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. From the analysis of the categorical variables from the dataset it could be inferred the total number of bike rental rates are increased in summer and fall season. Also, in the month of September and October rental rates are high. We could also inferred that in “clear” weather sit rental rates are higher. In 2019 the total number of bike rental rates increased with compared to 2018. We could also observe that total number of bike rental rates are higher on holidays.

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

Ans. drop_first is important to use because it helps to reduce extra column created during dummy variable creation. It also reduces correlations created among dummy variables.

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

Ans. Looking at the pair plot we can say temp variable has the highest correlation with the target variable.

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

Ans. Linear regression assumption 1 is all independent variables should not have any correlation so we check multicollinearity among independent variables using VIF. Linear regression assumption 2 is that residual should be normally distributed so we checked error distributions. Also, we checked that the relation between independent variable and target variable should be linear.

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

Ans. Top 3 features contributing significantly towards explaining the demand of the shared bikes are temp, season, windspeed and holiday.

General Subjective Questions

1. Explain the linear regression algorithm in detail?

Ans. Linear regression is a supervised ML algorithm. It helps to predict dependent variable based on independent variable. Dependent variable in regression problem is a continuous numerical variable. There are two types of Linear regression -simple linear regression and multiple linear regression. In Simple Linear regression only one independent variable. In Multiple linear regression there could be more than one independent variables. In linear regression model predict the o/p with a regression line this line is called best fit line if this makes minimum errors on predicting our data. The equation of line is **y=mx+b** , here y is target variable and x is input variable . So here algorithm need to find that value of m and b where the line makes minimum errors. This line is called best fit line which shows the relationship between input(x) and target(y).here m is slope and b is y-intercept. There are some assumption in Linear regression.

- Relation between dependent and independent variable should be linear.
- Independent variables should not be correlated with each other means there should not be multicollinearity.
- Residuals should be normally distributed.
- The Variance of error is constant for various values of independent variable x .This is known as homoscedasticity.

2. Explain the Anscombe’s quartet in detail?

Ans. Anscombe’s quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analysing it and the effect of outliers on statistical properties.

3. What is Pearson’s R?

Ans. The **Pearson correlation coefficient (*r*)** is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

Below is a formula for calculating the Pearson correlation coefficient (*r*):

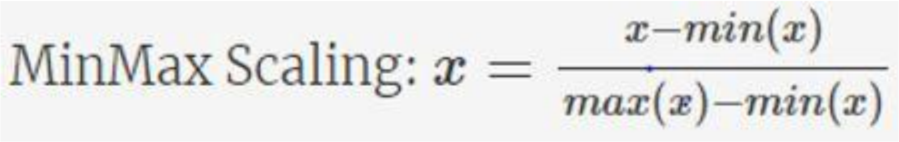
$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

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

Ans. 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.

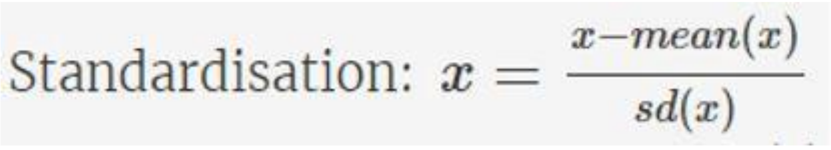
Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

Normalization/Min-Max Scaling: It brings all data in the range of 0 and 1.



$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling: Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).



$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

sklearn. preprocessing.scale helps to implement standardization in python. One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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

Ans. If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

Ans. Q-Q plots are also known as Quantile-Quantile plots. As the name suggests, they plot the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential.