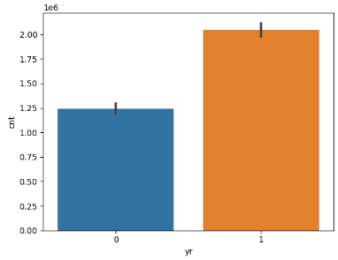
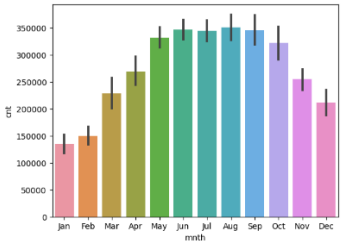
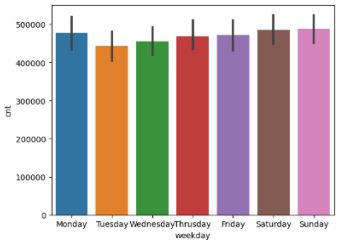
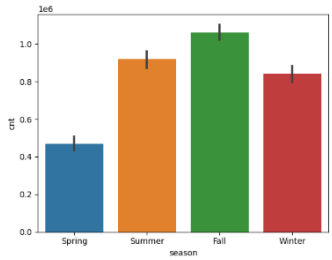
Assignment-based Subjective Questions

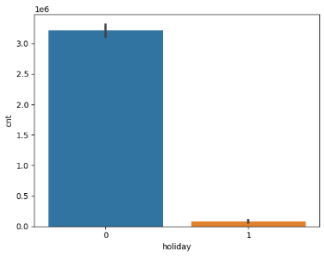
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

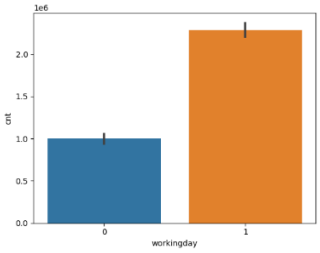


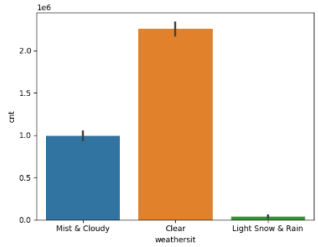












From the bar plots visualizing the relationship between categorical variables and the dependent variable cnt, we can infer the following:

1. **Year**: The year (2019) has a significantly higher count compared to the first year (2018), indicating an increase in the dependent variable over time.
2. **Season:** Fall and Summer show the highest counts, followed by Winter and then Spring, suggesting that the dependent variable is more influenced during warmer seasons (possibly due to favorable weather for outdoor activities like biking).
3. **Month:** June, July, and August (summer months) exhibit the highest counts, while colder months like January, February, and December show lower values. This suggests a strong seasonal trend in the dependent variable, with summer months being the most active.
4. **Holiday:** There is a noticeable drop in counts on holidays (1), while non-holidays (0) have significantly higher activity, implying that the dependent variable is less active on holidays.
5. **Weekday:** The counts are relatively consistent across all weekdays, with no significant spikes on any particular day. This suggests that the dependent variable does not vary much based on the day of the week.

6. **Working day:** Working days (1) show a substantially higher count than non-working days (0), implying that the dependent variable is more influenced by work-related activities (likely commuting).
7. **Weather Situation:** Clear weather results in the highest counts, followed by mist and cloudy conditions. Light snow and rain drastically reduce the counts, indicating that bad weather negatively impacts the dependent variable.

Overall Inference:

The categorical variables such as **year, season, month, working day, and weather situation** have a significant impact on the dependent variable, with activity peaking during favorable weather conditions, in warmer months, and on working days.

Holidays and adverse weather conditions tend to reduce the dependent variable considerably.

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

drop_first=True helps in reducing the extra column created during the dummy variable creation and hence avoid redundancy of any kind. Also, it helps in reducing the multicollinearity and also keeps the data simple, increasing interpretability.

If all are 0's then it would imply that it's a Friday. Hence, only 6 columns are created instead of 7 for Weekdays category.

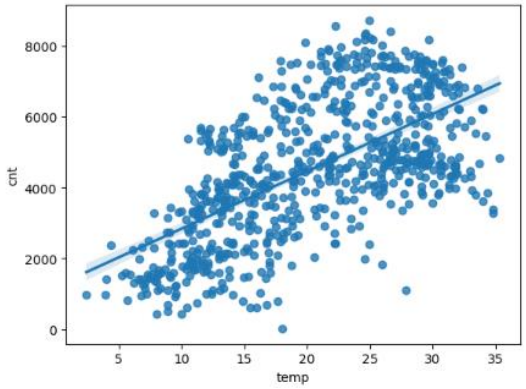
```
weekday_dummy.head()
```

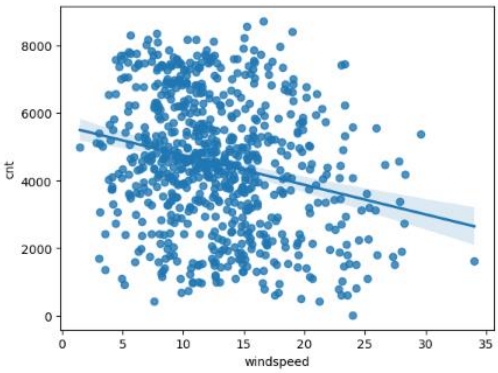
	Mon	Sat	Sun	Thu	Tue	Wed
0	0	1	0	0	0	0
1	0	0	1	0	0	0
2	1	0	0	0	0	0
3	0	0	0	0	1	0
4	0	0	0	0	0	1

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

From the visual analysis of the scatter plots, it appears that **temp & atemp** has the strongest positive correlation with the target variable **cnt** (count). The upward trend in the first plot indicates that as temperature increases, the count of rentals increases.

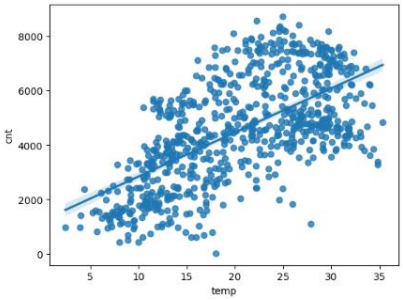
In contrast, **humidity (hum)** and **windspeed** do not show strong correlations. Humidity has a slight negative correlation, as indicated by the flat or slightly downward trend, while windspeed also shows a slight negative correlation.





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

- To determine if the assumption is met or not we create a scatter plot for X vs Y graph. The data points should have a straight line in the graph to confirm there is a linear relationship between the dependent and independent variables.



- There's no multicollinearity in the data. It can be checked via VIF and correlation matrix.

	Features	VIF
2	temp	4.60
3	windspeed	4.00
0	yr	2.06
6	Spring	1.65
9	Mist & Cloudy	1.51
7	Winter	1.40
4	Jul	1.35
5	Sep	1.20
8	Light Snow & Rain	1.08
1	holiday	1.04

- There Should be Homoscedasticity Among the Data. The data is said to be homoscedastic when the residuals are equal across the line of regression. In other words, the variance is equal. It can be checked

by plotting the residuals vs. fitted (predicted) values. If the spread of residuals is consistent across all fitted values, the homoscedasticity assumption is valid.

- Ensure that the residuals are normally distributed. This can be validated by plotting graphs of residuals like histogram or Q-Q plot.

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

The top 3 features contributing significantly to explaining the demand for shared bikes are:

- Temperature (temp)
    - Coefficient: 0.4515, P-value: 0.000
    - Temperature has the largest positive effect on the bike demand. For each unit increase in temperature, the demand increases by 0.4515, making it the most influential variable in the model.
- Year (yr)
    - Coefficient: 0.2341, P-value: 0.000
    - The second most important factor is the year. The positive coefficient suggests that in the second year (2019) the bike demand significantly increased by 0.2341 units on average compared to the first year.
- Windspeed:
    - Coefficient: -0.1398, P-value: 0.000
    - Windspeed has a negative effect on bike demand. As windspeed increases, the demand decreases by 0.1398 units. Though the effect is negative, it is still significant in influencing demand.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Regression analysis is a form of predictive modelling technique investigating the relationship between a dependent and independent variable. Regression is one of the types of Machine Learning Algorithms.

Linear Regression is one of the types of regression Models which is a statistical model used to determine the relationship between dependent variable (target) and one or more independent variables (features).

One needs to determine the best-fit line or a linear equation between observed data points.
The linear equation:

$$y=\beta0+\beta1x+\epsilon$$

Where:

- y is the predicted value (dependent variable/ target variable).
- x is the independent variable (predictor/ features).
- β0 is the intercept (value of y when x=0).
- β1 is the slope of the line (indicating how much y changes for a unit change in x).
- ϵ is the error term (the difference between actual and predicted values).

The main goal of linear regression is to minimize the sum of squared errors (differences between actual and predicted values). This is achieved using techniques like Ordinary Least Squares (OLS).

Here, there are two linear regression models, i.e., Simple Linear Regression and Multiple Linear Regression.

Simple Linear Regression models the relationship between two variables. Having an equation shown above i.e., $y=\beta0+\beta1x+\epsilon$

Assumptions made in simple linear regression:

- target & input variables are linear dependent i.e., linear relation between x and y.
- Error Terms are normally distributed
- Error Terms are independent of each other
- Error terms have constant variance (Homoscedasticity).
-

In **Multiple Linear Regression**, more than one independent variable is used, and the equation becomes:

$y=\beta0+\beta1x1+\beta2x2+\cdots+\beta nxn+\epsilon$

Interpretation of coefficients in the above equation:Change in the mean reasonce per unit increase in the variable when other predictors are held constant.

Assumptions made in multiple linear regression:

- Model now fits in hyperplave instead of a line.
- Coefficients are still obtained by maintaining the sum of squared errors.

- For Inferences, the assumptions of simple Linear Regression still holds

Few Considerations for MLR:

- Model may 'overfit' by becoming too complex
- Multicollinearity.
- Feature selection is important aspect

Linear regression, both simple and multiple, is widely used due to its simplicity and interpretability. It provides a clear understanding of how dependent and independent variables are related.

2. Explain the Anscombe's quartet in detail.

Anscombe's Quartet is an example used to highlight the importance of exploratory data analysis (EDA) and the limitation is that it solely relies on summary statistics. It displays that data visualization plays a critical role in uncovering patterns, outliers, and key insights that might otherwise be missed if one only focuses on statistical metrics.

Anscombe's Quartet consists of four distinct datasets, each having nearly identical summary statistics, such as mean, variance, and correlation. Although their statistical metrics are identical, they exhibit strikingly different distributions and relationships when visualized.

Let's take an example:

Dataset I:

- x-values: [10.0, 8.0, 13.0, 9.0, 11.0, 14.0, 6.0, 4.0, 12.0, 7.0, 5.0]
- y-values: [8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68]
- Summary statistics:
    - Mean of x: 9.0, Mean of y: 7.50
    - Variance of x: 10.0, Variance of y: 3.75
    - Correlation between x and y: 0.816
    - Linear regression equation:
        - $Y = 3.00 + 0.50 * x$
        - $y = 3.00 + 0.50 * x$
        - $y = 3.00 + 0.50 * x$

Dataset II:

- x-values: [10.0, 8.0, 13.0, 9.0, 11.0, 14.0, 6.0, 4.0, 12.0, 7.0, 5.0]
- y-values: [9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26, 4.74]
- Summary statistics:
    - Mean of x: 9.0, Mean of y: 7.50
    - Variance of x: 10.0, Variance of y: 3.75
    - Correlation between x and y: 0.816
    - Linear regression equation:
        - $y = 3.00 + 0.50 * x$

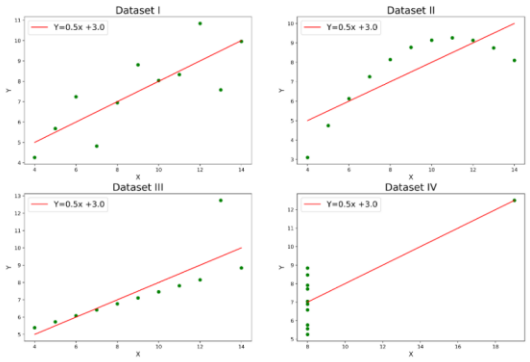
        - y = 3.00 + 0.50 * x
        - y = 3.00 + 0.50 ∗ x

Dataset III:

- x-values: [10.0, 8.0, 13.0, 9.0, 11.0, 14.0, 6.0, 4.0, 12.0, 7.0, 5.0]
- y-values: [7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73]
- Summary statistics:
    - Mean of x: 9.0, Mean of y: 7.50
    - Variance of x: 10.0, Variance of y: 3.75
    - Correlation between x and y: 0.816
    - Linear regression equation:
        - y = 3.00 + 0.50 ∗ x
        - y = 3.00 + 0.50 * x
        - y = 3.00 + 0.50 ∗ x

Dataset IV:

- x-values: [8.0, 8.0, 8.0, 8.0, 8.0, 8.0, 8.0, 19.0, 8.0, 8.0, 8.0]
- y-values: [6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.50, 5.56, 7.91, 6.89]
- Summary statistics:
    - Mean of x: 9.0, Mean of y: 7.50
    - Variance of x: 10.0, Variance of y: 3.75
    - Correlation between x and y: 0.816
    - Linear regression equation:
        - y = 3.00 + 0.50 ∗ x
        - y = 3.00 + 0.50 * x
        - y = 3.00 + 0.50 ∗ x



Anscombe's quartet Plot

Key Observations:

- All four datasets share identical or nearly identical summary statistics: mean, variance, correlation, and linear regression equations.
- Datasets reveal vastly different patterns:
    - Dataset I exhibits a near-perfect linear relationship.

- Dataset II displays a non-linear relationship, with a curve evident in the data.
- Dataset III contains a single influential outlier that distorts the correlation.
- Dataset IV shows a cluster of points with one extreme outlier, significantly affecting the regression line.

Lessons from Anscombe's Quartet:

- Importance of Visualization: Even though the datasets have the same summary statistics, visualizing them reveals different relationships, including linear, non-linear, and outlier-heavy patterns.
- Limitations of Summary Statistics: Summary statistics like mean and correlation can be misleading and fail to capture important data characteristics, such as outliers and non-linear trends.
- EDA Role: Anscombe's Quartet emphasizes that visualizing data during EDA is crucial for uncovering the true nature of data, helping to avoid faulty assumptions in analysis.

3. What is Pearson's R?

Pearson's Correlation Coefficient (r) quantifies the strength and direction of the linear relationship between two continuous variables. It ranges from -1 to 1, where:

- r = 1: Perfect positive linear relationship.
- r = -1: Perfect negative (inverse) linear relationship.
- r = 0: No linear relationship.

The formula for Pearson's R is:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Where:

- xi and yi are the data points for variables x and y.
- $\bar{x}$ and $\bar{y}$ are the means of x and y, respectively.

Few key characteristics:

- Measures Linear Relationships: Pearson's r only captures linear relationships, assuming the data follows a straight-line pattern.
- Strength and Direction: The magnitude of r shows the strength, with values near ±1 indicating a stronger relationship, while the sign (+/-) indicates the direction.
- Assumption and Limitation: Pearson's r assumes normally distributed data and may not detect non-linear relationships. Importantly, correlation does not imply causation.

Pearson's R is commonly used in statistics and machine learning to evaluate the strength and direction of the linear relationship between two continuous variables. A larger absolute value of R signifies a stronger linear correlation.

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

Scaling is a technique performed in pre-processing during building a machine learning model to standardize the independent feature variables in the dataset in a fixed range.

Scaling ensures that features contribute proportionally to the model. Without scaling, features with larger ranges could dominate those with smaller ranges, leading to biased model predictions. By rescaling the data, models can focus on the relationships between features rather than their absolute magnitudes, improving model performance, interpretability, and computational efficiency.

Two Key Scaling Techniques:

1. Normalization (Min-Max Scaling):
    - Definition: Normalization rescales data to fit within a specific range, typically between [0, 1].
    - Formula:

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

2. Standardization (Z-Score Scaling):

    Definition: Standardization transforms the data to have a mean of 0 and a standard deviation of 1.

$$X' = \frac{X - \mu}{\sigma}$$

Aspect	Normalized Scaling	Standardized Scaling
Definition	Rescales data to a fixed range, usually [0, 1].	Transforms data to have a mean of 0 and standard deviation of 1.
Range	Values are bounded within a specific range (e.g., 0 to 1).	Values are not bounded; they can be any real number.
Distribution	Retains the shape of the data distribution.	Centers the data around the mean and adjusts the spread.
Effect on Outliers	Highly sensitive to outliers, which can distort scaling.	Less sensitive to outliers due to reliance on mean and standard deviation.
Interpretability	Retains original units of the data.	Transforms data into unitless z-scores.

Sensitivity to Scale	Works best when the features are uniformly distributed and don't vary significantly.	Works better when the features have varied scales or non-uniform distributions.

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

The Variance Inflation Factor (VIF) is a measure used to detect and quantify multicollinearity in regression analysis.

Formula

The VIF for independent variable is calculated as follows:

$$\text{VIF} = \frac{1}{1 - R^2}$$

The value of the Variance Inflation Factor (VIF) becomes infinite when there is perfect multicollinearity between two independent variables, meaning they are perfectly correlated. In such cases, the R^2 value between these variables becomes 1. This suggests that one of these variables need to be dropped in order to define a working model for regression. Or other techniques such as Principal Component Analysis (PCA) or Regularization (Lasso, Ridge) can be used to reduce multicollinearity.

6. 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 method used to compare the quantiles of a sample distribution with those of a theoretical distribution to check if the data follows a particular distribution, such as normal, uniform, or exponential. It helps identify whether two datasets share the same distribution type and can reveal deviations.

A Q-Q plot is created by:

- Plotting two sets of quantiles against each other. 2.
- If both sets of quantiles came from the same distribution, the points should form a roughly straight line.

Use and importance in Linear Regression:

- Normality Assessment:
    - Use: A Q-Q plot is primarily used to assess whether the residuals of a linear regression model follow a normal distribution. By plotting the quantiles of the residuals against the quantiles of a theoretical normal distribution, analysts can visually inspect the alignment of points.
    - Importance: Since one of the key assumptions of linear regression is that the residuals should be normally distributed, confirming this assumption is crucial for valid statistical inference, such as hypothesis testing and constructing confidence intervals.

- Model Diagnostics:
    - Use: Q-Q plots are used to diagnose issues within the regression model. If the points deviate significantly from the straight line, it indicates that the residuals may not be normally distributed.
    - Importance: Identifying potential model specification errors or violations of assumptions allows for corrective measures to be taken, thereby improving the reliability and validity of the regression analysis.
- Outlier Detection:
    - Use: The Q-Q plot helps identify outliers and extreme values that can disproportionately influence the regression results. Outliers are typically represented as points that fall far from the expected normal line.
    - Importance: Recognizing and addressing outliers is essential for enhancing the accuracy of the model and ensuring that the results are representative of the underlying data.
- Visual Representation:
    - Use: Q-Q plots provide a clear visual representation of how the residuals compare to a theoretical normal distribution, making the analysis more interpretable.
    - Importance: This visualization aids communication with stakeholders, allowing them to understand the model's validity and the nature of the residuals, which is essential for informed decision-making.