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)

There were multiple categorical variables in the dataset, and below are inferential for different variables:

1) **Season**:
    a. Seasons have positive correlation to target variable, as ride counts were higher on pleasant seasons like summer and fall, while significantly lower during spring.
    b. However, during EDA it was discovered that seasons has shown multicollinearity with other variables like month.
    c. In final linear regression model dummy variables for Summer and Winter were selected while others were dropped due to higher P values, or higher VIF (>5) and lower coefficient value.
2) **Mnth (Month):**
    a. Months have high correlation with target variable as it's clearly following pattern as given below:
    b. From Jan to June, ride counts (cnt) is constantly going up, but post June it is going down.
    c. Months have multicollinearity with other variables, therefore only dummy variables for Dec, Jan, July, Nov and Sep were picked while others dropped from final model.
3) **Yr (Year):** Year has positive correlation with cnt, as ride counts going up in following year.
4) **Holiday:** Holiday does not show much of correlation and it was dropped from final model.
5) **Weekday:** Outside Sunday, weekday did not show much correlation and therefore only Sunday was included in final model.
6) **Workingday:** Working day also have minimal impact but was picked as it could take care of impact due to non-holiday/ weekday and also shown p value as 0 and low VIF 1.65.
7) **Weathersit:** Weathersit Clear has shown positive correlation while Snow has shown negative correlation to target variable.

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)

Drop_first=True drops first dummy variable, which is essential to handle multicollinearity in dummy variables as each variable value is related to all other dummy variables for same categorical variable.

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)

Temp and atmp have highest correlation to cnt variable.

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)

1) By predicting the value for test set and calculating R Squared. If R Squared for test set is withing 5% range of R squared for train set, then model is fit.
2) Another important part is to perform residual analysis and plot error terms to verify that error terms are equally distributed and mean is centered around 0.
3) Y_pred vs y_test scatter plot could be plotted to validate the error distribution.

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)
Below are top 3 features contributing to the demand of the bikes:

1) Temp
2) Year
3) windspeed

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)

<Your answer for Question 6 goes here>

The main objective of linear regression is to find the linear relationship between dependent variables y and independent variables (target variable) X. Below are Linear Regression equation:

$y = \beta 0 + \beta 1 * x1 + \beta 2 * x2 + \text{----------} + \beta n * xn + \varepsilon$

y: dependent variable
Xi: independent variable
βi: Coefficient
β0: Intercept value (value of y when all the independent variables are 0.)
ε: error term

There are some assumptions for linear regression:

1) Dependent variable should have linear relationship with independent variables.
2) All the observations are independent to each other.
3) Error term variance (y actual – y predicted) should be constant at all levels.
4) Error terms should be independent to each other's.
5) Error terms should be normally distributed and mean should be centered around 0.

6) No multicollinearity among independent variables.

There are some evaluations steps should be taken care to validate the Linear Regression model:

1) Higher the value of R-squared and adjusted R-squared, better the model.
2) P value for independent variable should be within -0.05, +0.05. Lower is better.
3) VIF analysis should be performed to verify multicollinearity and make sure VIF is not more than 10 for any independent variables.

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)

Anscomb's quarter is a set of 4 datasets which have nearly identical statistical properties like Mean, Variance, Correlation coefficient, Linear Regression line which gives the illusion of similar data, but differ significantly when observed in visual graphics.

Purpose of Andcombe's Quarter is to emphasize the importance of Visual inspection as relying solely on statistical summary can be misleading.

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 (Pearson's correlation coefficient) is a statistical measure to quantify the strength and direction of linear relationships between 2 continuous variables. It is used to validate how strong 2 variables are related.

Pearson's variables ranges between -1 to +1, as interpretation is done as given below:

0 **:** No Linear correlation
>0 **:** Positive Correlation (+1 means perfect positive correlation)
<0 **:** Negative Correlation (-1 means perfect negative correlation)

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

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 mechanism to transform independent variables to ensure that all features contribute equally to a model.

Scaling is performed to scale the values in a way that they retain their variations while fitting to a given scale so that coefficients could be comparable.

Normalized Scaling: scale data to a fixed range, typically between 0-1.

Formula = (X – min(X))/(max(x) -min(X))

Standardized Scaling: Standardized scaling centers data around mean and scales by standard deviation.

Formula = $(X - \mu)/\sigma$

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)

Value of VIF is representation of linear relationships between 2 variables, and higher VIF value means higher degree of multicollinearity between variables.

VIF value could reach infinite when a variables have perfect linear relationships with one or more variables. Infinite VIF signals that one or more variables are redundant and adds no extra value to the model.

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)

A Q-Q plot (Quantile-Quantile plot) is a plot to represent relationships between observed and predicted values from a linear regression model to assess whether residuals follow a normal distribution or not.

Q-Q plot is important to stablish if residuals are normally distributed to verify one of key assumptions for Linear regression model.
