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)

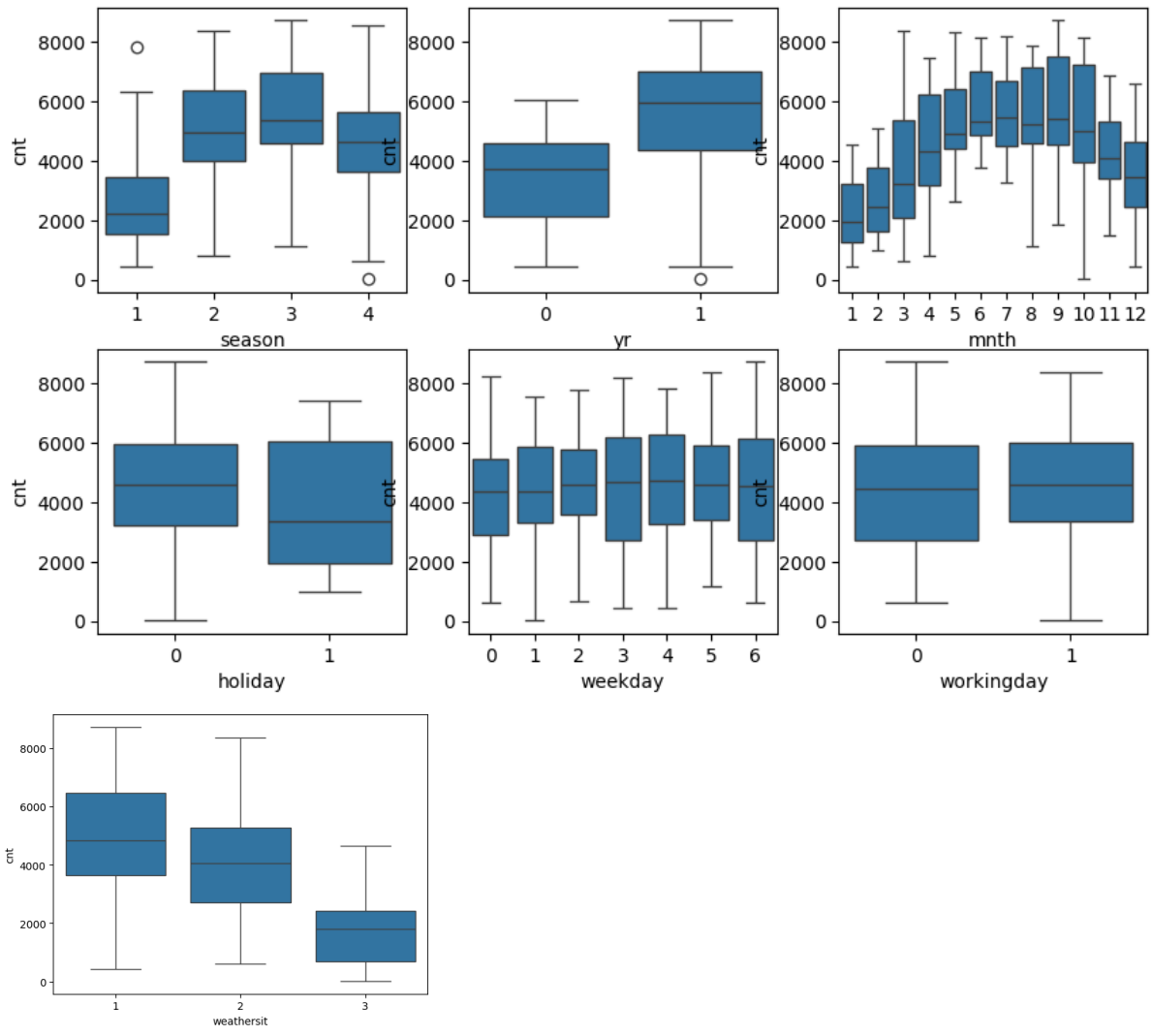
Total Marks: 3 marks (Do not edit)

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

We can observe a pattern in the 'season' variable. Specifically for spring the 'cnt' is low in comparison to other seasons. Similar pattern can be seen in 'mnth'. There is not much variation in terms of 'weekday' and 'workingday'. Almost same mid value of 'cnt' for all categories. However, for 'holiday', there is a pattern observed. There is variation in 'yr'.

weathersit is categorical variable and it can be seen there is a pattern when we plot a boxplot.

We can convert all these categorical variable data into numeric using dummy method and use in our analysis.



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)

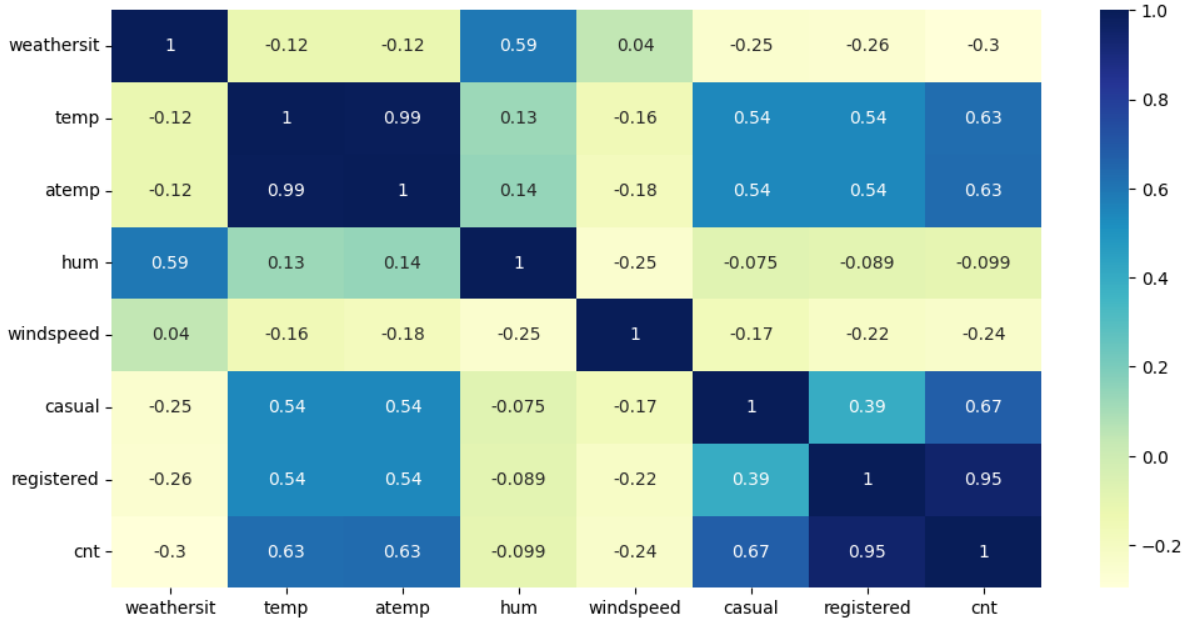
When creating dummy variables for a categorical feature, it's essential to use drop_first=True to:

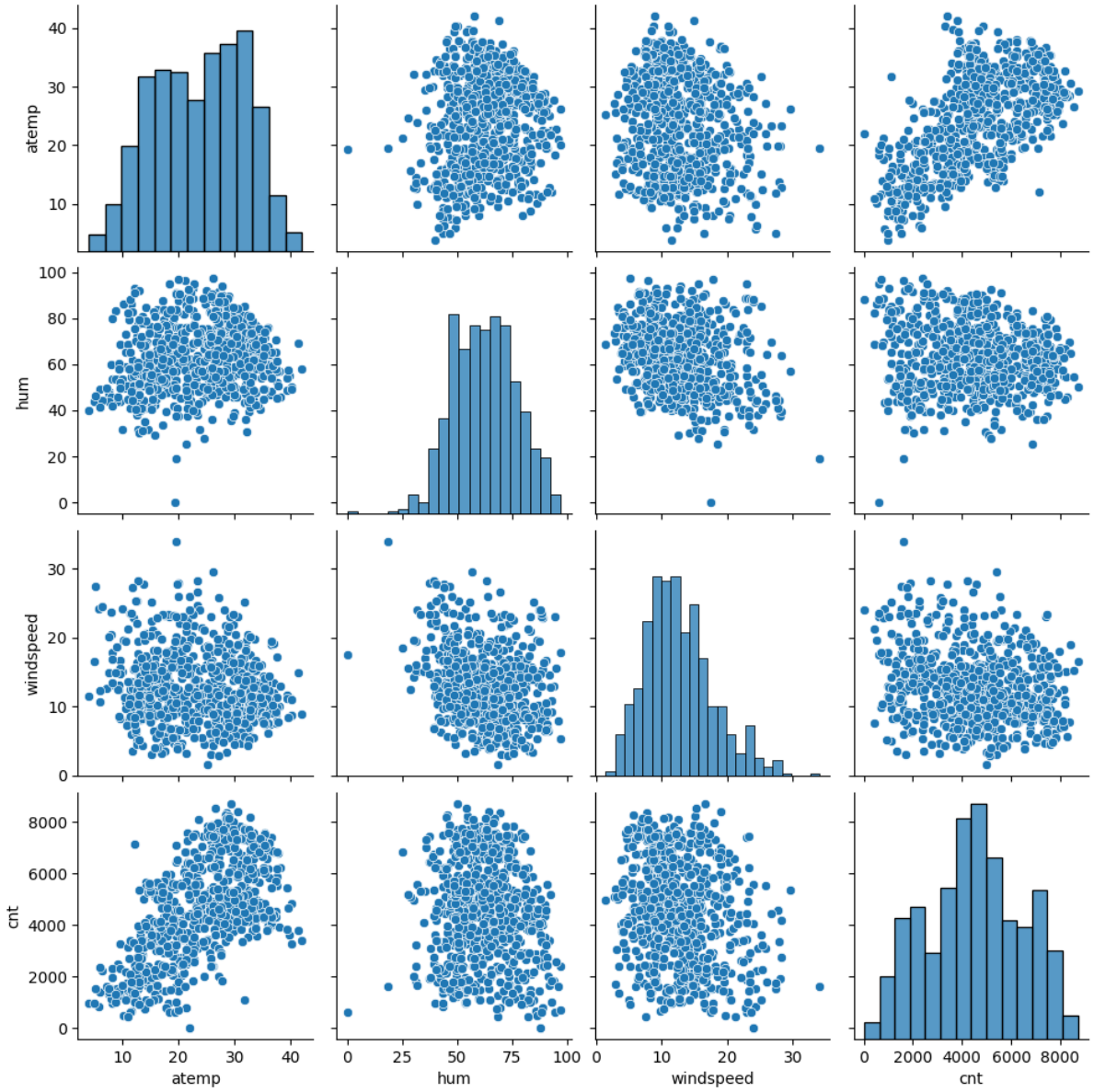
1. Avoid redundancy: Prevents perfect collinearity between dummy variables.
2. Set a reference category: Allows for clear comparisons between categories in the regression model.
3. Improve model interpretability: Makes it easier to understand the coefficients of the dummy variables.

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)

Among windspeed, humidity and temp, temp has the highest correlation with cnt, the target variable. Below is the heatmap plotted by calculating the correlation. Also, the pair plot has been shown to visualize the pattern.

'casual' and 'registered' can be dropped as these are constituents of the target variable 'cnt'. There is a positive correlation of 'temp', and 'atemp' with 'cnt'. However there is negative correlation with 'weathersit', 'windspeed' and 'humidity'.
There is strong positive correlation between 'temp' and 'atemp' as well. We can remove one of these.





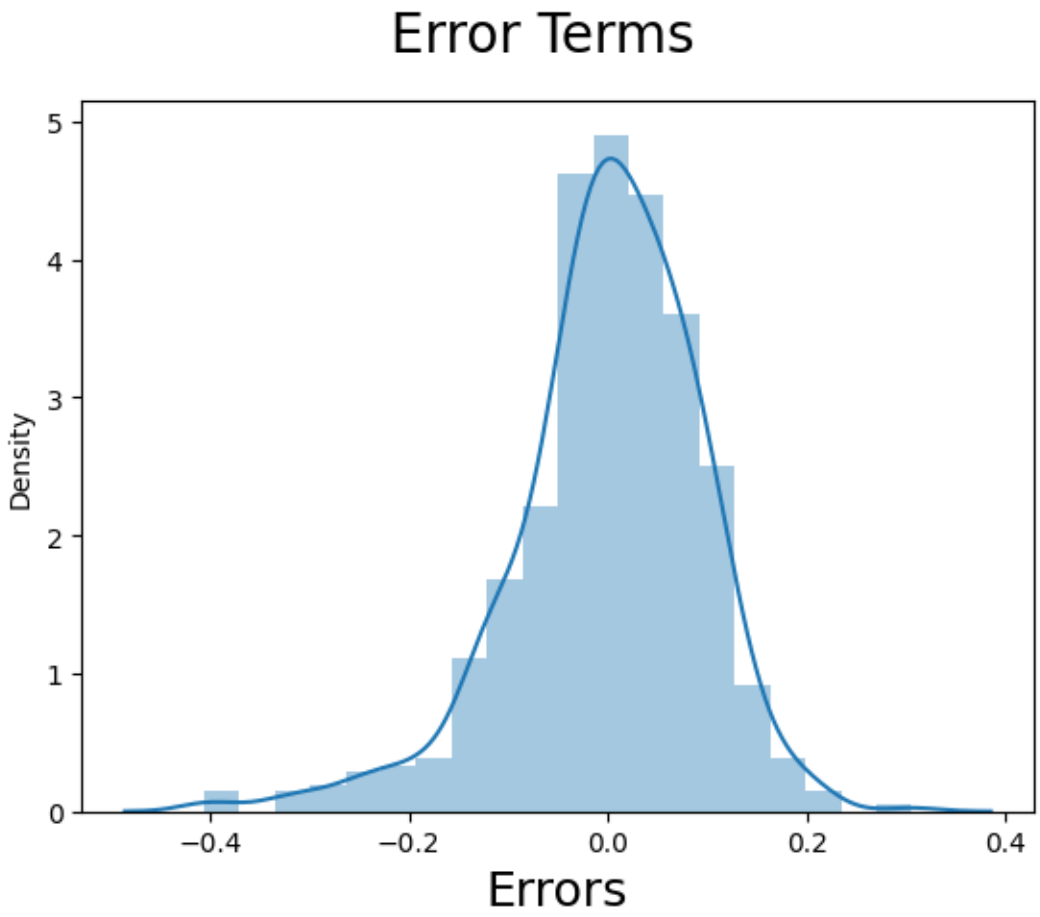
Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 4 goes below this line> (Do not edit)

This was done through residual analysis of the train data

To check if the error terms are also normally distributed (which is in fact, one of the major assumptions of linear regression), plotted the histogram of the error terms.



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)

Atemp, yr and to some extent, winter season, are the major contributors that explain the shared bike demand.

Higher the temperature, more is the demand.

Please refer to the coefs in the below summary.

OLS Regression Results

Dep. Variable:	cnt	R-squared:	0.823
Model:	OLS	Adj. R-squared:	0.819
Method:	Least Squares	F-statistic:	241.4
Date:	Tue, 28 Jan 2025	Prob (F-statistic):	1.11e-206
Time:	02:27:52	Log-Likelihood:	548.43
No. Observations:	584	AIC:	-1073.
Df Residuals:	572	BIC:	-1020.
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2058	0.028	7.249	0.000	0.150	0.262
yr	0.2234	0.008	27.825	0.000	0.208	0.239
atemp	0.6210	0.021	29.186	0.000	0.579	0.663
hum	-0.1913	0.037	-5.145	0.000	-0.264	-0.118
windspeed	-0.1272	0.023	-5.517	0.000	-0.172	-0.082
summer	0.0927	0.010	9.164	0.000	0.073	0.113
winter	0.1236	0.012	10.284	0.000	0.100	0.147
Mar	0.0380	0.017	2.278	0.023	0.005	0.071
Sep	0.1161	0.015	7.655	0.000	0.086	0.146
Oct	0.0563	0.017	3.389	0.001	0.024	0.089

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 outcomes. It models the relationship between a dependent variable (target) and one or more independent variables (features) by fitting a linear equation.

Linear Regression Equation

The linear regression equation takes the following form:

$y = \beta 0 + \beta 1x + \varepsilon$

where:

- y: dependent variable (target)
- x: independent variable (feature)
- β0: intercept or constant term
- β1: slope coefficient

- ε: error term (residual)

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 collection of four datasets created by statistician Francis Anscombe in 1973 to illustrate the importance of visualizing data before analyzing it. Each dataset consists of 11 points and appears to have similar statistical properties when viewed through summary statistics, but they reveal distinct patterns when plotted graphically.

Datasets

The quartet consists of four datasets, each with 11 data points:

1. Dataset 1: A simple linear relationship between x and y.
2. Dataset 2: A curved relationship between x and y.
3. Dataset 3: A linear relationship with an outlier.
4. Dataset 4: No clear relationship between x and y.

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 statistical measure that calculates the strength and direction of the linear relationship between two continuous variables.

Interpretation of Pearson's R:

- Values: Pearson's R ranges from -1 to 1.
- Direction: A positive value indicates a positive linear relationship, while a negative value indicates a negative linear relationship.
- Strength: The absolute value of Pearson's R indicates the strength of the linear relationship:
  - 0-0.3: Weak relationship
  - 0.3-0.6: Moderate relationship
  - 0.6-1: Strong relationship

Assumptions:

- Linearity: The relationship between the variables should be linear.
- Independence: Each observation should be independent of the others.
- Homoscedasticity: The variance of the residuals should be constant across all levels of the

independent variable.
- Normality: The residuals should be normally distributed.

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 a data preprocessing technique used to transform numerical data into a common range, usually between 0 and 1, to prevent differences in scales from affecting the analysis or modeling.

Scaling is performed to:

1. Prevent feature dominance: When features have different scales, those with larger ranges can dominate the analysis, masking the effects of other features.
2. Improve model performance: Many machine learning algorithms, such as neural networks and support vector machines, perform better when features are on the same scale.
3. Enhance interpretability: Scaled data can be easier to visualize and understand.

Normalized Scaling (Min-Max Scaling)

Normalized scaling, also known as min-max scaling, transforms data to a common range, usually between 0 and 1, using the following formula:

X_normalized = (X - X_min) / (X_max - X_min)

where X is the original value, X_min is the minimum value, and X_max is the maximum value.

Standardized Scaling (Z-Score Scaling)

Standardized scaling, also known as z-score scaling, transforms data to have a mean of 0 and a standard deviation of 1 using the following formula:

X_standardized = (X - μ) / σ

where X is the original value, μ is the mean, and σ is the standard deviation.

Key differences:

1. Range: Normalized scaling transforms data to a specific range (e.g., 0 to 1), while standardized scaling transforms data to have a specific mean and standard deviation.
2. Robustness to outliers: Standardized scaling is more robust to outliers, as it uses the mean and standard deviation, which are less affected by extreme values. Normalized scaling, on the other hand, uses the minimum and maximum values, which can be influenced by outliers.
3. Preservation of relationships: Standardized scaling preserves the relationships between

variables, while normalized scaling can distort these relationships.

In summary, normalized scaling is suitable when you want to transform data to a specific range, while standardized scaling is preferred when you want to preserve the relationships between variables and are concerned about the impact of outliers.

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>

An infinite VIF value indicates that:

1. We have redundant or duplicate predictor variables.
2. There is a linear dependency between two or more predictor variables.
3. We need to reassess the model and remove or transform the problematic variables to resolve the multicollinearity issue.

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 known as a Quantile-Quantile plot, is a graphical tool used to compare the distribution of two datasets. In the context of linear regression, a Q-Q plot is used to assess the normality assumption of the residuals.

Construction of a Q-Q plot:

---

1. Calculate the residuals from the linear regression model.
2. Sort the residuals in ascending order.
3. Calculate the quantiles (percentiles) of the residuals.
4. Plot the quantiles of the residuals against the quantiles of a standard normal distribution (Z-distribution).

Use of a Q-Q plot in linear regression:

---

1. Normality check: A Q-Q plot helps to verify if the residuals follow a normal distribution, which is a crucial assumption in linear regression. If the points on the Q-Q plot lie close to a straight line, it suggests that the residuals are normally distributed.
2. Detection of outliers: A Q-Q plot can help identify outliers or unusual observations in the data. Points that deviate significantly from the straight line may indicate outliers.

3. Assessment of skewness: A Q-Q plot can also reveal skewness in the residuals. If the points on the Q-Q plot curve upward or downward, it may indicate skewness.

Importance of a Q-Q plot in linear regression:

1. Ensures robustness: By verifying the normality assumption, a Q-Q plot helps ensure that the linear regression model is robust and reliable.
2. Prevents incorrect conclusions: A Q-Q plot can prevent incorrect conclusions that may arise from non-normal residuals.
3. Guides model selection: A Q-Q plot can inform the choice of regression model. For example, if the residuals are not normally distributed, alternative models like generalized linear models or transformation-based models may be more suitable.