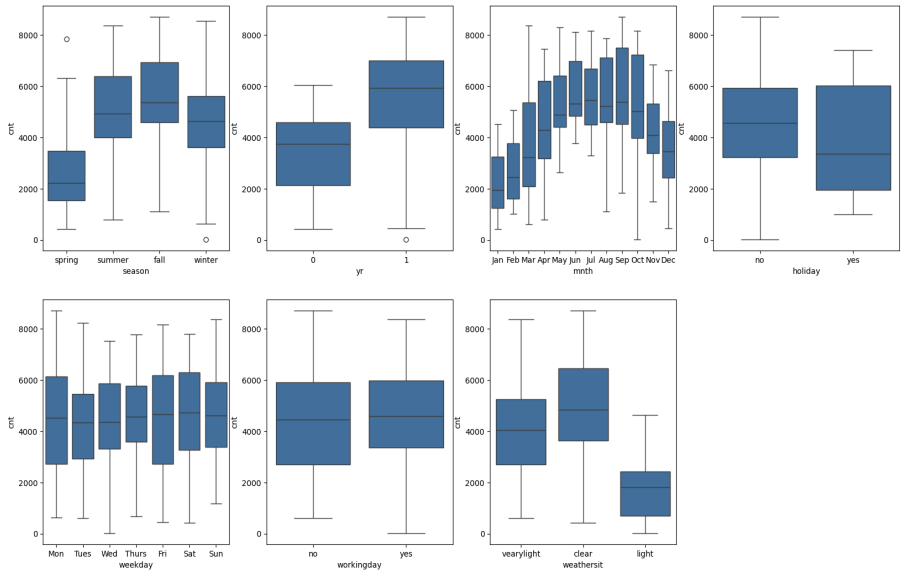
Assignment-based Subjective Questions

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)**

   Categorical Variables in the Data set are : 'season','yr', 'mnth', 'holiday', 'weekday','weathersit'



   Inference

   - fall has maximum demand and spring has lowest demand
   - 2019 has more demand when compared to 2018
   - On clear days (Few clouds, Partly cloudy, Partly cloudy) demand is high
   - It's observed that on Non Holidays demand is higher than holidays.

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

   To avoid Multicollinearity. I.e One of the variables can be defined as a linear equation of the other variables.
   So during the process of creating a dummy variable we need to change **drop_first setting to True.**

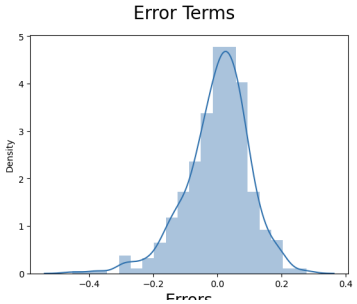
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**

   'Temp' and 'atemp' have a high correlation of 0.99 with target variable 'cnt'. It can be observed from heatmap.

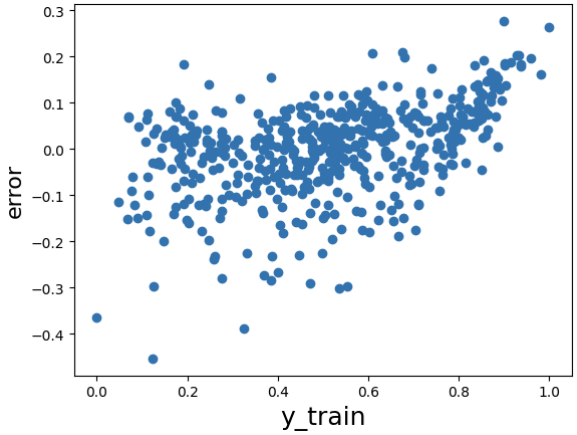
4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**

Below are list of assumptions validated in the model

1. Assumption "*Error terms are normally distributed*" is validated using the below graph. **Mean is centered at Zero .** Same graph can be used to validate "*Error terms has constant variance"*.

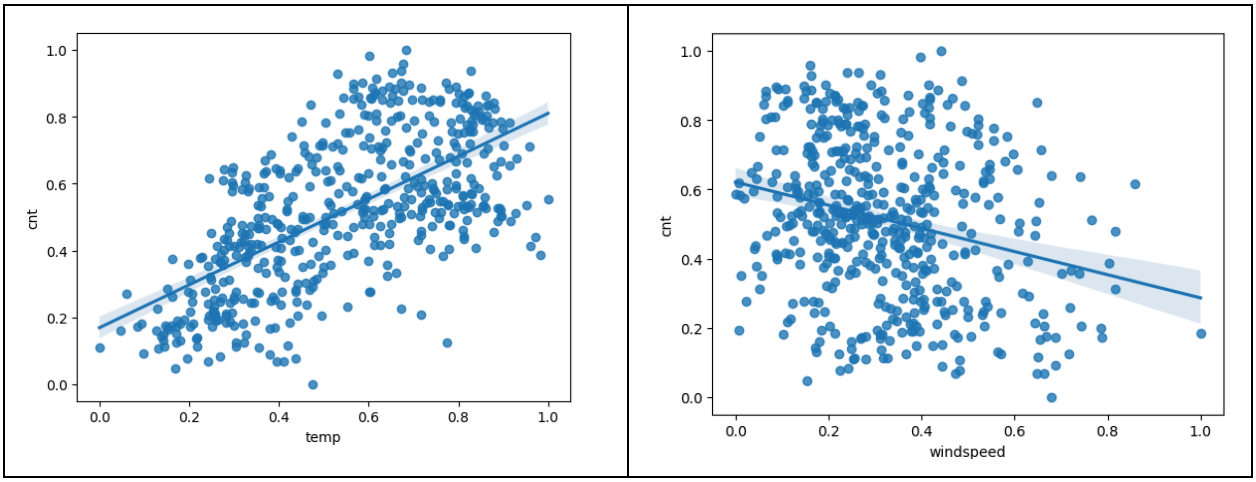


2. Assumption "*Error terms are independent of each other*" is validated using the below graph. Graph does **not have any pattern** which confirms error terms are independent.



3. Assumption "*There is linear relationship between Y and x's*" is validated using below plots.

Temperature and Windspeed has linear relationship with demand.



5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**

3 significantly contributing factors are from **'temperature','Weather situation-light'**(**negatively**) and **'Year'**.

- **Year** and **temperature** are impacting positively.
- The **Weather situation** of *Light (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds)* is impacting negatively.

Bike Demand(Y) = 0.0875+ 0.2334*yr -0.0867*holiday + 0.5682 *temp -0.1455 * windspeed -0.2535 *light + 0.0812 *summer + 0.1261*winter + 0.0895 *Sep

General Subjective Questions

1. **Explain the linear regression algorithm in detail. (4 marks)**

Linear regression is a supervised machine learning algorithm used to build a model to build a relationship between a dependent variable (Target-What we want to predict) and one or more independent variables (the predictors).

Building Blocks :
Model is represented as equation
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \varepsilon$

y : Target variable
x_i …xn : Predictors
B_0 : intercept/constant
β_1……β_n : multiplier which captures change in y for unit of change in x_i
ε : Error

4 assumption related to linear regression are

1. Linear relationship between y (target variable) and x's (predicting variables)
2. Error terms are normally distributed.
3. Error terms are independent of each other
4. Error terms have constant variance

Process : Algorithm arrives by fitting a line (y equation above) to the data(x's). Line is arrived at by minimizing the squares of residual errors. Residual is the difference between actual and predicted value. Algorithm iteratively adjusts the coefficients (β) to minimize error.

2. **Explain the Anscombe’s quartet in detail. (3 marks)**

Anscombe's Quartet is a collection of four data sets with identical descriptive statistics like mean, variance, R-squared, correlations, and linear regression lines. However, when plotted on a scatter plot, these four data sets exhibit visibly different patterns and distributions.

Anscombe's Quartet highlights the limits of relying on numerical metrics alone, as it consists of four datasets with identical summary statistics but visually distinct patterns.

It also emphasizes, data visualization is crucial for identifying trends, outliers, and other details not evident in summary statistics.Core concept emphasizes the importance of visualizing data, as graphs can reveal patterns and outliers that summary statistics might over look. It reminds us that graphs are not just a way to present data but also a powerful tool for understanding it.

3. **What is Pearson’s R? (3 marks)**

Pearson's R is correlation coefficient that quantifies the strength and direction of linear relationship between two continuous variables.

Range of R will be -1 to 1

- -1 (Perfect Negative) means the variables are inversely proportional. *I.e One increases other decreases*
- 0 means variables are not related
- +1 (Perfect Positive) means a strong linear relationship. *I.e One increases other also increases*

Assumptions : variables are linearly distributed and no significant outliers.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)**

Scaling, also known as data scaling, is a preprocessing technique used to standardize the range of independent variables or features of data.

During Linear regression building we have multiple independent variables with different units .i.e is height of building and number of doors. In order to reduce the impact of unit of measure in the model all the variables are scaled to 0 to 1 without losing the impact. Helps avoiding domination of variables with large numerical values

Difference :

- **Normalized scaling** (min-max) gets values of results in a range of 0 and 1. Works better when data is **NOT Normally distributed**.
- **Standardized Scaling** gets values of results ito have mean of ZERO and Standard deviation of 1. Data is close to normal distribution. Outliers are handled better.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**

When multicollinearity happens VIF will be infinite for that variable.

Logical Explanation: When one of the variables say x_i predicted perfectly by other variables x's then R_i^2 will be 1 and VIF_i will be infinite.

Mathematical explanation

VIF formula : $VIF_i = 1 / (1 - R_i^2)$

R_i^2 is coefficient to determine the regression of variable x_i on all other independent variables. When R_i^2 equals 1 or very close to 1 we get VIF as infinite.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)**

Quantile-Quantile (Q-Q) are plots where quantiles of the observed data to the quantiles of the expected distribution. After sorting quantiles are obtained.Q-Q plots summarize the distribution visually.

Q-Q plots will be of practical use during validation of models. I.e when you want to check if two data sets are of the same distribution or to check if residuals have normal distribution.

Q-Q plots can be used to check the assumptions of linear regression like normality, linearity, and homoscedasticity for the data.