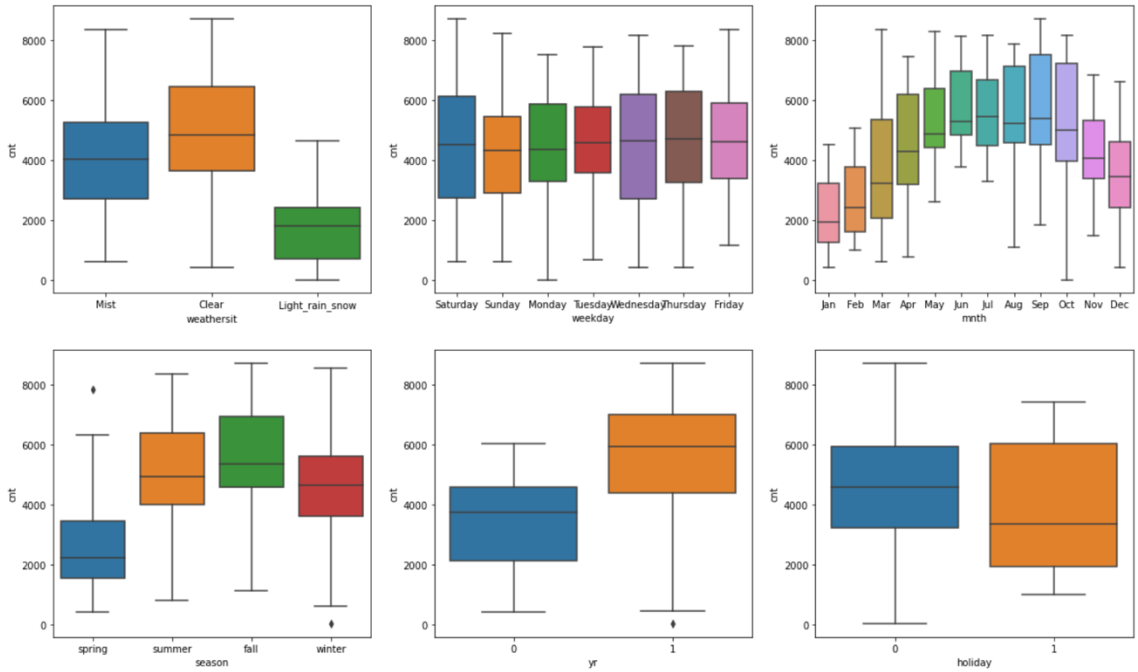
Assignment-based Subjective Questions

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

Ans . The categorical variables are : 'mnth', 'weekday', 'season' , 'weathersit' ,'yr' and 'holiday'. Based on the boxplot analysis we can infer following:



A. Weathersit : There is no data for Heavy rain/snow weather indicating this is extremely unfavourable weather for bike renting. The highest bike renting demand occurs in Clear weather while the Light rain/snow weather decreases the demand
B. Weekday : No significant conclusion can be made based on this
C. Month : The demand increases from Jan till Sep where it is maximum, henceforth the demand decreases.
D. Season : Spring has lowest demand while Fall has the highest demand. Summer and winter has intermediate demand.
E. Yr : Year 2019 has higher demand than 2018
F. Holiday : Holiday reduces the bike demand. It appears most people prefer renting bike on non-holiday day.

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

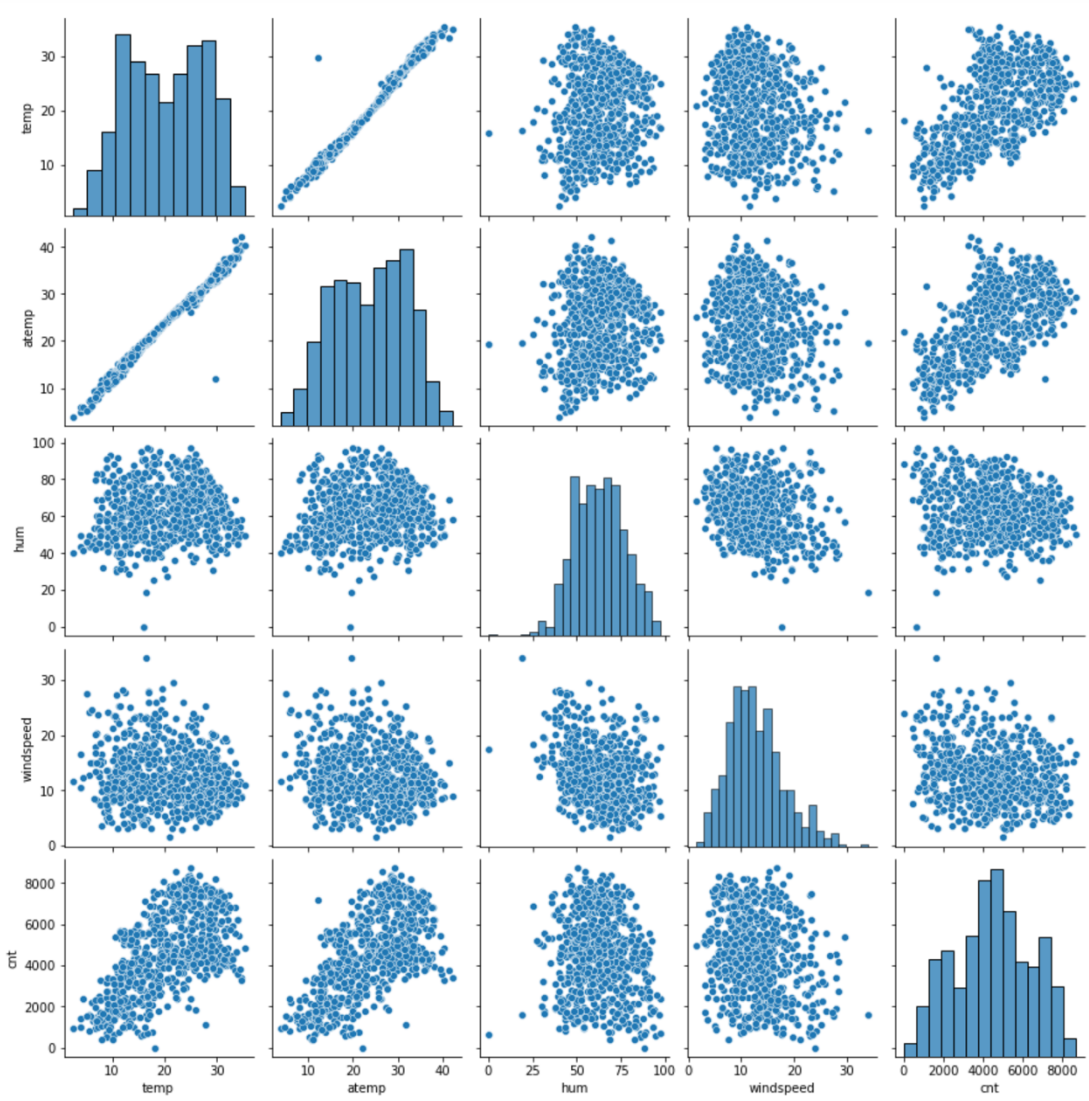
Ans. Reasons for dropping first dummy variable :

1. To reduce multicollinearity : If we do not drop the first column during dummy variable creation , it will lead to high multicollinearity between the dummy variables and will adversely affect the model.
2. To reduce redundancy. For eg for a variable gender , both male and female dummy are not required. Male=0 will be a female.

However , sometimes it depends on the number of values in a categorical variables. For a categorical variable with large number of values, the drop_first could be avoided to see the effect of all the values of the variable. For eg a categorical variable=Month , here all values Jan-Dec should be created as dummy column.

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

Ans : Based on the pair-plot among numerical variables temp has highest correlation with the target variable cnt.

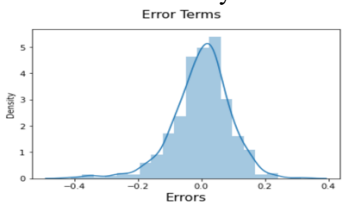


4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans

The assumptions of the Linear Regression model can be validated in following way :

A. Error terms are normally distributed with mean zero (not X, Y)

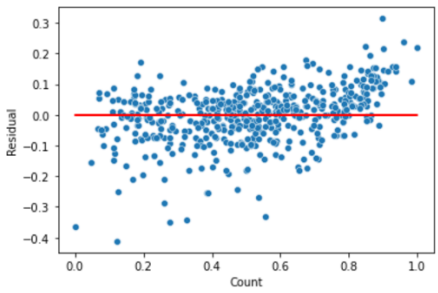


Residual distribution should follow normal distribution and centred around 0. We confirm this by plotting a distplot of the residuals.

B. Independence of residuals can be calculated via the Durbin-Watson value of the model

C. The Linear distribution of the independent variables(temp,hum,windspeed) can be scatter plot with the target variable(cnt)

D. Validating Homoscedasticity i.e the residuals should not follow a pattern of distribution.



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

Ans. The top 3 features are :

- Temperature (Temp) with a coefficient of 0.549892
  - It has the most significant impact on bike rentals
- Light Rain & Snow (weathersit =3) with a coefficient of -0.287
  - This creates a negative impact on bike renting as people do not opt to take a bike in such a weather.
- Year (yr) with a coefficient of 0.233
  - Rental increase with every year

General Subjective Questions

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

Ans .

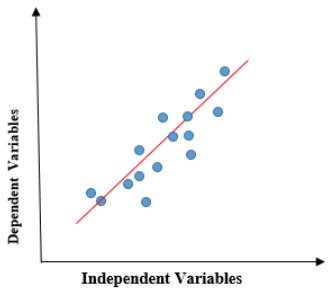
Linear regression

It is a simple statistical regression method used for predictive analysis and shows the relationship between the continuous variables. Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis).

Linear regression models can be classified into two types depending upon the number of independent variables:

- Simple linear regression: When the number of independent variables is 1
- Multiple linear regression: When the number of independent variables is more than 1

The linear regression model gives a sloped straight line describing the relationship within the variables.



The best fit regression line $Y = \beta_0 + \beta_1 X$

Here,

Y= Dependent Variable (Target Variable)
X= Independent Variable (predictor Variable)
β_0 = intercept of the line (Gives an additional degree of freedom)
β_1 = Linear regression coefficient (scale factor to each input value).

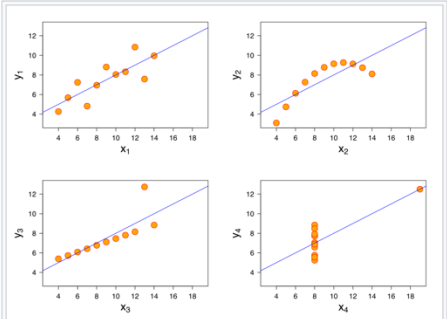
Finding the best fit line:

When working with linear regression, our main goal is to find the best fit line that means the error between predicted values and actual values should be minimized. The best fit line will have the least error.

Assumptions of Linear Regression

1. Linear relationship between the features and target
2. Small or no multicollinearity between the features:
3. Homoscedasticity Assumption: there should be no clear pattern distribution of data in the scatter plot.
4. Normal distribution of error terms: It can be checked using the q-q plot.
5. No autocorrelations: no dependency between residual errors.

2. Explain the Anscombe's quartet in detail. (3 marks)
Ans .
Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analysing it, and the effect of outliers and other influential observations on statistical properties.



- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.
- The second graph (top right) is not distributed normally; while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third graph (bottom left), the distribution is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

Importance
The quartet is used to illustrate the importance of looking at a set of data graphically before starting to analyse according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

3. What is Pearson's R? (3 marks)
Ans. It is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
Ans :
Scaling :
It is a data pre-processing technique which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Why to perform Scaling

The variables in the data set under analysis may have highly varying magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence results in incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

So we need to scale features because of two reasons:

- Ease of interpretation
- Faster convergence for gradient descent methods

Normalization/Min-Max Scaling:

It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

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

Standardization Scaling:

Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

$$x = \frac{x - mean(x)}{sd(x)}$$

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

Ans.

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

Using this formula , VIF=infinity implies R2=1, which implies a perfect correlation between two independent variables .To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

Ans.

Q-Q Plots (Quantile-Quantile plots)

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution. It helps to determine if two data sets come from populations with a common distribution.

This helps in a scenario of 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.

It compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.

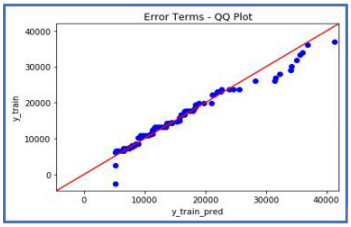
Interpretation:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

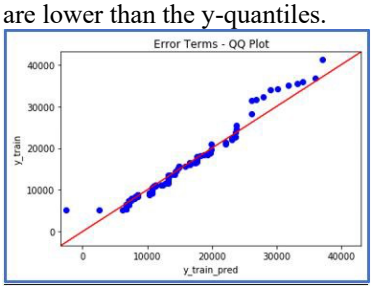
Below are the possible interpretations for two data sets.

a) **Similar distribution**: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x - axis

b) **Y-values < X-values**: If y-quantiles are lower than the x-quantiles.



c) **X-values < Y-values**: If x-quantiles are lower than the y-quantiles.



d) **Different distribution**: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis