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?

Inferences for the categorical variables are,

- The value of Cnt is comparatively less in the spring season, the bike-sharing company can do better promotion and marketing campaigns to boost the business during the Spring

- 2019 has more users compared to 2018, bike-sharing systems are slowly gaining popularity, and the demand for these bikes is increasing every year

- The median value of the number of users is more in the months of May to Oct, compared to other months

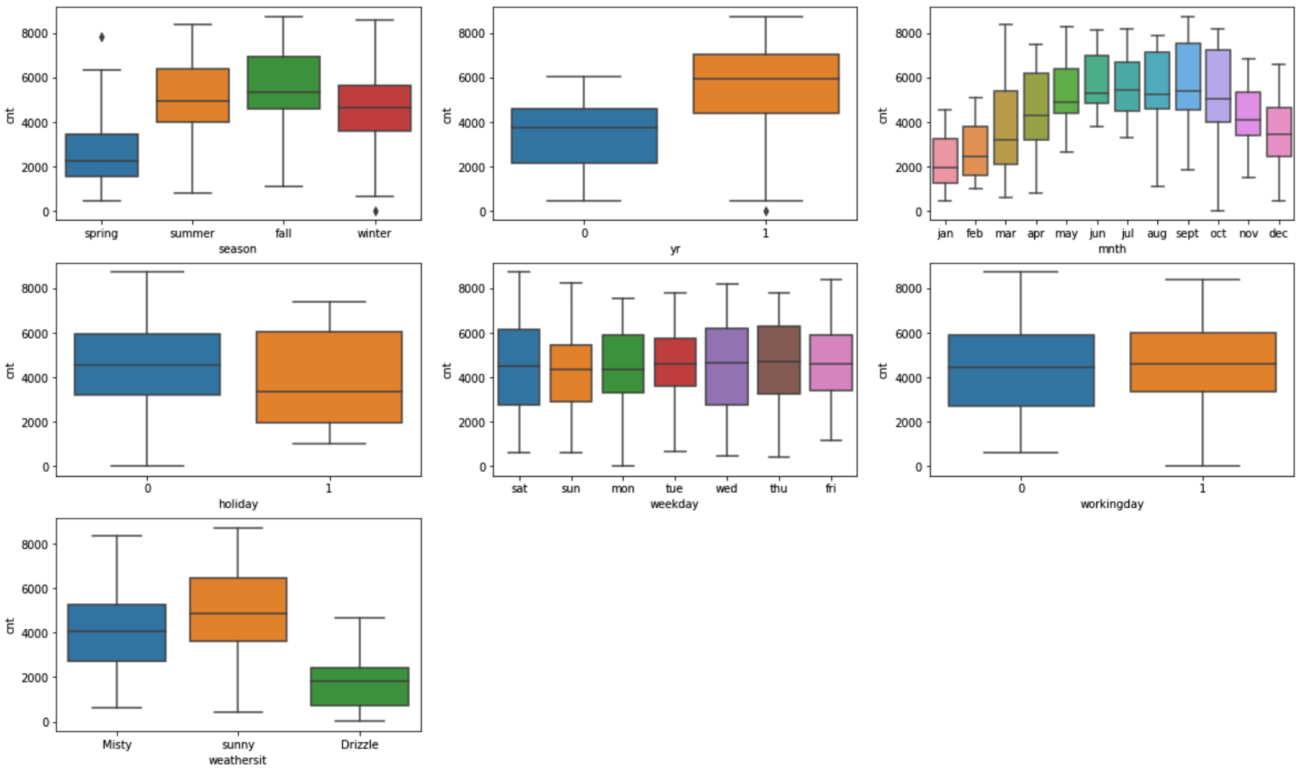
- When it's not a holiday, the number of users is comparatively high

- By looking at the upper whisker, Sat & Fri seems to be the highest. However median value is almost the same for all the days

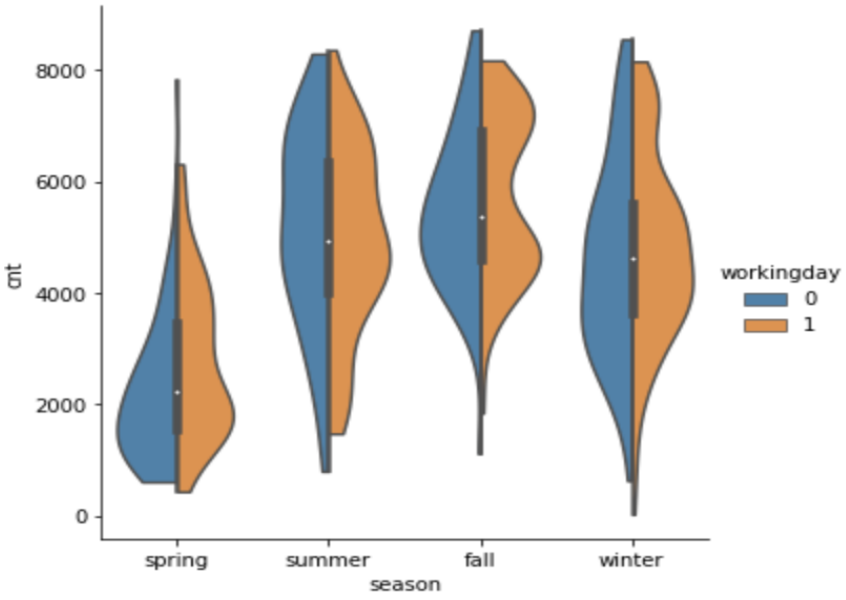
- The number of users is slightly higher when comparing the 25th, 50th and 75th percentile values during the working day

- During sunny, more users seem to use the bike and during Drizzle the number of users is drastically reduced, The bike system app can consider providing rainproof helmets which might help to increase the number of users

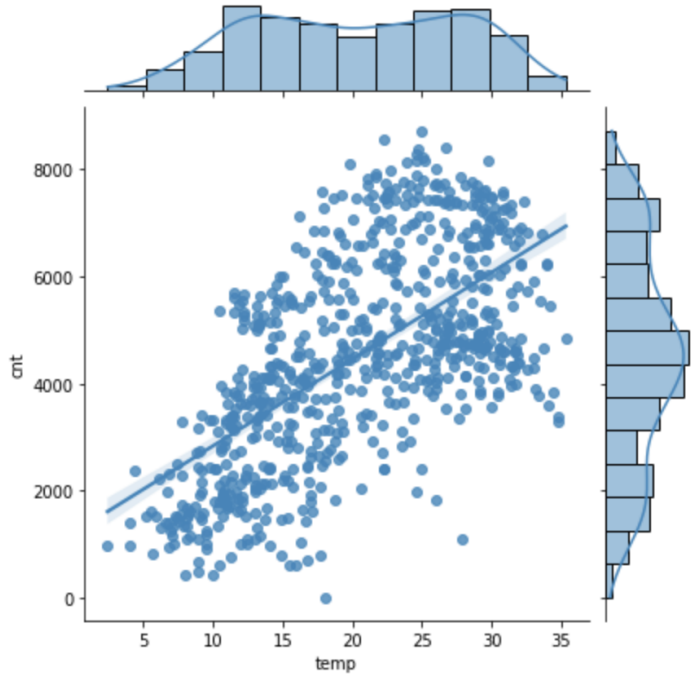
The inferences are derived from the below charts,

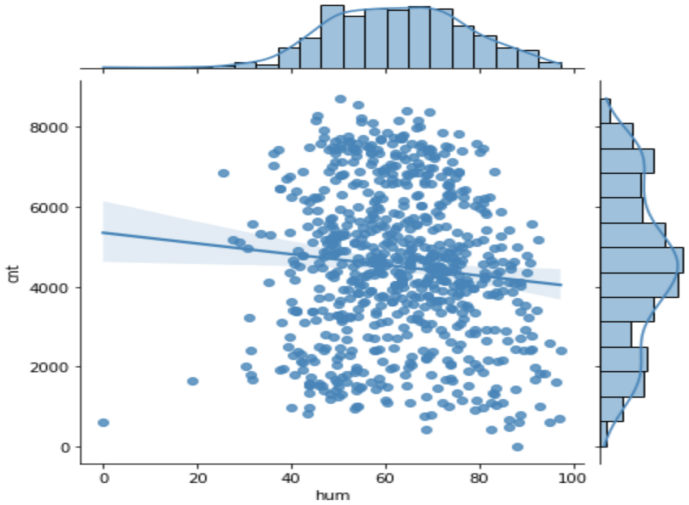


It's quite evident Spring has the least number of users and it's also interesting to observe, that users prefer to use the biking system during the working day.



The reg line for **temp** is perfect because of the high correlation , we attempted the same with **hum** but its not strongly correlated with **cnt** as like temp





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

The drop_first=True helps us to get N-1 levels which is sufficient to represent the data when we have N levels for the given categorical variables. Let's understand this with an example,

]:

	furnished	semi-furnished	unfurnished
0	1	0	0
1	1	0	0
2	0	1	0
3	1	0	0
4	1	0	0

Now, we don't need three columns. We can drop the furnished column, as the type of furnishing can be identified with just the last two columns where —

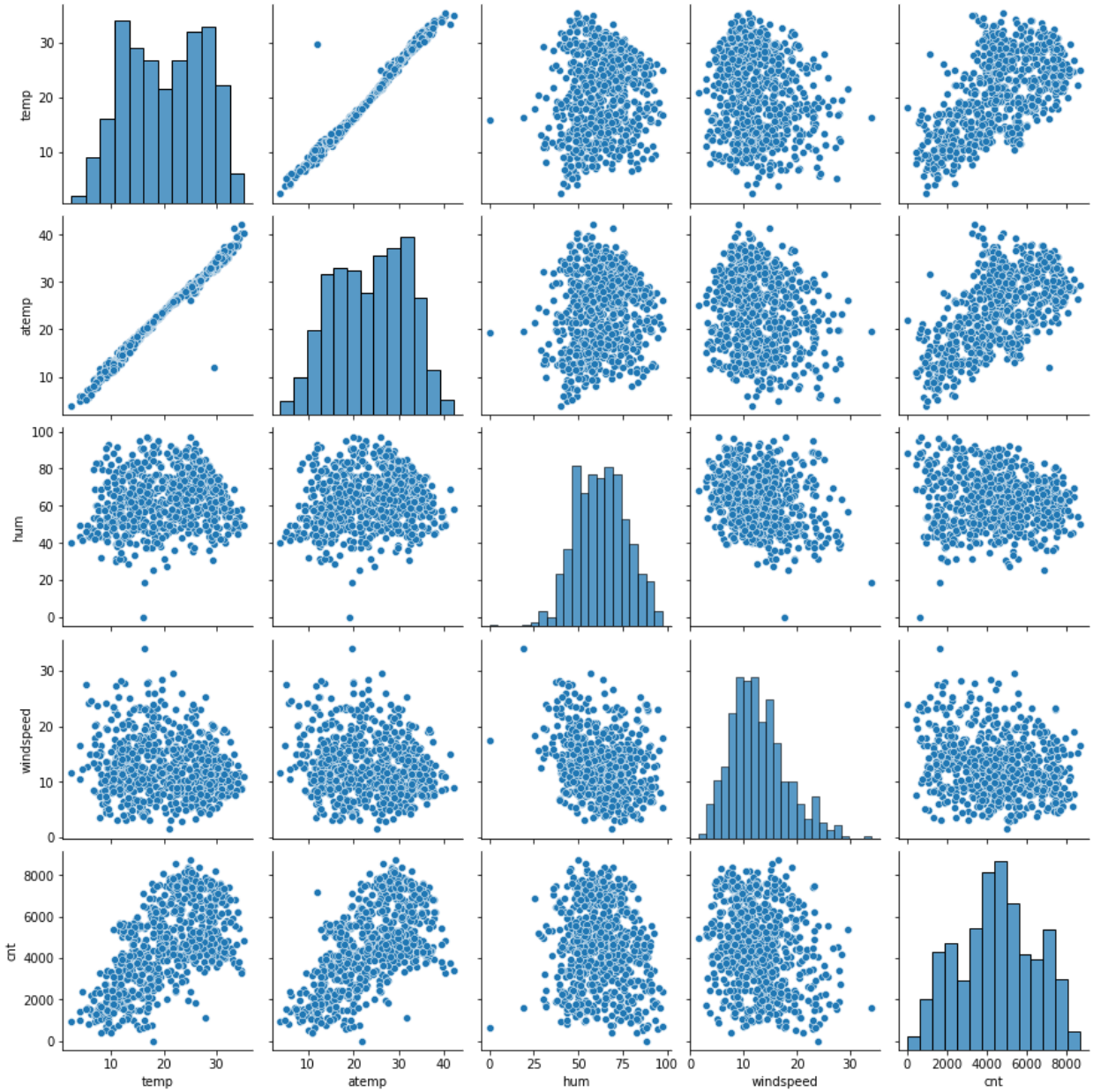
- 00 will correspond to the furnished
- 01 will correspond to unfurnished
- 10 will correspond to semi-furnished

So we have N-1 which is 2 levels to explain the above variable, the final dataset will look like below,

semi-furnished	unfurnished
0	0
0	0
1	0
0	0
0	0

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

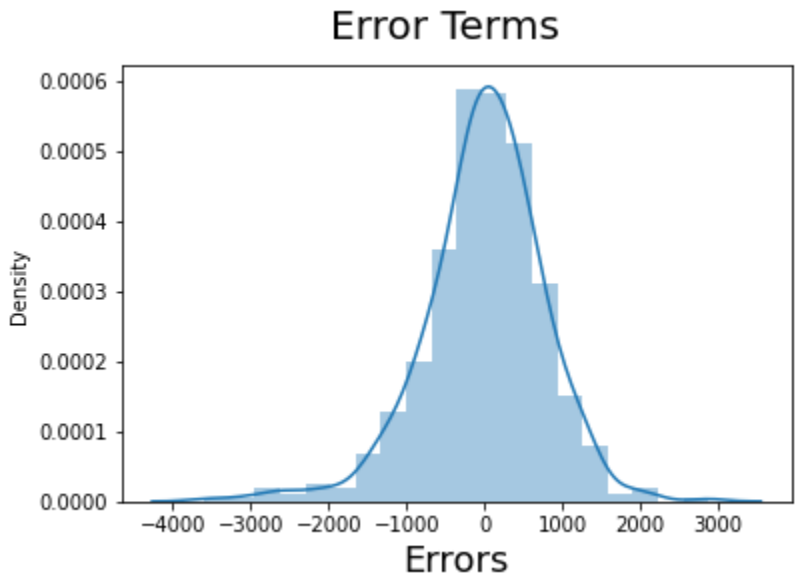
atemp and **temp** are the two variables which has high correlation with the target variable **cnt**

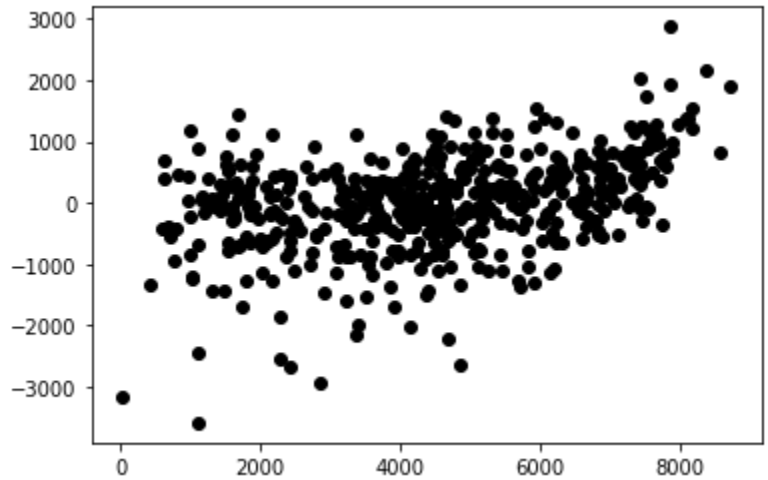


4. How did you validate the assumptions of Linear Regression after building the model on the training set?

The below assumptions are validated using the below charts ,

- Error terms are normally distributed with a mean zero (please refer chart 1)
- Error terms are independent of each other (please refer chart 2)
- Error terms have constant variance (homoscedasticity) (please refer chart 2)





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

The top 3 features based on the +ve coefficient value are,

- temp
- mnth_sept
- Season_winter

Note - Though yr has a high coefficient value, its not a business column which can influence the target other than every yr we expect an increase in the number of users which is business goal

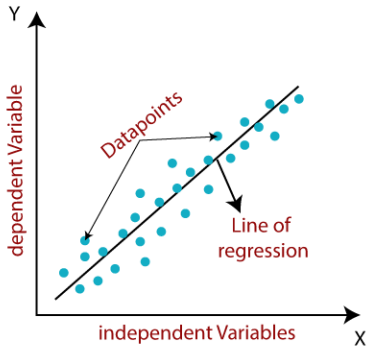
General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as sales, salary, age, product price, etc.

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable changes according to the value of the independent variable.

The linear regression model provides a sloped straight line representing the relationship between the variables. For example -



Equation of linear regression

$$y = c + m_1x_1 + m_2x_2 + \ldots + m_nx_n$$

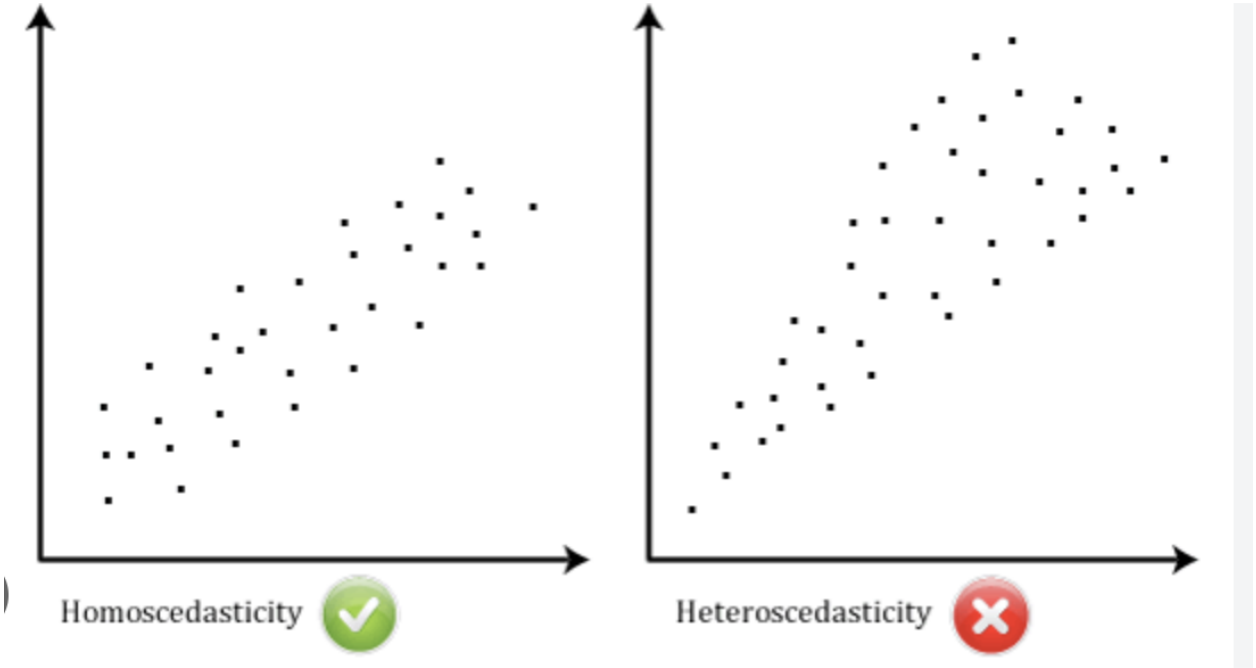
- y is the response
- c is the intercept
- m_1 is the coefficient for the first feature
- m_n is the coefficient for the nth feature

Assumptions of Linear Regression:

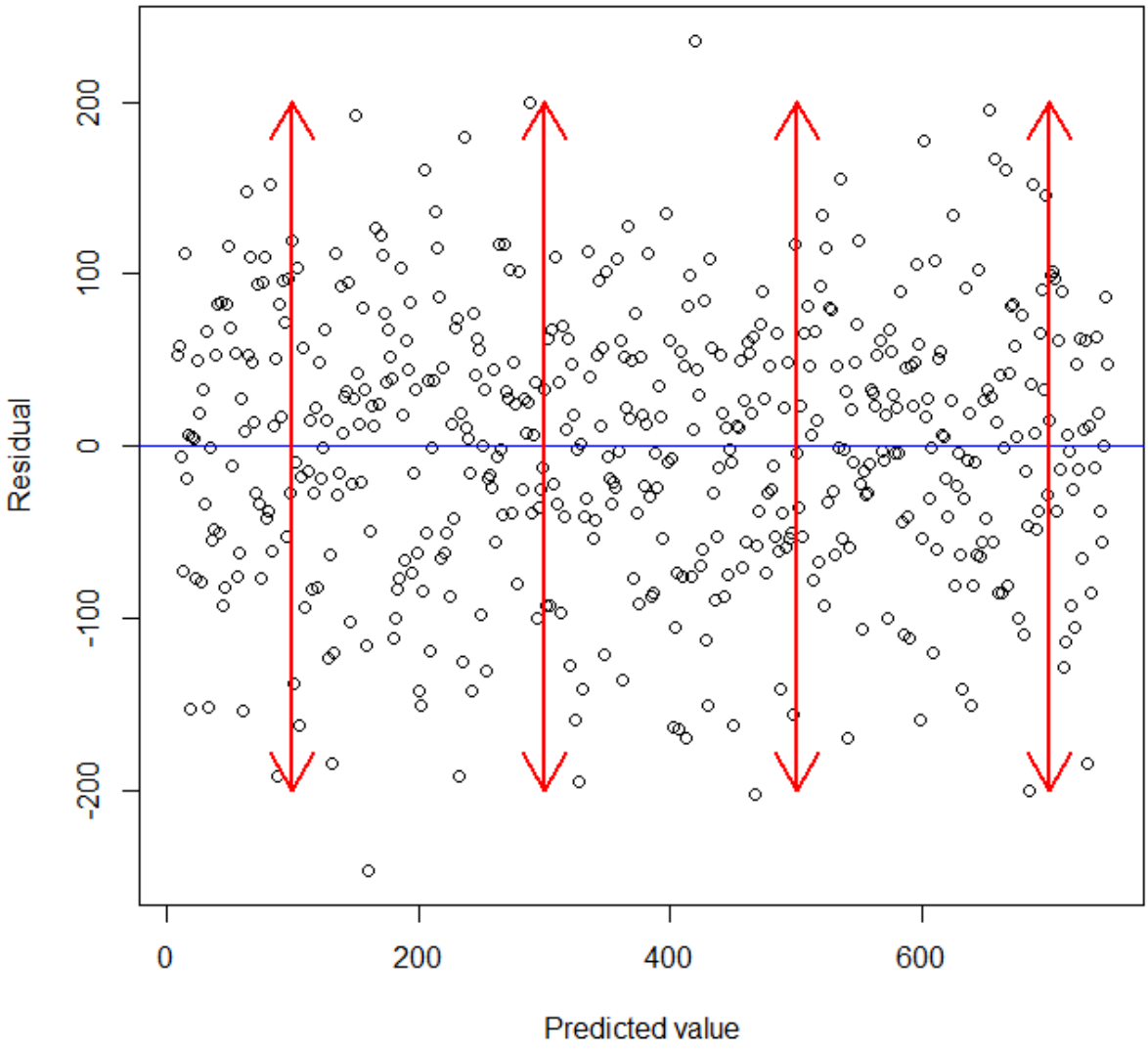
- ❖ Linear relationship - Relationship between X and Y variable should be linear
- ❖ Error terms should be normally distributed - The error terms which is distribution plot on the residuals will form normal distribution
- ❖ No or little multicollinearity - The predictor variables shouldnt have high correlation among themselves,ariance Inflation Factor (VIF) measures the severity of multicollinearity in regression analysis. It is a statistical concept that indicates the increase in the variance of a regression coefficient as a result of collinearity.

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

- Homoscedasticity - The assumption of homoscedasticity (meaning “same variance”) is central to linear regression models. Homoscedasticity describes a situation in which the error term (that is, the “noise” or random disturbance in the relationship between the independent variables and the dependent variable) is the same across all values of the independent variables



- Plot between Residual vs Preddicted value should not be any predictable pattern



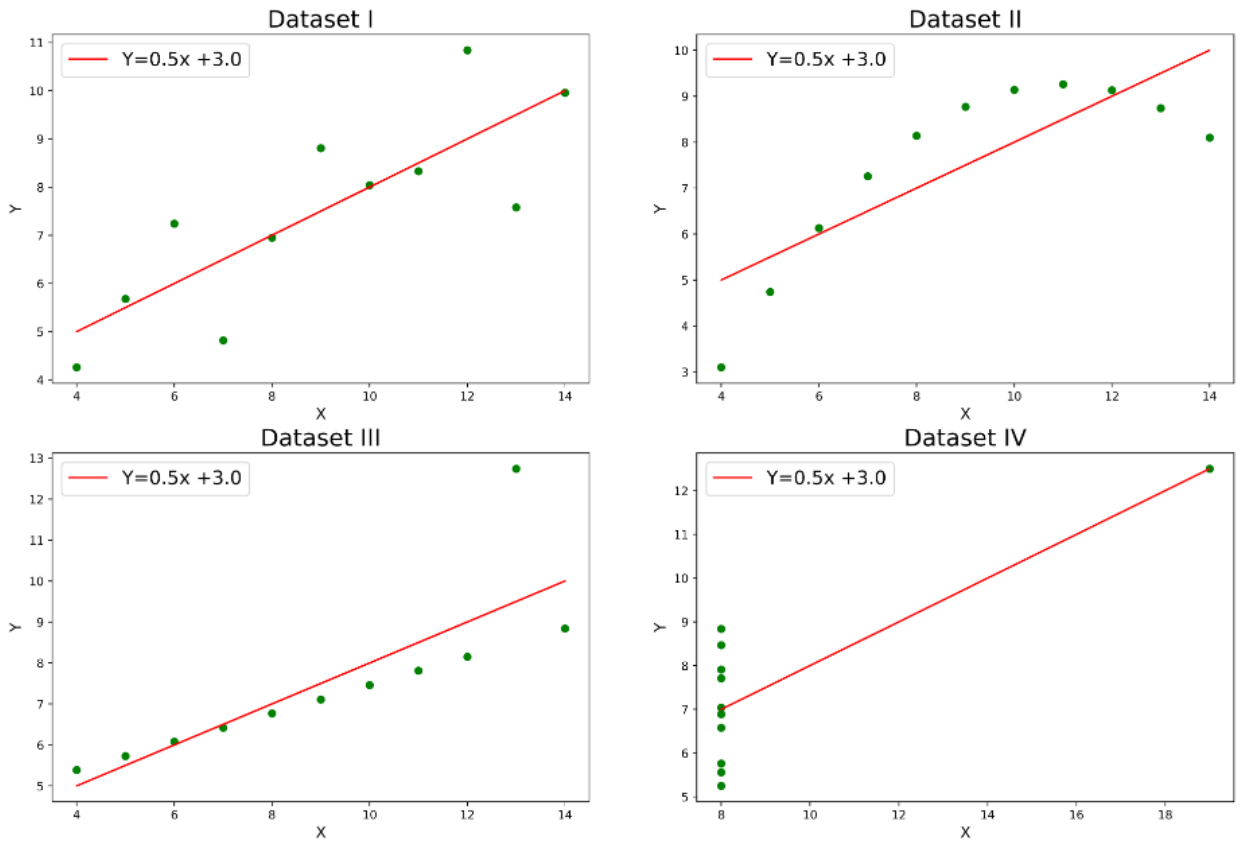
2. Explain the Anscombe's quartet in detail.

Anscombe's quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on graph. The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

In layman's terms, the summary statistics will give misleading proof that all datasets are the same but while we plot we can find what is the significant difference between the variables



If we observe the data points for the charts, they do not always form a linear relationship, every dataset has a different relationship, however, the mathematical

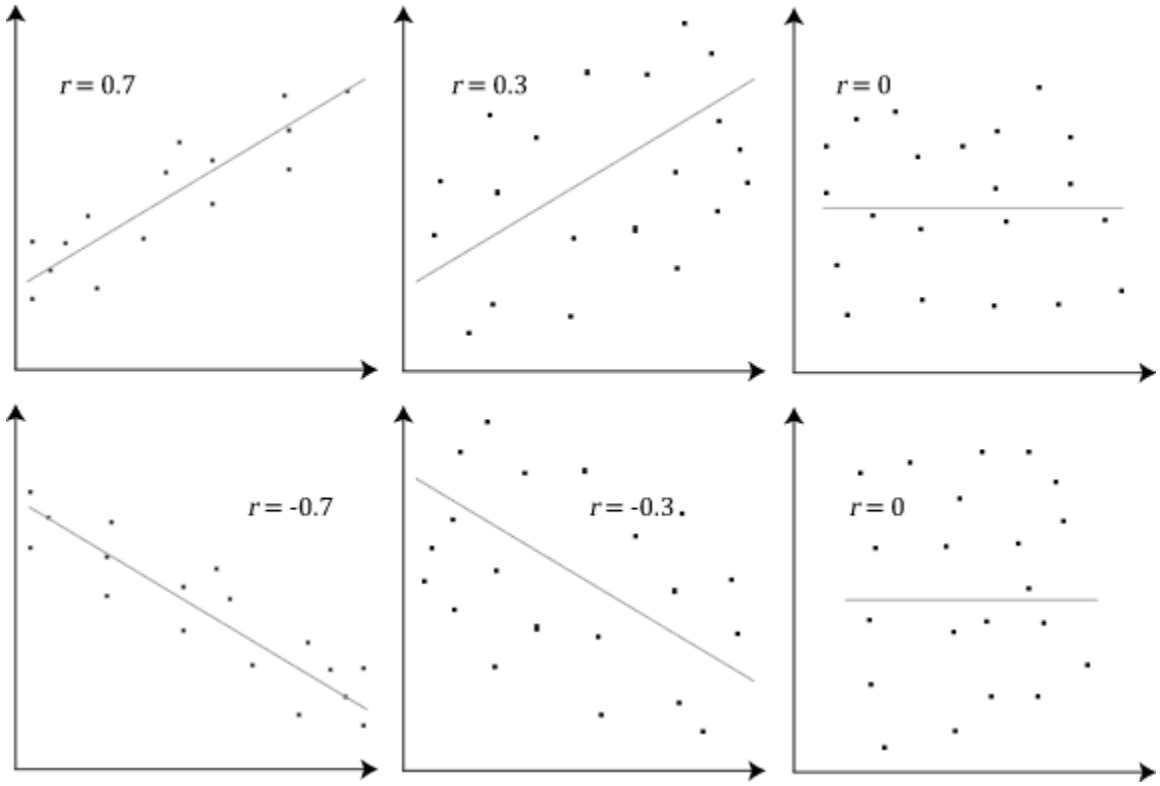
equation is same for all the dataset which means summary statistics doesn't have any significance. This also proves the power of visualisation.

3. What is Pearson's R?

The Pearson correlation coefficient (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.

Let us understand this with an example graph to understand the relationship between x and y,

r=0.7 => Strong Positive correlation
r=-0.7 => Strong negative correlation
r=0.3 => weak positive correlation
r=-0.3 => weak negative correlation
r=0 => Zero correlation or no relationship between two variables



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

Feature scaling is a data preprocessing technique used to transform the values of features or variables in a dataset to a similar scale. The purpose is to ensure that all features contribute equally to the model and to avoid the domination of features with larger values.

Feature scaling becomes necessary when dealing with datasets containing features that have different ranges, units of measurement, or orders of magnitude. In such cases, the variation in feature values can lead to biased model performance or difficulties during the learning process.

There are several common techniques for feature scaling, including standardization, normalization, and min-max scaling. These methods adjust the feature values while preserving their relative relationships and distributions.

By applying feature scaling, the dataset's features can be transformed to a more consistent scale, making it easier to build accurate and effective machine learning models. Scaling facilitates meaningful comparisons between features, improves model convergence, and prevents certain features from overshadowing others based solely on their magnitude.

Machine learning algorithms like linear regression, logistic regression, neural network, PCA (principal component analysis), etc., that use gradient descent as an optimization technique require data to be scaled. Take a look at the formula for gradient descent below:

Gradient descent formula
The presence of feature value X in the formula will affect the step size of the gradient descent. The difference in the ranges of features will cause different step sizes for each feature. To ensure that the gradient descent moves smoothly towards the minima and that the steps for gradient descent are updated at the same rate for all the features, we scale the data before feeding it to the model.

Normalization

Normalization is a data preprocessing technique used to adjust the values of features in a dataset to a common scale. This is done to facilitate data analysis and modeling, and to reduce the impact of different scales on the accuracy of machine learning models.

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

Here's the formula for normalization:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Here, Xmax and Xmin are the maximum and the minimum values of the feature, respectively.

- When the value of X is the minimum value in the column, the numerator will be 0, and hence X' is 0
- On the other hand, when the value of X is the maximum value in the column, the numerator is equal to the denominator, and thus the value of X' is 1
- If the value of X is between the minimum and the maximum value, then the value of X' is between 0 and 1

Standardization

Standardization is another scaling method where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero, and the resultant distribution has a unit standard deviation.

Here's the formula for standardization:

$$X' = \frac{X - \mu}{\sigma}$$

μ is the mean of the feature values and σ is the standard deviation of the feature values. Note that, in this case, the values are not restricted to a particular range.

Now, the big question in your mind must be when should we use normalization and when should we use standardization? Let's find out!

Normalization	Standardization
Rescales values to a range between 0 and 1	Centers data around the mean and scales to a standard deviation of 1
Useful when the distribution of the data is unknown or not Gaussian	Useful when the distribution of the data is Gaussian or unknown
Sensitive to outliers	Less sensitive to outliers
Retains the shape of the original distribution	Changes the shape of the original distribution
May not preserve the relationships between the data points	Preserves the relationships between the data points
Equation: (x – min)/(max – min)	Equation: (x – mean)/standard deviation

However, at the end of the day, the choice of using normalization or standardization will depend on your problem and the machine learning algorithm you are using. There is no hard and fast rule to tell you when to normalize or standardize your data. **You can always start by fitting your model to raw, normalized, and standardized data and comparing the performance for the best results.**

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

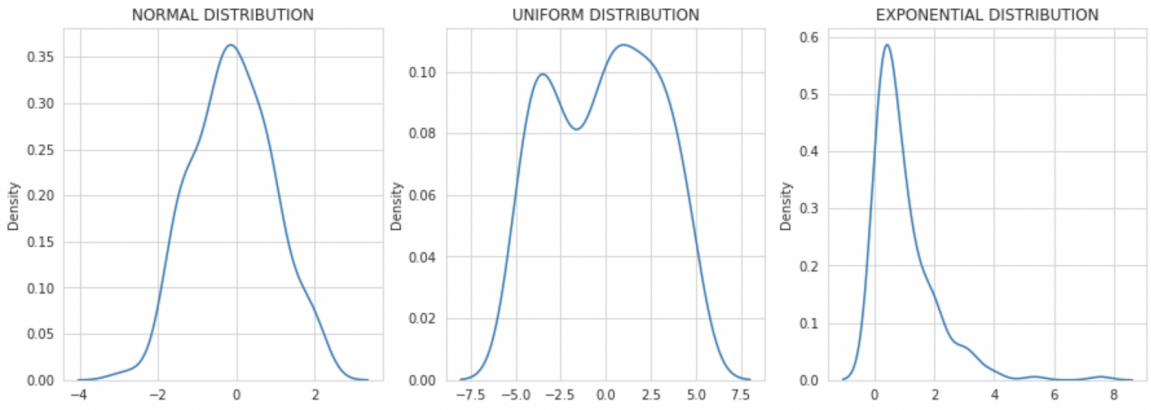
The Variance Inflation Factor (VIF) measures the severity of multicollinearity in regression analysis. It is a statistical concept that indicates the increase in the variance of a regression coefficient as a result of collinearity.

The common heuristic we follow for the VIF values is:
> 10: Definitely high VIF value and the variable should be eliminated.
> 5: Can be okay, but it is worth inspecting.
< 5: Good VIF value. No need to eliminate this variable.

When the VIF is infinity, it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential



In probability distributions, we represent data using charts where the x-axis represents the possible values of the sample and the y-axis represents the probability of occurrence.

There are various probability distribution types like Gaussian or Normal Distribution, Uniform distribution, Exponential distribution, Binomial distribution, etc.

QQ plots is very useful to determine

- If two populations are of the same distribution
- If residuals follow a normal distribution. Having a normal error term is an assumption in regression and we can verify if it's met using this.
- Skewness of distribution

In Q-Q plots, we plot the theoretical Quantile values with the sample Quantile values. Quantiles are obtained by sorting the data. It determines how many values in a distribution are above or below a certain limit.

If the datasets we are comparing are of the same type of distribution type, we would get a roughly straight line. Here is an example of normal distribution.