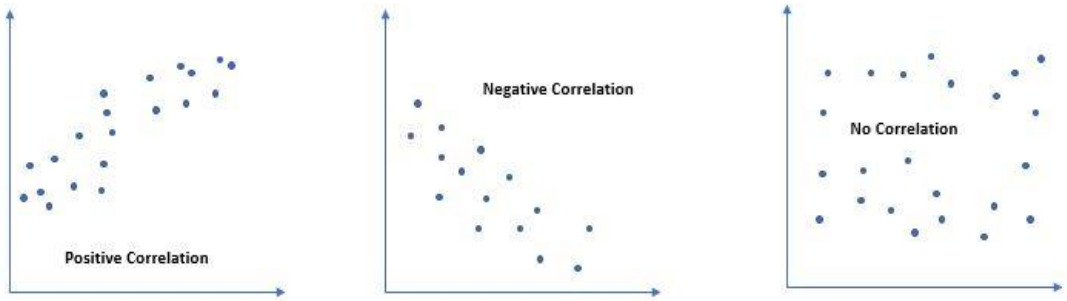
y_i = y variable sample

$\bar{x}$ = mean of values in x variable

$\bar{y}$ =mean of values in y variable

Values of Pearson's Correlation are:

- Value of 'r' ranges from '-1' to '+1'. Value '0' specifies that there is no relation between the two variables.
- A value greater than '0' indicates a positive relationship between two variables where an increase in the value of one variable increases the value of another variable.
- Value less than '0' indicates a negative relationship between two variables where an increase in the value of one decreases the value of another variable.



- Pearson correlation attempts to draw a line of best fit through the spread of two variables.
- Hence, it specifies how far away all these data points are from the line of best fit.
- Value of 'r' equal to near to +1 or -1 that means all the data points are included on or near to the line of best fit respectively.
- Value of 'r' closer to the '0' data points is around the line of best fit.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Answer:-

- 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.
- It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Normalization/Min-Max Scaling:

It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

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

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).

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

sklearn.preprocessing.scale helps to implement standardization in python.

One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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

Answer:-

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 multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables. For example, we would fit the following models to estimate the coefficient of determination R1 and use this value to estimate the VIF:

$$X\_1=C+ \alpha\_2 X\_2+\alpha\_3 X\_3+\cdots$$

$$〖VIF〗\_1=1/(1-R\_1^2)$$

Next, we fit the model between X2 and the other independent variables to estimate the coefficient of determination R2:

$$X\_2=C+ \alpha\_1 X\_1+\alpha\_3 X\_3+\cdots$$

$$〖VIF〗\_2=1/(1-R\_2^2)$$

If all the independent variables are orthogonal to each other, then VIF = 1.0. If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables.

If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. This would mean that that standard error of this coefficient is inflated by a factor of 2 (square root of variance is the standard deviation).

The standard error of the coefficient determines the confidence interval of the model coefficients. If the standard error is large, then the confidence intervals may be large, and the model coefficient may come out to be non-significant due to the presence of multicollinearity.

A general rule of thumb is that if VIF > 10 then there is multicollinearity. Note that this is a rough rule of thumb, in some cases we might choose to live with high VIF values if it does not affect our model results such as when we are fitting a quadratic or cubic model or depending on the sample size a large value of VIF may not necessarily indicate a poor model.

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

Answer:-

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 behaviour

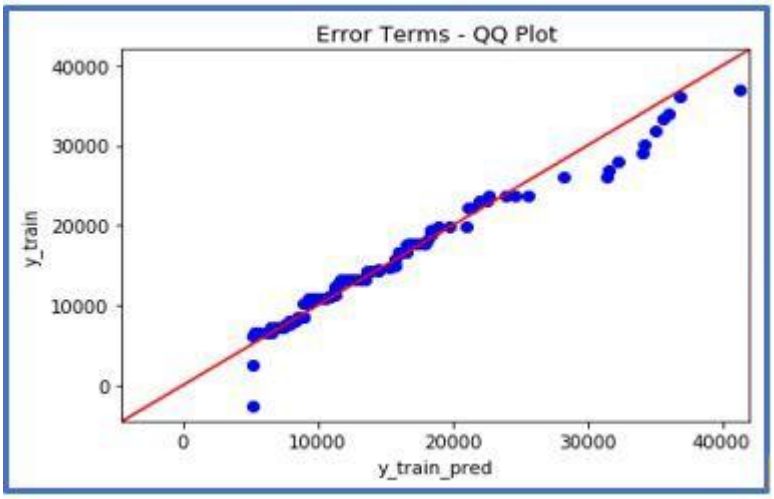
Interpretation:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

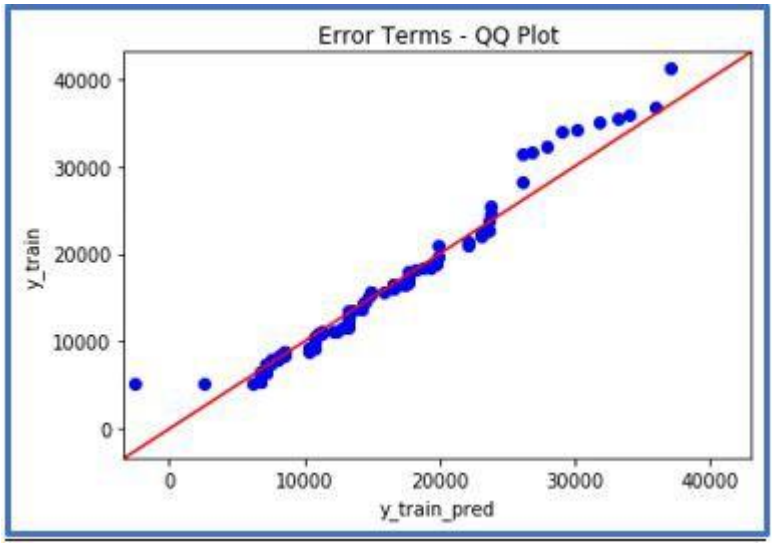
Below are the possible interpretations for two data sets.

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