Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 1 goes below this line> (Do not edit)

1. There is an increase in the bike rental count in spring and summer seasons , and then a decrease in
the bike rental count in fall and winter season.
2. The demand for rental bikes increased in the year 2019 when compared with the year 2018.
3. Bike demand is less in holidays in comparison to when not being holidays.
4. There is no significant change in bike demand with working days and non working days.
5. During clear, partly cloudy weather, the bike rental count is the highest, second-highest during misty
cloudy weather, and followed by 3rd highest, during light snow and light rain weather.

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: <Your answer for Question 2 goes below this line> (Do not edit)

it helps in reducing the extra column created during dummy variable creation. Hence it reduces multicollinearity.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

Total Marks: 1 mark (Do not edit)

Answer: <Your answer for Question 3 goes below this line> (Do not edit)

The temp and atemp variables are highly positively correlated to each other.

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 4 goes below this line> (Do not edit)

R-squared is 0.81 on average in our case . It means that the predictor is only able to predict 81% of the variance in the target variable which is contributed by independent variables.

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: <Your answer for Question 5 goes below this line> (Do not edit)

1. Temperature
2. Weathersit
3. Year

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Linear Regression is fundamental algorithm in machine learning and statistics used to model the relationship between a dependent variable and one or more independent variables.

Mathematically, we can write a linear regression equation as:

$$y = b0 + b1x1 + b2x2 + ... bnxn + e$$

b0 is intercept.
b1,b2 ...bn are slopes for the independent variables.
y is dependent variable
x1,x2, ...xn are independent variables
e is error term

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Anscombe's quartet is a set of four datasets that were created by the statistician Francis Anscombe in 1973. These datasets are famous for having nearly identical simple descriptive statistics, such as mean, variance, and correlation, yet they appear very different when graphed. The quartet was designed to demonstrate the importance of visualizing data before analyzing it and to show how statistical properties can be misleading without graphical representation

statistical analysis should always be accompanied by graphical analysis to avoid misleading interpretations.

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Person's coefficient, Pearson's R – the Pearson Product-moment correlation coefficient, the bivariate correlation or simply as correlation coefficient – is measure of linear correlation between two sets of data.

Below is the formulae I took from google.

$$r = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

Where:

- N = number of pairs of scores
- Σxy = sum of the products of paired scores
- Σx = sum of x scores
- Σy = sum of y scores
- Σx^2 = sum of squared x scores
- Σy^2 = sum of squared y scores

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: Please write your answer below this line. (Do not edit)

Scaling is part of Pre-processing step applied on independent variables to normalize the data within a particular range. It helps also in speeding up the algorithm.

Most of the times, the collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then the algorithm only takes magnitude in account and not units hence incorrect modeling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude. Scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Normalized Vs Standardized scaling:

MinMax scaling = (x - min(x))/(max(x) - min(x))

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). sklearn.preprocessing.scale helps to implement standardization in Python.

Standardization = (x - mean(x))/std deviation(x)

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: Please write your answer below this line. (Do not edit)

If VIF(Variance Inflation Factor) is infinite, it means perfect correlation between 2 independent variables.

If R2 (R-square) = 1 then, 1/(1-R2) = infinity.

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)

Sol – we need to drop one of the variables which is causing the perfect multicollinearity.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

(Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: Please write your answer below this line. (Do not edit)

Q-Q plot or quantile-quantile plots, are plots of 2 quantile against each other. (median is quantile where 50% of data fall below that point and 50% lie above it).

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

If the data follows the theoretical distribution, the points on the Q-Q plot will lie approximately along a straight line.

Use and Importance in Linear Regression:

1. Assumption Checking: Linear regression assumes that the residuals (errors) are normally distributed

2. Identifying Deviations

3. Model Diagnostics: A Q-Q plot is a diagnostic tool that helps in understanding the fit of the model

---