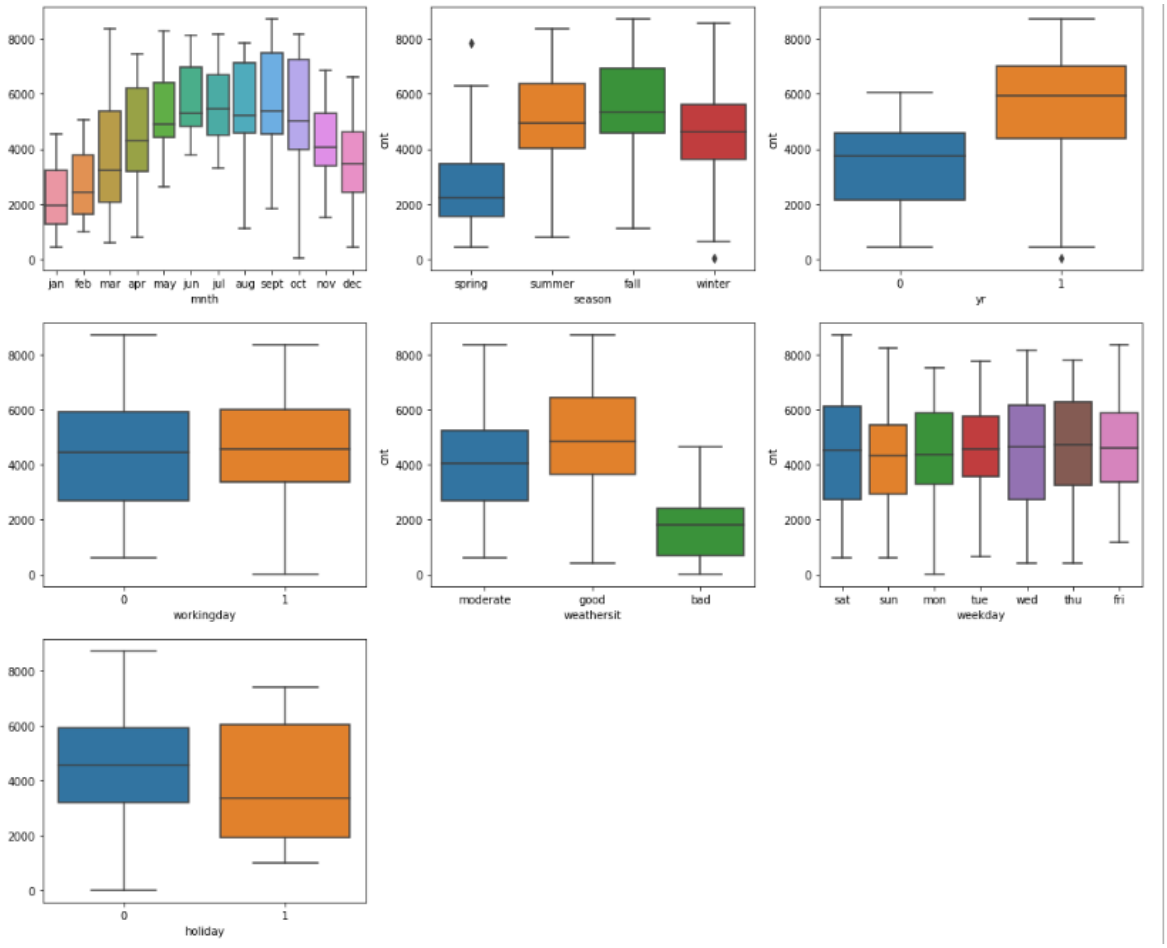
## Assignment-based Subjective Questions

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




From categorical variables like month, season, yr, workingday, weathersit, weekday, holiday give many inferences about their effect on dependent variable cnt . such as

1.highest demand /cnt of bike sharing is highest during month of September and lowest in January and in December.

2.During fall season bike sharing count is high and spring is least.

3.During year 1/2019 has more bike sharing than in the previous year 2018.

4.During working days demand is more

5.demand is more during good weather and lowest in bad.

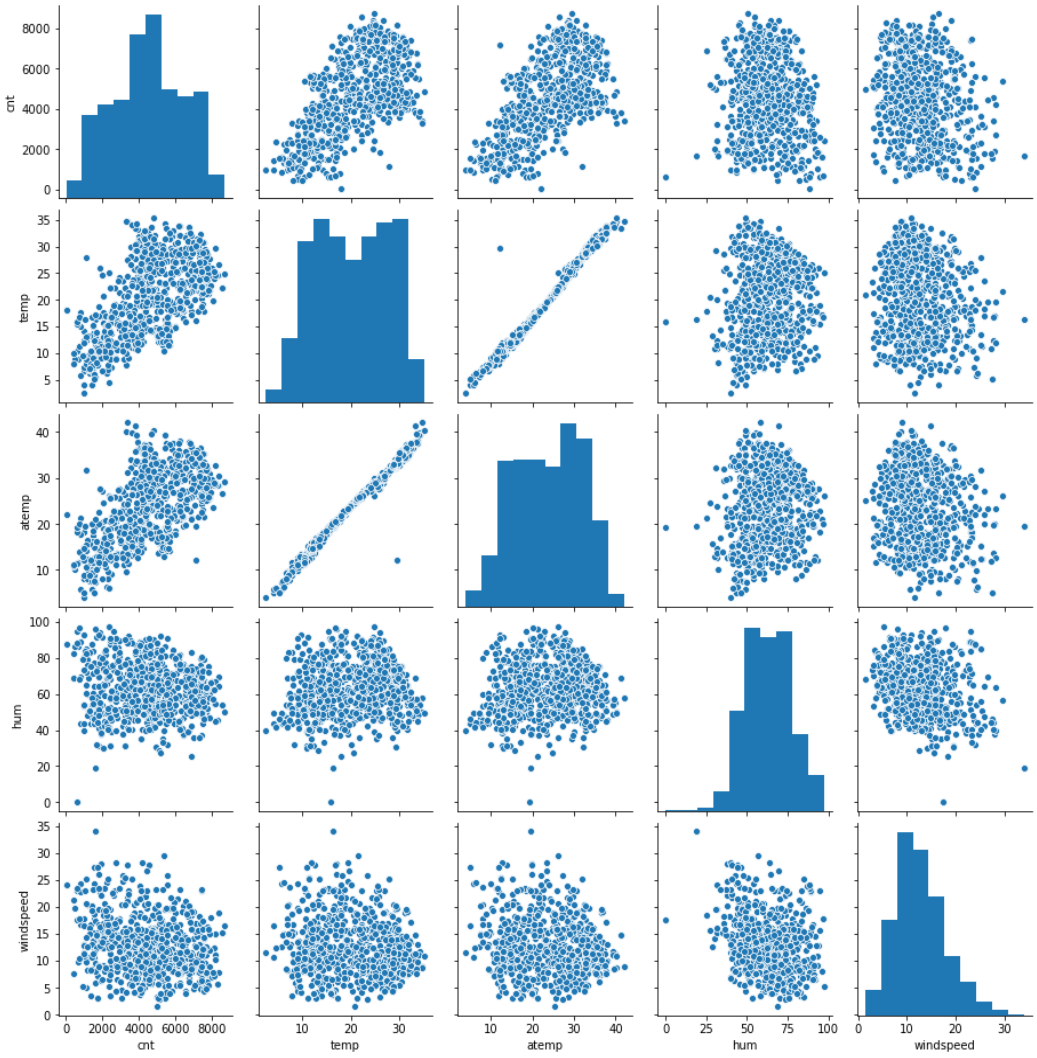
6.During weekdays, Saturday has most bike sharing and Sunday has lowest but not much differences.

7.holiday demand is less.

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

We create dummy variables from categorical variables to numerical variables to make it useful for model building. For n number of levels, we create n-1 new columns by setting drop_first=True so that the resultant can match up n-1 levels. Hence it reduces the correlation among the dummy variables and helps avoid multicollinearity by creating n-1 dummy variables instead of n .

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




Variable temp and atemp share highest correlation with the target variable cnt (1 and 0.99 respectively ) and with each other i.e.0.63 .

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

After building linear regression models, we validate the model by

1.Residual analysis: difference between actual and predicted values and checking the pattern against training model's pattern.

2.Normality distribution of error term with mean 0.

3.Homoscedasticity which ensures data points are spread out , error is consistence throughout the prediction.

4.Multicolinearity is checked among predictor variables by calculating variance inflation factor. High VIF indicates multicollinearity which can distort parameter effects and model interpretability. For model selection to be good enough here, selected VIF should be < 5.

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

Top 3 features contributing significantly are temp, season, weathersit.

# General Subjective Questions

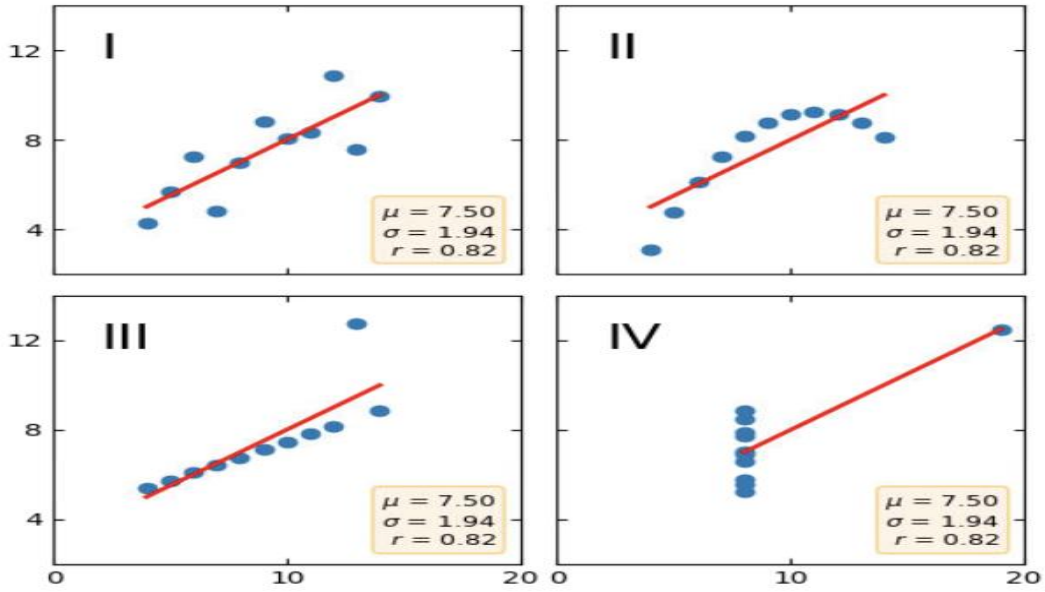
**1. Explain the linear regression algorithm in detail.**

Linear regression models the relationship between a dependent variable or target variable and one or more independent variables by fitting a straight line through the data points.

1. Assumptions:
    - Linearity: The relationship between the independent variables and the dependent variable is assumed to be linear.
    - Independence: Observations are assumed to be independent of each other.
    - Homoscedasticity: The variance of the errors (residuals) is assumed to be constant across all levels of the independent variables.
    - Normality: The residuals are assumed to be normally distributed.
2. Model Fitting:
    - The model coefficients are estimated using a method called Ordinary Least Squares (OLS). For this assignment used OLS from stats model api.
    - OLS minimizes the sum of the squared differences between the observed and predicted values.
    - It involves finding the values of coefficients that minimize the cost function.
3. Model Evaluation:
    - After fitting the model, its performance is evaluated using metrics like Mean Squared Error or R-squared which measures the average squared difference between observed and predicted values.
    - Also by checking the coefficients .
    - Finally selecting the model with the features which gives high R squared value , P value <0.05 ,and VIF < 5 .
4. Verifying the assumptions made.

**2. Explain the Anscombe's quartet in detail.**

In statistics Anscombe's quartet demonstrates the importance of visualizing data before drawing conclusions by regression analysis. It consists of four datasets that have nearly identical summary statistics (mean, variance, correlation, etc.) but shows different graphical patterns.




- Dataset I shows linear relationships between X independent variable and y target variable.
- Dataset II shows linear relationships, but dataset is heavily influenced by an outlier.
- Dataset III exhibits a clear quadratic relationship.
- Dataset IV displays two distinct groups with a significant outlier.

**3. What is Pearson's R**?

The Pearson correlation coefficient ,also referred to as Pearson's r, is the most common way of measuring a linear correlation .It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables. When one variable changes, the other variable changes in the same direction. Pearson's r draws a line of best fit through two variables, indicating the distance of data points from this line. A 'r' value near +1 or -1 implies all data points are close to the line and 'r' value close to '0' suggests data points are scattered around the line.

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

$r$ = correlation coefficient
$x_i$ = values of the x-variable in a sample
$\bar{x}$ = mean of the values of the x-variable
$y_i$ = values of the y-variable in a sample
$\bar{y}$ = mean of the values of the y-variable

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

Scaling means transforming the data so that it fits within a specific scale. Scaling is performed to ensure that the variables are comparable and to improve the performance algorithms and is data pre-processing step where we will fit data in specific scale. Variables may have different units or scales, making direct comparisons challenging. Scaling standardizes the units and ranges of variables, allowing for fair comparisons. If scaling is not performed, then algorithm tends to weigh high values magnitudes and ignore other parameters which will result in incorrect modeling.

Difference between Normalizing Scaling and Standardize Scaling:

**->Normalized scaling**, also known as min-max scaling, rescales the values of variables to a range between 0 and 1. Normalized scaling preserves the shape of the distribution but compresses it into the range [0, 1].

$$X_{\text{norm}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}$$

- Where:
  - $X$ is the original value of the variable.
  - $X_{\text{min}}$ and $X_{\text{max}}$ are the minimum and maximum values of the variable,

**-> Standardize Scaling,** also known as z-score normalization or standardization, transforms the values of variables to have a mean of 0 and a standard deviation of 1. Standardized scaling centers the distribution around the mean and scales it based on the standard deviation.

$$X_{\text{std}} = \frac{X - \mu}{\sigma}$$

Where:

- $X$ is the original value of the variable.
- $\mu$ is the mean of the variable.
- $\sigma$ is the standard deviation of the variable.

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

Variance Inflation factor (VIF) explains how the relationship of one independent variable is explained by all other independent variables combined.

A VIF values

->greater than 10 is definitely high and variable should be eliminated

-> greater than 5 , can be ok but is worth inspecting

-> less than 5 , Good ,no need to eliminate this variable.

VIF= 1/(1-Rsquared value) , If R squared value is 1 i.e. a perfect correlation, the denominator becomes zero, resulting in an infinite VIF value.

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

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution such as the normal distribution.

It compares the quantiles of the dataset to the quantiles of a theoretical distribution, typically the normal distribution, on a scatter plot.

QQ plot can also be used to determine whether two distributions are similar or not. If they are quite similar you can expect the QQ plot to be more linear. The linearity assumption can best be tested with scatter plots. Secondly, the linear regression analysis requires all variables to be multivariate normal. This assumption can best be checked with a histogram or a Q-Q-Plot.

Q-Q plot for linear regression:

**1**.Q-Q plots are commonly used in linear regression to check the normality assumption of the residuals (errors).

**2.**In linear regression, the residuals should ideally follow a normal distribution with a mean of zero.

**3.**By examining the Q-Q plot of the residuals, you can assess whether the assumption of normality is met.

**4.**If the points on the Q-Q plot fall approximately along a straight line, it suggests that the residuals are normally distributed.

**5.**Deviations from the straight line indicate departures from normality, which may affect the validity of statistical inference and confidence intervals in linear regression analysis.

**Importance**:

1. The Q-Q plot provides a visual assessment of the normality assumption, which is crucial for the validity of statistical inference in linear regression.
2. It complements formal statistical tests for normality by offering a graphical representation of the distributional properties of the residuals.
3. Q-Q plots allow for easy identification of departures from normality, such as skewness or heavy tails, which may require further investigation or transformation of the data.
4. By examining the Q-Q plot, we can make informed decisions about the appropriateness of linear regression assumptions and take corrective actions if necessary.