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

Answer :

> season : In summer & fall the cnt is more than winter and that more than spring, so season have some impact on cnt
> weathersit: snow has smallest count vs clear has the highest cnt
> year : in 2019 we have cnt is much more than 2018, so year will have impact on the cnt
> mnth : in the mid of the year in month of june-oct there is highest cnt vs other mnth
> holiday : In holiday people don't prefer to go for ride, that why cnt is low for value 1
> weekday : weekday does not have much impact on cnt , so may be we can drop this in our model building
> workingday : working day does not have much impact on the higher side of the cnt , so may be we can drop this in our model building

Question : Why is it important to use drop_first=True during dummy variable creation?

Answer :

> drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

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

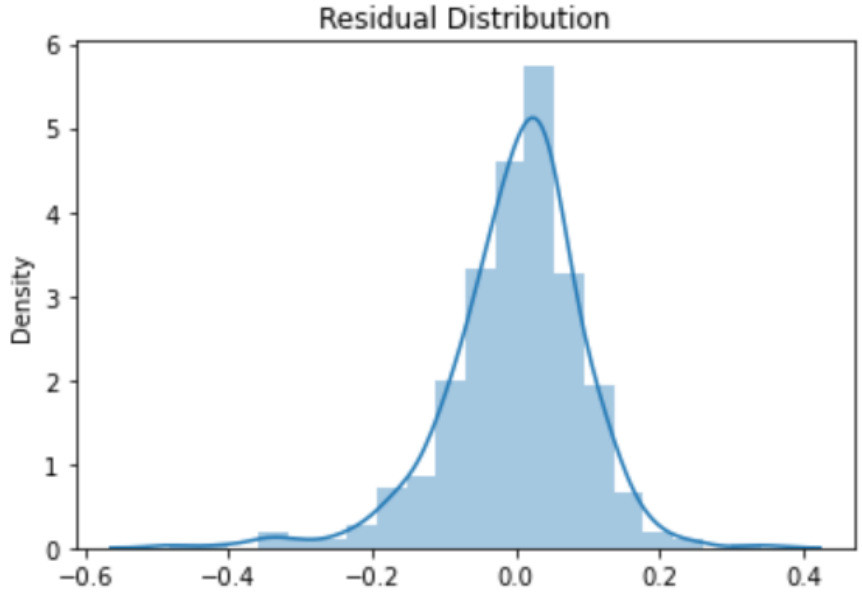
Answer :

> temp & atemp (having correlation : 0.63)

Question : . How did you validate the

assumptions of Linear Regression after building the model on the training set?

Answer :

> Error (Residual) analysis : it is normally distributed with mean 0
>
> 
>

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

Answer :

> Feature contributing to the model is temp (0.432146 positive impact) , year (0.230864 positive impact) , weathersit_Snow (-0.339344 negative impact)

```
temp                0.432146
yr                  0.230864
const               0.157514
mnth_9              0.090304
mnth_3              0.063676
mnth_10             0.061741
weekday_6           0.055632
season_winter       0.046058
workingday          0.045649
mnth_5              0.041010
weathersit_Mist    -0.074587
season_spring      -0.121536
weathersit_Snow    -0.339344
dtype: float64
```

General Subjective Questions

Question : Explain the linear regression algorithm in detail

Answer :

Linear regression model is the linear relation between the target & independent variable.
It is supervised learning, where we have defined set of target variable value.
It is used for predection or projection.
Features having high correlation have greater chances of been in the linear regression model.

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 is changing according to the value of the independent variable.

Simple Linear Regression : if there is one predictor or one independent variable its called Linear Regression.

> Multiple Linear Regression : if there is more than one predictor / independent variable its called Multiple Linear Regression.

Question : Explain the Anscombe's quartet in detail.

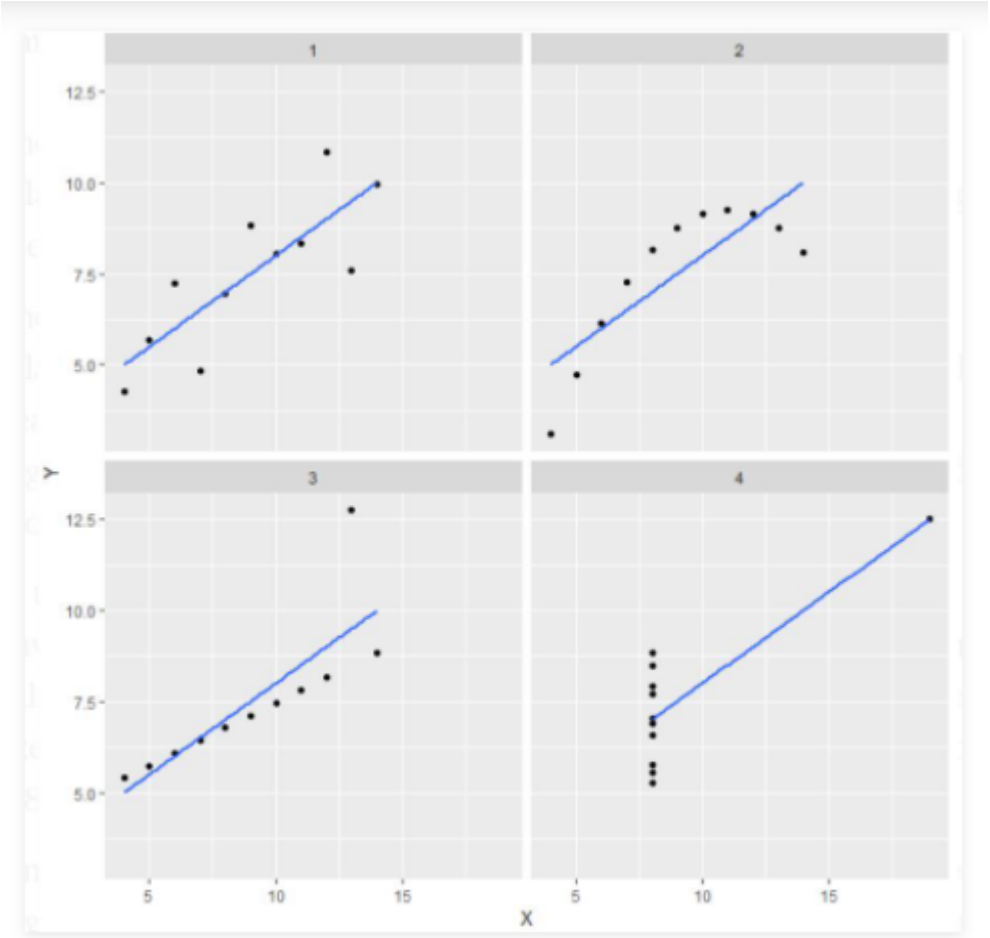
Answer :

> Anscombe's quartet is a set of 4 dataset that have same descriptive statistic (mean/mode/variance) but when we graph the plot there is lot of differences in the distribution

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Summary

Set	mean(X)	sd(X)	mean(Y)	sd(Y)	cor(X,Y)
1	9	3.32	7.5	2.03	0.816
2	9	3.32	7.5	2.03	0.816
3	9	3.32	7.5	2.03	0.816
4	9	3.32	7.5	2.03	0.817



Question: What is Pearson's R?

Answer:

Measure of linear correlation between two data set, it help to understand the linear relation and form the linear regression model on the data set.

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

Question: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer:

Scaling is done to bring the feature on the normal distribution and in range of -1 to 1

it is been performed to bring all the feature on the same scale as some variable may be measured in KM (distance) other in KG (weight) .. but when we scale those all comes in range of -1 to 1

Normalization (Min Max Scaling) : Minimum and maximum value of features are used for scaling. It is really affected by outliers. It is used when features are of different scales.

```
X_new = (X - X_min)/(X_max - X_min)
```

Standarization (Z-Score Normalization) : Mean and standard deviation is used for scaling.It is much less affected by outliers. It is used when we want to ensure zero

mean and unit standard deviation.

```
X_new = (X - mean)/Std
```

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

Answer:

If there is perfect correlation, then VIF = infinity. This 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.

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).

in case of bike sharing dataset, if we have feature casual,registered , cnt and calculate vif we will get infinite as casual + Registered = cnt, perfect correlation exist between them.

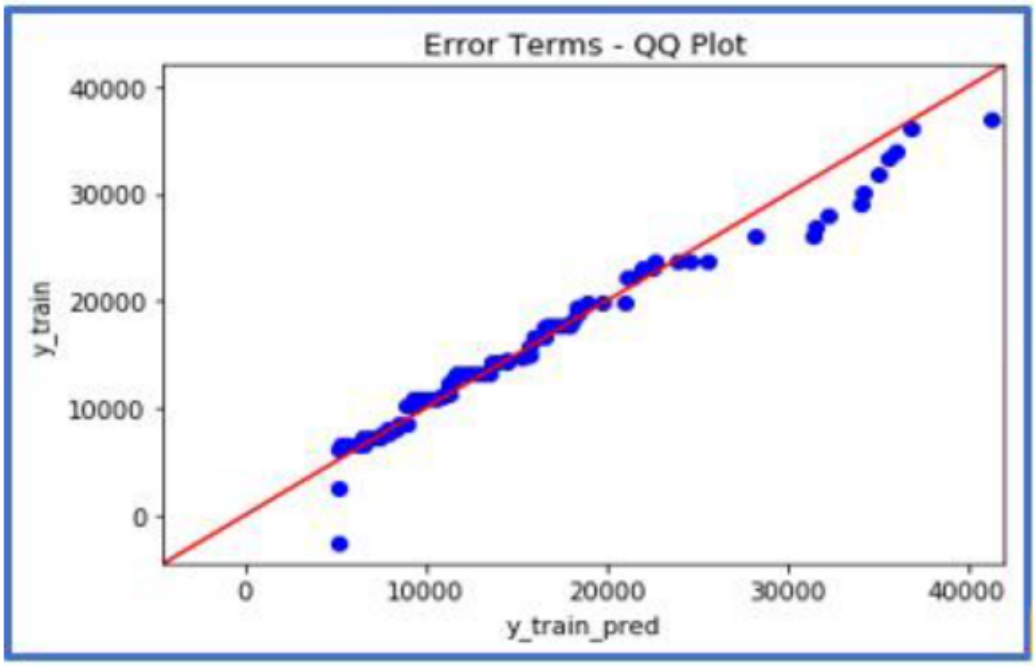
Question : What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression?

Answer :

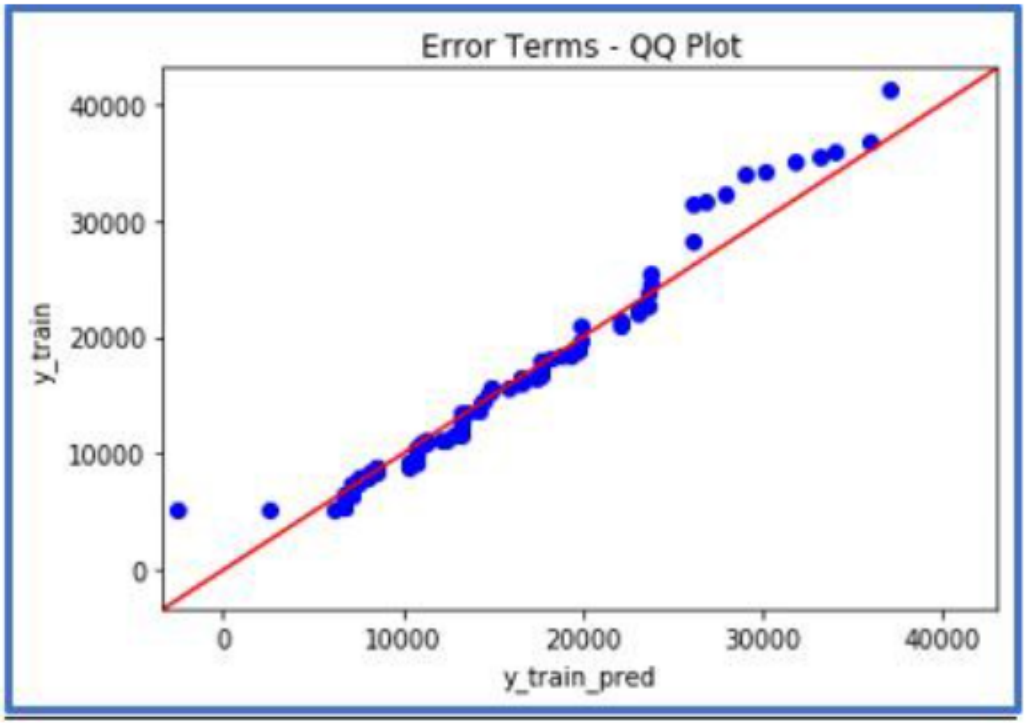
In Statistics, Q-Q(quantile-quantile) used to graphically analyze and compare two probability distributions by plotting their quantiles against each other. If the two distributions which we are comparing are exactly equal then the points on the Q-Q plot will perfectly lie on a straight line y = x.

This helps 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 or not.

possible interpretations for two data sets.
1. Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis
2. Y-values < X-values: If y-quantiles are lower than the x-quantiles.



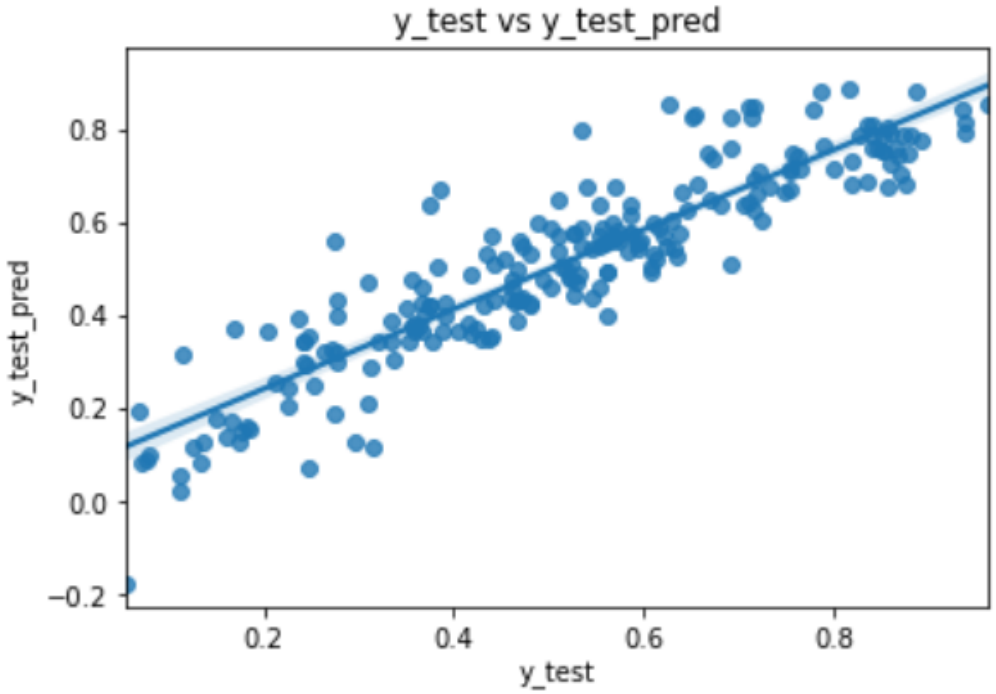
3. X-values < Y-values: If x-quantiles are lower than the y-quantiles.



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

in bike sharing model used this Q-Q plot to check y-test & y-test-pred, and validity of

the model


y_test vs y_test_pred
y_test_pred
y_test
0.8
0.6
0.4
0.2
0.0
−0.2
0.2
0.4
0.6
0.8

In [ ]: