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 are 7 categorical variables, out of which season and month were inter corelated, hence not mentioning month.

Yr and Season had high effect on the dependent variable (count)

Holiday has high effect on the dependent variable (count)

Weathersit has high effect on the dependent variable (count)

Weekday and workingday has not much effect on the dependent variable (count)

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)

When we create dummy variables from a categorical feature, we convert each category into a separate column. Each of these new columns will have values of 0 or 1, indicating whether the category is present or not.

However, if we include all these new columns in a model, we can run into a problem called multicollinearity. This happens when some of these columns are too closely related to each other. It makes it hard to figure out how much each feature is actually contributing to the prediction, because they overlap too much in the information they provide.

To prevent this problem, we use the drop_first=True option when creating dummy variables. This means we drop one of the columns, usually the first one. By doing this, we ensure that no column can be perfectly predicted by the others, which keeps our model simpler and more reliable

---

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 columns have highest correlation – atemp and temp, as we have choosen temp variable in our model we can say that variable has highest correlation.

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)

Assumptions of a Linear Regression Model can be validated using

1. Linearity - To Check if the relationship between the predictors and the target variable is linear. Residuals vs Predicted values, there should be no clear pattern
2. Homoscedasticity - To check whether residuals have constant variance, residuals vs predicted values. The spread of residuals should be constant across all the levels of predicted values
3. Normality of Residuals - To Check whether the residuals are normally distributed using a histogram plot

4. Independence - To check if residuals are independent, use Durbin-Watson test
5. Multicollinearity - To check the predictor variables are not highly correlated with each other using VIF

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)

Features positively contributing significantly towards explaining the demand of the shared bikes are

1. Temperature with coefficient of 0.4442 impacting **positively** towards the demand
2. Year with coefficient of 0.2338 impacting **positively** towards the demand
3. Winter season with coefficient of 0.0534 impacting **positively** towards the demand

Note: Light Snow / Rain with coefficient of -0.2949, Spring with coefficient of -0.1204, holiday with coefficient of -0.0906 and misty weather with coefficient of -0.0727 impacts **negatively** towards the demand. These variables impacts the demand in a negative way

---

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>

Linear Regression is one of the most commonly used algorithms in Machine Learning, aim of Linear Regression is to find the best fit line through the data points that minimized the difference between the predicted and the actual values.

There are two types of Linear Regression
Simple Linear Regression – Involves one / single predictor
Multiple Linear Regression – Involves multiple predictor

Equation for a linear regression model is
$y= B0 + B1x1+ B2x2+ B3x3+……. +Bnxn + e$
Where y is the dependent variable, x1,x2,x3….xn are independent variables
B0 is the intercept
B1,B2,B3 are coefficients of the predictors (independent variables) x1,x2,x3 and so on
e is the Error term

Linear Regression makes few assumptions
1. Linearity
2. Homoscedasticity
3. Normality of Residuals
4. Independence

5. No Multicollinearity

Steps Involved
1. Data Reading and Data Preparing (EDA, encoding, scaling etc)
2. Model Building
3. Model Fitting
4. Model validating
5. Interpretation

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)

<Your answer for Question 7 goes here>

Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics, yet appear very different when graphed, this was founded by a person named Anscombe and hence the name. This is used to illustrate the importance of data visualization and the effect of outliers on statistical properties.
For example, 4 datasets might have similar, in fact same statistical properties like Mean, variance, correlation and even Linear Regression model. How ever of the 4 datasets 1 might be simple and well behaved, second one might have an outlier which significantly effects correlation, third one might have completely non-linear relationship and fourth one might be linear but with significant outliers in the dataset
This emphasizes that summary statistics alone can be misleading and Data visualization plays a critical / key role in data analysis and helps in

1. Detecting outliers and anomalies
2. Understanding the data distribution patterns
3. Ensures statical analysis and conclusions are accurate and meaningful

---

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)

<Your answer for Question 8 goes here>

Pearson's R, also known as the Pearson correlation coefficient, is a measure of the linear relationship between two variables. It is a widely used statistical tool to understand the strength and direction of the linear association between two continuous variables. It is represented by r.
r = covariance of variables X, Y / SD of X * SD of Y

Value of r will always be between -1 and 1
r =1 signifies, perfect positive linear correlation
r = -1 signifies perfect negative linear correlation
r = 0 signifies No linear correlation

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)

<Your answer for Question 9 goes here>

It is extremely important to rescale the variables so that they have a comparable scale. If we don't have comparable scales, some coefficients obtained by fitting the regression model might be very large or very small compared to others. This can lead to difficulties during model evaluation, making it hard to interpret the contribution of each predictor accurately. Therefore, it is advised to use standardization or normalization to ensure that the units of the coefficients are all on the same scale

To summarize, Scaling is the process of transforming the features in your dataset so that they fall within a specific range or follow a specific distribution.
There are two types of scaling Normalized scaling (Mix-Max Scaling) and Standardized scaling

Normalized scaling is useful when features need to be on a common scale without distorting differences in ranges. It's particularly helpful when the input variables are on different scales and you want them to be within a common range. This Normalizes the data to fit with in the range 0 to 1.

Formula for Normalized scaling = (x-xmin)/ (xmax – xmin)

Standardization scaling is suitable when features have different scales but need to be standardized to have the same mean and variance. This is often used for algorithms that assume normally distributed data. This transforms data to have a mean of 0 and SD of 1

Formula for Standardization scaling = (x-Mu(Average)) / SD

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)

<Your answer for Question 10 goes here>

When the value of the VIF becomes infinite, it indicates a situation of perfect multicollinearity among the predictor variables in your regression model. Perfect multicollinearity occurs when one predictor variable can be expressed as an exact linear combination of other predictor variables
VIF = 1/(1-R^2)

When R^2 for a given variable is 1. VIF which quantifies how much the variance of a regression coefficient is inflated due to multicollinearity. R^2 = 1 means the predictor variable is perfectly

predictable from other predictors, thus creating a situation of 1/0 which leads to infinity VIF.

This happens when there is perfect multicollinearity or due to dummy variable trap as we saw in Question 2 while creating dummy variables without dropping one category leads to this scenario

---

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)

<Your answer for Question 11 goes here>

A Q-Q plot also called Quantile -Quantile plot is a tool used to assess if the dataset follows normal distribution. It's like a diagnostic tool which checks the assumptions of linear regression are accurate.

In Q-Q plot, Quantiles are calculated at regular intervals from a theoretical cumulative distribution functional of a random variable and then quantiles of dataset are plotted against the quantiles from theoretical distribution. If points lie along the straight line the dataset follows the theoretical distribution, deviation from the line indicates the deviation from the theoretical distribution

Importance of Q-Q plot in linear regression

Q-Q Plot helps in validating Residual Normality, also helps in identifying Outliers and validates the model by checking if residuals follow normal distribution

---