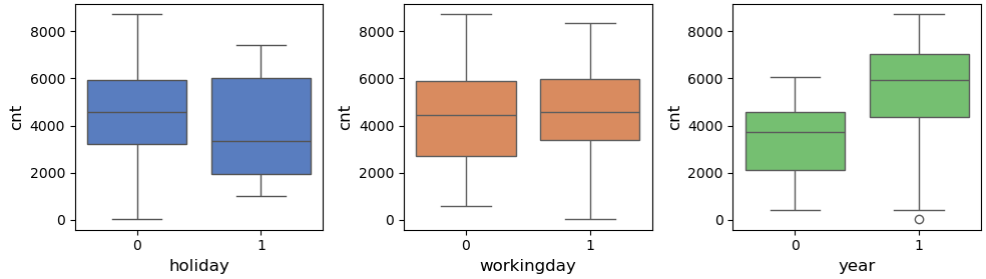
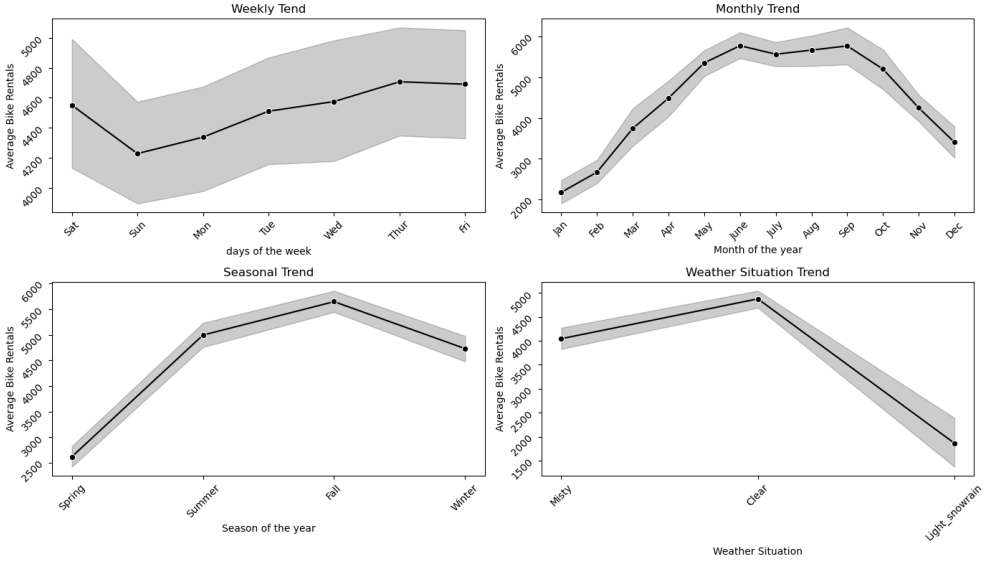
Assignment-based Subjective Questions

Question 1: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?



- **Working Day**: Bike users are consistent across working and non-working days, with median usage around 4,000 rides.
- **Year**: Bike usage rose significantly in 2019 compared to 2018, indicating growing users over the years.
- **Holiday**: Bike usage drops noticeably on holidays.



- **Season**: Fall has the highest bike usage, while spring has the least.
- **Month**: Bike usage occur from May to September. Usage rises early in the year, peaks mid-year, and declines towards year-end.
- **Weekend**: Sundays have the lowest users, with a steady increase from Monday through Saturday.
- **Weather Situation**: Clear weather sees the highest bike rentals, while light snow/rain significantly reduces usage.

Question 2: Why is it important to use drop_first=True during dummy variable creation?

- It is important to use `drop_first=True` during dummy variable creation to **avoid multicollinearity**.
- Multicollinearity occurs when two or more predictor variables in a regression model are highly correlated, leading to unstable and unreliable estimates of the model coefficients.
- By dropping the first dummy variable, we ensure that the remaining dummy variables are linearly independent, preventing multicollinearity and improving the model's stability and interpretability

$$n - 1$$

Where,

n = Number of categories in a variable

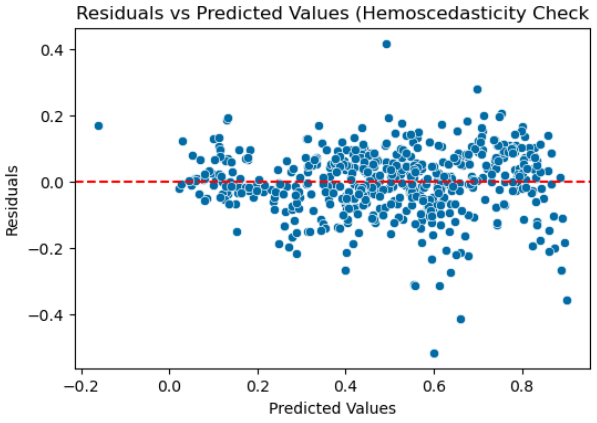
Question 3: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

From the pair plot, the variable that has the highest correlation with the target variable `cnt` appears to be `temp` and `feels_like_temp` based on the scatter plot.

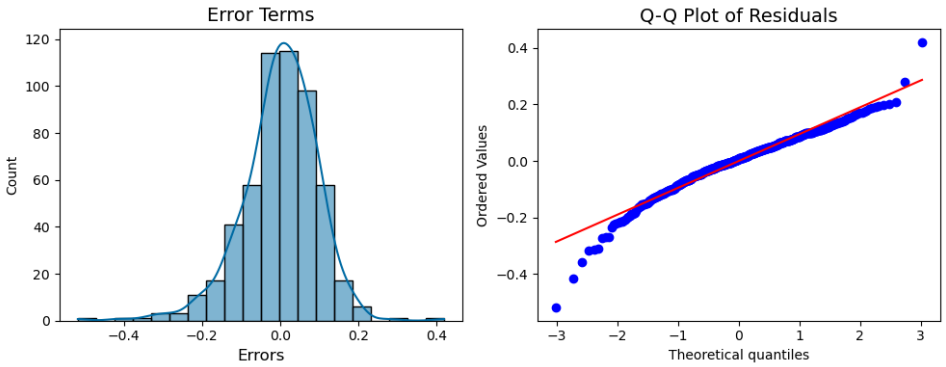
Question 4: How did you validate the assumptions of Linear Regression after building the model on the training set?

To validate the assumptions of Linear Regression on the training set, I checked the following:

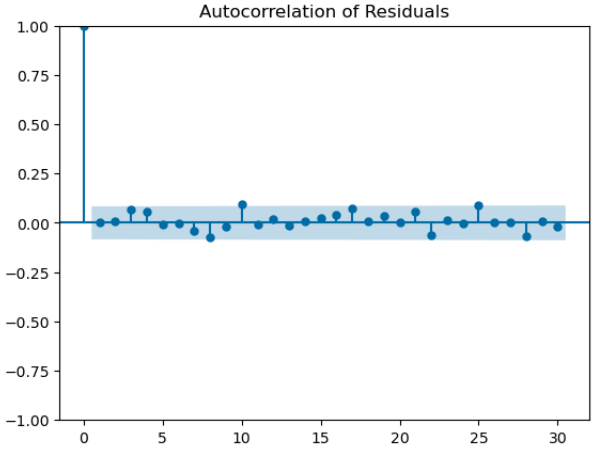
1. **Linearity**: Plotted residuals vs. predicted values to confirm a random distribution, indicating that the linear model captures the relationship correctly.
2. **Homoscedasticity**: Checked residuals for constant variance across predicted values, using a residuals vs. fitted plot; patterns here indicate heteroscedasticity.



3. **Normality of Residuals**: Normality of residuals is **the assumption that the residuals in a regression model are normally distributed**. Residuals are the errors in the relationship between the independent and dependent variables in a regression model. Using histplot, checked if residuals follow a normal distribution, validating the normality assumption. and Using Q-Q plot examining data distributions.



4. **Checking Independence of Residuals:** To check if residuals are independent (uncorrelated), use an autocorrelation plot. Autocorrelation refers to the correlation between observations within a time series. We can understand it as how each data point is related to lagged data points in a sequence.



Question 5: Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes:

- **Previous Day's Demand:** Both models consistently identified previous day's demand (previous_day_diff) as a primary driver, suggesting that usage trends and demand continuity play a key role in predicting future demand.
- **Temperature and Heat Index:** Model 4 highlighted temperature, while Model 5 introduced the heat index as a strong indicator of demand, indicating that comfortable weather conditions (moderate temperatures) are conducive to higher usage.
- **Year-on-Year Increase:** The year variable was significant in both models, reflecting potential growth in market acceptance or expanding user adoption over time.

General Subjective Questions

Question 6: Explain the linear regression algorithm in detail.

Linear regression is a data analysis technique that predicts the value of unknown data by using another related and known data value. It mathematically models the unknown or dependent variable and the known or independent variable as a linear equation. The basic formula for linear Algebra for straight line is:

$$Y = mX + c$$

Y = Dependent variable
X = independent variable
m = Slope of the linear line.
C = is a constant, Known as intercept

It is one of the easiest and most popular Machine learning algorithms.

Types of Linear Regression

Simple Linear Regression

This is the simplest form of linear regression, and it involves only one independent variable and one dependent variable. The equation for simple linear regression is:

$$y = \beta_0 + \beta_1 X$$

Where:
Y = Dependent variable
X = Independent variable
β1 = Slope of the regression line
β0 = is a constant, Known as intercept

Multiple Linear Regression

This involves more than one independent variable and one dependent variable. The equation for multiple linear regression is:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

where:
Y is the dependent variable
X1, X2, ..., Xn are the independent variable
β0 is the intercept
β1, β2, ..., βn are the slopes

The goal of the algorithm is to find the best Fit Line equation that can predict the values based on the independent variables.

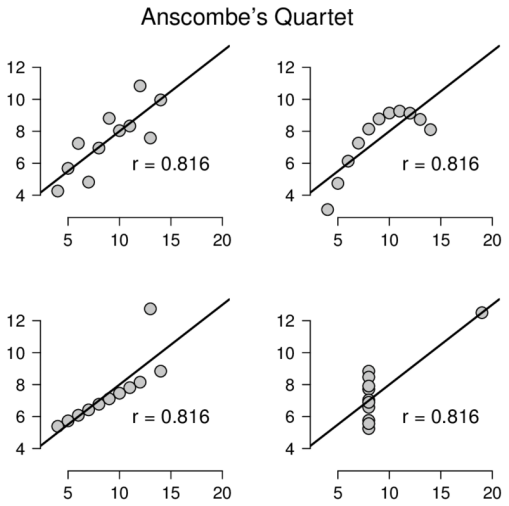
Question 7. Explain the Anscombe's quartet in detail.

Anscombe's Quartet is a collection of four datasets that have nearly identical statistical properties (such as mean, variance, correlation, and linear regression line) but appear very different when graphed. This set was created by statistician Francis Anscombe in 1973 to illustrate the importance of data visualization.

1. **Dataset Properties**: Each dataset has the same mean and variance for both X and Y values, the same correlation between X and Y, and an identical linear regression line equation.

	I	II	III	IV
Mean_x	9.000000	9.000000	9.000000	9.000000
Variance_x	11.000000	11.000000	11.000000	11.000000
Mean_y	7.500909	7.500909	7.500000	7.500909
Variance_y	4.127269	4.127269	4.122620	4.127269
Correlation	0.816421	0.816237	0.816287	0.816521

1. **Visualization Differences**: Despite the identical statistics, the datasets exhibit very different patterns when plotted. For example, one dataset forms a perfect line, another has a clear curve, one has a distinct outlier, and one forms a scatter without linear correlation. These variations show that relying solely on summary statistics can be misleading.



2. **Key Insight**: Anscombe's Quartet emphasizes the need for graphical analysis to understand data structure and detect outliers or patterns that summary statistics alone cannot reveal. This underscores the importance of combining statistical analysis with visualization for accurate data interpretation.

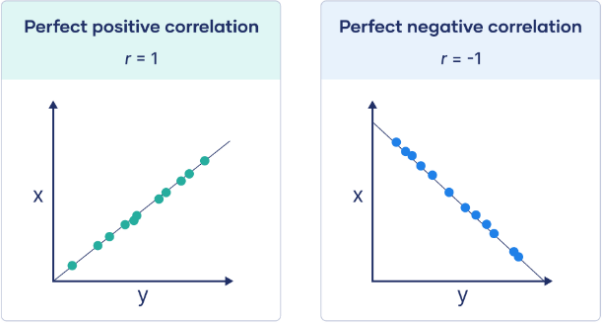
Question 8. What is Pearson's R?

The Pearson correlation coefficient (R) is the most widely used correlation coefficient. The Pearson correlation coefficient is a descriptive statistic, meaning that it summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables.

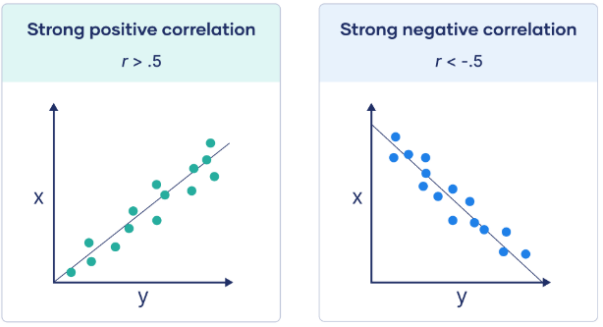
Pearson correlation coefficient (*r*) value	Strength	Direction
Greater than .5	Strong	Positive
Between .3 and .5	Moderate	Positive
Between 0 and .3	Weak	Positive
0	None	None
Between 0 and –.3	Weak	Negative
Between –.3 and –.5	Moderate	Negative
Less than –.5	Strong	Negative

Visualizing the Pearson correlation coefficient

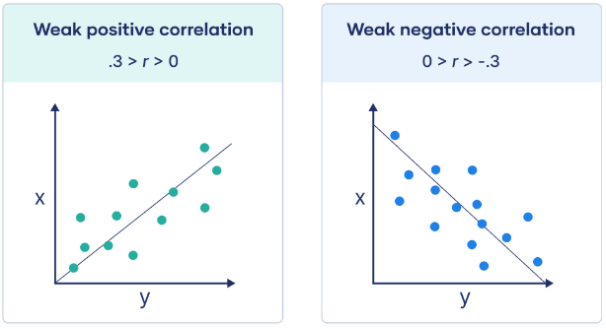
When R is 1 or –1, all the points fall exactly on the line of best fit:



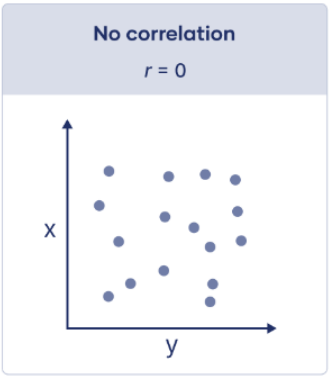
When R is greater than .5 or less than –.5, the points are close to the line of best fit:



When R is between 0 and .3 or between 0 and –.3, the points are far from the line of best fit:



When R is 0, a line of best fit is not helpful in describing the relationship between the variables:



Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling is the process of transforming data features so they fit within a specific range or distribution. In machine learning, scaling is essential because algorithms that rely on distance calculations (such as k-nearest neighbors, support vector machines, and neural networks) can be sensitive to the magnitude of the features. Unscaled features can lead to biased or inaccurate model performance if one feature dominates due to its scale.

Why Scaling is Performed

1. **Improving Model Performance**: Scaling ensures that all features contribute equally to the model, which is especially crucial for models that calculate distances or gradients.

2. **Faster Convergence**: In algorithms like gradient descent, scaled features help achieve faster convergence since steps are more uniform across dimensions.
3. **Better Interpretability**: Scaled data makes it easier to interpret and compare feature effects, which is helpful in certain analysis techniques.

Types of Scaling

1. **Normalized Scaling (Min-Max Scaling)**:
   - **Definition**: This method scales the data to a fixed range, usually [0,1] or [−1,1]. Each value xi is transformed as:

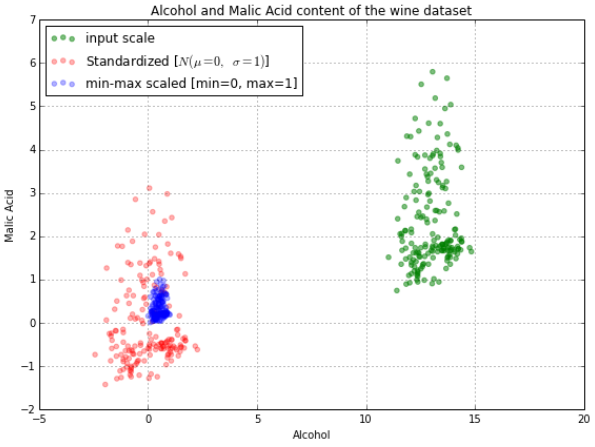
$$x' = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

   - **When to Use**: Useful when you want data within a fixed boundary (e.g., pixel values in image processing). It is often used when you know the minimum and maximum values are consistent across training and testing data.
   - **Effect on Outliers**: Normalization is sensitive to outliers, as they can skew the range.
2. **Standardized Scaling (Z-score Normalization)**:
   - **Definition**: This method scales data based on the mean and standard deviation of each feature. Each value xi is transformed as:

$$x' = \frac{x_i - \mu}{\sigma}$$

     where,
     μ = mean
     σ = standard deviation of the feature.
   - **When to Use**: Commonly used in algorithms that assume normally distributed data or when working with features with different units. Standardization does not bound values to a specific range.
   - **Effect on Outliers**: Less sensitive to outliers than min-max scaling, though extreme outliers can still influence the mean and standard deviation.

difference between normalized scaling and standardized scaling



One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers. scaling clusters all the data very close together, which may not be what you want. It might cause algorithms such as gradient descent to take longer to converge to the same solution they would on a standardized data set, or it might even make it impossible.

Summary of Differences

Scaling Type	Range	Sensitive to Outliers	Best for
Normalization	[0, 1] or [-1, 1]	Yes	Bounded values, known range
Standardization	Unbounded	Less sensitive	Normally distributed data, features with varying units

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

The **Variance Inflation Factor (VIF)** measures how much the variance of a regression coefficient is inflated due to multicollinearity in the data. When the VIF of a feature is `infinite`, it indicates `perfect multicollinearity` between that feature and one or more other features.

Formula for calculating the VIF:

$$VIF_i = \frac{1}{1 - R_i^2}$$

Why VIF is Infinite:

	Feature	VIF
10	weathersit_2	inf
15	weekday_2	inf
2	holiday	inf
3	workingday	inf
18	weekday_5	inf
17	weekday_4	inf
6	weathersit_2	inf
7	weathersit_2	inf
8	weathersit_3	inf
9	weathersit_3	inf
16	weekday_3	inf
11	weathersit_2	inf
12	weathersit_3	inf
13	weathersit_3	inf
14	weekday_1	inf
0	const	17.373457
5	windspeed	1.054732
4	atemp	1.048255

1. **Perfect Linear Dependence**: An infinite VIF occurs when a feature is an exact linear combination of other features. In other words, one feature can be perfectly predicted by the others.
2. **Deterministic Relationships**: If two or more features have a deterministic relationship (like total and sub-totals or redundancy), VIF cannot be computed reliably, as the regression matrix becomes **singular**, causing division by zero in the VIF formula.
3. **Implication**: Infinite VIF signals that the model cannot estimate unique regression coefficients for the perfectly collinear features, indicating a need to remove or transform some features to resolve multicollinearity.

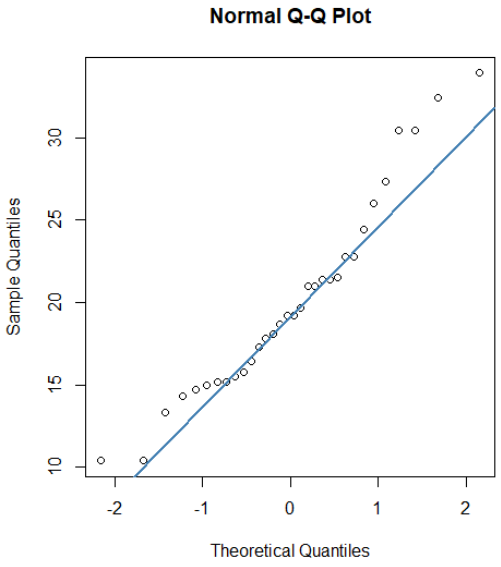
Observation:

Perfect Multicollinearity (Infinite VIF):

- Many features, particularly `weathersit_2`, `weathersit_3`, and various `weekday` indicators, have **infinite VIF** values. This implies **perfect multicollinearity** among these features, meaning that each of these features can be exactly or nearly exactly predicted by a linear combination of other features in the dataset.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

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**. In a Q-Q plot, the quantiles of the sample data are plotted against the quantiles of the theoretical distribution.



Use and Importance in Linear Regression:

1. **Assessing Normality**: In linear regression, residuals are ideally normally distributed. A Q-Q plot helps visualize whether the residuals deviate from normality. Points should ideally fall along a straight line if they follow a normal distribution.
2. **Detecting Outliers and Skewness**: Any substantial deviations from the line indicate issues such as skewness, heavy tails, or outliers, which can impact the reliability of regression estimates.
3. **Model Validity**: Since many statistical tests in linear regression (e.g., confidence intervals, p-values) assume normality, checking residuals with a Q-Q plot is essential to validate the model assumptions and ensure accurate interpretation of results.