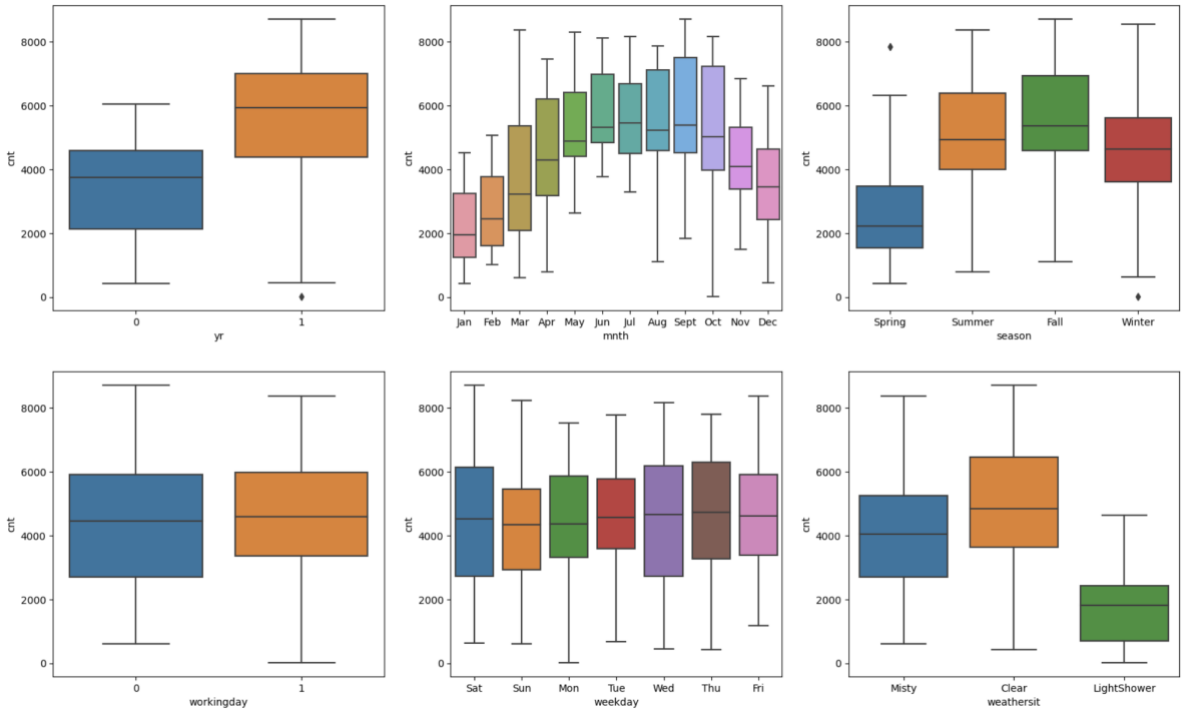
# Assignment-based Subjective Questions

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




Observations from above box plots

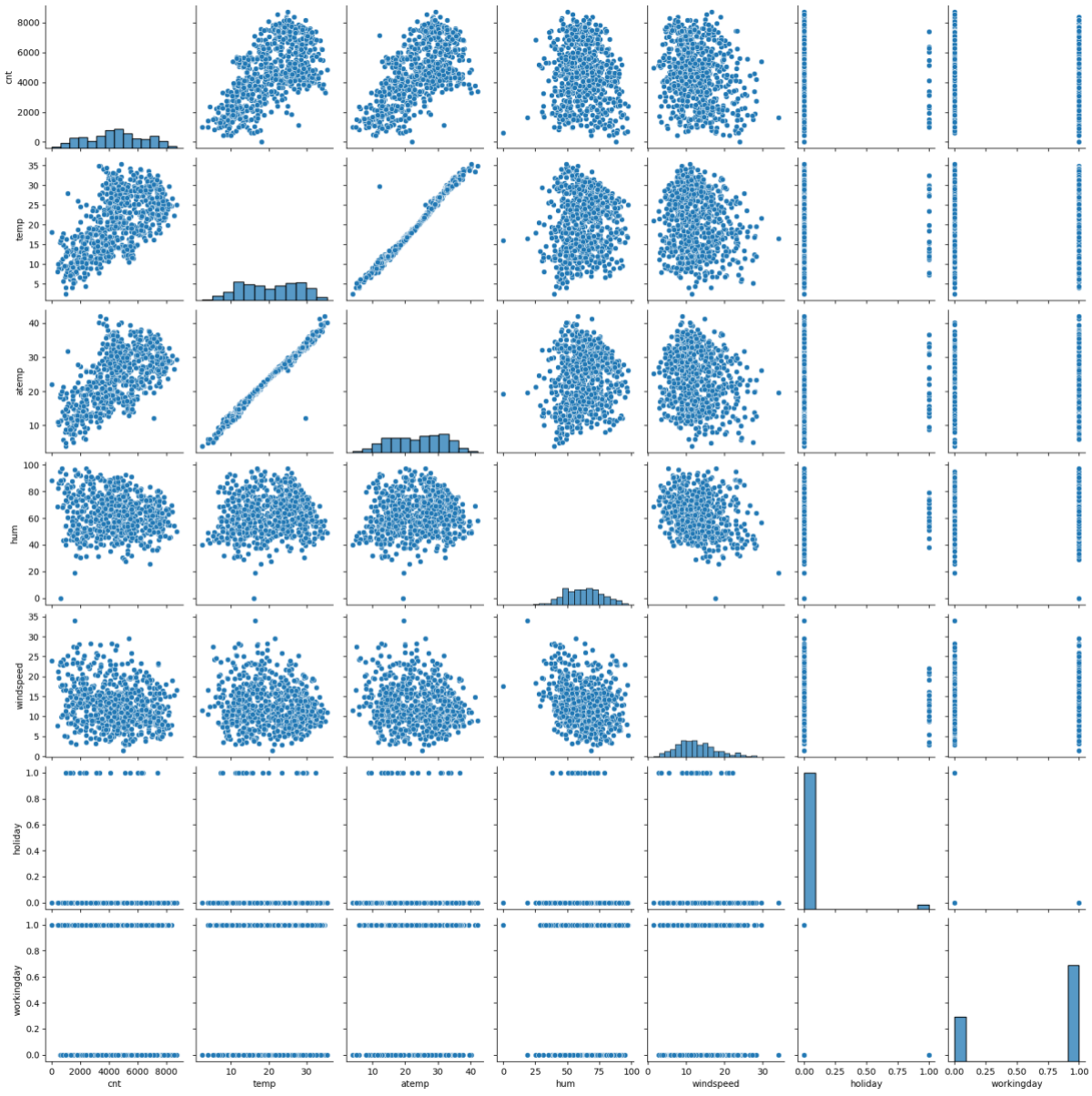
- There was higher demand for the bikes in 2019 compared to 2018
- Demand is high during the mid of the year.
- Fall has the highest demand followed by Summer and then winter.
- Trend looks same across all week days & looks like not much impact on holidays
- More demand if the weather is Clear or slightly misty.

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

   "drop_first=True" drops the first column during the dummy variable creation from categorical variables. For representing N distinct values it requires only N-1 variables; the Nth variable will be represented any way when all other variable values are 0(say example). It also reduces the correlation created among dummy variables and cause any cyclic effect.

   For example, when we created dummy columns from 'mths' column, it dropped mths_Apr columns and kept rest 11 variables.

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




Pairplot above shows that temp & atemp variables has highest correlation to target variable 'cnt'.

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

   Assumptions of Linear Regression Model is validated based on

   - linearity(R-Square values 0.8352749595695671)
   - mean residual (mean residual 4.318985256581001e-16),
   - Homoscedasticity,
   - multicollinearity(low multicollinearity between predictors)
   - Normality of error terms(histogram plot)
   - No correlation of residuals(y_pred vs residuals)

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 towards bike demand are temperature, weather and year.

```
temp                        0.5761
yr                          0.2573
weathersit_LightShower     -0.2649
```

# General Subjective Questions

1. Explain the linear regression algorithm in detail.

   Linear regression is a statistical method which helps to provide a linear relationship between an independent variable (predictors) and a dependent variable(target variable) to predict the outcome of future events. It makes predictions based on continuous/real numeric values only.

   There are 2 types of linear regression – simple(single input variable) and multiple(more than one input variable) linear regression.

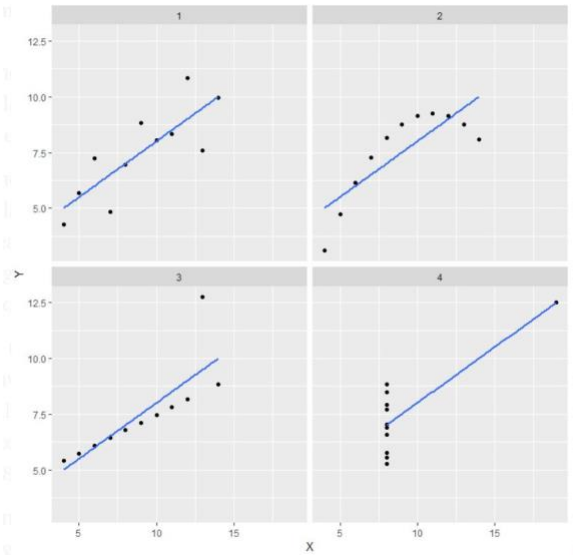
   Hypothesis function for linear regression is y = (intercept) + (slope)*x

   Steps involved in algorithm includes train the model, residual analysis, predictions and model validation(in addition to initial data preparation & analysis)

   It is used in the fields like Sales, Sports, Production etc.

2. Explain the Anscombe's quartet in detail.

   Anscombe's quartet comprises four data sets that have nearly identical simple statistical properties, yet have very different distributions and appear very different when graphed. This is done to illustrate the importance of plotting graphs before model building; sometimes its very deceiving otherwise. To understand this better, refer the scatterplots below




- From first graph, there seems to be a linear relationship between x & y
- From second one, its clear that there are is a non-linear relationship between x & y.
- Third one shows a linear relationship between x & y, but there are outliers.
- Forth one shows that it produces high correlation coef.

3. What is Pearson's R?

   Pearson's correlation coefficient(or simply Pearson's R), is a measure of linear correlation between two data sets.

   It's the ratio of covariance vs product of standard deviation of two variables.

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

   Or in detail

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

   If R is between 0 & 1 then it shows a positive correlation, 0 means no correlation & 0 to -1 means negative correlation.

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

   Scaling is a kind of data pre-processing where we fit data in a specific scale. This step is important in linear regression algorithm since we will be comparing values across columns and hence the values should be in same scale. If variables are not scaled properly then its high chance of ranking variables with high magnitude high and ignore other variables.

   There are 2 types of scaling – Normalized (minmax) and Standardized

In Normalized scaling minimum and maximum values are used for scaling. In Standardized scaling mean and standard deviation is used for scaling.

Normalized scaling is bound to range(normally 0 to 1); but standardized scaling is not bound to a certain range.

Normalized scaling is used when we don't know about the distribution whereas standardized scaling is used when distribution is normal.

Normalized scaling is affected by outliers(since it is based on min & max) whereas standardized scaling is not having any effect by outliers (since based on mean & std).

Normalization is used when the data doesn't have Gaussian distribution whereas Standardization is used on data having Gaussian distribution.

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

   Variance Inflation Factor aka VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity.

   VIF = 1/(1-R^2)

   From the equation its clear that

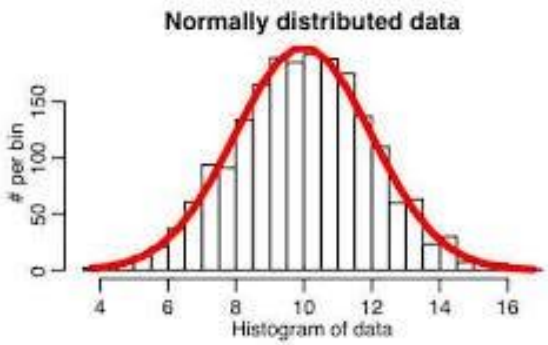
   - If R is 0, i.e variables are orthogonal to each other, then VIF will be 1. i.e No multicollinearity
   - If R is 1, i.e variables are perfectly correlated, then VIF will be infinity (1/0). i.e variables are perfectly correlated

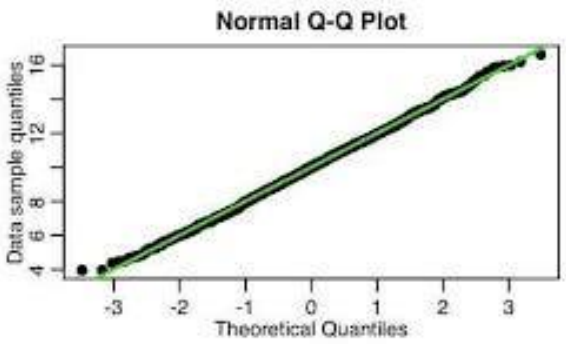
   As explained above, VIF value will be infinity if variables are perfectly correlated.

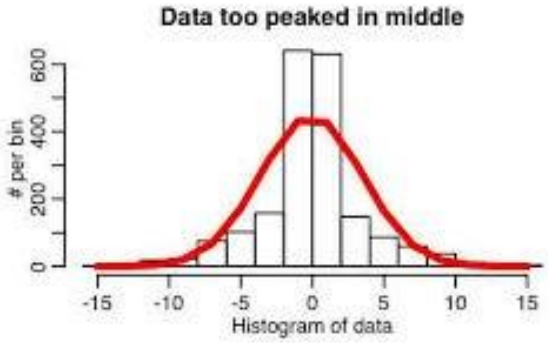
   Note: VIF values upto 5 is considered to be Ok.

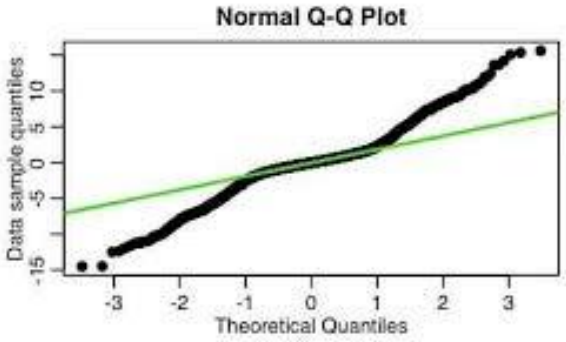
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

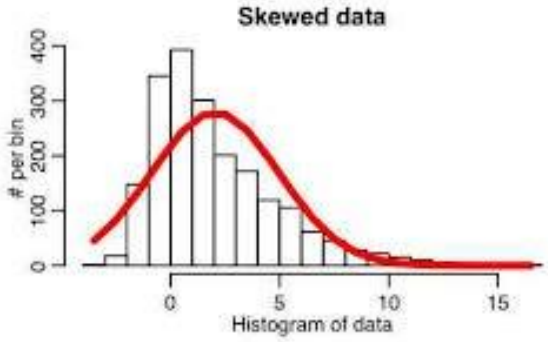
   Quantile-Quantile plot(or Q-Q plot) is a probability plot for comparing two probability distributions by plotting their quantiles against each other. The purpose of Q Q plots is to find out if two sets of data come from the same distribution.

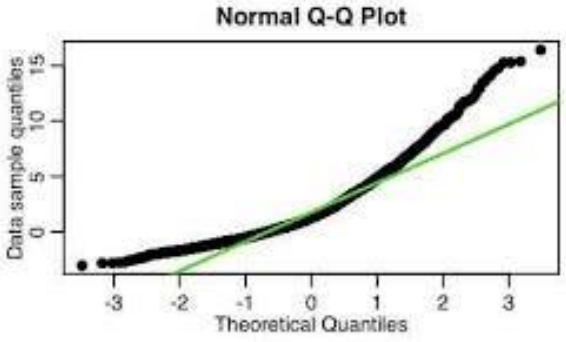



If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the identity line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

## Importance of a Q-Q plot in linear regression

- It helps to understand if both datasets came from population with common distribution
- It can be used with sample sizes also
- Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.
- Whether both datasets have similar type of distribution shape
- Whether both datasets have common location and scale