LINEAR REGRESISON SUBJECTIVE QUESTIONS

PART-I: 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:

Let us start with the results of the analysis of the target variable (“cnt”) with each categorical variable

a. Target variable vs Season variable:
    i. The median value and IQR of Fall season is the highest among all the seasons.
    ii. The median value and IQR of Spring season is the least among all the seasons.
b. Target variable vs month:
    i. January has the lowest median count, whereas July has the highest median count.
    ii. October has the highest IQR value, and October has the lowest starting count value.
c. Target variable vs weekday:
    i. Wednesday has the highest IQR value.
    ii. The median values are almost the same every day, but among them Thursday has the highest.
d. Target variable vs weather:
    i. "Light Snow Rain" weather situation seems to have the lowest median count value, where as "Clear" weather situation has the highest median count value. i.e., the people prefer to take or rent a bike on a clear day as compared to a snowy or a cloudy day.
    ii. The IQR of clear weather is the highest.
e. Target variable vs year:
    i. From the year 2018 to the year 2019, the median value of count has increased.
    ii. The IQR of year 2019 is far greater than the IQR of 2018. This increase is a good sign for the business because people are showing an interest in renting the bicycles.
f. Target variable vs holiday:
    i. As expected the median count value on a holiday is less than the median count on the other days.
g. Target variable vs working day:
    i. The median value remains almost the same whether it is a working day or not.

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

Ans:

Generally during the dummy variable creation, we create a dummy variable for each level inside a categorical variable.

So, if there are “n” levels inside a categorical variable, we initially create “n” dummy variables. But we will be able to identify each of the level of a given categorical variable with only “n-1” dummy variables. For example, if we have 3 levels inside a category say Low, Medium, and High. If a value is neither Low nor Medium, then it is obviously high. So, we do not need the third variable to identify the High value. In such cases, the usage of ***drop_first = “True”*** comes into play. It ***(drop_first = “True”)*** is important to use because it helps in reducing an extra column generated during the creation of dummy variables.

Also, it helps in reducing the redundancy and the correlations created among the dummy variables.

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

Ans:
Among the numerical variables, both “temp” and “atemp” have the highest correlation of(0.63) with the target variable.

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

Ans:
I have run the linear model on the final X_train data frame to make the predictions, and stored the data regarding the same in a variable named y_train_bike. After that, I have plotted a distplot of the residuals. The residuals can be calculated by taking the difference of y_train and y_train_bike (“y_train” – “y_train_bike”). The distplot shows that the error terms are normally distributed with a mean zero, and it also shows that the sum of the errors terms is equal to zero. In the model evaluation step, after we plot the scatter plot we have obtained a linear model, which can be used to say that the error terms have a constant variance, i.e., the error terms are homoscedastic.

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

Ans:
Based on the final model the columns yr(“Year”), temp(“Temperature”), and LightSnowRain (one of the dummy variables for the weather situation) are the 3 top factors contributing significantly towards explaining the demand of the shared bikes.

PART-II: General Subjective Questions.

1. Explain the linear regression algorithm in detail.

Ans:
Regression is one of the techniques used for establishing a relationship between two or more variables. Here we will analyse the past data and try to establish the relation between the different variables in the data. Regression generally involves two types of variables : Dependent Variable (or Response variable), and Independent Variables (or Predictor variables).
A dependent variable is the one that we are trying to predict using the other variables. The dependent variable will also be the target variable for our analysis. The independent variables are the ones that are used to explain the dependent variables.
If the dependent variable is explained using one independent variable, then we call it simple regression. If the dependent variable is explained using multiple independent variables, then we call it multiple or multivariate regression.

Linear Regression:
Linear regression is a type of regression in which the relationship between the dependent and the independent variables is linear. The regression model that we build attempts to explain the dependent variable using a linear combination of the independent variables.
There are two types of linear regression:

a) Simple Linear Regression.
b) Multivariate Linear Regression.

Simple Linear Regression:

It is a basic linear regression model, which explains the relationship between the dependent and the independent variable using a straight line. This straight line is plotted on the scatterplot between the dependent and independent variable.
The correlation between the dependent and the independent variable can be obtained by using either Pearson's R or by plotting a scatterplot.

The general equation of a simple linear regression model is : Y = b0+b1*X
Where, Y is the dependent variable and X is the independent variable. b0 is the constant and b1 is the slope. The Regression coefficients are still obtained by minimising the sum of squared errors, that is, the Ordinary Least Squares method.
Before building a linear regression model we need to check whether there is a linear relationship between the dependent and independent variable.

The best fit line over the dataset will have the minimum difference between the actual values and the values predicted using the model. This difference between the two values is termed as **Residual**. The strength of the linear regression model can be assessed using R-Squared or Coefficient of Determination.

A simple linear regression model should satisfy the following four basic assumptions:

- There should be a linear relationship between the independent and dependent variable to build a model between the two variables.
- The independent variable and the error terms should be independent of each other.
- The error terms are normally distributed, with mean as zero.
- The error terms should have a constant variance.

Multivariate Linear Regression:
A linear regression model with more than one independent variable is known as a **Multivariate Regression model** or **Multiple Regression model.** The general equation of the multivariate linear regression model or MLM is:
Y = b0+ b1*X1 + b2*X2 + …. +bn*Xn
Where, Y is the independent variable, and X1, X2, … Xn are the independent variables, and b0, b1, ..bn are the regression coefficients.
In Simple Linear Regression, R-Squared is used to assess the strength of the model. The R-Squared calculation does not penalise the introduction of a new variable. This should not be the case, as increasing the number of variables in the model makes the model complex and challenging to interpret. Therefore, a new parameter should be used to check the model fit, which penalises the introduction of a new variable. Hence, in MLR we use the Adjusted R-Squared. The adjusted R-square value attempts to penalise the addition of new variables in the model.
When moving from a simple linear regression model to a multiple linear regression model, there are a few things that remain the same and some new points that we must consider. Most of the concepts in multiple linear regression are quite similar to those in simple linear regression.
The following points remain the same between the two models:

i. The model now fits a hyperplane instead of a line. The dimensions of the hyperplane depend on the number of independent variables.
ii. Regression coefficients are still obtained by minimising the sum of squared errors, that is, the OLS method.
iii. The assumptions from simple linear regression still hold for the error terms.

However, with the introduction of new parameters, A multivariate regression model may face the problems of

i. Overfitting
ii. Multicollinearity
iii. High Computation Time.

The following are the steps that are to be performed while building a Linear Regression Model:

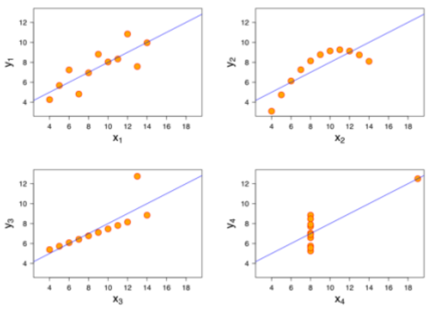
- **Identifying the Problem Definition**: This is the first step of the model building process where we are required to identify the problem that needs to be solved.
- **Data collection**: Once the problem is identified, we need to gather the data required for analysis.
- **Data Visualisation**: In this step we visualise the data to capture any early trends of multicollinearity. Here we can also identify whether some predictors have a strong association with the outcome variable.
- **Data Cleaning and Data Preparation**: After the data is collected, we need to select the necessary variables that are required for the model building. In this step we convert the various categorical variables to numeric variables by creating dummy variables for each level of a categorical variable.
- **Model Specification**: In this step, from the list of available parameters, we need to select the independent variables that we will use to establish a relationship with the dependent variable.
- **Model Building:** In this step, we build a linear regression model over the provided dataset. Here, we will fit a model over the training dataset, which is usually 70% of the total dataset in most cases. This process of train-test split is important as it would validate our model. Here we will split the data into train and test data, with test data being 30 percent and train data being 70 percent. Here we will also scale the numeric feature variables because when we have a lot of independent variables in a model, a lot of them might be on very different scales which will to lead a model with very weird coefficients that might be difficult to interpret. There are 2 types of scaling. They are standardised scaling and minmax scaling. After the scaling is done we build the model, and eliminate the features one by one by checking the significance of each variable(the p-value) and also by checking the VIF of each of the variable.
- **Residual Analysis of the train data:** So, now we check whether the error terms are also normally distributed across the mean (which is one of the major assumptions of linear regression).
- **Model Evaluation**: After a model is fit over the training dataset, it is evaluated on various parameters. These checks help us validate whether the model is good or not. Finally, we also test the model on the testing dataset (remaining 30%) and check if the model performs well. If the model satisfies all the criteria, then we can proceed further, else we will be expected to select a different set of variables or the fitting method which can provide a better model.
- **Using the model to solve the problem**: Once we have obtained a model that fits the dataset in the best possible manner, we can use it to solve the business problem.

2. Explain the Anscombe's quartet in detail.

Ans:

Anscombe's quartet is a quartet of four datasets that were created to describe the importance of **data visualisation**, and how any regression algorithm can be fooled by the same. It comprises four datasets, each containing eleven (x,y) pairs. These four data set plots have nearly **same statistical observations**, and provide same statistical information such as **variance**, and **mean** of all (x,y) points in all four datasets, but when plotted they depict an entirely different picture. This tells us about the importance of visualising the data before applying various algorithms to build models out of them. It also 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, etc. Also, the Linear Regression can be only be considered a fit only for the data with linear relationships.

Consider the following image :



The above image describes the Anscombe's quartet.
The four datasets 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

3. What is Pearson's R ?

Ans:
Correlation coefficients are used in statistics to measure how strong a relationship is between two variables. There are several types of correlation coefficient, and the is Pearson's R is one among them. Pearson's correlation (also called Pearson's R) is a correlation coefficient commonly used in linear regression. Correlation coefficient formulas are used to find how strong a relationship is between data. The formulas return a value between -1 and 1, where:

- 1 indicates a strong positive relationship.
- -1 indicates a strong negative relationship.
- A result of zero indicates no relationship at all.

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

Ans:
Scaling is a process used in the model building to convert all the independent variables, which are in different scales into a similar scale, so that the data is easy to interpret, and it is faster to convert. Scaling just affects the coefficients and none of the other parameters like t-statistic, F statistic, p-values, R-square, etc. There are two popular methods for scaling. They are Standardized Scaling and Normalized Scaling.
Standardized Scaling:
The variables are scaled in such a way that their mean is zero and standard deviation is one.

Normalized Scaling:
The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data. It is also known as Min-Max scaling.

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

Ans:
VIF is given by 1/(1-R^2).
If VIF is infinite, it means that (1-R^2) = 0.
This implies that R^2 =1. This means that there is perfect correlation. An infinite VIF value indicates that the corresponding variable or feature is perfectly correlated with other features or may be expressed exactly by a linear combination of other variables.

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

Ans:

The Quantile-Quantile (Q-Q) plot is a graphical technique used to determine whether two data sets come from the same distribution or not. A 45 degree angle is plotted on the (Q-Q) plot; if the two data sets come from a common distribution, the points will fall on that reference line. The greater the deviation from this reference line, the greater is the evidence to conclude that the two data sets have come from different distributions.

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 of Q-Q plots :

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.