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?

Ans: data set contains 4 categorical varaiables and they are Season, month, weekends, weathersit and cnt is count of bikes

- from my analysis bike demand is high in fall season and low in spring season
- jun and september bike demand is high and low in jan
- weekends and thu and fri are having similar bike demand
- bike demand is high in clear to partly cloudy days and low in light snow and light rain

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

Ans: 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.

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

Ans: from the plots target variable is directly proportional to year, temp, atemp, casual and registered that means they have positive correlation with target variables but target is sum of registered and casual so we don't consider them in model.so from pair plots **temp and atemp** have highest correlation with target variables

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

Ans: By doing Residual analysis and multicollinearity checking we can validate the model.

1.using dist plot we can see the residual error distribution. If the distribution having mean is at zero means the assumptions are correct

2.if VIF is less than 5 means there no vaiable are related with other or combination of other variables .that means there is no multicollinearity

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

Ans: 1. Year (positively related)

2. Temperature (positively related)

3. Weather sit _ light rain and snow (Negatively related)

General Subjective Questions

1.Explain the linear regression algorithm in detail.

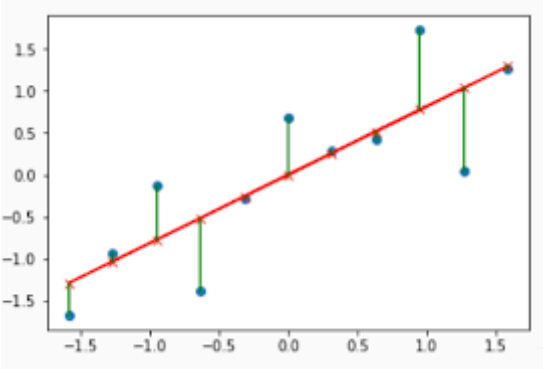
Ans: **Linear Regression** is a machine learning algorithm based on **supervised learning**. It performs a **regression task**. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression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

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression. Mathematically, we can write a linear regression equation as:

Y=mx+c ,where m is slope and c is y intercept

The goal of regression analysis is to create a trend line based on the data. The line is considered best fit if the predicted values and the observed values is approximately same. In simple words, the sum of distance of data points from the line is minimum then it is a best fit line.

The Line is also called the regression line and the errors are also known as residuals which are shown below. It can be visualized by the vertical lines from the data point to the regression line.



error ,in this case, is the sum (mean or standard deviation) of the point from the line chosen.

Model Performance

After the model is built, We need to check the difference between the values predicted and actual data, if it is not much, then it is considered to be a good model. Below is a metric tool we can use to calculate errors in the model.

R — Square (R2) score:

$$R^2 = \frac{TSS - RSS}{TSS}$$

Where

Total Sum of Squares (TSS): The measure of how a data set varies around a mean. The TSS tells us the variation in the dependent variable.

TSS = Σ (Y — Mean[Y])2

Residual Sum of Squares (RSS): sum of the squared differences between the actual Y and the predicted Y. The RSS tells us how much variation of the dependent variable is not explained by our model.

RSS = Σ (Y — f[Y])2

(TSS — RSS) measures the amount of variability in the response that is explained by performing the regression.

R2 score can be used to check all regression model's performance.

Steps in building linear regression model

1.data reading and understanding(data cleaning and understanding the data tpyes)

2.performing EDA on the Data(Uni variate and Bivariate analysis)

3.Data preparation(creating dummy variables for categorical variables)

4.splitting Data into train and test data(70% of data train data and 30% as test data)

5. Build a linear model

6.Residual analysis for checking error distribution

7.making predictions and evaluating the final model using r2 score.

1. 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

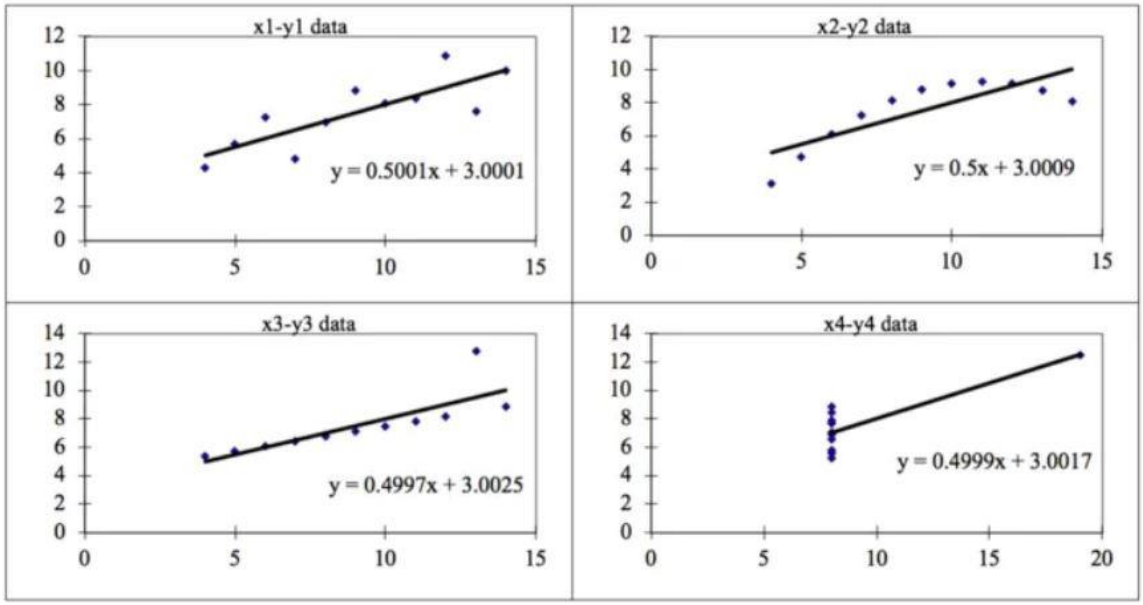
fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc.

These four plots can be defined as follows:

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

When these models are plotted on a scatter plot, all datasets generates a different kind of plot that is not interpretable by any regression algorithm which is fooled by these peculiarities and can be seen as follows:



These can be described as

Dataset 1: this **fits** the linear regression model pretty well.

Dataset 2: this **could not fit** linear regression model on the data quite well as the data is non-linear.

Dataset 3: shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model.

Dataset 4: shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model.

Hence all the important features are visualised before implementing any model.

2. What is Pearson's R?

Ans: Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

The Pearson's correlation coefficient varies between -1 and +1 where:

r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)

r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)

r = 0 means there is no linear association

r > 0 < 5 means there is a weak association

r > 5 < 8 means there is a moderate association

r > 8 means there is a strong association

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

Ans: 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 typically means rescales the values into a range of [0,1] .for this we use MinMaxscaler from skllear library.Here we lose the data if outliers are present,and formula is

$$x = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).for this scaling we use standard scaler from skllearn library. here we don't lose the data and the formula is

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

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

Ans: An infinite value of VIF for a given independent variable indicates that **it can be perfectly predicted by other variables in the model**.

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.

Most research papers consider a VIF (Variance Inflation Factor) > 10 as an indicator of multicollinearity, but some choose a more conservative threshold of 5 or even 2.5.

5. 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.

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

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.