Assignment-based Subjective Questions

Q1 : From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans: Season, weather situation, holiday, month, working day and weekday were categorical variable , box plot was use to analyze them against target variable.
These influence dependent variable as below:

Season : During Summer/Fall season there is considerably more distribution and median of data
Month : We can see there is more Bike rentals taken during months of those seasons and hence months June - October are mostly months in which demand is higher. Demand is increasing each month till June. September month has highest demand.
Weather Situation: During heavy rain there is less demand, however its more during clear weather
Holiday: Holiday has less demand
Weekday: Weekend have more demand then weekday
Working day: Little impact on dependent variable

Q2. Why is it important to use drop_first=True during dummy variable creation?

Ans: Machine learning algorithms mostly works with numeric variables to understand relationships. Categorical variables, like season (spring, summer, fall, winter), may have important information for building a machine learning model but would have to be encoded to its numeric representations.
Dummy variable creation is a process of converting or encoding categorical data to binary representation. (i.e., each variable is converted in as many 0/1 variables as there are different values.)
Specifically, the Categorical variable with ‘n’ levels, dummy variable creation process reduces extra levels (n-1) each indicating whether that level exists or not using a zero or one.

Q3. Looking at the pair plot among the numerical variables, which one has the highest correlation with the target variable?

Ans: The ‘temp’ variable has the highest correlation with the target variable.

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans: 1) There must be a linear relationship between the outcome variable and the independent variables. Scatterplots can show whether there is a linear or curvilinear relationship.

2) Multivariate Normality–Multiple regression assumes that the residuals are normally distributed. Here we can plot the bar graph of errors between observed and predicted values (i.e., the residuals of the regression) and it should be normally distributed.
3) No Multicollinearity—Multiple regression assumes that the independent variables are not highly correlated with each other. This assumption is tested using Variance Inflation Factor (VIF) values.
4) Homoscedasticity–This assumption states that the variance of error terms are similar across the values of the independent variables. A plot of standardized residuals versus predicted values can show whether points are equally distributed across all values of the independent variables.

Q5: Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans: Temperature : With coefficient of 0.5677 which is highest, so if temp will increase then bike rental will also increase by 0.5677 unit
Year: With coefficient of 0.2296 , bike rental will increase with unit of 0.2296 with year
Weather situation Light snow: With coefficient of -0.2103 , unit increase in this weather will decrease demand of bike.

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

Ans: In Machine learning broadly there are two variables to work with:
a. Independent variable (X) –the predictors or features used to predict or explain changes in dependent variable.
b. Dependent variable (Y) – or the target or predicHon expected out of the model.
Example: Marketing budget (X) used to predict Sales (Y)
An equation with highest degree of 1 is called as a Linear equation. The graph of Linear equation forms a straight line.
In algebraic terms, a Linear equation could be written as:
$y = c + mx$
(c = y-intercept & m = slope)
Similarly, the equation of a Linear regression line is given by:
$Y = \beta 0 + \beta 1 * X$
($\beta 0$ = y-intercept & $\beta 1$ = slope)
The Linear regression algorithms establishes a linear relationship between independent variables and dependent variable by finding the best fit line using observed data.
The best fit line is arrived by minimizing the sum of squared errors (RSS – Residual Sum of squares) by taking each data point in the plot.

The Ordinary Least Squares (OLS) method is employed by calculating the residuals at each data point in the plot.
Residual = Actual value - Predicted value
e(i) = y(i) - y(pred)

RSS = e(1)^2 + e(2)^2 + . . . + e(n)^2
= (y(1) - y(pred1))^2 + (y(2) - y(pred2))^2 + . . .
= (y(1) - (β_0 + β_1.X_1)) ^2 + . . .
= Sum of ∑(y(i) - (β_0 + β_1 * Xi))^2 (where i = 1 to n)
- Goal is Minimize RSS for best value of β_0 & β_1

Once the model or best value of Linear regression quesHon (best fit line) is arrived, The strength of the Linear regression model is assessed using:

a. R2 or Coefficient of determination – Higher the R-squared value, the better fit model
b. Residual Standard Error (RSE) – Lower RSS means Hghter model fit.

The following assumptions are also validated after building the model on training set:

a. Error terms to be normally distributed.
b. Multicollinearity among predictor variables to be insignificant.
c. Linear relationship among variables to be observed.
d. Homoscedasticity or the spread of errors (between observed & predicted values) in a regression model to be consistent without any visible pattern which will make prediction results more statistically accurate.

There are two types of Linear regression:

a. Simple Linear Regression (one independent variable)
b. Multiple Linear Regression (several independent variables)

Q2. Explain the Anscombe's quartet in detail.

Ans: Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.
As we know, statistics have long been used to describe data in general terms. For example, things like variance and standard deviation allow us to understand how much variation there was in some data without having to look at every data point individually. They give us a rough idea as to how consistent data is. However, knowing variance alone does not give you the full picture as to what the data truly is in its native form. So in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties, Anscombe's quartet was constructed.
There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets. This tells us about the importance of visualizing the data before applying various algorithms out there to build models out of them which suggests that the data features must be

plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can be only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets.

Q3. What is Pearson’s R?

Ans: Pearson's R, also known as Pearson correlation coefficient or Pearson's product-moment correlation coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It measures the degree of association between the variables on a scale ranging from -1 to 1.

Few key points about Pearson's R:

Range and interpretation: Pearson's R ranges from -1 to 1. A value of -1 indicates a perfect negative linear relationship, meaning that as one variable increases, the other variable decreases in a consistent manner. A value of 1 indicates a perfect positive linear relationship, where both variables increase or decrease together. A value of 0 indicates no linear relationship between the variables.
Linear relationship: Pearson's R specifically measures the linear association between variables. It assumes that a straight line can approximate the relationship between the variables. If the relationship is nonlinear, Pearson's R may not accurately capture the association.
Symmetry: Pearson's R is symmetric, meaning that the correlation between variable X and variable Y is the same as the correlation between variable Y and variable X. The order of the variables does not affect the magnitude or interpretation of the correlation coefficient.
Strength of association: The magnitude of Pearson's R indicates the strength of the association between the variables. A value close to -1 or 1 suggests a strong linear relationship, while a value closer to 0 indicates a weaker association.
Interpretation of magnitude: There is no definitive threshold for determining what constitutes a "strong" or "weak" correlation, as it can depend on the context and field of study. However, commonly used guidelines consider correlations around ±0.3 to ±0.5 as moderate, ±0.5 to ±0.7 as strong, and above ±0.7 as very strong.
Statistical significance: In addition to the magnitude, the statistical significance of Pearson's R is important. Hypothesis tests can determine if the observed correlation coefficient is significantly different from zero, indicating a meaningful relationship between the variables.
Assumptions: Pearson's R assumes that the relationship between the variables is approximately linear, the variables follow a bivariate normal distribution, and there are no outliers or influential observations. Violations of these assumptions can impact the accuracy and validity of the correlation coefficient.
Limitations: Pearson's R measures only the linear relationship between variables and does not capture other types of relationships (e.g., nonlinear, curvilinear). It is sensitive to outliers and can be influenced by extreme observations.

Q4: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans: Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.
Example: If an algorithm is not using feature scaling method then it can consider the value 3000 meter to be greater than 5 km but that's actually not true and in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans: Variance inflation factor (VIF) is a measure (esHmate) of severity of multicollinearity in regression. VIF explains the relationship of an independent variable with all other independent variables.
Multicollinearity could have negative effects when building regression models of prediction and interpreting the results. This could also make it difficult to identify the contribution of significant variables.
The formula is:
VIF (i) = 1 / (1 – R-squared (i))
VIF > 10 is Very High and has to be investigated and corrected
VIF > 5 is High and has to be investigated and corrected
1 < VIF < 5 (moderately correlated)
VIF=1 (No multicollinearity)

When VIF is infinite (∞),
There is perfect multicollinearity in the regression model which suggests that the independent variables are not suitable for prediction in the model. This could be corrected by removing variables and stabilizing the model.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans: A Q-Q plot, short for quantile-quantile plot, is a graphical tool used to assess the distributional similarity between a sample of data and a theoretical distribution. It compares the quantiles of the observed data against the quantiles expected under a particular distribution, such as the normal distribution. The Q-Q plot allows visual inspection of whether the data deviates from the expected distribution and provides insights into the goodness-of-fit between the data and the assumed distribution.

Below are few of the uses and importance of a Q-Q plot in linear regression:

Distributional assessment: In linear regression, it is often assumed that the residuals (the differences between the observed and predicted values) follow a normal distribution. The QQ plot helps assess the validity of this assumption by visually comparing the quantiles of the residuals against the quantiles of a normal distribution. If the residuals approximate a straight line on the Q-Q plot, it suggests that the residuals follow a normal distribution.
Detecting departures from normality: The Q-Q plot can reveal departures from normality in the residuals. If the plotted points deviate from a straight line, it indicates a departure from the assumed normal distribution. Departures may include skewness (asymmetric tails) or heavy-tailedness (excess kurtosis) in the residuals, which can affect the reliability of the linear regression model.
Assessing model assumptions: Linear regression relies on several assumptions, including linearity, independence, and homoscedasticity. Violations of these assumptions can lead to biased or inefficient coefficient estimates. The Q-Q plot helps diagnose departures from the normality assumption, which is a crucial assumption in linear regression analysis. If significant deviations from normality are observed, it may indicate violations of other assumptions as well.
Model refinement and diagnostics: The Q-Q plot provides insights for model refinement and diagnostics. If the Q-Q plot shows deviations from the expected straight line, it suggests that the model assumptions need further examination. It can guide the selection of appropriate transformations or suggest the need for robust regression techniques to account for nonnormality or outliers.
Comparing alternative distributions: Besides assessing normality, the Q-Q plot can be used to compare the observed data against other theoretical distributions. This allows researchers to explore whether a different distribution may provide a better fit to the data, potentially leading to more accurate and reliable regression modeling.
By visually examining the patterns in the Q-Q plot, we can evaluate the distributional assumptions in linear regression and make informed decisions about the model's validity and reliability. It helps identify potential issues, guide model diagnostics, and support the selection of appropriate modeling techniques to improve the accuracy and interpretation of regression results.