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:

Below are the few points that could be inferred from the analysis of the categorical variables from the dataset.

- Most number of bike rentals were in Fall and Summer seasons respectively.
- Number of bike rentals were higher in 2019 compared to 2018.
- While considering months, most number of bike rentals were in the months of May to October.
- Number of bike rentals were higher on holidays than on regular days.
- Most number of bike rentals were on Thursday to Sunday towards the end of the week as compared to the beginning of the week.
- Number of bike rentals were almost similar on whether it's a working day or not.
- Most number of bike rentals were in clear weather.

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

Ans:

drop_first=True helps in reducing the extra column created during the dummy variable creation and thus by avoiding redundancy of any kind. If we do not use drop_first = True, then n dummy variables will be created, and these predictors (n dummy variables) will be correlated to themselves which is known as multicollinearity.

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 A and B, then It is obvious C. So, we do not need a 3rd variable to identify the C. Hence if we have categorical variable with n levels, then we need to use n-1 columns to represent the 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 pair-plot among the numerical variables, temp variable 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?

Ans:

Residuals of the final train model were analysed by checking for the normality using a histogram plot of the error terms and checked whether the error terms are following a normal distribution. i.e. the centre of the curve is at zero. The histogram of the error terms turned out to be in a normal distribution and thus the appropriateness of the model was assured.

In addition to this, validated the assumptions of linear regression by checking the VIF and linear relationship between the dependent variable and feature variables.

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 top 3 features contributing significantly towards explaining the demand of the shared bikes are:

1. temp (Coefficient value: 0.5977 - A unit increase in temp variable increases the number of bike rentals by 0.5977 units.)
2. light_snowrain (Coefficient value: -0.2318 - A unit increase in Light_snowrain variable decreases the number of bike rentals by 0.2318 units)
3. yr (Coefficient value: 0.2280 - A unit increase in yr variable increases the number of bike rentals by 0.2280 units.)

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans:

Linear Regression is one of the most fundamental algorithms in Machine Learning which is based on supervised learning category. Basically, it performs a regression task. Regression models predict a dependent (target) value based on independent variables. The regression technique tends to establish a linear relationship between a dependent variable and the other given independent variables. 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.

There are two types of linear regression - simple linear regression and multiple linear regression. Simple Linear Regression explains the relationship between a dependent variable and only one independent variable; a single independent variable is used to predict the value of the target variable. Multiple Linear Regression shows the relationship between one dependent variable and several independent variables; multiple independent variables are used to predict the value of the target variable.

A linear line showing the relationship between the dependent and independent variables is called a regression line. There is a positive linear correlation when the variable on the x-axis increases as the variable on the y-axis increases. This is shown by an upwards sloping straight regression line. There is a negative linear correlation when one variable increases as the other variable decreases. This is shown by a downwards sloping straight regression line.

2. Explain the Anscombe's quartet in detail.

Ans:

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (*x*,*y*) points. These four datasets have nearly the same statistical observations, which provide the same information (involving variance and mean) for each x and y point in all four datasets. However, when you plot these datasets, they look very different from one another.

Anscombe's quartet was constructed in 1973 by statistician Francis Anscombe to demonstrate both the importance of graphing data when analysing it, and the effect of outliers and other influential observations on statistical properties.

3. What is Pearson's R?

Ans:

The Pearson correlation coefficient also known as Pearson's R is the most common way of measuring a linear correlation. Pearson's R is the test statistics that measures the statistical relationship, or association, between two continuous variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship.

Pearson's R was developed by Karl Pearson and it is denoted by 'r'. It has a value between +1 and −1, where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation.

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

Ans:

Scaling is the process to normalize the data within a particular range. It is a step of data pre-processing which is applied to independent variables to normalize the data within a particular range.

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 it will end up in incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude. It also helps in speeding up the calculations in an algorithm.

The Two most discussed scaling methods are Normalization and Standardization. Normalization, also known as min-max scaling typically scales the values into a range of 0 and 1. On the other hand, Standardization or Z-score normalization typically scales data to have a mean of 0 and a standard deviation of 1 (unit variance).

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

Ans:

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.

The value of VIF is calculated by the following formula:

$$VIF_i = \frac{1}{1-R_i^2}$$

Where, 'i' refers to the ith variable.

If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite. It denotes perfect correlation in variables. It suggests that is there is a problem of multi-collinearity and one of these variables need to be dropped in order to define a working model for regression.

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

Ans:

The Q-Q plot or quantile-quantile plot is a graphical technique for determining if two data sets come from populations with a common distribution. It is a graphical tool to help assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution.

The Q-Q plot helps in a scenario of linear regression where the training and test data set received separately and then can confirm using Q-Q plot that both the data sets are from populations with same distributions.

Q-Q plot is used to check if two data sets come from populations with a common distribution, have common location and scale, have similar distributional shapes and have similar tail behaviour. It can be used with sample sizes also. Many distributional aspects such as shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from Q-Q plot.