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)

In this dataset, the categorical variables include season, weathersit, yr (year), mnth (month), weekday, holiday, and workingday. Here's an inference about their effect on the dependent variable (number of bike rentals) based on the analysis:

1. **Season (season):**
   The season variable divides the data into four categories: Spring, Summer, Fall, and Winter. During the warmer months, such as Spring and Summer, bike rentals are typically higher, which is consistent with people's preference to ride bikes in pleasant weather. In Winter, rentals are generally lower, likely due to colder temperatures and less favorable conditions for outdoor activities. This variable helps explain seasonal trends in bike rentals.
2. **Weather Situation (weathersit):** The weathersit variable is categorical and can take values such as "Clear", "Mist", "Light Snow", etc. On days with clear weather, bike rentals are generally higher because people are more likely to rent bikes in good weather conditions. Misty or rainy weather (e.g., "Mist", "Light Snow") is associated with lower bike rentals, as fewer people would want to ride bikes in these conditions. This highlights the importance of weather conditions on the number of rentals.
3. **Year (yr):**
   The yr variable is binary, indicating the year (either 0 for 2011 or 1 for 2012). There might be a small effect of year on rentals, with variations possibly due to changes in bike-sharing popularity, operational capacity, or external factors. This categorical variable could capture any annual trends in the data.
4. **Month (mnth):**
   The mnth variable represents the month of the year (from 1 to 12). Months like May, June, and September (typically warmer months) tend to see more bike rentals, while colder months such as December, January, and February witness fewer rentals. This supports the idea that rental demand is higher during specific months (usually warmer months).
5. **Day of the Week (weekday):**
   The weekday variable represents the day of the week (0 for Sunday, 6 for Saturday).Weekdays (Monday to Friday) tend to have a higher number of rentals due to commuting needs and regular use, while weekends (Saturday and Sunday) may have varying levels of rentals, depending on factors like weather and social activities. The weekend effect may vary by city, but generally, weekday rentals could be higher due to work commute.
6. **Holiday (holiday):**
   The holiday variable is binary, indicating whether the day is a holiday (1) or not (0).On holidays, bike rentals may increase due to more leisure activities and people being free from work. However, in some cases, holidays might also lead to fewer rentals due to people traveling or not needing bikes. The overall effect of holidays on rentals can vary based on the context of the city or region.
7. **Working Day (workingday):**
   The workingday variable is binary, with 1 indicating a working day and 0 indicating a non-working day (weekend or holiday).On working days, rentals are typically higher due to commuter use. People tend to use bikes as a mode of transport to get to work or run errands.On non-working days (weekends or holidays), rental usage might vary but can be lower for commuting purposes, although leisure and recreational bike rentals may still occur.

Summary of Effects:

Season: Strong effect with higher rentals in warmer months (Spring, Summer).
Weather: Strong effect, with clear weather resulting in more rentals.
Year: Mild effect, potentially reflecting trends in bike rental adoption over time.
Month: Strong effect, with higher rentals during warmer months.
Weekday: Weekdays (especially workdays) tend to have higher rentals due to commuting.
Holiday: Holidays can slightly increase or decrease rentals depending on the region.
Working Day: Non-working days may see slightly fewer rentals, but this can vary.

In conclusion, season, weather, and day-related factors (month, weekday, working day) have a significant effect on the number of bike rentals. These variables capture time-based trends that affect people's decision to rent bikes.

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)
Using **drop_first=True** during dummy variable creation is important to avoid multicollinearity. When all categories are included as dummy variables, they become perfectly correlated, which can lead to unreliable regression results. By dropping the first category, we prevent this correlation, with the dropped category acting as the reference. This allows the model to interpret the effect of the other categories relative to the reference category, ensuring valid and interpretable results.

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)
From the pair-plot among the numerical variables, temperature (temp) has the highest correlation with the target variable (number of bike rentals). This can be observed visually as the temp variable shows a clear positive relationship with the target variable , where higher temperatures generally correspond to more bike rentals.

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)
To validate the assumptions of Linear Regression after building the model on the training set, I performed the following steps:

- **Linearity**: The assumption of linearity states that there is a linear relationship between the independent variables and the dependent variable. This was visually validated by plotting the residuals (the differences between actual and predicted values). In a well-fitting linear regression model, residuals should be randomly distributed without any clear pattern.The Residual Plot was examined to ensure there were no patterns indicating a non-linear relationship.

- **Homoscedasticity** (Constant Variance of Errors): Homoscedasticity assumes that the variance of the residuals is constant across all levels of the independent variables. This was validated by plotting a scatter plot of residuals against the predicted values (or independent variables). If the plot shows random scattering without any funnel shape, it suggests that the assumption holds true.
- **Normality of Residuals**: The residuals should be normally distributed for the model to provide reliable results. This assumption was validated by plotting a histogram or kernel density estimate (KDE) of the residuals. A normal distribution of residuals would appear as a bell curve. Additionally, a Q-Q plot (Quantile-Quantile plot) could be used to check if the residuals align with a straight line, indicating normality.

These steps helped me ensure that the assumptions of linear regression (linearity, homoscedasticity, and normality) were met, thereby confirming the validity of the model and its predictions.

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)

Based on the final linear regression model, the top 3 features contributing significantly towards explaining the demand for shared bikes (cnt) are:

- **Temperature (temp):** As expected, temperature has a strong positive correlation with bike rentals. Higher temperatures lead to more favorable conditions for cycling, thus increasing bike rentals.
- **Apparent Temperature (atemp)**:Apparent temperature, which accounts for both the actual temperature and the humidity, also influences bike rentals. Higher apparent temperatures tend to correlate with increased bike usage, as people feel more comfortable outdoors.
- **Weather Situation (weathersit):** The weather situation, including conditions like clear skies or mild temperatures, directly affects the demand for shared bikes. Favorable weather (clear or mild) boosts rentals, while rainy or snowy conditions reduce them.

These features are the most influential in predicting the number of bike rentals and drive the demand for shared bikes.

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 a statistical method used to model the relationship between a dependent variable and one or more independent variables. It is one of the simplest and most widely used techniques in both machine learning and statistics. The main objective of linear regression is to find the best-fitting line (or hyperplane in the case of multiple variables) that represents the relationship between the independent and dependent variables.

In simple linear regression, there is one independent variable (predictor) and one dependent variable (target). The algorithm aims to find a straight line that best fits the data. In multiple linear regression, there are multiple independent variables, and the goal is to find a hyperplane that

explains the relationship between all the predictor variables and the target variable.

The linear regression algorithm works by finding the coefficients (also called weights) of the linear equation. These coefficients represent the impact of each independent variable on the dependent variable. The algorithm minimizes the difference between the observed data points and the predicted values by adjusting the coefficients. This process is done by minimizing the sum of squared errors or residuals, which is the difference between the actual and predicted values.

The key assumptions in linear regression are that the relationship between the dependent and independent variables is linear, the residuals (errors) should be normally distributed and homoscedastic (i.e., having constant variance), and the observations should be independent of each other. These assumptions are important because violating them can lead to unreliable results.

After fitting the model, the performance is evaluated using metrics such as R-squared, Mean Absolute Error (MAE), and Mean Squared Error (MSE). R-squared measures the proportion of the variance in the dependent variable that is explained by the independent variables. MAE and MSE give insights into the average magnitude of errors in the model's predictions.

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 collection of four datasets created by the statistician Francis Anscombe in 1973. The main purpose of the quartet is to show that datasets with identical descriptive statistics, such as mean, variance, and correlation, can still have very different distributions and relationships between variables. This serves as a reminder of the importance of data visualization in understanding the structure of the data.

Each dataset in the quartet consists of 11 pairs of data points, where the x and y variables share identical simple statistical properties. Despite this, when plotted, the datasets reveal vastly different relationships between the variables. In the first dataset, the relationship between x and y is linear, meaning the data points lie close to a straight line. The linear regression model fits this data well, as expected. The second dataset, however, shows a quadratic relationship. The points form a curve, and although the linear regression model might still fit, it does not capture the true nature of the data.The third dataset features a strong linear relationship with one extreme outlier, which significantly affects the correlation and the best-fit line. Visualizing this dataset clearly highlights the outlier's impact, which is not apparent through simple statistical measures alone. The fourth dataset appears linear at first glance, but all the points are clustered around the same value of x, with one outlier. The presence of this outlier again distorts the results of a linear regression model, making it appear to have a strong relationship when, in reality, it does not.

Anscombe's Quartet demonstrates that relying solely on statistical summaries like means and correlations can be misleading. It emphasizes that visualizing data is essential for uncovering its true structure, as the visual patterns can reveal information that summary statistics cannot.

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, also known as the Pearson correlation coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It ranges from -1 to 1, where a value of 1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear relationship at all.

Pearson's R is calculated by dividing the covariance of the two variables by the product of their standard deviations. This formula standardizes the covariance, making it a unitless measure, so it can be applied regardless of the units of the variables being analyzed. A positive value of Pearson's R indicates that as one variable increases, the other also tends to increase, while a negative value suggests that as one variable increases, the other decreases.

It is important to note that Pearson's R only measures linear relationships. This means that even if two variables have a strong non-linear relationship, Pearson's R may not detect it, and thus, other methods might be more appropriate for those cases. Additionally, Pearson's R assumes that the data is normally distributed and that there are no significant outliers, as these can heavily influence the coefficient and the conclusions drawn from it.

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 refers to the process of transforming the features of a dataset into a consistent range or distribution. This is important because many machine learning algorithms are sensitive to the scale of the input data, meaning that features with larger magnitudes may dominate the learning process, leading to biased or incorrect models. Scaling ensures that each feature contributes equally to the model's performance, preventing one feature from unduly influencing the outcome.

Scaling is performed to make the data more suitable for algorithms that rely on distance metrics or gradient-based optimization methods. For instance, algorithms like k-nearest neighbors, support vector machines, and linear regression often perform better when the data is scaled because they compute distances between points or rely on gradient descent, where large values can cause slow convergence or suboptimal results.

There are two common types of scaling: normalized scaling and standardized scaling. Normalization typically rescales the data so that the values fall within a fixed range, often between 0 and 1. This is done by subtracting the minimum value and dividing by the range (maximum value minus minimum value). Standardization, on the other hand, transforms the data to have a mean of 0 and a standard deviation of 1, achieved by subtracting the mean and dividing by the standard deviation of each feature. The key difference is that normalization compresses the data into a specific range, while standardization adjusts the data to follow a standard normal distribution, making it more suitable for algorithms that assume the data is normally distributed.

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)

The variance inflation factor (VIF) measures the extent to which the variance of a regression coefficient is inflated due to multicollinearity among the independent variables. In essence, it quantifies how much the variance of a coefficient is increased because of linear dependence with other predictors in the model. A high VIF indicates that one or more predictors are highly correlated with other predictors, which can lead to issues in the model's interpretation and stability.

When the value of the VIF is infinite, it typically occurs when there is perfect multicollinearity, meaning that one of the independent variables is a perfect linear function of one or more other independent variables in the model. This could happen when one variable is directly derived from another, or if the independent variables are highly redundant in some way. In such cases, the regression model cannot uniquely estimate the coefficient for that variable, leading to an infinite VIF. This is problematic because it makes the model unstable, with unpredictable or unreliable coefficient estimates.

In practical terms, infinite VIFs are often a signal that the model suffers from severe multicollinearity, and some predictors need to be removed, combined, or modified to ensure a stable and interpretable 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 (Quantile-Quantile) plot is a graphical tool used to assess whether a dataset follows a particular theoretical distribution, most commonly the normal distribution. It compares the quantiles of the sample data with the quantiles of a specified distribution, typically the normal distribution. If the data points in the Q-Q plot lie along a straight line, it indicates that the data approximately follows the theoretical distribution. If the points deviate significantly from the line, it suggests that the data does not follow the expected distribution.

In the context of linear regression, the Q-Q plot plays a crucial role in validating one of the key assumptions of the model: that the residuals (errors) are normally distributed. This assumption is important because linear regression models rely on normally distributed errors for accurate predictions and reliable hypothesis testing. The Q-Q plot helps in visually identifying deviations from normality, such as skewness or heavy tails, which could indicate issues with the model's validity.

By using a Q-Q plot to check the normality of residuals, analysts can ensure that the assumptions of linear regression are met, or they can take corrective actions if needed, such as applying transformations to the data or choosing a different modeling approach.
