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)

Created box plots to understand the relationship between dependent variable 'cnt'. Here are the details.

- During season 3 and 4, demand is significantly higher. Median value during Spring is significantly low.
- Year 2019, there is a significant growth in the demand of bikes. That means demand is growing over the years.
- The median demand between the 5^{th} and 10^{th} months is higher.
- Median demand on all weekdays is similar on all days.
- Median demand on working day or non-working day is similar.
- Median demand of bikes reduces significantly during the snow and light rain. It is significantly higher during the clear days or partly cloudy days.
- Median demand is significantly lower on holidays compared to non-holidays.

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 is important to use **drop_first=True** during dummy variables creation to reduce the multicollinearity. If there are n values for a categorical variable, using the statement will create n-1 dummy variables. This reduces the correlation between the new variables created. This also makes the model more interpretable.

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)

Temperature variables 'temp' and 'atemp' have the most correlation with the target variable 'cnt'.

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)

You use the Residual Analysis and predict the target variable value based on the training data set. Residual is the error between the actual and predicted value by the model. Then create a plot distribution plot (sns.displot()) function to check the distribution of the differences. The graph should show the normal distribution of the residual values, and the mean is close to 0. This proves that the assumption for linear regression is true.

Question 5. Based on the final model, which are the top 3 features contributing significantly

towards explaining the demand for 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)

Top 3 coefficients significantly impacting the target variables are

- `temp`: - Coef value is 0.5738, which indicates that a unit increase in the variable increases the bike hire numbers by 0.5738 unit.
- `yr`: - Coef value is 0.2436, which indicates that a unit increase in the variable increases the bike hire numbers by 0. 2436 unit.
- `weathersit_3`: - Coef value is -0.2622, which indicates that a unit increase in the variable decreases the bike hire numbers by 0.2622 unit.

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 one of the basic statistical algorithms, which is used to find the relationship of a dependent variable with one or more independent variables. With the help of a linear regression, you try to find the best fitting straight line through the data point. Linear regression can be categorized into Simple linear regression and Multiple Linear regression.
Simple linear regression has one independent variable and can be represented as $y=\beta0+\beta1x+\epsilon$. Here y is a dependent variable, x is an independent variable, β1 is the coefficient or slope, β0 is the intercept and ϵ is the error term (difference between the actual and predicted value).
Multiple linear regression has multiple independent variables, which can impact the value of the target variable. This is represented as $y=\beta0+\beta1x1+\beta2x2+\ldots+\beta nxn+\epsilon$. Here x1, x2…xn are the independent variables and β1, β2… βn are the coefficients for the respective independent variables. β0 is intercept and ϵ is the error term.

Below are the assumptions of the Linear Regression algorithms.

- There is a linear relationship between the independent variables and the target dependent variable.
- For Multiple linear algorithms, independent variables should not have multicollinearity, meaning that variables should not be highly correlated.
- The errors are normally distributed and should be constant for all values.

Here are the high-level steps which you can use to implement Linear Regression in Machine Learning.

1. Collect the data
2. Perform Exploratory Data Analytics (EDA) and find if linear relationship exists.
3. Prepare data for the linear regression model
4. Train the model & perform –
    a. Variance Inflation Factor (VIF) analysis to determine multicollinearity
    b. p-value analysis to check statistical significance of independent variables for target variable
    c. F-Statistics analysis to see if the model explains the variance in the target variable.

5. Perform Residual Analysis to determine the assumptions for linear regression.
6. Test the model and generate R-square and adjusted R-square values for test set.
7. Publish interpretation on the model.

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 4 datasets of 11 x and y data pair, which have nearly identical summary statistics, but these appear differently when put into the graphs. This emphasizes the importance of using the graphs before analyzing the data. Just relying on numeric statistics does not represent the convey all the information about the data.

Below are the identical statistical properties of Anscombe's quartet.

- Mean of x and y
- Standard variance of x and y
- Correlation between x and y
- Regression line between x and y

When these datasets are plotted, these shows different qualities mentioned below.
Dataset 1 – this shows simple linear relationship.
Dataset 2 – Shows linear relation but with one outlier
Dataset 3 – shows non-linear relation (a quadratic curve)
Dataset 4 – shows linear relation with one high leverage point outlier

It is very important to plot the graph before analyzing the data to understand the data points better.

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)

Pearson's R is called Pearson correlation coefficient which represents the strength and direction of the linear regression between 2 continuous variables and is denoted a'r'. It ranges between -1 to 1, where 1 means a perfect positive linear relationship and the value of second variance increases when the first variable value increases. In case of negative value, the second variable value decreases when and increase in the first variable value, while 0 means no linear relation.

Formula for Pearson's R is
r = sum((xi – xm)(yi – ym)/SQRT (sum((xi – xm)^2)* sum((yi – ym)^2))

Here xi and yi are the value of data point
xm and ym are the mean of variables.
Below are the assumptions of the Pearson correlation.

- Variables should be continuous and follow normal distribution
- Variance should be constant

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 a technique which is used to adjust or rescale the values of numerical variables/features. This helps to make the values comparable and suitable for Machine Larning algorithms. Example, if you have 2 variables with one having value in range 0 and 1, while the other has values in 1000s or more, in this case scaling helps to convert the values and make them comparable.

Normalized scaling is also called Min and Max scaling, which rescales the data to a fixed range, typically between 0 and 1. Normalized scaling is used when the data needs to be uniformly distributed within a specific range. This technique preserves the relation between data and ensures that all features have the same scale. Formula for Normalized/Min-Max scale is
= (X – Xmin) / (Xmax – Xmin)
Here X is the data point, Xmin is the minimum of the data set and Xmax is maximum of the dataset.

Standardized scaling, also known as z-score scaling, transforms the data to have a mean of 0 and standard deviation of 1. This method ensures that the features follow the normal distribution. This method works well when the data is normally distributed.
Formula for standard or z-score scale is
= X – mue/sigma
Here 'mue' is the meaning of feature values and sigma is the standard deviation.

Differences between these two types of scaling techniques are

- Normalized scaling scales data to a fixed range normally between 0 to 1, while Standardized scaling keeps the value around the mean with a unit standard deviation.
- Normalized scaling is used when you need all the features to have the same range, while Standardized scaling assumes that the data is normally distributed, and units of the features are different.
- Normalization preserves the relationship between the data points, but distribution may be distorted. Standardized ensures the normal distribution with mean as 0 and standard deviation of 1.

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)

VIF means variance inflation factor, which represents the multicollinearity between the independent or predictor variables. High value of VIF means that there is very high correlation between 2 or more variables. To address the multicollinearity, you need to

- Remove one of the highly correlated variables

- Create another variable by combining the correlated variables
- Use regularization techniques like Lasso or Rodge regression.

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 is a type of graph, which is used to check if dataset follows a particular fixed distribution. It plots the quantiles of the data against the quantiles of the suggested theoretical distribution. If the point lies along the 45-degree reference line, it indicates that the data follows the specific distribution, which is mostly normally distributed.

As linear regression assumption is that the residual values are normally distributed, Q-Q plot helps to verify this assumption. Q-Q plots also help to identify the skewness or Kurtosis, representing outliers' presence in dataset or model inaccuracy.
