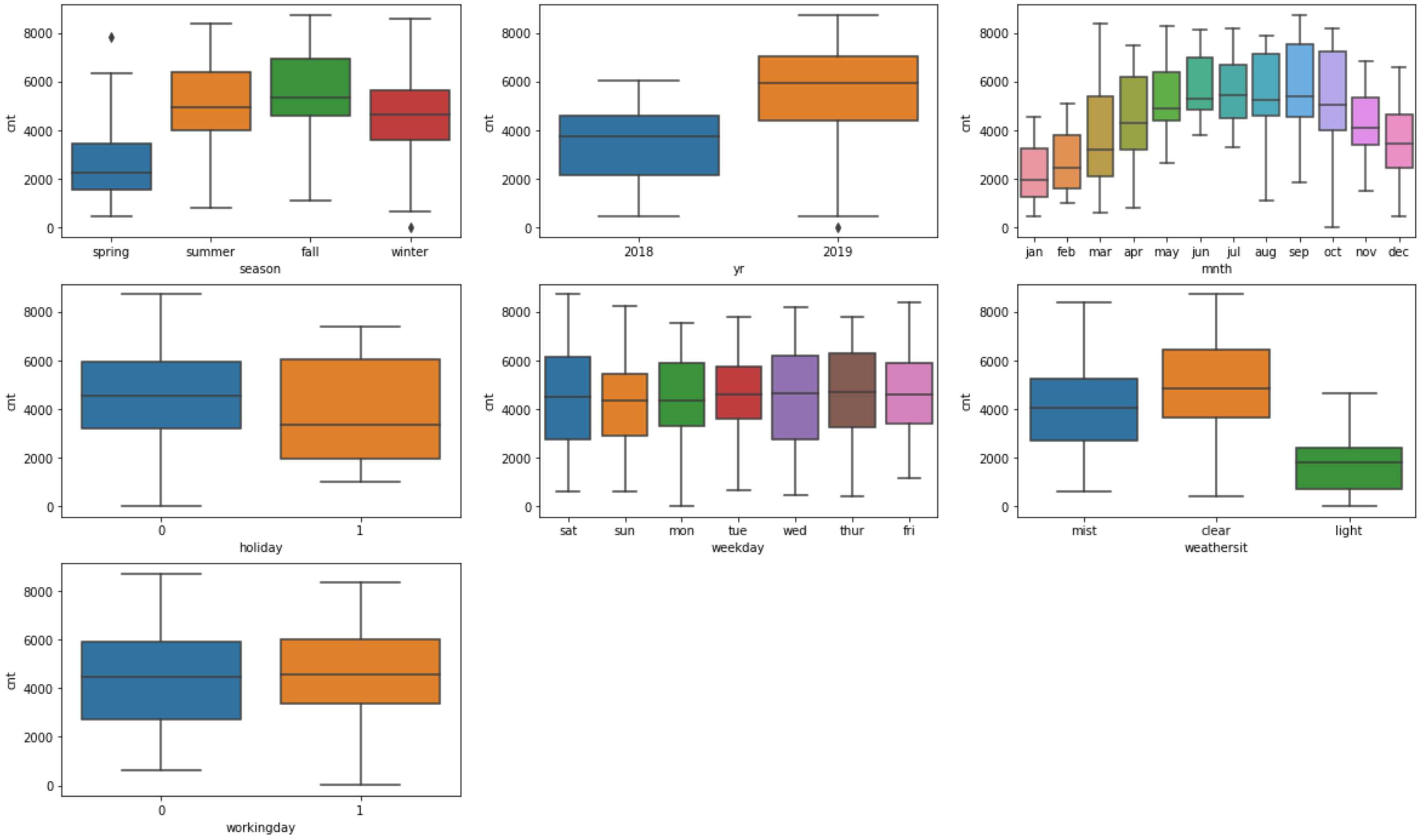
## Assignment-based Subjective

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



From my analysis, I can infer below items:

1. Fall season has more demand than other seasons.
2. 2019 Yr has more demand compared to 2018.
3. Non-holiday days has more demand.
4. Weather should be clear for demand.
5. Months (may to sep) has good demand compared to other months.
6. working days - Demandy is slightly more in nworking days.
7. Weekday - Could not get much correlation of demand with days.

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

drop_first=True drops the extra column created during dummy variable creation. it reduces the correlations created among dummy variables.

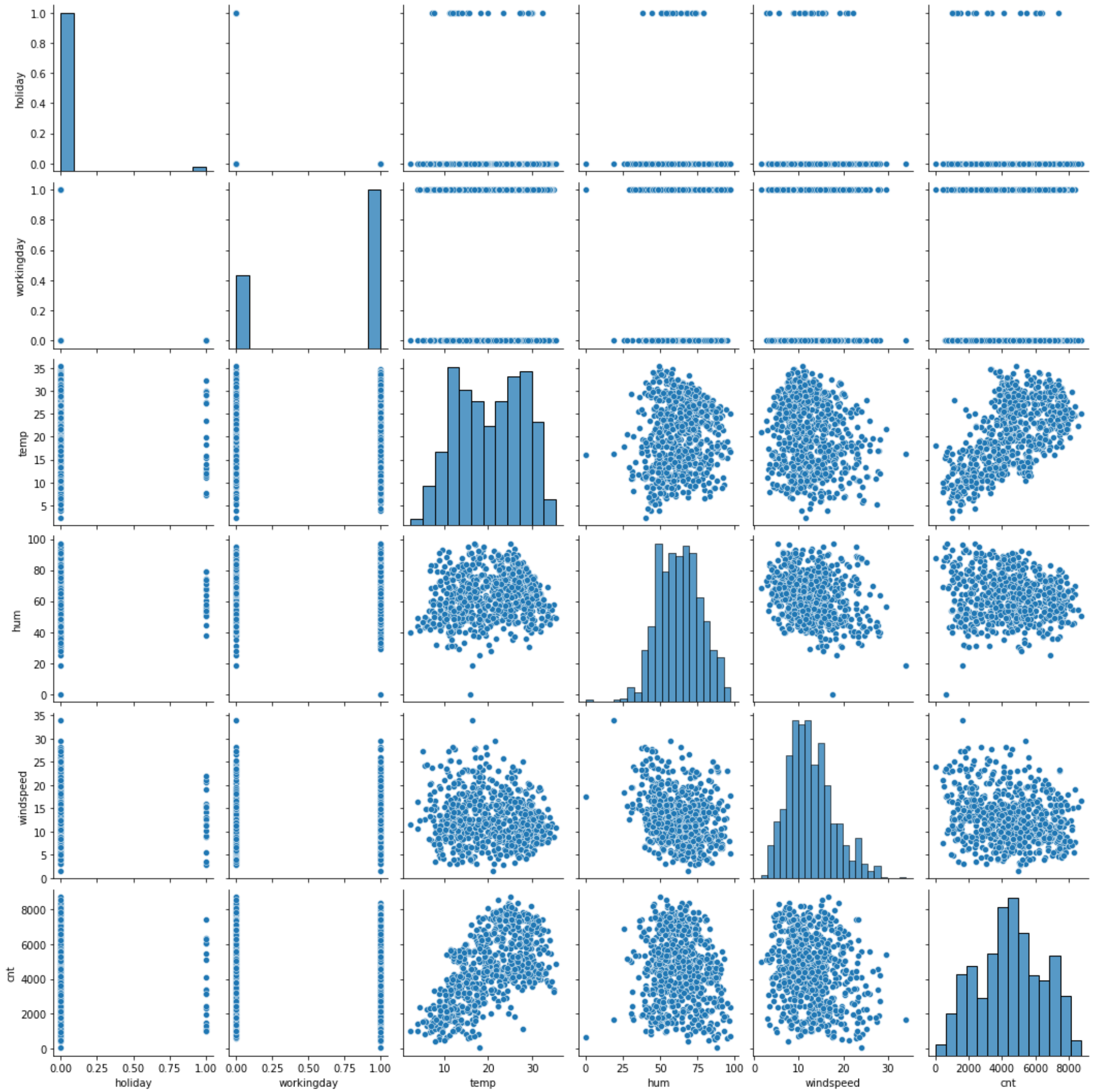
For ex: For seasons Sumer, winter, fall and Spring – Fall is dropped. And it can be represented with 0 value of all seasons.

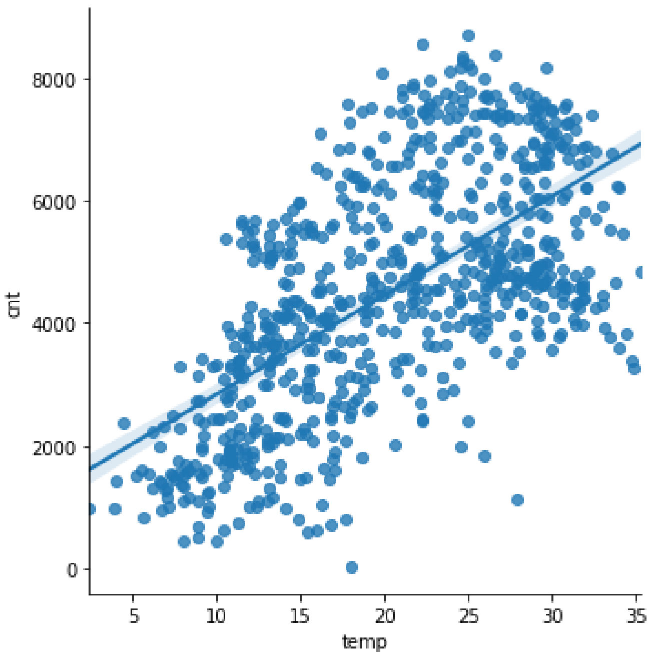
| Season | Summer | Spring | Winter |
|---|---|---|---|
| Summer | 1 | 0 | 0 |
| Spring | 0 | 1 | 0 |
| Winter | 0 | 0 | 1 |
| Fall | 0 | 0 | 0 |

Combined values of Summer, Spring and Winter can represent Fall. Hence by dropping Fall, co-relation is reduced.

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

Temp has the highest corelation with target variable cnt.

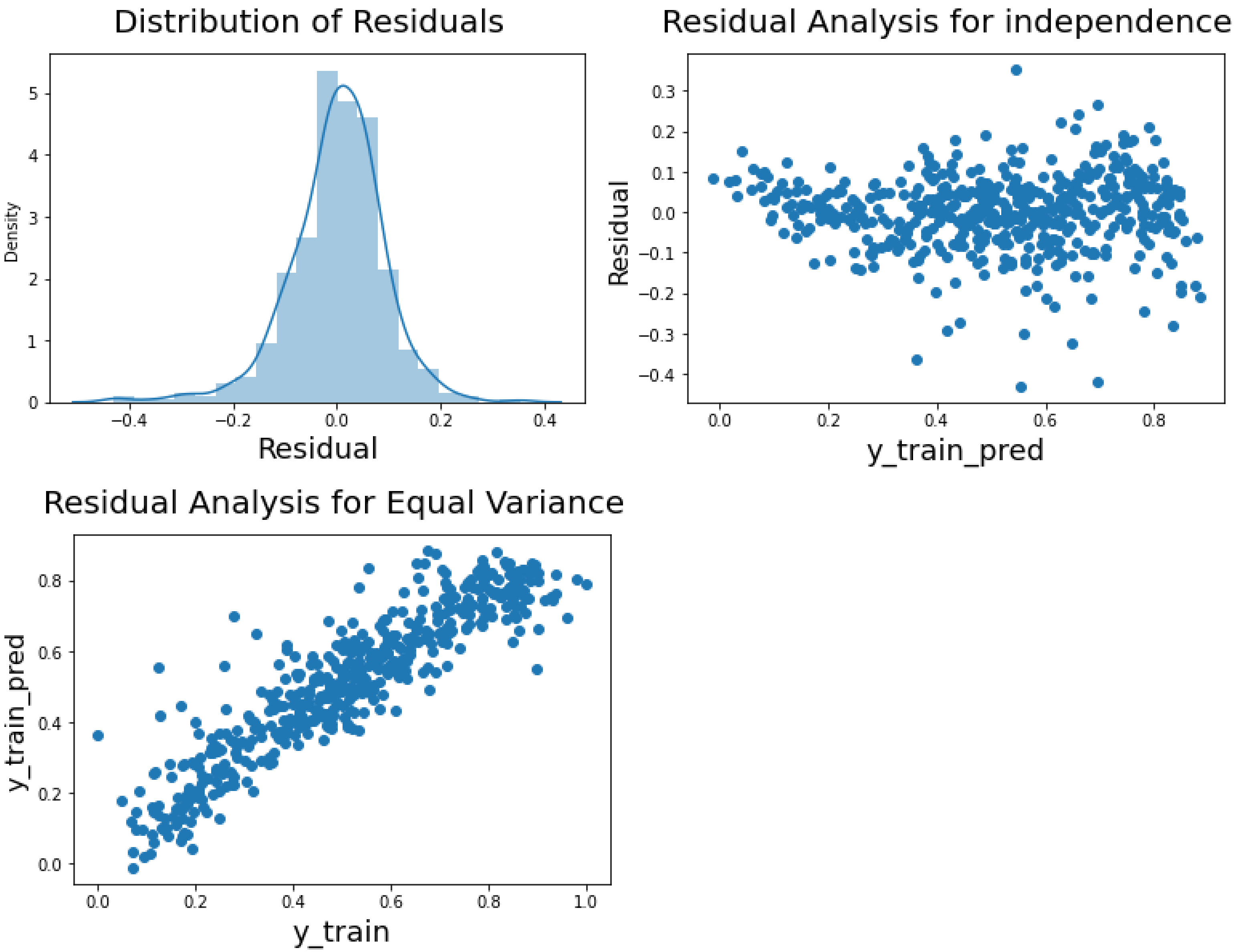

holiday
workingday
temp
hum
windspeed
cnt
0.0
0.2
0.4
0.6
0.8
1.0
5
10
15
20
25
30
35
0
20
40
60
80
100
0
2000
4000
6000
8000
0.00
0.25
0.50
0.75
1.00
0
25
50
75
100


cnt
0
2000
4000
6000
8000
temp
5
10
15
20
25
30
35

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

Below assumptions are full filled by the training data set.

1. Error terms are normally distributed with mean zero.
2. Error terms are independent of each other
3. Error terms have constant variance (homoscedasticity)



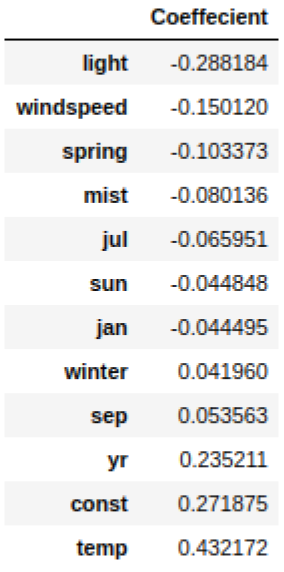
Distribution of Residuals

Residual Analysis for independence

Residual Analysis for Equal Variance

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

Based on below model coefficients top 3 features are - **temp, yr, weathersit**



| | Coeffecient |
|---|---|
| light | -0.288184 |
| windspeed | -0.150120 |
| spring | -0.103373 |
| mist | -0.080136 |
| jul | -0.065951 |
| sun | -0.044848 |
| jan | -0.044495 |
| winter | 0.041960 |
| sep | 0.053563 |
| yr | 0.235211 |
| const | 0.271875 |
| temp | 0.432172 |

**General Subjective Questions**

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

Linear Regression is a statical model to analyse the linear relationship between independent and dependent variables.

Key – points:

1. It is Supervised machine learning method.
2. The dependent variable to be predicted is a continuous variable.
3. The independent variables are also known as the predictor variable. And the dependent variables are also known as the output variables.

4. It is of two types -
    a. Simple Linear Regression - One independent variable
    b. Multiple Linear Regression – Multiple independent variables

**It is represented by formula:**

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_pX_p+\epsilon$$

**Strength of Simple Linear Regression:**

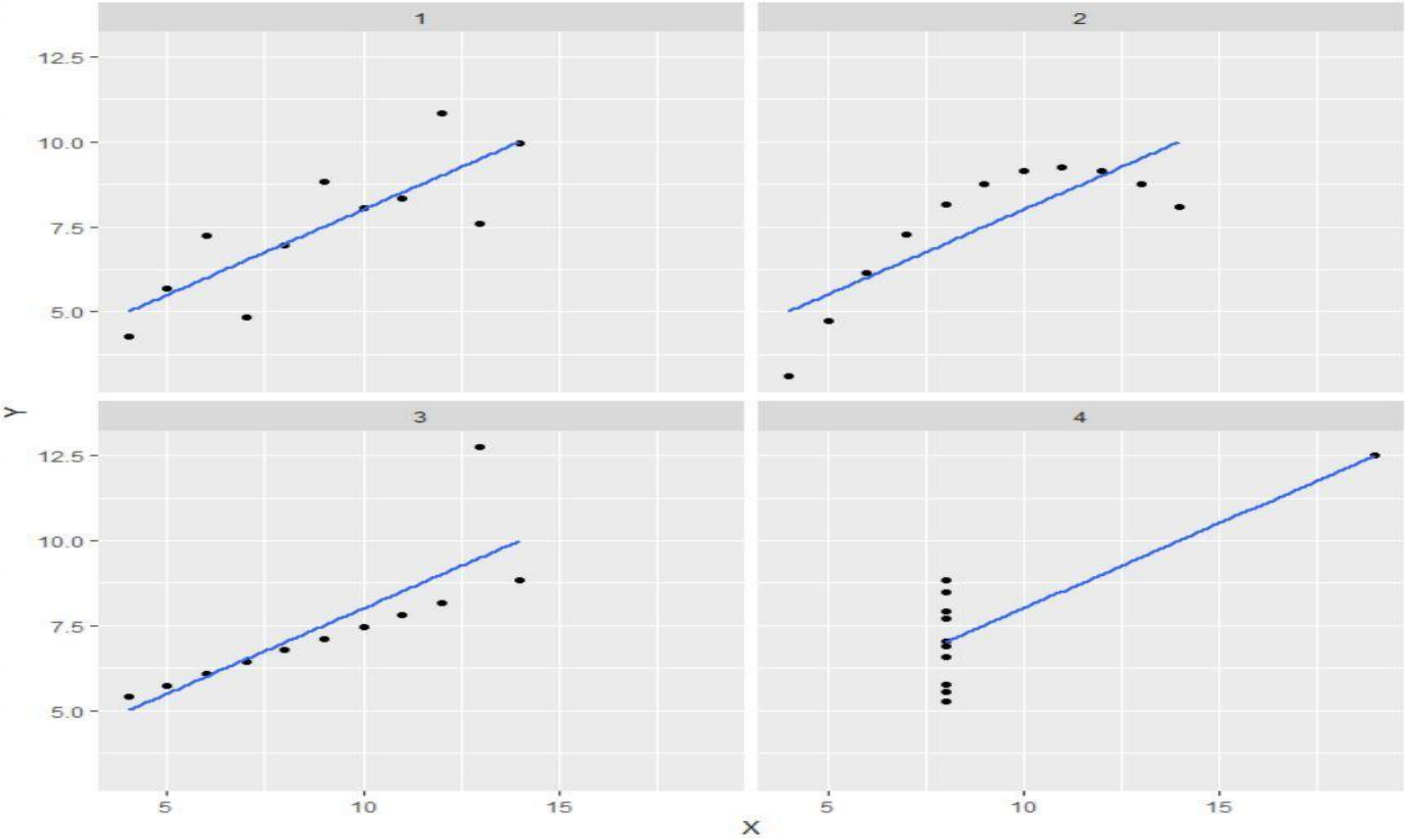
1. Accuracy – How well the best fit linear line represents the data. (It is measured by r-squared)
2. Prediction- How well best fit linear line predict the new data. (It is measured by P-value)

**Assumptions of Linear Regression:**

1. Multi-collinearity − Linear regression model assumes that there is very little or no multi-collinearity in the data. Basically, multi-collinearity occurs when the independent variables or features have dependency in them.
2. Relationship between variables − Linear regression model assumes that the relationship between output and feature variables must be linear.
3. There is no dependency between residual errors. It can be checked by below
    a. Error terms are normally distributed with mean zero
    b. Error terms are independent of each other
    c. Error terms have constant variance (homoscedasticity)
4. Overfitting - the model should not be too complex such that the training accuracy is high while the test accuracy is very low. Basically – model should be able to generalize new data points as well.

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

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties (mean, variance, corelation coefficient and best-fit linear regression line), yet appear very different when graphed.

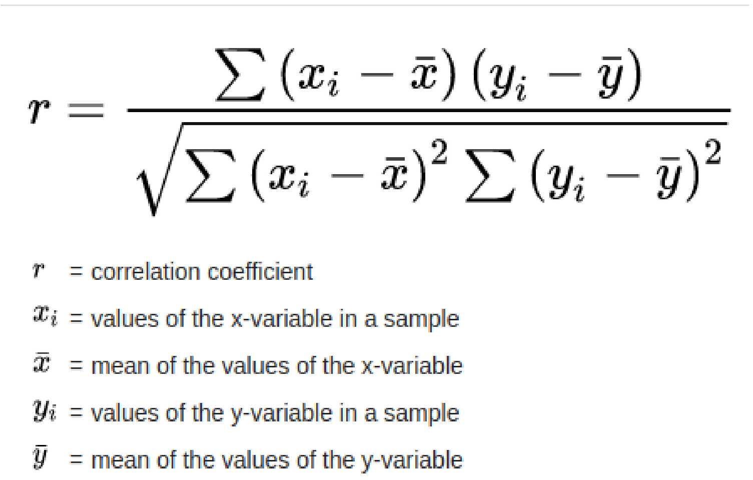


Data set of all 4 graphs have same statistical properties. But their plot is very different. **Hence, we should always plot the graph and not only rely on statistical properties.**

**3. What is Pearson's R? (3 marks)**

It measures the strength of linear relationship between two variables.

**Formula:**



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

$r$ = 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

**Its value lies between –1 to 1.**

a. 1 means perfect positive relationship.
b. -1 means perfect negative relationship.
c. 0 means no linear relationship.

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

Scaling is a process to convert the features of different magnitude into single scale.

Ex: Feature A has values between 1000 and 5000

Feature B has value between .0001 and 1.

Scaling will convert the data of both feature in single scale. (ex: value between 0 and 1.)

**it helps in speeding up the calculations in an algorithm.**

**Standardisation** basically converts all of the data into a standard normal distribution with mean zero and standard deviation one.

Standardisation: $X = x - \text{mean}(x)/\text{sd}(x)$

**MinMax scaling (normalized scaling)** converts all of the data in the range of 0 and 1.

MinMax Scaling: $X = x - \text{min}(x)/\text{max}(x) - \text{min}(x)$

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

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

**Formula:**

$$VIF_i = \frac{1}{1 - {R_i}^2}$$

If R squared is 1 then VIF becomes infinity. And R squared is 1 means model is able to interpret all values (Complete Fit). Hence, perfect collinearity.

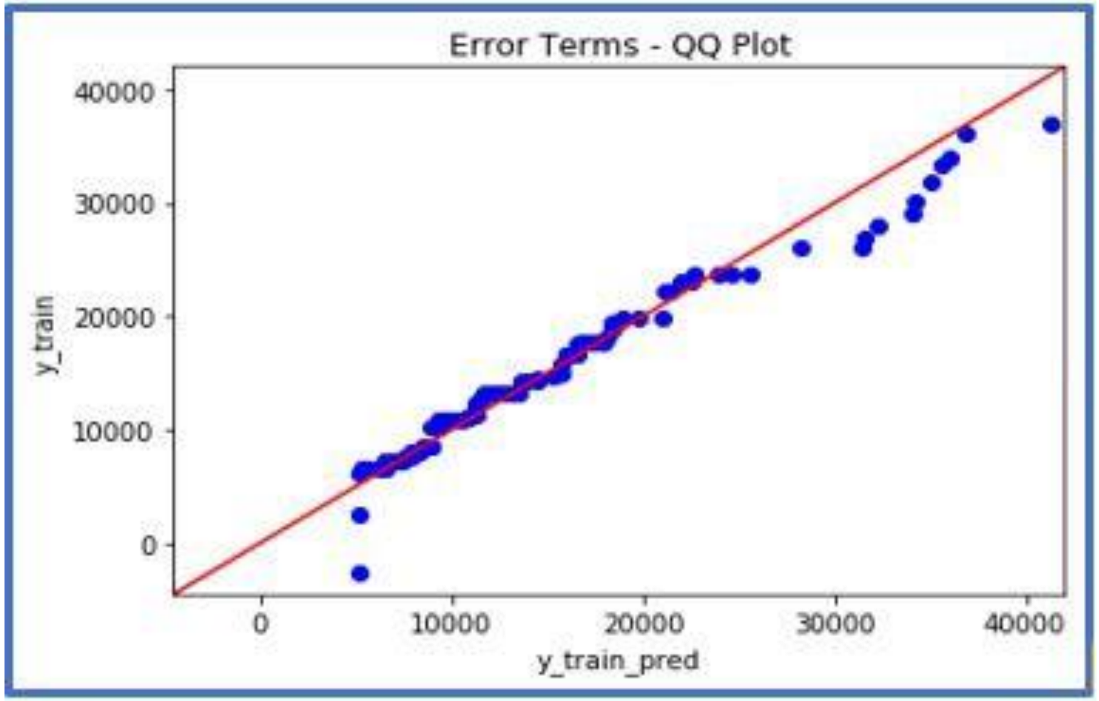
**Conclusion:**

| VIF | Conclusion |
|---|---|
| 1 | No multi collinearity |
| <5 | Moderate |
| >10 | Severe |
| Infinite | Perfect collinearity |

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

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from th e same distribution, we should see the points forming a line that's roughly straight.

In Linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.


Error Terms - QQ Plot
y_train
y_train_pred