**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?

**Answer**:

1. Fall season has the highest demand for bike rentals
2. The demand in September and October months is the highest
3. Thursday and Friday has the highest demand for bike rentals in a week
4. The demand for bikes is higher when the weather is clear
5. The demand has increased from the year 2018 to 2019

Q.2 Why is it important to use drop\_first=True during dummy variable creation?

**Answer:**

drop\_first=True is used to reduce the redundancy of the dummy variables. And it helps in reducing the extra column as it drops the first column of the dummies. For example: if the column is year column where the values can be either 2018 or 2019. Then we don’t have to retain both the dummy variables, only one will be fine.

If the year is 2019 then it will depict 1 and if the year is 2018 it will depict 0.

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

**Answer:**

The columns temp and atemp have the highest correlation among each other.

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

**Answer:**

**The assumptions of the linear regression model can be validated by looking at the distplot. We can see a normal distribution of the error terms which shows the gaussian distribution and it can be validated.**

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

**Answer:**

**The top three features and there coefficients are as follow:**

1. **season\_Spring -** 0.4920
2. **yr\_2019 -** 1.0239
3. **weathersit\_Light\_Snow - -**0.9518

**General Subjective Questions**

**Q.1** Explain the linear regression algorithm in detail.

**Answer:**

Linear regression model is a type of supervised machine learning modelling technique. In the linear regression model we can have one ore more then one independent variable or features and only one target variable or the dependent variable.

In this technique we ought to find out the correlation between the independent variables (s) and the dependent variable.

If the number of independent variables is more then one and target variable is onlt one then it is called Multiple linear regression but if the number independent variables is only one with one target variable, then it is known as simple linear regression.

The equation for simple linear regression is:

***y= βx + α + ϵ***

Where:

y= dependent variable

= slope

X = independent variable

= y-intercept

The equation for multiple linear regression is:

***y=β0 + β1X1 + β2X2 +...+ βpXp +*** ***ϵ***

where:

y= dependent variable

β0= y- intercept

X1-Xp  = independent variables

Q.2 Explain the Anscombe’s quartet in detail.

Answer:

Anscombe’s quartet can be defined as a group of four data sets. It was first designed by statistician Francis Anscombe in 1973. The four data sets are nearly identical in shape and statistics. But when they are plotted on a scatter plot then they show different variations.

It shows the importance of plotting graphs and plots before performing analysis and building model. We can not predict whether the different data-sets are identical or not without plotting them on a graphs. So, all different features must be visualised first before building any model.

Q.3 What is Pearson’s R?

Answer:

Pearson’s r is a statistics measure which depicts the correlation between any two variables in the data-set. It has a numerical value which can be between -1 to 1. It only measures the linear relationship between the two variables.

Suppose we have to variables X and Y. If the correlation between them is 1 or near to one then they have a positive correlation and if the X increases then Y will also increase. And if X and Y have negative correlation them Y will decrease when X increases.

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

Answer:

Whenever we work on a data set , then different variables can have values in different range ,magnitude or units. We must scale all the variables of varying magnitudes and range into an similar range , which will help in building a precise model. We have to bring all the values to same level of magnitude. It does affect the coefficients but it doesn’t affect any of the statistical values such as F-statistics, p-value or T-statistics.

It is mainly of two types:

1. **Standardisation**: It is helpful when the data follows a normal distribution. It converts data into the mean vector of original data. It doesn’t get affected by the outliers in the data because there is no predefined range of converted features. It can be used when we want mean to be 0 and a unit standard deviation. It is also called as Z-score normalization.

X=

1. **Normalisation** or **Min-max scaling**: When the features are of different scales it converts all the data on a scale of 0,1 or -1 ,1. It is useful when the data has no outliers. Features like age can be scaled by Min-max scaling but no the features like salary because some of the people can have high incomes and can lead to outliers.

X=

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

Answer:

VIF stands for variance inflation factor. It is the amount of multicollinearity between the different independent variables in a data set. It can be calculated by the below formula:

VIF=

If the value of Ri is one then that means there is a perfect correlation between the two variables. Which depicts that one variable is perfectly explained by the other variable. Then the value of VIF will be infinte.

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

Answer:

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.

A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.