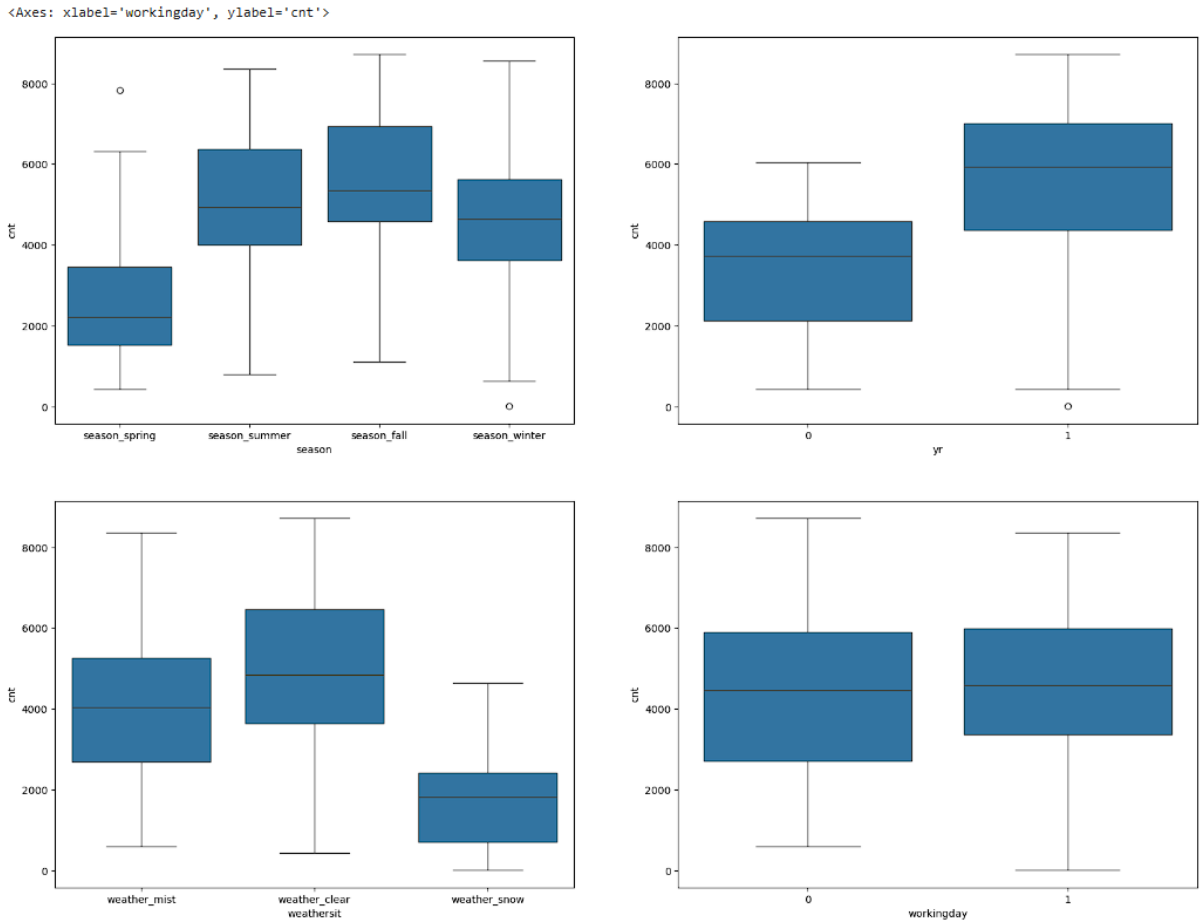
Assignment-based Subjective Questions

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)



From the figure above we can note that in season, spring has the lowest demand while fall has highest demand on average. There is growth in 2019 compared to 2018. Also, clear conditions are most favourable for bike rental.

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)

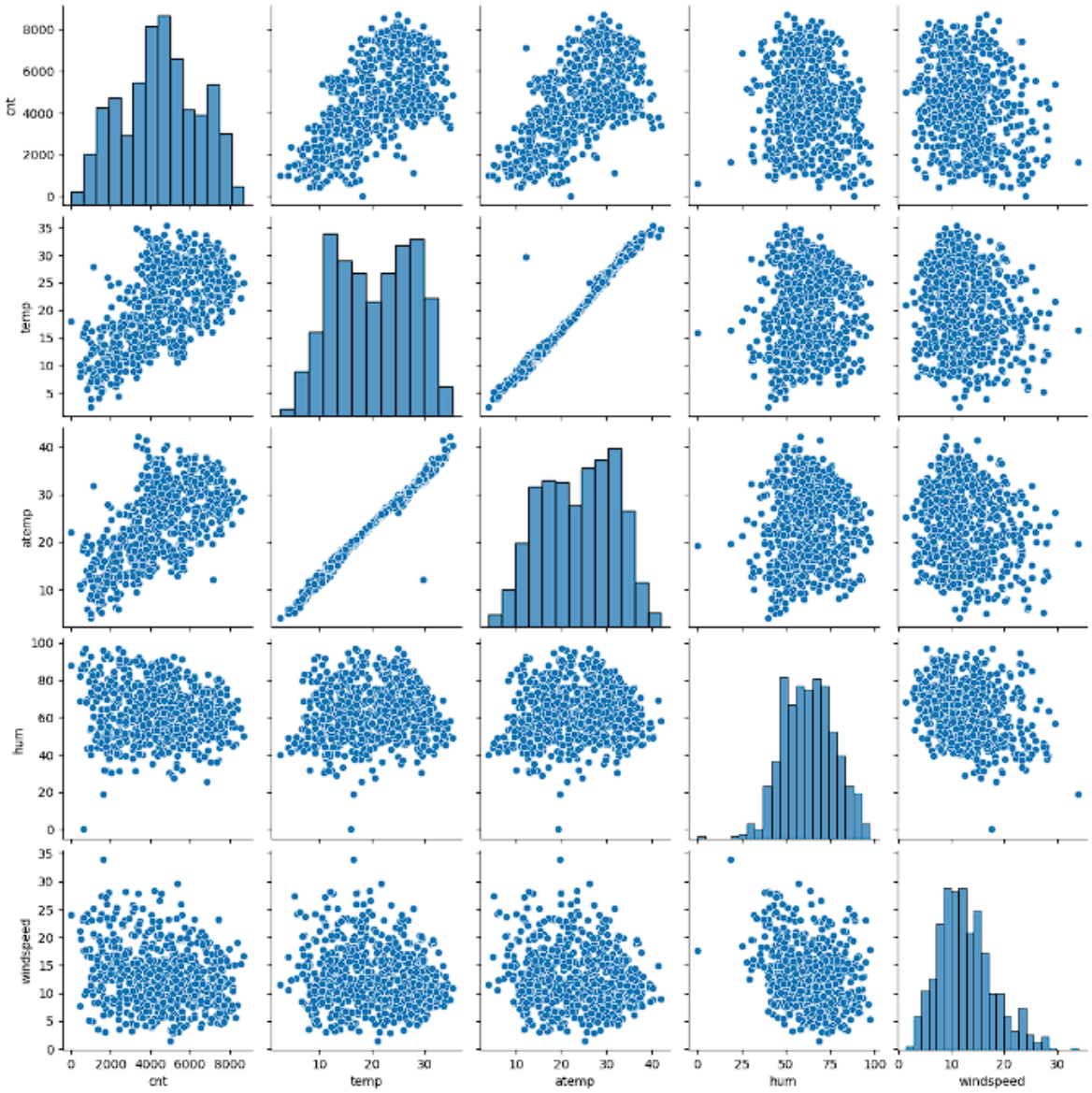
Answer: <Your answer for Question 2 goes below this line> (Do not edit)

drop_first=True should be used since it reduces the number of unnecessary columns that are created when creating dummy variables. As a result, it lessens the correlations that are formed between 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)



Temp and atemp look like they have the highest correlation to target variable 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)

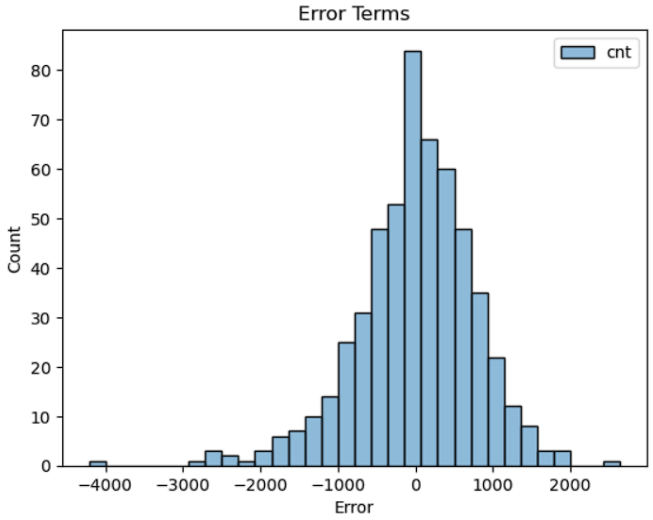
OLS Regression Results

Dep. Variable:	cnt	R-squared:	0.840
Model:	OLS	Adj. R-squared:	0.836
Method:	Least Squares	F-statistic:	200.1
Date:	Wed, 30 Oct 2024	Prob (F-statistic):	1.90e-201
Time:	18:01:26	Log-Likelihood:	-4415.2
No. Observations:	547	AIC:	8860.
Df Residuals:	532	BIC:	8925.
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2552.3667	222.770	11.457	0.000	2114.750	2989.983
yr	2038.1948	67.733	30.092	0.000	1905.138	2171.251
holiday	-714.4835	202.232	-3.533	0.000	-1111.754	-317.213
temp	3382.5089	273.532	12.366	0.000	2845.174	3919.844
windspeed	-770.8180	190.039	-4.056	0.000	-1144.137	-397.499
season_spring	-1321.2239	134.850	-9.798	0.000	-1586.127	-1056.321
season_winter	696.0610	115.037	6.051	0.000	470.078	922.044
weather_mist	-702.2614	72.628	-9.669	0.000	-844.933	-559.589
weather_snow	-2288.2411	200.649	-11.404	0.000	-2682.403	-1894.079
day_sun	-373.0120	94.122	-3.963	0.000	-557.909	-188.115
month_dec	-591.4729	145.316	-4.070	0.000	-876.936	-306.010
month_jul	-437.4979	135.005	-3.241	0.001	-702.706	-172.290
month_mar	403.3722	144.207	2.797	0.005	120.088	686.657
month_nov	-693.0204	151.377	-4.578	0.000	-990.391	-395.650
month_sep	427.3799	125.845	3.396	0.001	180.165	674.594

Omnibus:	73.042	Durbin-Watson:	2.011
Prob(Omnibus):	0.000	Jarque-Bera (JB):	166.849
Skew:	-0.720	Prob(JB):	5.88e-37
Kurtosis:	5.291	Cond. No.	16.0

	Features	VIF
2	temp	5.47
3	windspeed	5.04
5	season_winter	2.45
0	yr	2.08
4	season_spring	1.82
12	month_nov	1.82
6	weather_mist	1.57
10	month_jul	1.45
9	month_dec	1.34
13	month_sep	1.23
8	day_sun	1.19
11	month_mar	1.17
7	weather_snow	1.11
1	holiday	1.06



In the result obtained, the p-values are all below 0.05, VIF values are close to or below 5. Also, the

error terms are normally distributed.

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 model, the top 3 features contributing significantly are –

1. Temperature
2. Year
3. Season

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 fundamental statistical method used to model the relationship between a dependent variable (target) and one or more independent variables (predictors). The goal is to find the best-fitting linear equation that describes how changes in the independent variables affect the dependent variable.

Model Equation: In simple linear regression with one independent variable, the model is represented as: y=mx+b

where:

- y is the predicted value,
- mmm is the slope (coefficient),
- xxx is the independent variable,
- b is the y-intercept.

In multiple linear regression, the equation extends to multiple predictors:

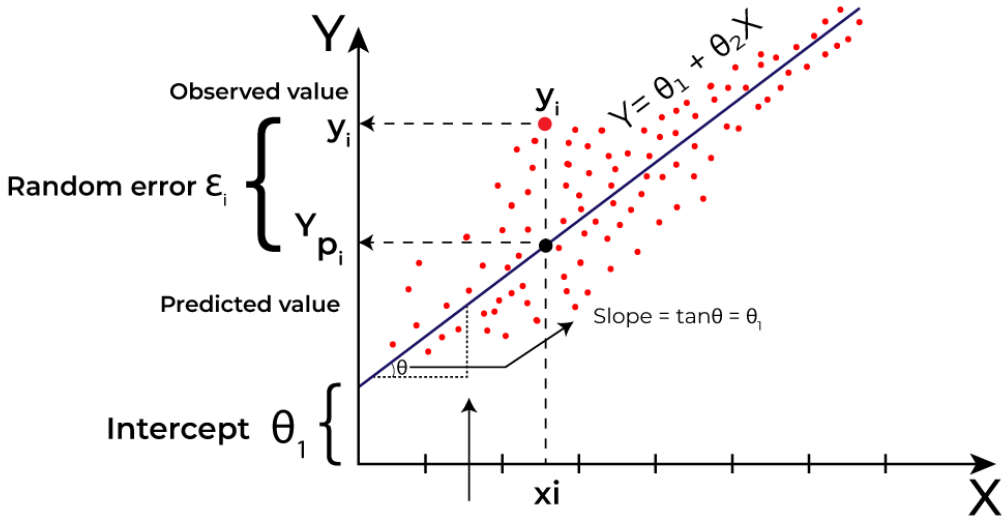
y=b0+b1x1+b2x2+…+bnxny

Objective: The algorithm aims to minimize the sum of the squared differences (residuals) between the observed and predicted values. This is achieved using the least squares method.

Linear regression relies on several key assumptions:

- Linearity: The relationship between the dependent and independent variables is linear.
- Independence: Residuals are independent of each other.
- Homoscedasticity: Constant variance of residuals across all levels of the independent variable.
- Normality: Residuals are normally distributed.

After fitting the model to training data, the coefficients are estimated, allowing for predictions on new data. The model's effectiveness is evaluated using metrics like R-squared, which indicates the proportion of variance explained by 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)

Anscombe's quartet is a set of four datasets. Each dataset consists of 11 pairs of x and y values and has nearly identical statistical properties, including the same means, variances, and correlation coefficients.

Datasets:

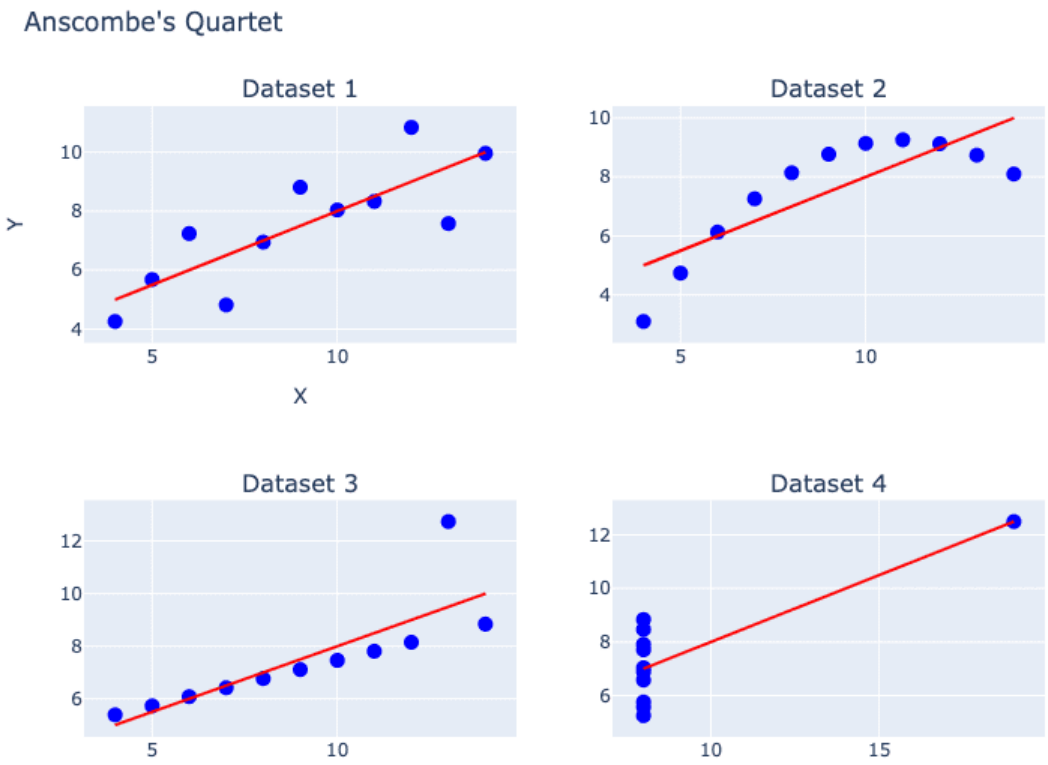
- Dataset I: A classic linear relationship.
- Dataset II: A quadratic relationship.
- Dataset III: A linear relationship with an outlier that heavily influences the slope.
- Dataset IV: A vertical line indicating no variation in xxx but a wide spread in y.

Statistical Similarities: All four datasets have:

- Similar means of x and y (mean of x is 9, mean of y is 7.5).
- Identical variances for x and y.
- The same correlation coefficient (r≈0.816r).

When plotted, the datasets reveal dramatically different relationships, showcasing that statistical summaries can obscure underlying patterns. Anscombe's quartet emphasizes the necessity of visualizing data to uncover insights that summary statistics might hide. It serves as a cautionary tale against the misuse of statistical methods without thorough exploratory data analysis, reminding

statisticians and analysts that context and data visualization are crucial for understanding relationships within data.



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, or Pearson correlation coefficient, is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It ranges from −1 to 1:

- r=1: Perfect positive correlation; as one variable increases, the other also increases.
- r=−1: Perfect negative correlation; as one variable increases, the other decreases.
- r=0: No correlation; changes in one variable do not predict changes in the other.

Calculation:

Pearson's R is calculated using the formula:

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,][\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

where n is the number of paired observations, x and y are the individual data points.

To accurately use Pearson's r, certain assumptions must be met:

1. Linearity: The relationship between the variables should be linear.
2. Normality: Both variables should be approximately normally distributed.
3. Homoscedasticity: The variance of residuals should be constant across all levels of the independent variable.

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 the process of transforming features in a dataset to a similar range or distribution, which is essential for improving the performance of machine learning algorithms. It helps prevent features with larger values from dominating the model's influence. Common scaling methods include normalized scaling (Min-Max scaling), which rescales data to a specific range (usually [0, 1]), and standardized scaling (Z-score scaling), which centers data around the mean with a standard deviation of 1. Scaling enhances model accuracy and convergence speed, especially for distance-based algorithms.

Normalized Scaling	Standardized Scaling
Rescales features to a specific range, typically [0, 1].	Centers data around the mean with a standard deviation of 1.
[0, 1] (or any specified range)	Mean = 0, Standard Deviation = 1
Sensitive; outliers can skew the scaling.	Less sensitive; outliers affect the mean and standard deviation but not as dramatically.
Best for bounded data, like image pixel values.	Best for normally distributed data, especially in algorithms like PCA.

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)

When creating the dummy variables in one hot encoding a categorical variable a VIF of infinite was observed. Variance Inflation Factor (VIF) measures how much the variance of a regression coefficient is inflated due to multicollinearity among independent variables.
A VIF value is considered infinite when one of the independent variables is a perfect linear combination of other variables. This happens because in case of perfect correlation, we get R2=1 which makes 1/(1-R2) tend to infinity. This indicates perfect multicollinearity.
So, the presence of all one hot encoded columns causes the VIF for all the columns to become infinite. To solve this, any one column can be dropped. This does not lead to a data loss as all the other columns having 0 as their values show that the last column would have a one.

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 specific theoretical distribution, most commonly the normal distribution. It plots the quantiles of the observed data against the quantiles of the expected distribution. If the points on the plot fall approximately along a straight diagonal line, it indicates that the data adheres to the specified distribution.

In the context of linear regression, a Q-Q plot is particularly useful for checking the normality of residuals, which is one of the key assumptions of the model. By visualizing the quantiles of the residuals against those of a normal distribution, analysts can identify deviations from normality, such as skewness or kurtosis. Points that significantly deviate from the diagonal line suggest that the residuals are not normally distributed, potentially compromising the validity of hypothesis tests and confidence intervals derived from the model. Ensuring that residuals are normally distributed helps validate the regression model, leading to more reliable predictions and inferences. Additionally, if a Q-Q plot indicates non-normality, it can guide the analyst toward necessary data transformations or alternative modeling techniques. Overall, Q-Q plots provide a clear visual representation of the distribution of residuals, enhancing the robustness of the linear regression analysis.