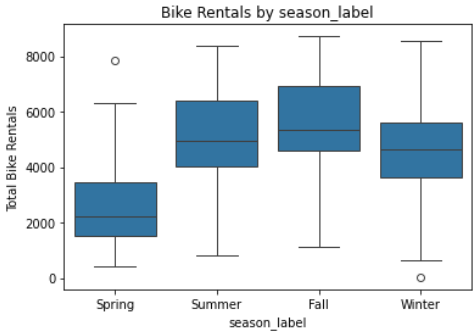
Assignment-based Subjective Questions

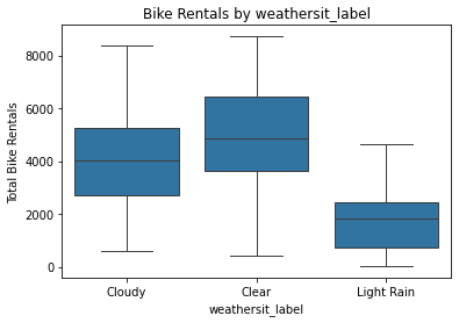
Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

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

The analysis which I've done examines categorical variables like season_label and weathersit_label and by using the box plots compared the distribution with the dependent variable which is 'cnt'





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)

To create dummy variables for the categorical columns using pg.get_dummies with drop_first=True and this eliminates one dummy variable per category to avoid dummy variable trap where multicollinearity arises in the linear regression due to the redundant information.

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)

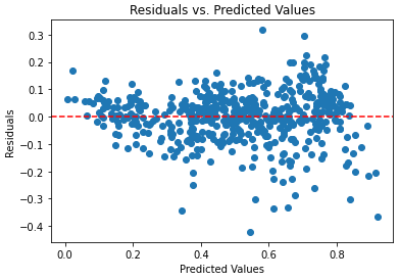
Using the pairplot it's visible that the temp and atemp are the numerical variables which are highly correlated with cnt

---

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

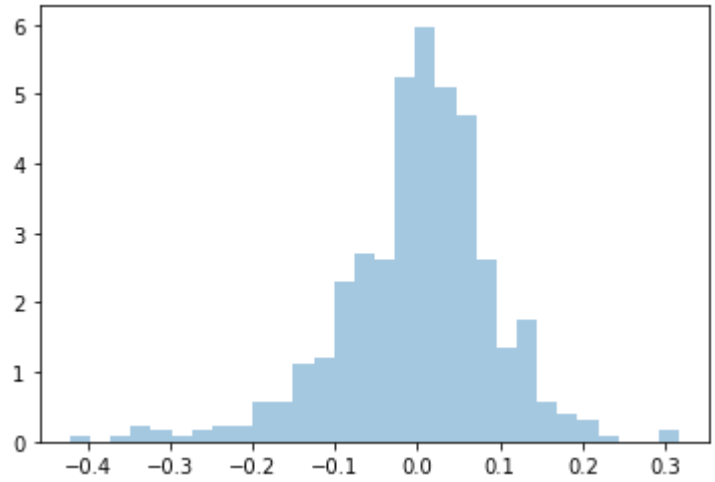
Total Marks: 3 marks (Do not edit)

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



As observed the residuals are randomly scattered around the red line but still there apprears to be a slight funnel shape or increase as the predicted values grow

Also The residuals should follow a normal distribution and be centered around zero (mean = 0). This assumption was validated by examining whether the residuals follow a normal distribution. The diagram below demonstrates that the residuals are indeed distributed around a mean of zero, satisfying this assumption.



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)
Thee top 3 features which are showing the most significant impact are
1.temp (coeff =0.5281)
2.hum (coeff = -0.2455)
3.yr (coeff = 0.2286)

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 simple and popular algorithm used to predict a continuous output variable (dependent variable) based on one or more input variables (independent variables). It tries to find a straight-line relationship between them.
The equation for linear regression looks like this:
$y=\beta0+\beta1X1+\beta2X2+\cdots+\beta nXn+\epsilon y$
y is the thing you want to predict
β0 is the intercept (the value of y when all X's are 0)
β0, β1, β2…. Bn are the coefficients
X1,X2,X3….Xn are the independent variables
The goal of linear regression is to find the best values for β0,β1,…,βn so that the line (or plane, in case of multiple variables) fits the data as closely as possible.
Steps to Perform Linear Regression

1. **Collect Data**: Gather the data you want to analyze. For example, let's say you're predicting house prices.
2. **Explore Data**: Check if there's a linear relationship between the variables by plotting them.
3. **Split Data**: Divide your data into a training set and a test set.
4. **Fit the Model**: Use the training data to calculate the coefficients (β0,β1,…) that minimize the error.
5. **Check Assumptions**: Validate assumptions like linearity, normality of errors, and no multicollinearity.
6. **Test the Model**: Use the test data to see how well the model predicts new values.

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 show the importance of visualizing data rather than relying only on summary statistics. Each dataset looks very different when plotted, but they have almost identical statistical properties.

The datasets illustrate how statistics like the mean, variance, correlation, and regression line can be misleading if you don't visualize your data.

The key characterstics are
Mean of x: same for all the datasets(mean = 9)
Mean of y: same for all the datasets(mean = 7.5)
Variance of x: same for all datasets
Variance of y: same for all datasets.
Correlation between x and y: same for all datasets
Regression line: same for all datasets

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 called the Pearson correlation coefficient, is a number that tells us how strong the relationship is between two variables and the direction of that relationship. It's like a score that shows how well two things are connected.

If R =1 it means that perfect positive relationship

If R=-1 perfect negative Relationship

R close to 1 or -1 strong relationship

R close to 0 = weak or no relationship

It is a handy tool to understand the connection between two variables and gives quick idea of whether the variables move together

---

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 is a data preprocessing technique used to adjust the range of features (variables) in a dataset. It ensures that all features have comparable scales, which is especially important for machine learning algorithms that are sensitive to the magnitude of data

Scaling is performed because of the Fair Comparison as this helps to ensure all features to contribute equally and improving the algorithm performance and avoiding bias in the distance metrics

The key difference between normalized scaling and standardized scaling is that normalization (min-max scaling) rescales data to a fixed range, typically [0, 1], based on the minimum and maximum values of the feature, making it sensitive to outliers, while standardization (z-score scaling) transforms data to have a mean of 0 and a standard deviation of 1, focusing on the spread of the data and being less sensitive to 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)

A Variance Inflation Factor (VIF) becomes infinite when there is perfect multicollinearity among independent variables, meaning one variable can be exactly predicted by a linear combination of others (R2=1).

Since VIF is calculated as 1/(1−R2), a perfect correlation makes the denominator zero, resulting in an infinite value. This typically occurs when redundant or duplicate variables exist in the dataset, and it can be resolved by removing one of the perfectly correlated variables, combining them, or

using techniques like Ridge Regression to mitigate multicollinearity.

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 plot (quantile-quantile plot) is a graphical tool used to compare the distribution of a dataset to a theoretical distribution, often a normal distribution. It plots the quantiles of the dataset against the quantiles of the reference distribution. If the points lie approximately along a straight diagonal line, it indicates that the data follows the reference distribution.
In a Q-Q plot, the quantiles of the residuals are plotted against the quantiles of a theoretical normal distribution. If the points in the plot fall approximately along a straight line, it suggests that the residuals are normally distributed, supporting the validity of the linear regression model. Deviations from the straight line indicate potential issues, such as non-normality, which could affect the reliability of the model's results

---