Assignment-based Subjective Questions

Que 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. From the analysis of the categorical variables, we can infer the following.

- **Season:** Most number of bike bookings are happening in FALL season and least number of bike bookings in SPRING season. SUMMER and WINTER seasons have decent number of bookings.
- **Weathersit:** Highest number of bike bookings are seen in CLEAR weather and low number of bookings in LIGHT SNOW weather conditions.
- **Weekday:** All the weekdays are performing almost equally well with not much difference in number of bookings.
- **Mnth:** JUNE, JULY, AUGUST and SEPTEMBER have high number of bookings while JANUARY and FEBRUARY have low bookings.
- **Holiday:** Non-holiday days seems to be performing well than holidays.
- **Workingday:** Both working days and non-working days are performing equally well.
- **Yr:** Number of bookings have increased significantly in 2019 compared to 2018.

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

Ans. It is important to use **drop_first=True** while dummy variable creation because it drops the extra column created during dummy variable creation and therefore, it reduces the correlation among the dummy variables.

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

Ans. By looking at the pair-plot among the numerical variables we can see that 'temp' variable has the highest correlation with the target variable 'cnt'.

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

Ans. I validated the assumptions of Linear Regression after building the model on the training set by performing the following steps.

- **Error terms distributions:** By plotting the histogram of residuals. Plot showed the normal distribution curve with mean 0.
- **Independent error terms:** By plotting the scatterplot between 'y_train_pred' and 'residuals'. Plot showed that the residual terms are well distributed and does not follow any pattern.
- **Homoscedasticity:** By plotting the regplot between 'y_train' and 'y_train_pred'. Plot showed that the residuals are equally distributed on both the sides of the line.

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

Ans. Top 3 features contributing significantly towards explaining the demand of the shared bikes are

- **Temperature (temp):** temp has a coefficient 0.5717 which indicates that a unit increase in temperature increases the bike rentals by 0.5717 units.
- **weathersit_Light Snow:** weathersit_Light Snow has a coefficient of -0.2362 which indicates that a unit increase in weathersit_Light Snow variable decreases the bike rentals by 0.2362 units.
- **Year (yr):** yr has a coefficient of 0.2289 which indicates that a unit increase in yr variable increases the bike rentals by 0.2289 units.

General Subjective Questions

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

Ans. Linear Regression is a machine learning algorithm based on Machine learning, it performs a regression task. Regression models a target prediction value based on independent variables. Linear Regression is generally used for finding out the relationship between variables and target. There are different regression models based on the kind of relationship between dependent and independent variables and the number of independent variables being used. Types are:

1. **Simple Linear Regression:** This type have only one independent variable.
2. **Multiple Linear Regression:** This type have more than one independent variables.

Linear regression performs the task to predict a dependent variable value (y) based on given independent variable/s (x). So, linear regression finds out a linear relationship between x and y. Hence, it is called Linear Regression.

The regression line is the best fit line for the model. The generalised equation for the regression line is:

$$\mathbf{y = \theta_1 + \theta_2 . x_1 + ... + \theta_{n+1} . x_n}$$

x: input training data (independent variable)

y: target variable (dependent variable)

θ_1: intercept

θ_2 , θ_{n+1}**:** coefficient of x, x_n

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

Ans. Anscombe's quartet consists of four data sets that have almost identical simple descriptive statistics, yet have very different distributions and appear very different when plotted on a graph. Each dataset consists of eleven (*x*, *y*) points. Anscombe's quartet was discovered by Francis Anscombe to demonstrate both the importance of graphing data before analysing it and the effect of outliers and other influential observations on statistical properties.

The quartet is still used to illustrate the importance of looking at a set of data graphically before starting to analyse.

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

Ans. The Pearson's R or Pearson correlation coefficient 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 normalised 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 relationship or correlation. It is also known as Pearson product-moment correlation coefficient or the correlation coefficient.

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

Ans. Scaling is one of the process of data Pre-Processing which is applied to the independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Generally, the data in the data set contains features highly varying in magnitudes, units and range. If scaling is not done on the data then, the algorithm only takes magnitude in account and not units therefore, leading to incorrect predictions by the model. To overcome this, we need to do scaling to bring all the variables to the same unit or level of magnitude.

Normalized Scaling: This scaling method brings all of the variables in the range of 0 and 1. **Sklearn.preprocessing.MinMaxScaler** from **sklearn** library is used to implement normalization. It is also known as *Min-Max Scaling.*

$$x = x - \min(x) / \max(x) - \min(X)$$

Standardization Scaling: This scaling method replaces the values by their Z scores. It brings all of the variables into a standard normal distribution which has mean (**μ)** zero and standard deviation one (**σ**). **Sklearn.preprocessing.scale** from **sklearn** library is used to implement standardization.

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

Disadvantage of using normalization over standardization is that normalization loses some information in the data, especially about outliers.

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

(3 marks)

Ans. Infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables. This happens in the case when R-squared value is 1. VIF is given by formula 1 / 1 – R-squared which tends to infinity as denominator is equal to 0. To avoid this, we need to drop one of the variable from the two which are causing the multicollinearity.

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

(3 marks)

Ans. Quantile - Quantile (Q-Q) plot, is a type of graph which is used to assess if a set of data possibly came from some theoretical distribution such as a Normal, exponential or uniform distribution. It also helps to determine if two data sets come from populations with a common distribution.

Q-Q plot helps in a scenario of linear regression when we have training and test data set received separately. In that case, we can confirm using Q-Q plot that both the data sets are from populations with same distributions.