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)

Ans:

The Categorical variables like yr,season(winter),month(sep) ,weathersit are linearly dependent on the dependent variable-'cnt'.

We can infer that in the month of September, the demands of bikes are higher compared to other months. The bike's demand was more during the year 2019 as compared to 2018, so we can foresee the increase in upcoming year too. The weather situation 1 , which is Clear, Few clouds, Partly cloudy, Partly cloudy , is expected to be good weather for bike's usage and demand may raise during this weather, also during other weather situation , Bike's demand/usage is expected reduce. Holiday , weekday or weekend have no much effect on the number of bikes used.

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

Ans:

drop_first=True – will drop one of the labels of the dummy variable , which is needed to avoid high correlation between these dummy variables itself.

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

Ans :

The numerical variable 'temp' is highly correlated (of 0.627) with 'cnt' target variable.

However casuals and registered are even more correlated with corelation of (0.67 and 0.945 respectively; but , these are not used for analysis to avoid data leakage. These variables are directly dependent on the target variable and is its derivative.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans:

Residual analysis - Residuals are normally distributed with a mean of 0 and variance σ

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

1. Temp
2. yr
3. Mnth_sep
4. Season_winter

General Subjective Questions

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

Ans:

It is the simplest form of regression. It is a technique in which the **dependent variable is continuous** in nature. The relationship between the dependent variable and independent variables is assumed to be linear in nature.

Residual sum of squares (RSS)

This gives information about how much the target value varies around the regression line (predicted value).

$$\text{Error} = \sum_{i=1}^{n} (Actual\_output - predicted\_output)\text{^}2$$

The best-fit line is obtained by minimizing a quantity called Residual Sum of Squares (RSS) which could be optimized using Gradient Descent to get parameters of the best fit line.

Total sum of squares (TSS)

This tells how much the data point move around the mean.

$$\text{Error} = \sum_{i=1}^{n} (\text{Actual\_output} - \text{average\_of\_actual\_output})\text{^}2$$

R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination

$$R^2 = 1 - \frac{RSS}{TSS} = \frac{Explained\ Variation}{Total\ Variation}$$

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

Ans:

nscombe'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.

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

Ans:

It is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationships or correlations.

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

Ans:

Scaling *(also called **min-max scaling**)*, you transform the data such that the features are within a specific range e.g. [0, 1].

Scaling is important in the algorithms such as support vector machines (SVM) and k-nearest neighbors (KNN) where distance between the data points is important. For example, in the dataset containing prices of products; without scaling, SVM might treat 1 USD equivalent to 1 INR though 1 USD = 65 INR.

Standardization *(also called **z-score normalization**)* transforms your data such that the resulting distribution has a mean of 0 and a standard deviation of 1

You need to normalize our data if you're going use a machine learning or statistics technique that assumes that data is normally distributed e.g. t-tests, ANOVAs, linear regression, linear discriminant analysis (LDA) and Gaussian Naive Bayes.

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

Ans:

It means the correlation is really too high and it has be dropped

(3 marks)

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

(3 marks)

Ans:

A q-q plot is plot of error terms on y_train and y_pred

Error terms are normally distributed with mean zero

The fourth assumption is that the error(residuals) follow a normal distribution. One particular repercussion of the error terms not being normally distributed is that the p-values obtained during the hypothesis test to determine the significance of the coefficients become unreliable.

Normal distribution of the residuals can be validated by plotting a q-q plot.