Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans1.

- The demand bike increased in the year 2019 when compared with year 2018.
- The demand of bike is less in the month of spring when compared with other seasons
- Month Jun to Sep is the period when bike demand is high. The Month Jan is the lowest demand month.
- Bike demand is less in holidays in comparison to not being holiday.
- The demand of bike is almost similar throughout the weekdays.
- There is no significant change in bike demand with workign day and non-working day.
- The bike demand is high when weather is clear and Few clouds however demand is less in case of Light-snow and light rainfall.

Q2. Why is it important to use drop_first=True during dummy variable creation?

Ans 2. drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. It reduces the correlations created among dummy variables.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans 3. temp has the highest correlation with the target variable – cnt

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans 4. Residual distribution should follow normal distribution and centred around 0. The plot confirms this.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans 5. temp, yr, weathersit_bad

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

Ans1. **Linear Regression** is a machine-learning algorithm based on **supervised learning**. It performs a **regression task**. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used. 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.
In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.

Q2. Explain the Anscombe's quartet in detail.

Ans 2. Anscombe'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.

Q3. What is Pearson's R?

Ans 3. 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 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

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans4. It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

Normalization/Min-Max Scaling brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python. Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean 0 and standard deviation 1

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans 5. VIF gives how much the variance of the coefficient estimate is being inflated by collinearity. VIF = 1/(1 – R^2). If there is perfect correlation VIF is infinity. Where R is the R square value of that independent variable which we want to check how well this independent variable is explained well by other independent variables – if that independent variable can explained perfectly by other independent variables. The R squared will be 1 in this case and if we plug the value of 1 in the above formula we get infinity as the value of VIF

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans 6. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. 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