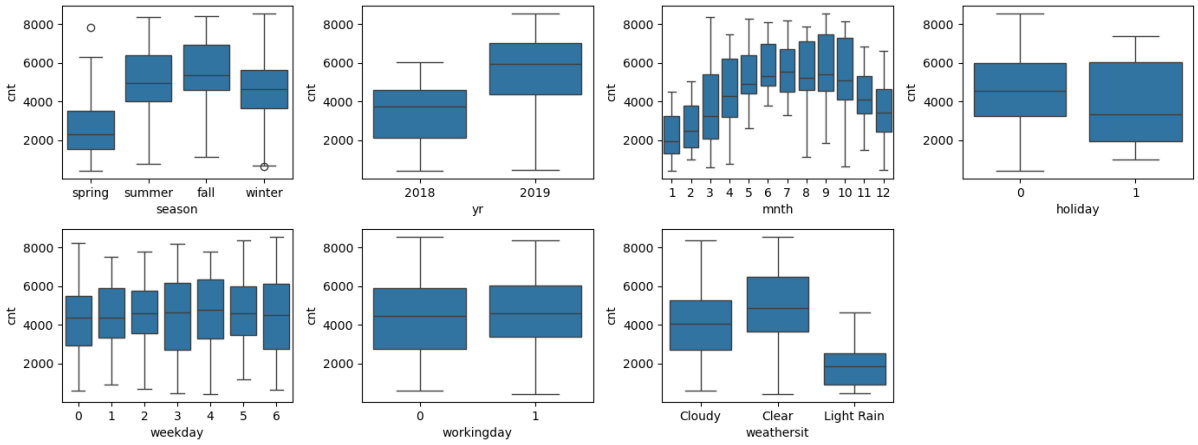
Assignment-based Subjective Questions

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

Answer:



Observations from bivariate analysis:

- In fall season, demand is high
- usage increased with year (increased from 2018 to 2019)
- Mid months has higher usage
- Non holidays have higher usage
- No effect of weekday and working day on usage.
- More usage with clear and cloudy weather situation

Observations from regression model:

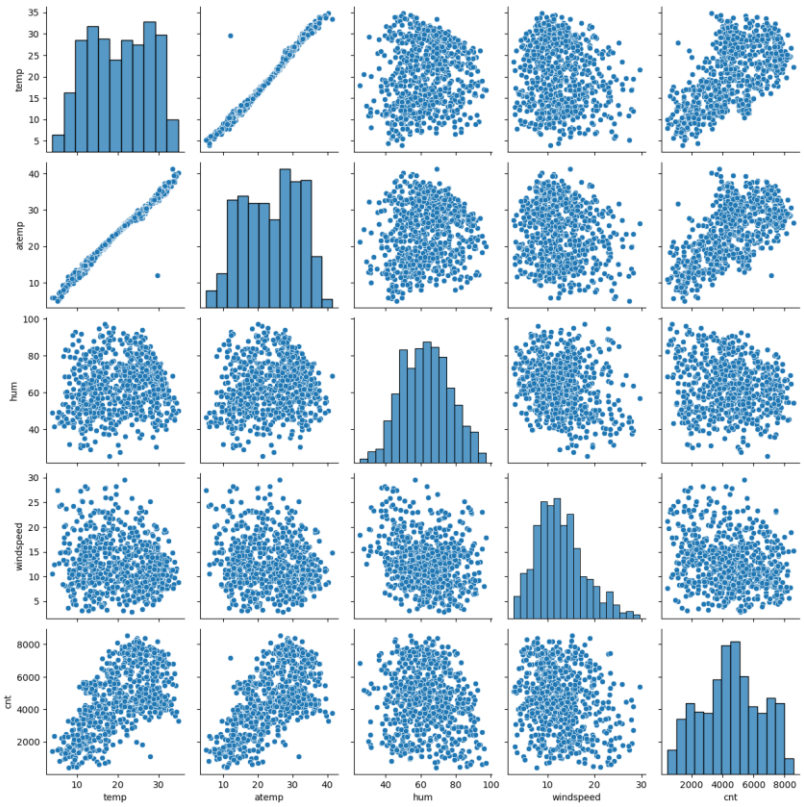
```
temp                    0.587136
yr_2019                 0.244342
const                   0.198394
season_winter           0.148751
mnth_5                  0.129213
mnth_9                  0.125724
mnth_4                  0.112094
mnth_3                  0.080959
mnth_6                  0.073022
mnth_10                 0.068575
mnth_8                  0.065016
weathersit_Light Rain  -0.115322
holiday                -0.119628
windspeed              -0.200412
hum                    -0.257064
dtype: float64
```

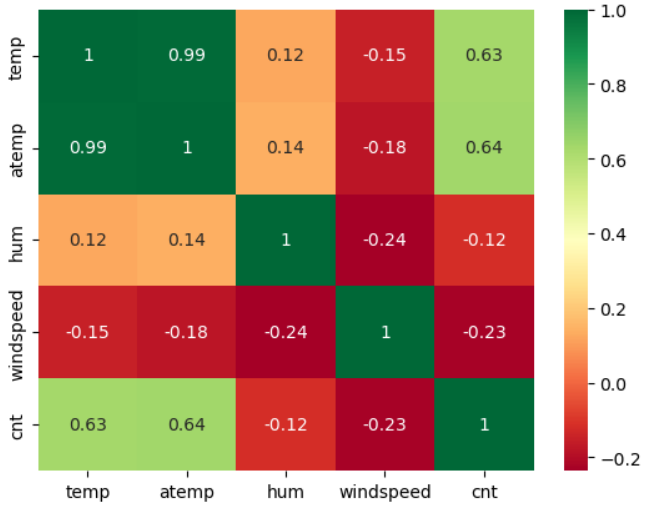
- Year is important - demand increases with each year
- months 3 to 10 are important, winter season (months 10-12) is also important - that means month contributes to demand.
- Light rain weather situation decreases the demand
- Holiday has less demand

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

If we do not use drop_first = True, then n dummy variables will be created, and these predictors (n dummy variables) are correlated which is known as multicollinearity. It makes it difficult to make valid inferences from the model. if there are n dummy variables, n-1 dummy variables will be able to predict the value of the nth dummy variable, so one dummy variable should be dropped to avoid multicollinearity.

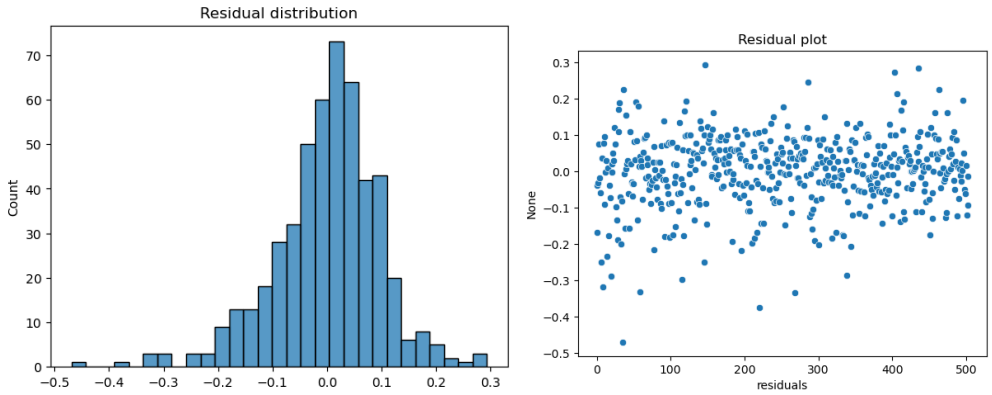
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?





Temperature has highest correlation with target variable ‘cnt’.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
   - R square and adjusted R-sq are close, indicates no curve fitting with additional parameters.
   - All terms p values are low – indicates their significance in model.
   - All terms VIF values are low – indicates there is low multicollinearity in final model.
   - Residual analysis is performed to check for assumptions of linear regression model.
     - They follow normal distribution
     - They are random and there is no pattern in them.



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

```
temp                    0.587136
yr_2019                 0.244342
const                   0.198394
season_winter           0.148751
mnth_5                  0.129213
mnth_9                  0.125724
mnth_4                  0.112094
mnth_3                  0.080959
mnth_6                  0.073022
mnth_10                 0.068575
mnth_8                  0.065016
weathersit_Light Rain  -0.115322
holiday                -0.119628
windspeed              -0.200412
hum                    -0.257064
dtype: float64
```

Above pictures shows important features which contributes significantly in the model.

Top 3 coefficients are:

1) Temperature
2) Humidity
3) Year

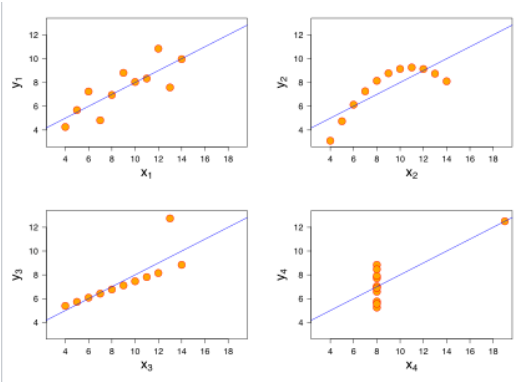
General Subjective Questions

1. **Explain the linear regression algorithm in detail.**
   - In linear regression, the output variable to be predicted is a continuous variable, such as scores of a student.
   - It falls under supervised learning methods.
   - It is a form of predictive modelling technique which tells the relationship between the dependent (target variable) and independent variables (predictors).
   - Two types are Simple linear regression, Multiple linear regression.
   - It takes the form of $Y = \beta_0 + \beta_1 X$, where β_0 is the intercept and β_1 is the slope.
   - Multiple linear regression has one coefficient for each term.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

   - The assumptions of simple linear regression were:
     - Linear relationship between X and Y
     - Error terms are normally distributed (not X, Y)
     - Error terms are independent of each other
     - Error terms have constant variance (homoscedasticity)
   - The parameters to assess a model are:
     - t statistic: Used to determine the p-value and hence, helps in determining whether the coefficient is significant or not
     - F statistic: Used to assess whether the overall model fit is significant or not. Generally, the higher the value of F statistic, the more significant a model.
     - R-squared: After it has been concluded that the model fit is significant, the R-squared value tells the extent of the fit, i.e. how well the straight line describes the variance in the data.
   - there are a few considerations before building the model:
     - Overfitting: Adding more isn't always helpful. Check for Adj. R-sq, it should be close to R-sq. Also, test set should show similar R-sq.
     - Multicollinearity: Associations between predictor variables. Srop variables with high correlation or looking at Variance Inflation Factor (VIF).
     - Feature scaling to avoid large coefficients.

2. **Explain the Anscombe's quartet in detail.**
   - Anscombe's quartet comprises four datasets

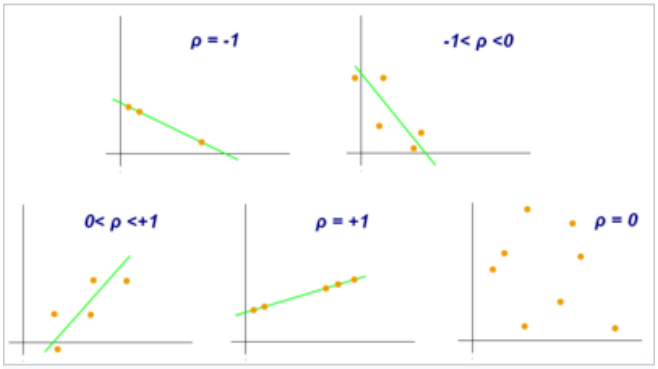


(Source: Wikipedia)

- All four sets have identical statistical parameters (such as mean of x and y, variance of x and y, correlation of x and y, linear regression equation), but the graphs show them to be considerably different.
- It shows the importance of graphing data when analysing it, and the effect of outliers and other influential observations on statistical properties.
- He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough".
- So, it is always required to understand data apart from looking at regression equations and statistical summaries.

3. What is Pearson's R?

- The Pearson correlation coefficient (r) is the most widely used correlation coefficient
- it describes the strength and direction of the linear relationship between two quantitative variables.



Examples of scatter diagrams with different values of correlation coefficient (ρ)

(Source: Wikipedia)

- Given a pair of random variables (X,Y) (for example, Height and Weight), Pearson correlation coefficient is defined as

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

  - Cov is the covariance between X and Y.
  - SigmaX is the standard deviation of X.
  - SigmaY is the standard deviation of Y.
- The Pearson correlation coefficient is a good choice when all the following are true:
  - Both variables are quantitative.
  - The variables are normally distributed.
  - The data have no outliers.
  - The relationship is linear.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**
   - When lot of independent variables are in a model, a lot of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret.
   - So, we need to scale features because of two reasons:
     - Ease of interpretation
     - Faster convergence for gradient descent methods
   - Scaling the features using two very popular method:
     - **Standardizing**: The variables are scaled in such a way that their mean is zero and standard deviation is one.

$$x' = \frac{x - \bar{x}}{\sigma}$$

     - **Normalized Minmax Scaling**: The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data.

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

   - It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F statistic, p-values, R-square, etc.

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

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

where:

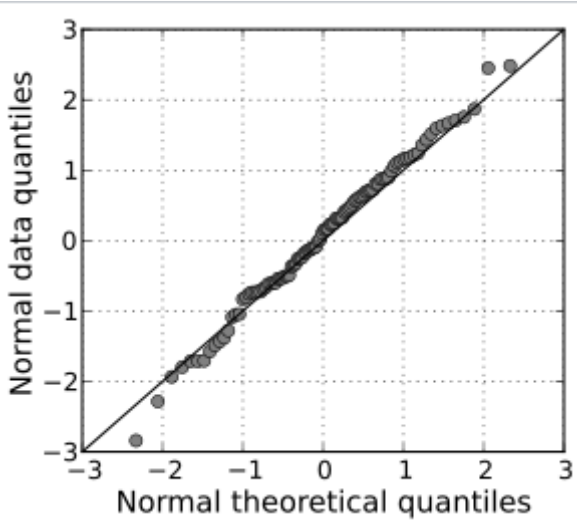
R_i^2 = Unadjusted coefficient of determination for regressing the ith independent variable on the remaining ones

(source: Investopedia)

   - VIF measures the strength of the correlation between the independent variables in regression analysis. This correlation is known as multicollinearity, which can cause problems for regression models.
   - **If R sq is 1, VIF becomes infinite**. That means there is a perfect multicollinearity with one of the features.
   - Good thumb rule is to have VIF below 5.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**
   - Q-Q plots are also known as Quantile-Quantile plots.
   - As the name suggests, they plot the quantiles of a sample distribution against quantiles of a theoretical distribution.
   - It is a graphical method for determining if a dataset follows a certain probability distribution
   - Q-Q plots are particularly useful for assessing whether a dataset is normally distributed.

- It can be used in regression models to check if the residuals of the model are normally distributed, which is an assumption.



Q-Q plot (source: Wikipedia)