Assignment-based Subjective Questions

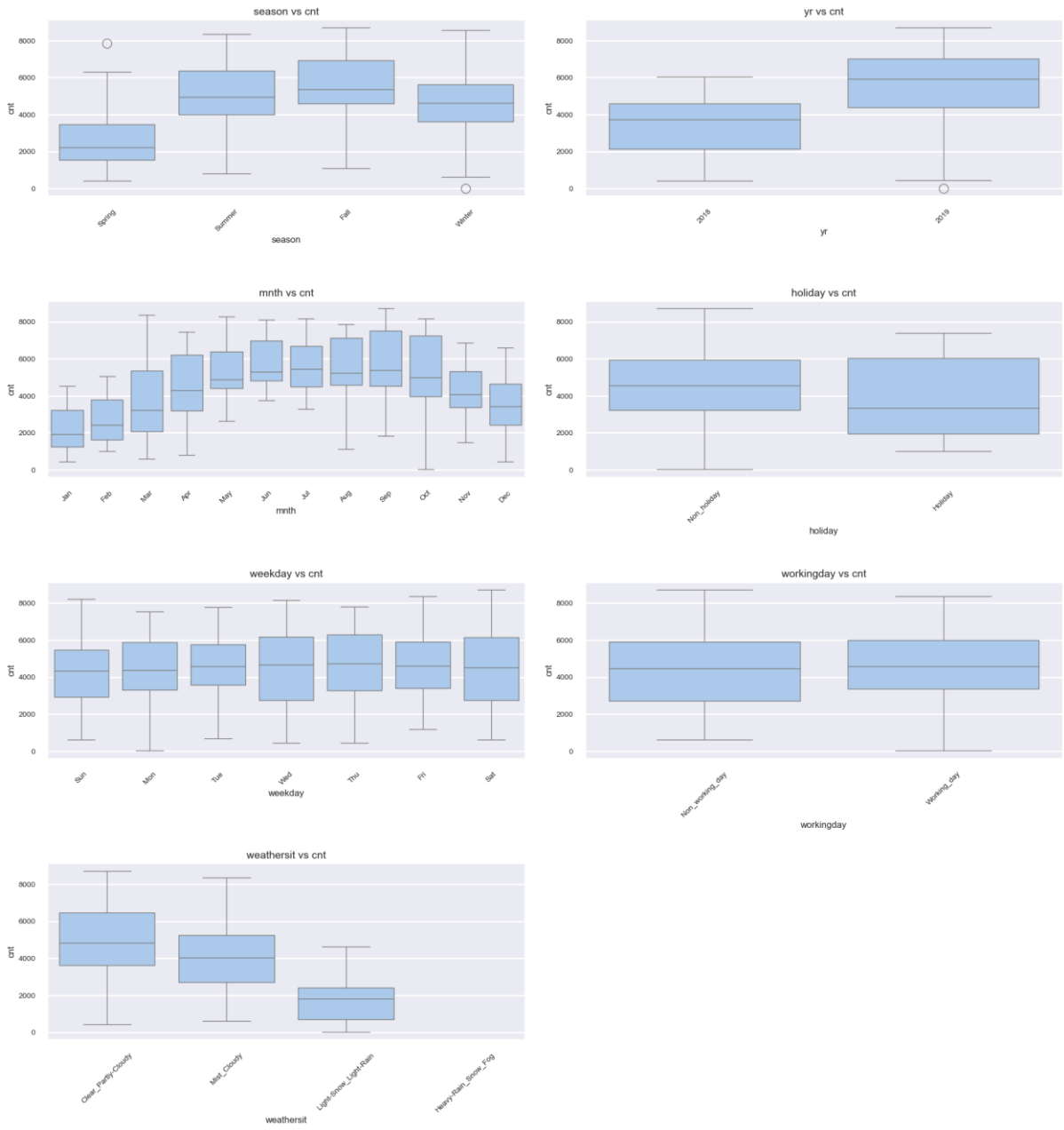
Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 1 goes below this line> (Do not edit)

The target variable / dependent variable in the dataset is 'cnt' which implies the demand. The categorical independent variables in the dataset are, ['season', 'yr', 'mnth', 'holiday', 'weekday', 'workingday', 'weathersit']

- 'season': Demand is highest in Fall, followed by Summer and Winter, with a drop in Spring, indicating strong seasonality.
- 'yr': Demand increased from 2018 to 2019, reflecting greater adoption of bike rentals.
- 'mnth': Peak demand occurs from May to October, especially in August, June, and September, aligning with warmer months.
- 'holiday': Rentals slightly drop on holidays, likely due to reduced commuting, but leisure usage balances the decline.
- 'weekday': Rentals remain high on Fridays, Saturdays, Sundays, and Thursdays, suggesting both commuting and leisure usage.
- 'workingday': Demand is steady regardless of working days, showing a balance between commuting and leisure.
- 'weathersit': Clear weather boosts demand, while adverse conditions like heavy rain or snow reduce it.



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 creating dummy variables, setting drop_first=True can help avoid multicollinearity by dropping the first level of the categorical variable. If a category has *n* levels, only *n-1* dummy variables are created. When all dummy variables are 0, it implies the dropped category, ensuring the model remains stable and 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)

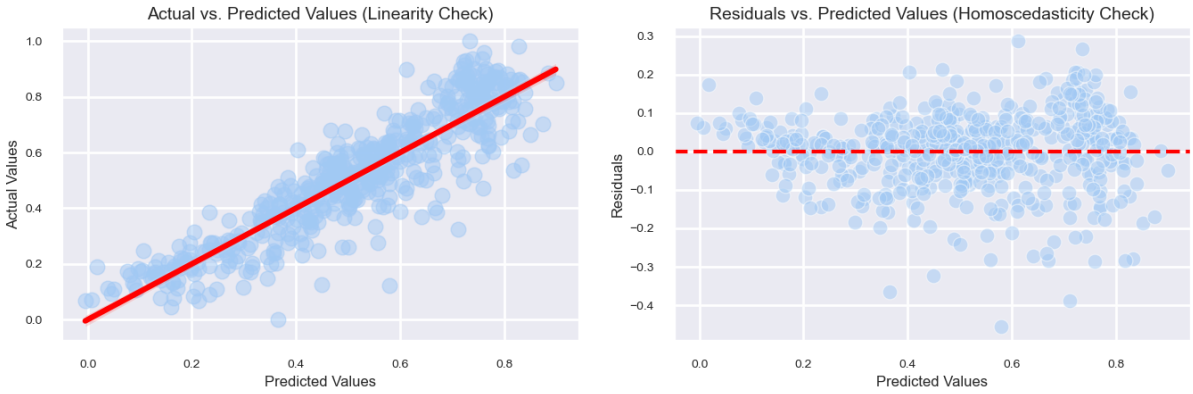
The target variable in this case is `cnt` (total bike rentals). Looking at the correlation values, the variable with the highest correlation with `cnt` is `registered` with a correlation of 95%.

This indicates that the number of registered users strongly influences the total bike rentals, which makes sense as registered users are likely more consistent renters compared to casual users.

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)

Linearity: The actual values plotted is checked against the predicted values in the blow graph, are well aligned with the regression line marked in solid red. This pattern shows the linearity between the prediction and the actual data.

Homoscedasticity: In linear regression, one of the key assumptions is that the variance of the residuals (errors) remains constant across all levels of the predicted values. Observing the scattered plot in the below image, the residuals are randomly scattered compared to the prediction proving the linear regression assumption is true.



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)

The top 3 features significantly contributing to bike demand are:

1. temp (Temperature): Positive impact, coefficient = 0.5660, p-value = 0.000.
2. yr (Year): Positive impact, coefficient = 0.2264, p-value = 0.000.
3. hum (Humidity): Negative impact, coefficient = -0.2848, p-value = 0.000.

These features have a significant statistical relationship with bike demand.

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 a supervised learning algorithm used for predicting continuous target variable based on input features.

In our case we must predict the demand 'cnt', which is our target variable and is given in the dataset. Hence, we call it supervised learning.

Linear regression models the relationship between the target variable and the independent variables as a linear equation,

Y = mx + b (Equation of line)
y = B0 + (B1 * X) + E

y is the target, X is the list of all independent variables, B0 is the intercept (constant), B1 is the slope (coefficient), E is the error term.

The model finds the best-fit line by minimizing the sum of error squares (Ordinary Least Squares method).

Linear regression is started with assumptions, including, linearity, homoscedasticity and residuals are normally distributed.

Finally model performance is evaluated based on the adjusted r2. The r2 score is a parameter which explains the percentage of variance in 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)

<Your answer for Question 7 goes here>

Anscombe's quartet consists of four datasets with identical mean, variance, correlation and regression line, but different distributions when plotted.

1. Linear relationship
2. Non-linear relationship
3. Outlier influence

4. Leverage point.

Though the summary of statistics is identical, the visual differences show why relying only on numerical metrics can be misleading in data analysis.

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 is a representation of correlation. It measures the linear relationship between two variables ranging between -1 and +1, where 1 being perfect positive correlation, -1 is perfect negative correlation and 0 represents no correlation.

It is calculated as the covariance of the variables divided by the product of their standard deviations. Outliers should be removed upfront which can impact the results.

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>

Scaling is the process of transforming numerical features to a common range, 0 to 1 or -1 to 1. This will help improve model performance and stability.

Numeric input features will have values in different ranges. It can be in millions, thousands or even in fractions.

Scaling ensures no features dominate due to difference in magnitudes.
Reduces numerical instability in calculations.

Normalized scaling is also called min-max-scaling, which computes the values and converts them to range between 0 and 1.

$X' = (X - \min(X))/(\max(X) - \min(X)$

Standardized scaling centers the data around mean 0 with standard deviation of 1, implying the values can range between -1 and +1

$X' = (X - \text{mean}) / \text{standard deviation}$

Normalized scaling is done when the data is not normally distributed and standardized scaling is done when 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)

<Your answer for Question 10 goes here>

The VIF becomes infinite when two or more features in the model are perfectly correlated. This means that one feature can be exactly predicted from the others, causing the model to struggle in separating their individual effects.

As a result, the variance (uncertainty) in estimating the coefficients of these features becomes very large, making the VIF go to infinity. Simply put, it happens when there's redundancy between the features, and there's no way to explicitly read which feature has better explanation for the target variable.

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 (Quantile-Quantile) plot is a graphical tool to compare the distribution of a dataset with a theoretical distribution (often the normal distribution). It plots the quantiles of the dataset against the quantiles of the theoretical distribution. If the points in the plot lie along a straight line, the dataset follows the theoretical distribution.

In linear regression, a Q-Q plot is used to check the normality of residuals. Normality of residuals is an important assumption in linear regression, as it ensures that the model's predictions are unbiased and valid.
