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

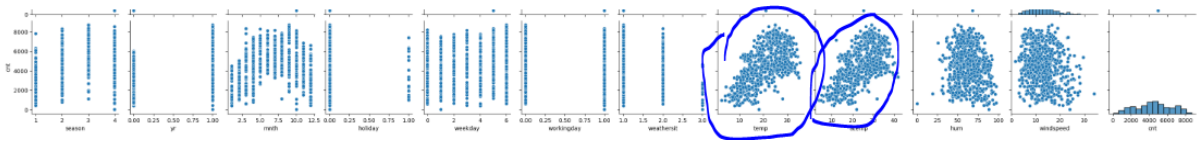
Ans: Here are a few inferences

1. There is a statistically significant decrease in bike rentals on holidays.
2. There is a statistically significant increase in bike rentals during fall and Spring
3. There is a statistically significant increase in bike rentals in August and September.
4. There is a statistically significant decrease in bike rentals as windspeed increases.
5. There is a statistically significant decrease in bike rentals as the weather situation is bad.

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

Ans: When encoding categorical values, we can create or add columns based on the number of values present under the category. However, the same thing can be done using n-1 columns (dropped column can be inferred). In this approach, reducing one column would take less time for the model to compute without affecting the accuracy.

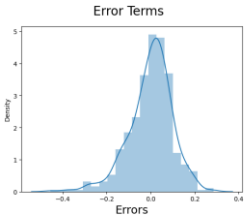
Q. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?



Looks like temp and atemp has the highest correlation with the target variable.

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

Ans: By doing residual analysis, I have validated the assumptions. The mean is zero. So it is normally distributed.



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

Ans : Top 3 features are Temperature, Year(2009) and Season 4(fall)

General Subjective Questions

Q. Explain the linear regression algorithm in detail.

Linear regression is a fundamental statistical and machine learning technique used to model the relationship between a dependent variable (often called the target or outcome) and one or more independent variables (predictors or features). It's particularly useful for predicting a continuous outcome based on input variables.

The core idea of linear regression is to find a line (in the case of one predictor) or a hyperplane (in the case of multiple predictors) that best fits the data. This line represents the best linear approximation of the relationship between the predictors and the target variable.

Mathematical Formulation

For simple linear regression, with one predictor variable x and one target variable y, the relationship is modelled as:

$$y = B0 + B1x + e$$

y is the dependent variable.

x is the independent variable.

B0 is the intercept, representing the value of y when x is zero.

B1 is the slope, representing the change in y for a one-unit change in x.

E is the error term, capturing the difference between the actual data and the model's predictions.

Q. Explain the Anscombe's quartet in detail.

Ans : Anscombe's quartet, created by statistician Francis Anscombe in 1973, is a group of four datasets that illustrate the importance of visualizing data before relying solely on summary statistics. Each dataset consists of eleven (x, y) data points, and they all share nearly identical statistical properties:

- Mean of x: 9 for all datasets.

- Mean of y: 7.5 for all datasets.

- Variance of x: 11 for all datasets.

- Variance of y: ≈4.12 for all datasets.

- Correlation between x and y: ≈0.816 for all datasets.

- Linear regression line: (y = 3 + 0.5x) for all datasets.

Despite these similarities, plotting each dataset reveals very different patterns:

1. Dataset-1 shows a typical linear relationship, where the linear model is appropriate.
2. Dataset-2 exhibits a clear nonlinear relationship, which the linear model fails to capture.
3. Dataset-3 features a linear relationship with an outlier that significantly distorts the regression line.
4. Dataset-4 consists mostly of a vertical line of points with one extreme outlier in the x-direction, making the linear model misleading.

Lessons from Anscombe's Quartet:

- Visual Inspection: Graphical representations are crucial, as they reveal patterns, outliers, and relationships that summary statistics can miss.
- Outliers: These can distort statistical measures and models, highlighting the need to identify and understand them.
- Model Appropriateness: Not all relationships are linear; visualization helps in choosing the right model.

Q: What is Pearson's R?

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

- +1 indicates a perfect positive linear relationship.
- -1 indicates a perfect negative linear relationship.
- 0 indicates no linear relationship.

Pearson's R is calculated by dividing the covariance of the two variables by the product of their standard deviations. It is widely used to assess the degree to which changes in one variable predict changes in another.

Q: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling adjusts the range of features in a dataset to ensure they contribute equally to a model. It's crucial for algorithms like gradient descent, k-nearest neighbours, and SVMs, which are sensitive to the magnitude of input features.

- **Normalized Scaling** (Min-Max Scaling) transforms data to a fixed range, typically [0, 1]. It's useful when you know the minimum and maximum bounds of the data.
- **Standardized Scaling** (Z-score Scaling) adjusts data to have a mean of 0 and a standard deviation of 1. It's ideal when the data follows a normal distribution or when different features have different units or scales.

Key Difference: Normalization scales data within a specific range, while standardization centers data around zero with unit variance.

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

Ans : The value of VIF becomes infinite when there is perfect multicollinearity in a regression model, meaning one predictor variable is a perfect linear combination of one or more other predictors. In this case, the denominator of the VIF calculation, which involves the coefficient of determination R2, becomes 1 (indicating perfect correlation), causing the VIF formula:

VIF=1/ 1−R2(r Square)

to result in division by zero, leading to an infinite VIF value. This tells that the predictor is redundant because it's completely explained by other predictors in the model.

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

A Q-Q plot (Quantile-Quantile plot) compares the quantiles of a dataset against the quantiles of a theoretical distribution, typically the normal distribution. In linear regression, it's used to assess if residuals are normally distributed, a key assumption for valid model inference. Points should align along a diagonal line if the residuals are normally distributed. Deviations from this line can indicate non-normality or outliers, suggesting the need for model adjustments or transformations to improve fit and validity. The Q-Q plot is crucial for ensuring that the assumptions of linear regression are met.