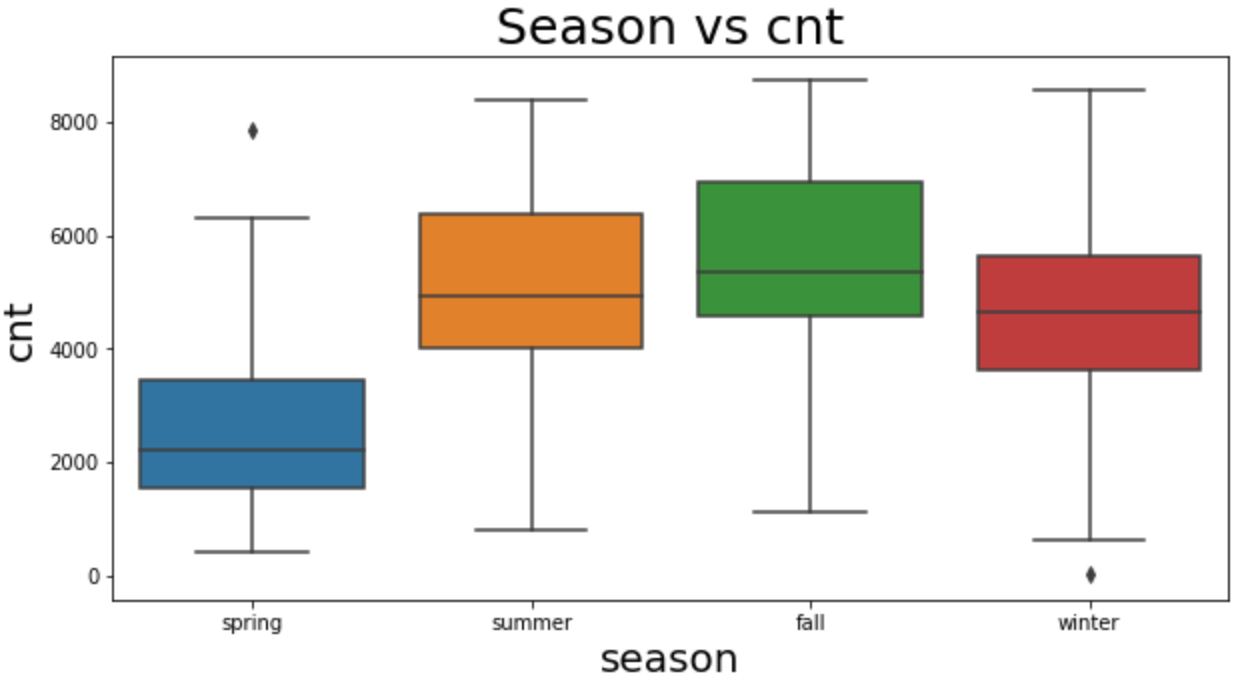
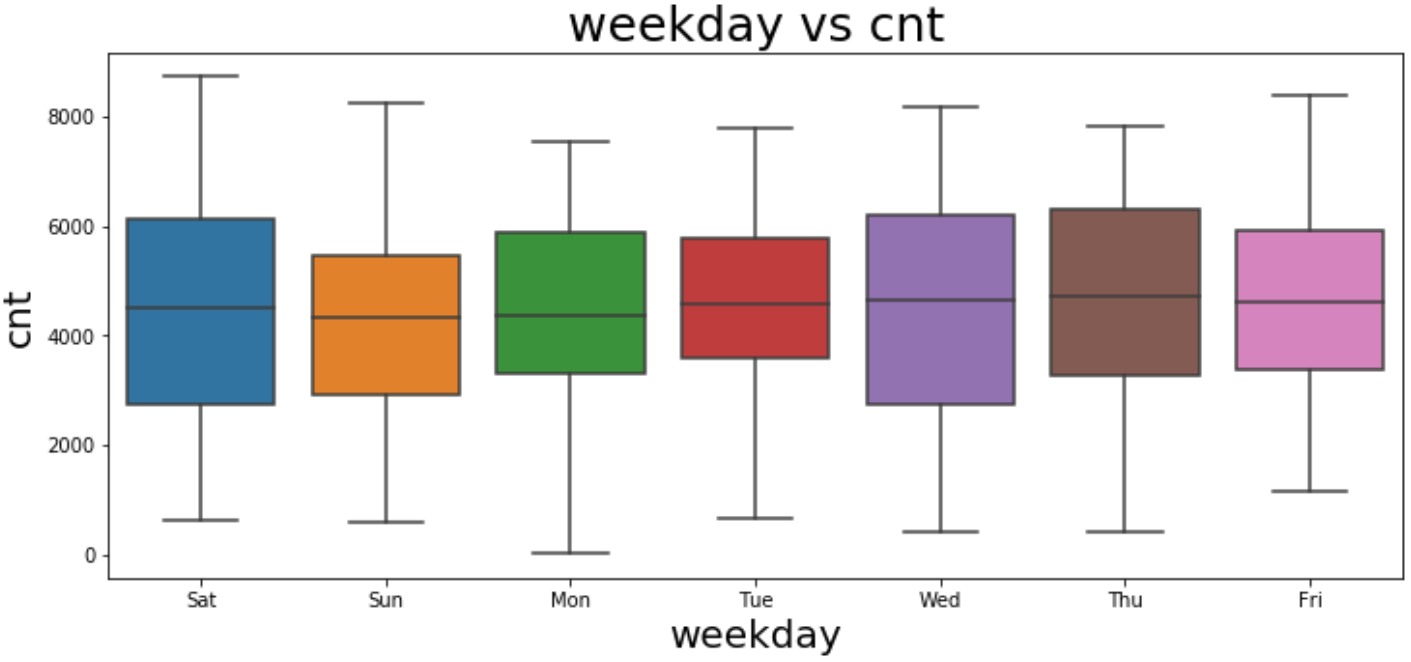
Assignment-based Subjective Questions

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

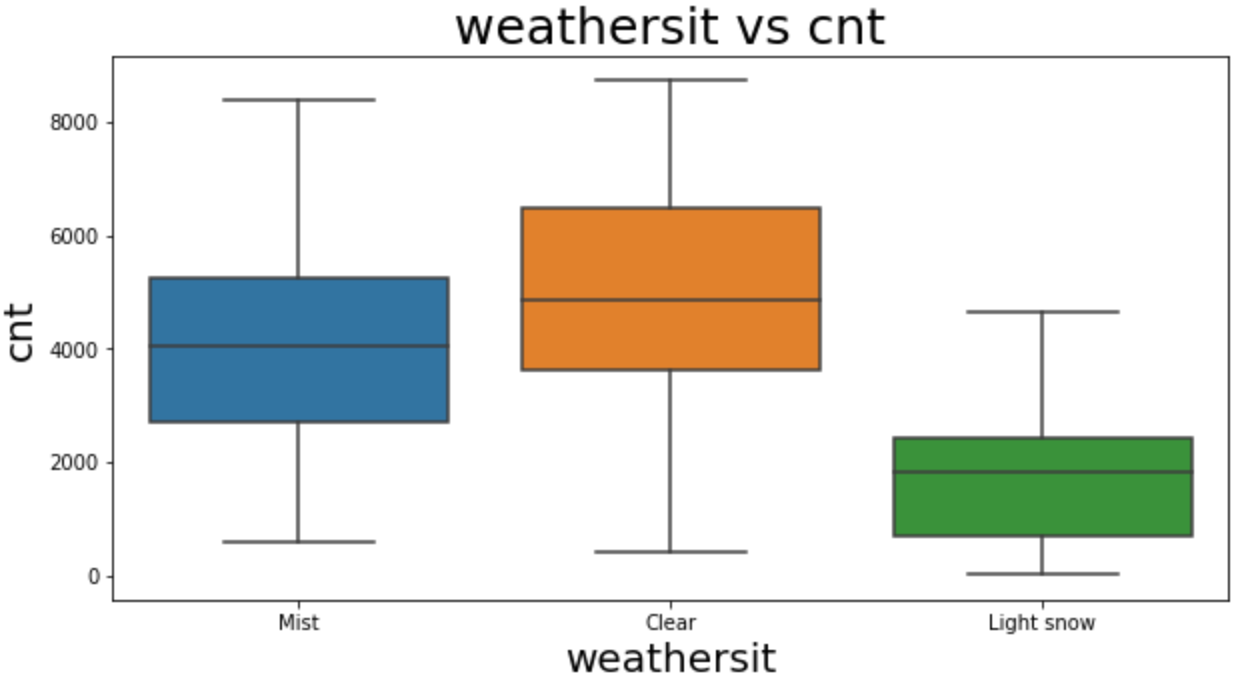


As observed from the above plot,
- bike rentals are significantly low during the season of spring
- It is highest in the month of fall

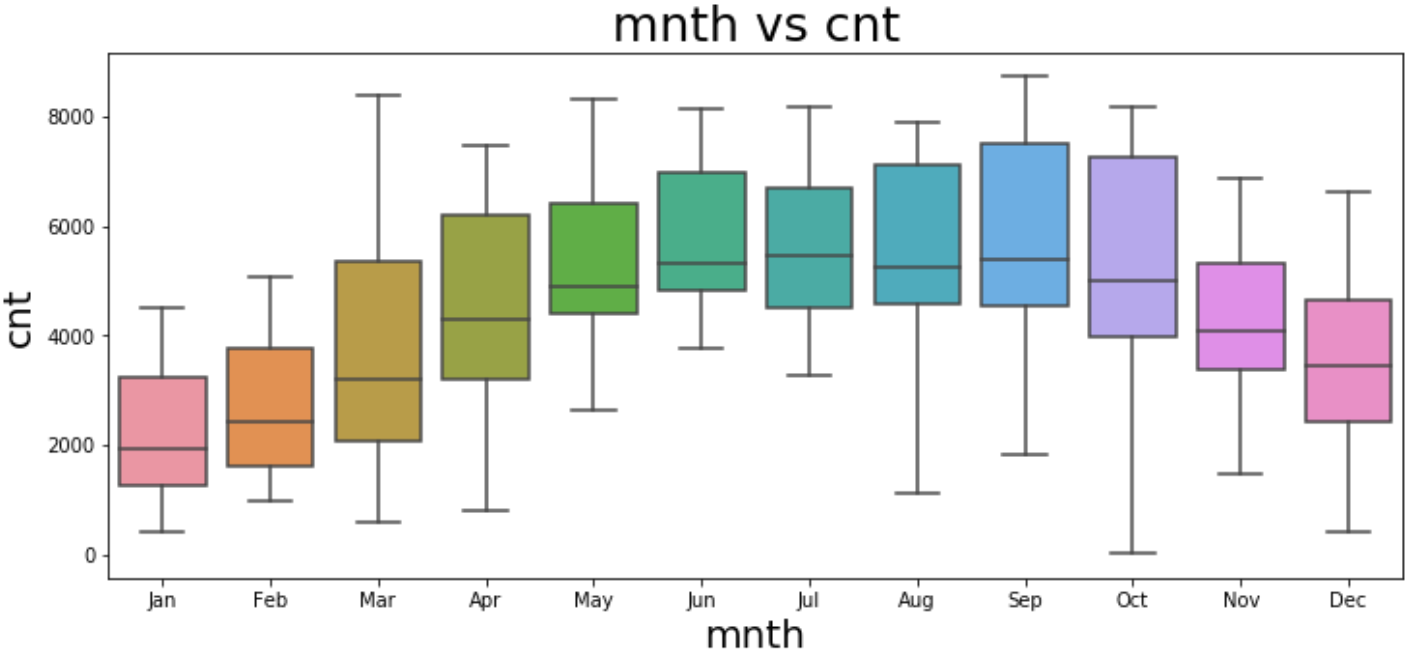


As observed from above plot,

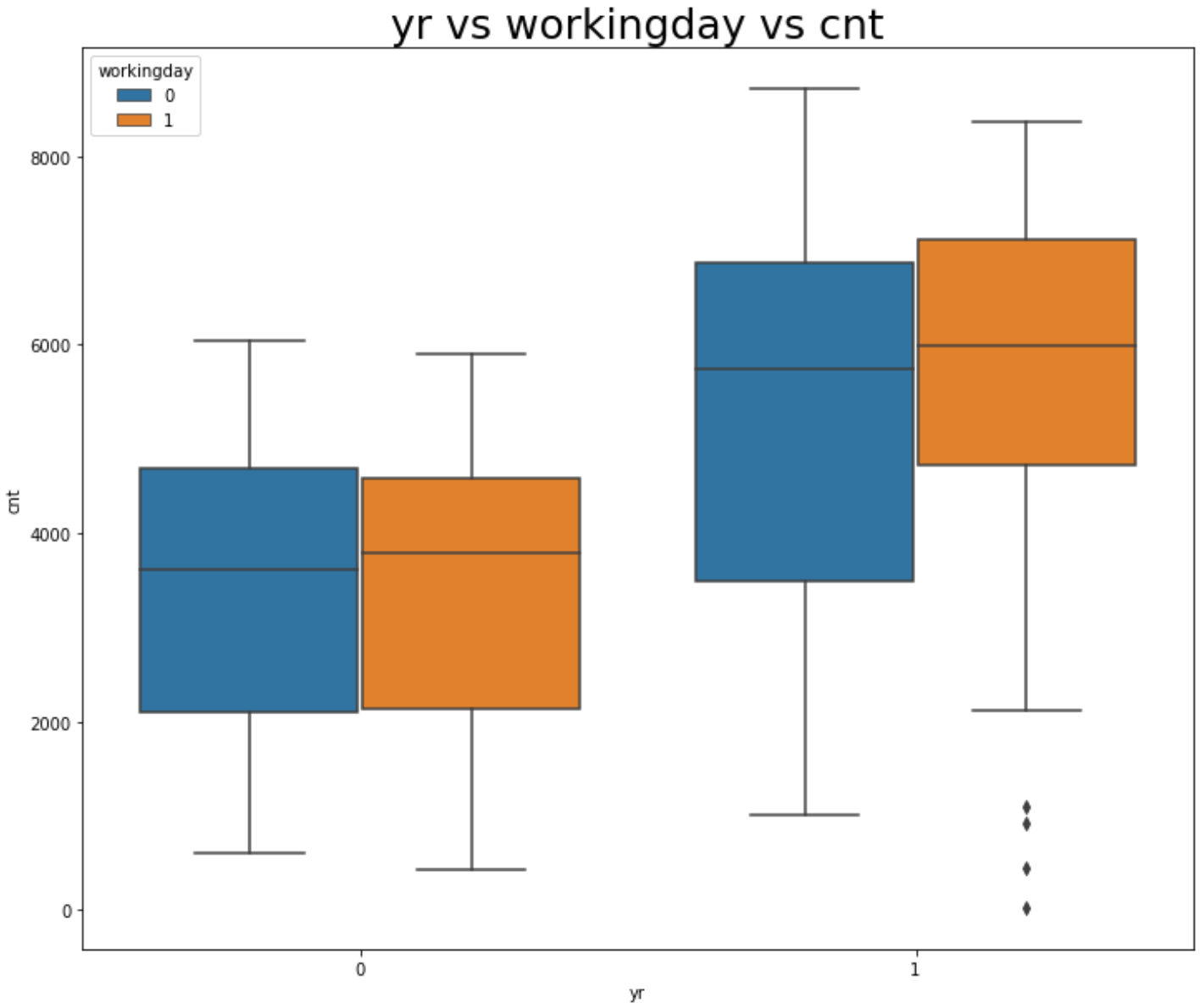
- Bike rentals are relatively low for the day - Monday, probably because most people are working after holiday on Sunday
- Bike rentals are significantly high for the weekends, Saturday and Sunday, and also for Wednesday
- Median is relatively the same for all days of the week. So this business is running all 7 days of week



As observed from above,
- Bike rentals is significantly low for weather conditions of light snow, which makes sense
- It is relatively high for clear weather as compared to Mist



As observed from above,
- Bike rentals is significantly high for the mid of the year, when the temp is high, which is also explained by positive correlation between temp and cnt



As observed from above,
- Bike rentals have increased significantly for the year 1 (2019) as compared to year 2018. It might be because more people are taking health seriously and want to exercise.
- The median of cnt of Bike rentals is relatively similar for working and non working days for both the years. But for 2019, the variation between 0 and 99percentile is higher for non working days as compared to working days.

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

Ans:- A Dummy variable or Indicator Variable is an artificial variable created to represent an attribute with two or more distinct categories/levels. The dummy variables act like 'switches' that turn various parameters on and off in an equation.
In python, to create dummy variables, we have a function in pandas, get_dummies that takes in categorical columns and gives out the corresponding data frame with dummy columns.
For example, suppose you have a column, “season” in your data set and it corresponds to Summer, Winter, Fall or Spring. To proceed with linear regression, we will have to convert this column to corresponding dummies, which can be done by command, pd.get_dummies(bikeRental['season'])
Output:

	fall	spring	summer	winter
0	0	1	0	0
1	0	1	0	0
2	0	1	0	0
3	0	1	0	0
4	0	1	0	0

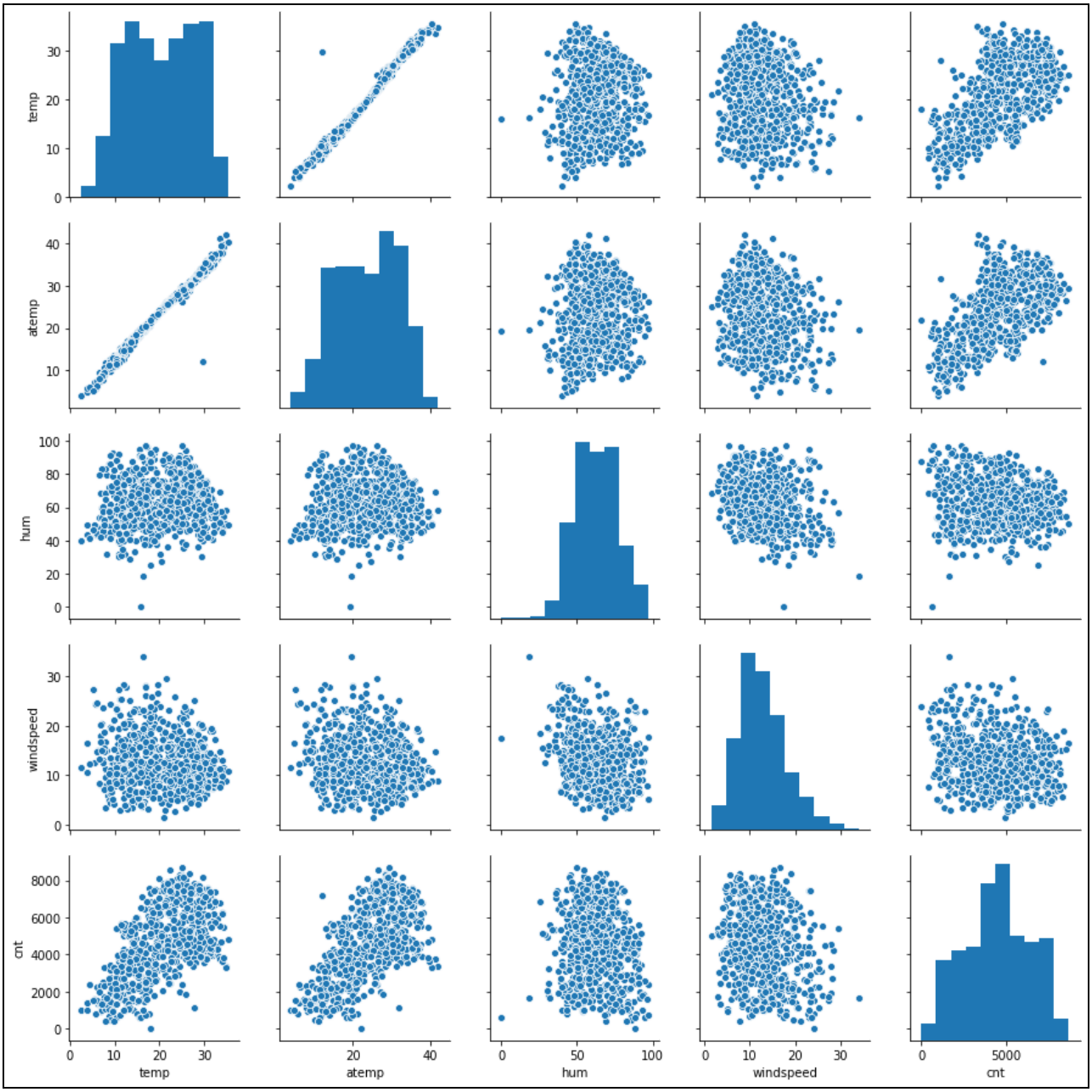
The value in any column, say spring, is set to 1 whenever that data row has season as spring, otherwise 0, and similarly for all the columns.
Now (refer to above image), to get fall season, values should be 1 0 0 0
For spring, 0 1 0 0
For summer, 0 0 1 0
And for winter 0 0 0 1

Notice here, that the season winter can also be defined as 0 0 0, beacuse season column had 4 possible values and it is not fall, summer or spring, it should be winter. Hence, we only require 3 dummy variables to define seasons columns. To generalise, we require k - 1 dummy variables to explain a column with k values.
Since, we do not require the 4th column, we can use drop_first = True argument in python function get_dummies, as follows:-
pd.get_dummies(bikeRental['season'], drop_first = True)
Output:-

	spring	summer	winter
0	1	0	0
1	1	0	0
2	1	0	0
3	1	0	0
4	1	0	0

By default, python drops the first dummy variable created, fall in this case. Hence, drop_first=True argument is necessary to prevent data set to have redundant columns which makes the dataset complex and reduces readability.To conclude, since one of the columns can be generated completely from the others, and hence retaining this extra column does not add any new information for the modelling process, would it be good practice to always drop the first column

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



Ans:-

From the above plot, we can conclude that temp has the highest correlation with the target variable, cnt. Also, temp and atemp have the highest correlation among all numerical variables in the data set, around 0.99.

Temp is the temperature on the day when the bike was hired, and cnt is the total count of bikes that were hired.

Since, temp and cnt have a positively strong correlation, we can say that as the temperature increases, more people hire bikes. In words, more bikes are hired on hot days as compared to cool days.

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

Ans:-

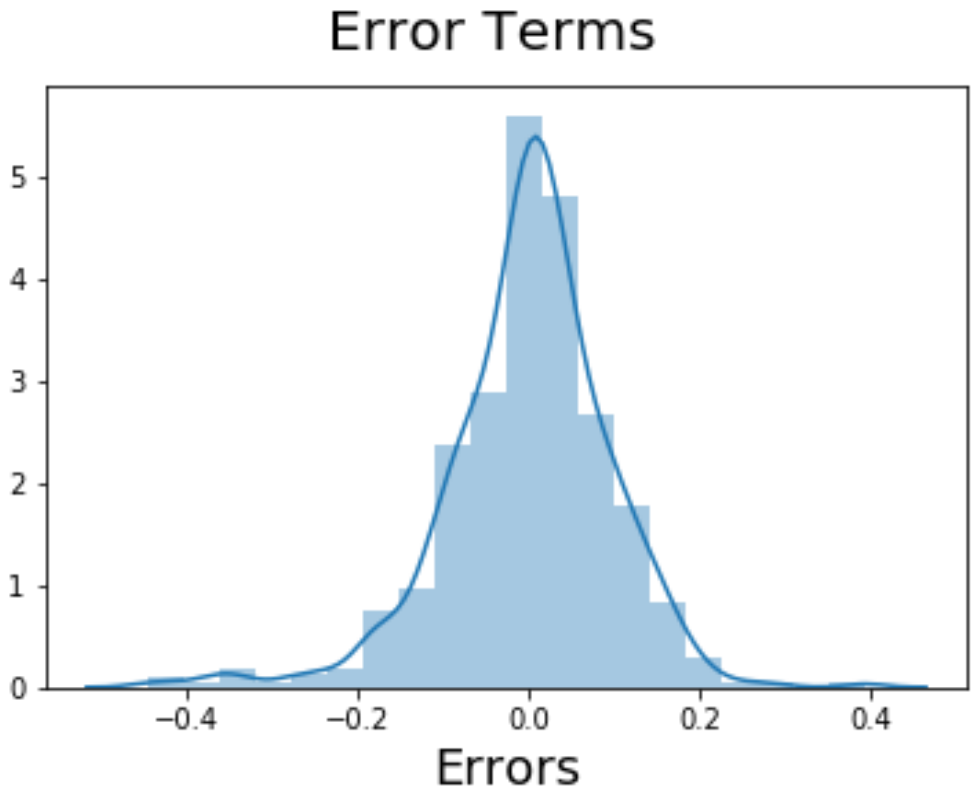
Assumptions of linear regression

1. Linear relationship between predictors and y.
2. Normal distribution of error terms with mean 0.
3. Independence of error terms.
4. Constant variance of error terms.

1. We checked for the correlation of the dependent variables with all numerical variables present in the dataset. The correlation between cnt (dependent) and temp (independent) came out to be 0.63, which suggests that there exists some sort of linear relationship between cnt and temp. Furthermore, after building the linear model on train we tried to predict the value of cnt on the unseen test data. The model predicted the values of cnt with an R-squared of 0.81 on train data and 0.803 on test data. This suggests that there exists a linear relationship between the dependent and the independent variables.

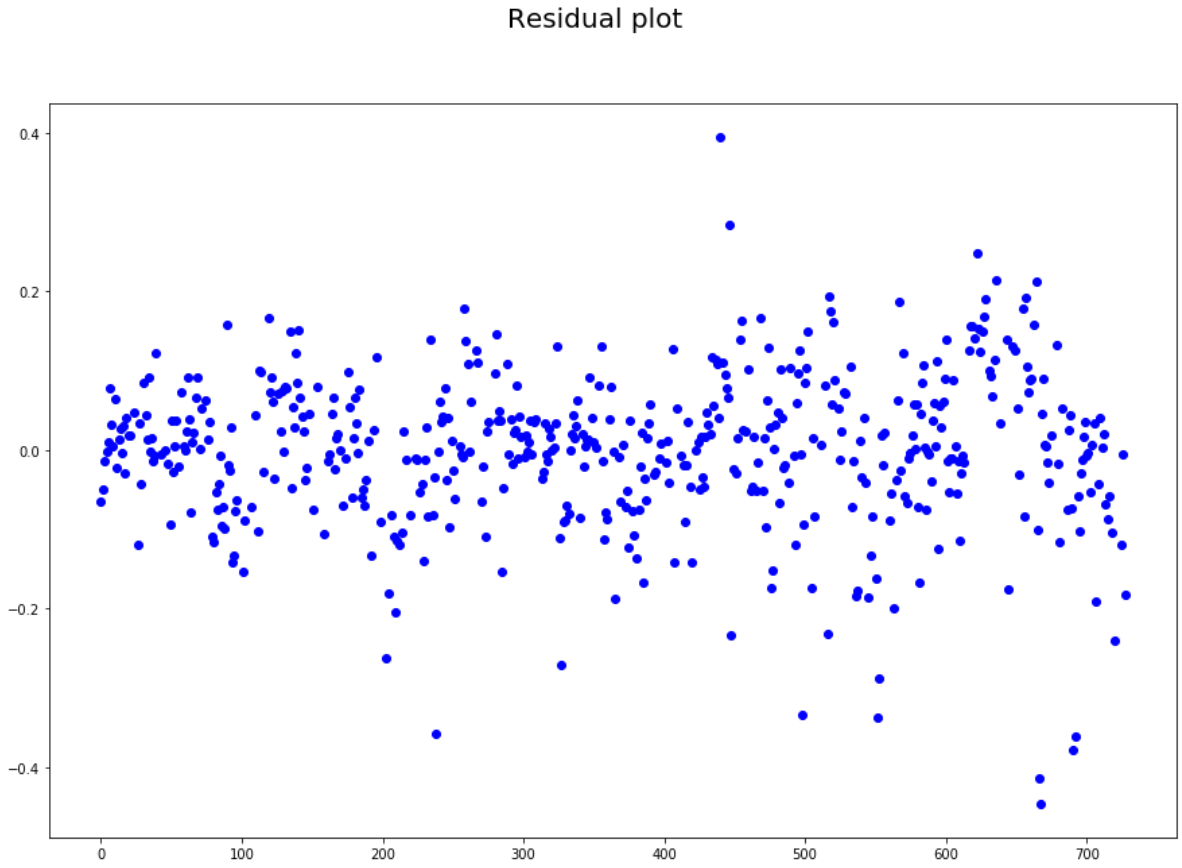
2. We calculated the predicted values of cnt for the unseen test data using the linear model we created and compared it to actual value of cnt we had in the test data set, to calculate the residual using the formula, res = y_pred - y_test

We then plotted a distribution graph for the residuals calculated above.



As we can see, the above graph is a normal distribution graph with mean equal to 0.

3. We plotted the individual residual points on a graph to observe their spread.



As we can see, the residual points are randomly spread and we do not observe any pattern between them. Hence, we can say that the errors (or residuals) are independent of each other.

4. From the above plot, we can see that all the data points of errors are spread between 0.2 and -0.2 and there are some outliers too. But for the majority of the points the range is constant, that is, 0.2 to -0.2. Hence, we can say that the errors have a constant variance.

Moreover, if we pass a line y = 0, we will observe that all the data points are independently above and below the line and will have a constant variance.

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

Ans:- The equation to final model is as follows:-

cnt = 0.2132 + (0.2341)* yr + (0.4446)* temp + (-0.1221)* season_spring + (0.0527)* season_winter + (-0.0476)* weekday_Sun + (-0.2964)* weathersit_Light snow + (-0.0733)* weathersit_Mist

From the above equation we can see that the predictors with descending order of their coefficients(without the sign) is as follows:-

1. Temp
2. weathersit_Light snow
3. Yr
4. Season_spring
5. weathersit_Mist
6. season_winter
7. weekday_Sun

The top 3 features that contribute significantly towards explaining the demand of shared bikes are,

a. Temperature of the day
b. Weather condition - Light_snow
c. Year of business

The demand of shared bikes seems to be positively dependent upon the temperature of the day, i.e., higher the temperature, higher is the demand for bikes, which might probably be due to more people coming out from their houses at higher temperatures for exercise and hiring bikes for that.

The demand of bikes seems to have a negative relationship with the weather condition - Light snow which means that the weather condition, if of Light snow, the demand of bikes reduces, which makes sense.

The demand for shared bikes seems to have a positive relationship with the year of business. We have observed that the demand for shared bikes has increased significantly in the year 2019 as compared to the year 2018. This might be due to the fact that people in 2019 are more aware about the importance of health benefits of exercise and are hiring more bikes for that.

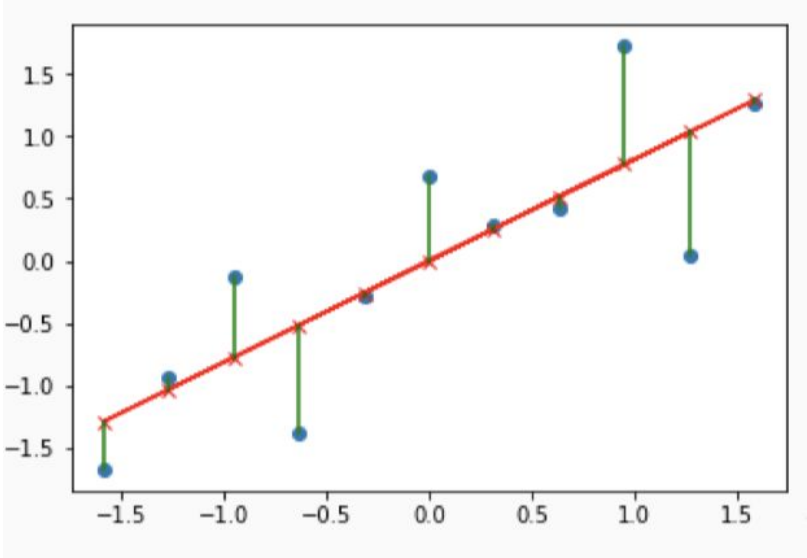
General Subjective Questions

1. Explain the linear regression algorithm in detail**.**

Ans:- Linear regression is a linear model, e.g. a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x).
When there is a single input variable (x), the method is referred to as simple linear regression. When there are multiple input variables, literature from statistics often refers to the method as multiple linear regression.
The linear equation assigns one scale factor to each input value or column, called a coefficient and represented by the capital Greek letter Beta (B). One additional coefficient is also added, giving the line an additional degree of freedom (e.g. moving up and down on a two-dimensional plot) and is often called the intercept or the bias coefficient.
For example, in a simple regression problem (a single x and a single y), the form of the model would be:
$y = B0 + B1*x$

Similarly, for multiple linear regression problem, the form of model would be:
$Y = B0 + B1*x1 + B2*x2 + B3*x3 + ….. + Bn*xn$

Learning a linear regression model means estimating the values of the coefficients used in the representation with the data that we have available.
With simple linear regression when we have a single input, we can use statistics to estimate the coefficients.
This requires that you calculate statistical properties from the data such as means, standard deviations, correlations and covariance. All of the data must be available to traverse and calculate statistics.
When we have more than one input we can use Ordinary Least Squares to estimate the values of the coefficients.
The Ordinary Least Squares procedure seeks to minimize the sum of the squared residuals. This means that given a regression line through the data we calculate the distance from each data point to the regression line, square it, and sum all of the squared errors together. This is the quantity that ordinary least squares seeks to minimize.



Given the representation is a linear equation, making predictions is as simple as solving the equation for a specific set of inputs.

Let's make this concrete with an example. Imagine we are predicting weight (y) from height (x). Our linear regression model representation for this problem would be:

$$y = B0 + B1 * x1$$

or

$$weight = B0 + B1 * height$$

Where B0 is the bias coefficient and B1 is the coefficient for the height column. We use a learning technique to find a good set of coefficient values. Once found, we can plug in different height values to predict the weight.

For example, let's use B0 = 0.1 and B1 = 0.5. Let's plug them in and calculate the weight (in kilograms) for a person with the height of 182 centimeters.

$$weight = 0.1 + 0.5 * 182$$

$$weight = 91.1$$

Considerations in linear regression:-

- Linear Assumption. Linear regression assumes that the relationship between your input and output is linear. It does not support anything else.
- Remove Noise. Linear regression assumes that your input and output variables are not noisy. This is most important for the output variable and you want to remove outliers in the output variable (y) if possible.
- Remove Collinearity. Linear regression will over-fit your data when you have highly correlated input variables. Consider calculating pairwise correlations for your input data and removing the most correlated.
- Rescale Inputs: Linear regression will often make more reliable predictions if you rescale input variables using standardization or normalization.

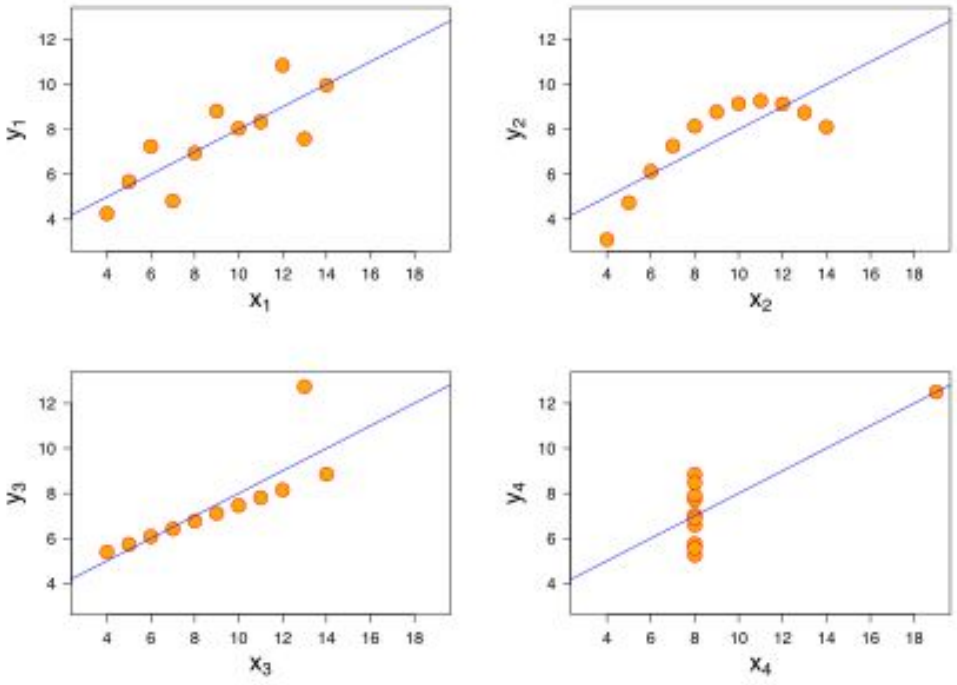
Assumptions in linear regression:-

1. Linear relationship between predictors and y.
2. Normal distribution of error terms with mean 0.
3. Independence of error terms.
4. Constant variance of error terms.

2. Explain Anscombe's quartet in detail.

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 demonstrate both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties.
For all four datasets:

Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	y = 3.00 + 0.500x	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression :	0.67	to 2 decimal places



The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y is 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..
In the third graph (bottom left), the distribution is linear, but should have a different regression line. 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.

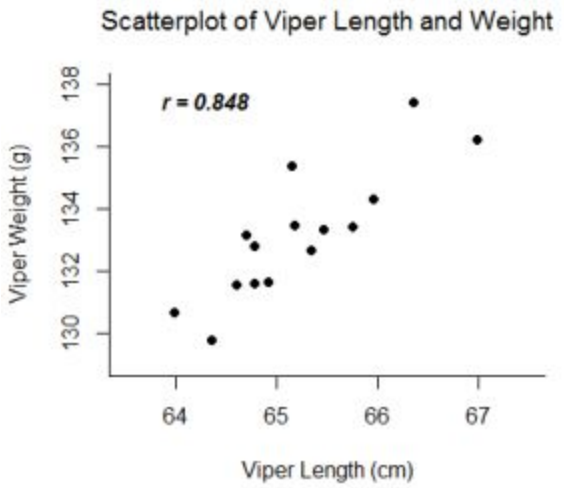
The datasets are as follows. The x values are the same for the first three datasets.

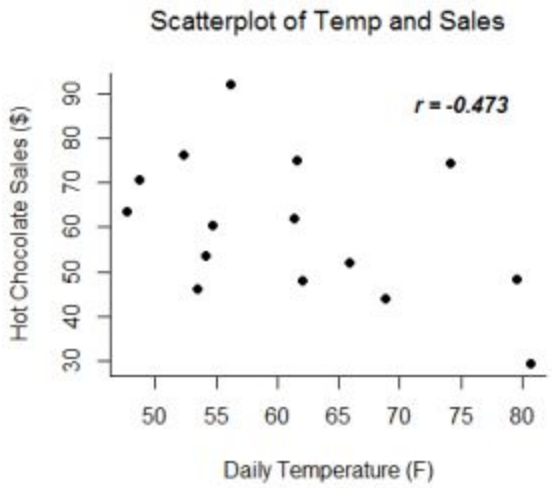
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	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

3. What is Pearson's R?

Ans:- A correlation or simple linear regression analysis can determine if two numeric variables are significantly linearly related. A correlation analysis provides information on the strength and direction of the linear relationship between two variables, while a simple linear regression analysis estimates parameters in a linear equation that can be used to predict values of one variable based on the other.

The Pearson correlation coefficient, r, can take on values between -1 and 1. The further away r is from zero, the stronger the linear relationship between the two variables. The sign of r corresponds to the direction of the relationship. If r is positive, then as one variable increases, the other tends to increase. If r is negative, then as one variable increases, the other tends to decrease. A perfect linear relationship (r=-1 or r=1) means that one of the variables can be perfectly explained by a linear function of the other.





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

Ans:- It is a step of Data Pre Processing which is applied to independent variables or features of data. It basically helps to normalise the data within a particular range. Sometimes, it also helps in speeding up the calculations in an algorithm.

Real world dataset contains features that highly vary in magnitudes, units, and range. Normalisation should be performed when the scale of a feature is irrelevant or misleading and not should Normalise when the scale is meaningful.
The algorithms which use Euclidean Distance measure are sensitive to Magnitudes. Here feature scaling helps to weigh all the features equally.
Formally, If a feature in the dataset is big in scale compared to others then in algorithms where Euclidean distance is measured this big scaled feature becomes dominating and needs to be normalized.

Examples of Algorithms where Feature Scaling matters
1. K-Means uses the Euclidean distance measure here feature scaling matters.
2. K-Nearest-Neighbours also require feature scaling.
3. Principal Component Analysis (PCA): Tries to get the feature with maximum variance, here too feature scaling is required.
4. Gradient Descent: Calculation speed increase as Theta calculation becomes faster after feature scaling.
Algorithm which is Not Distance based is Not affected by Feature Scaling.

Normalization usually means to scale a variable to have a values between 0 and 1, while standardization transforms data to have a mean of zero and a standard deviation of 1. This standardization is called a z-score, and data points can be standardized with the following formula:

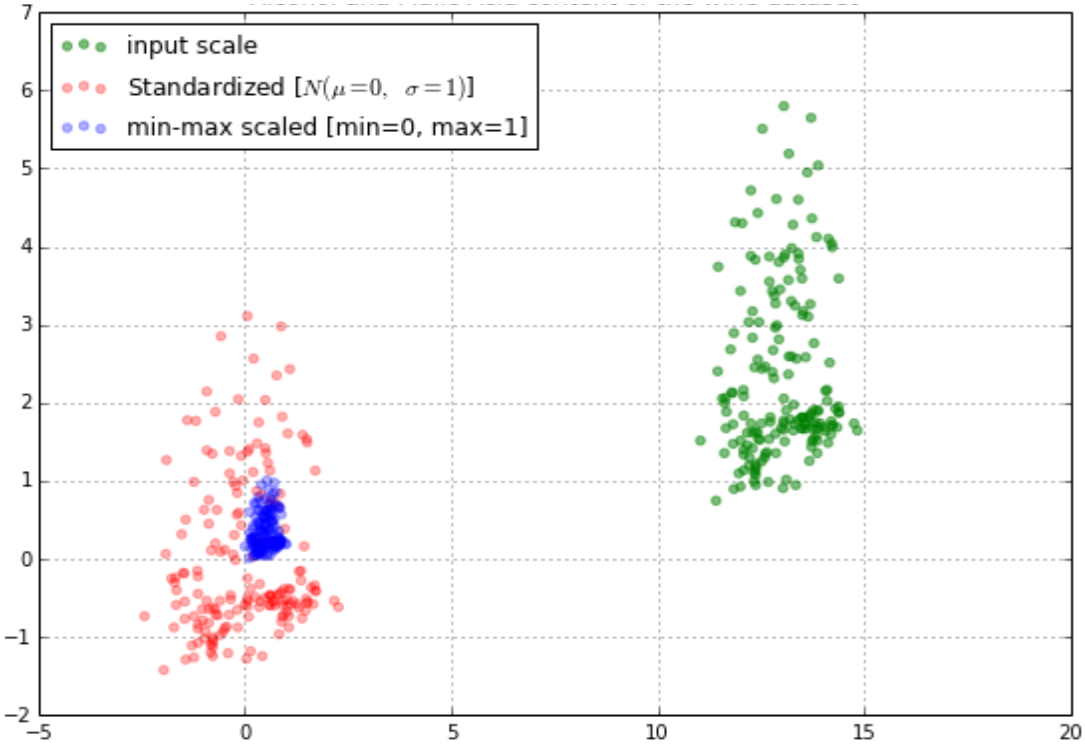
$$z_i = \frac{x_i - \bar{x}}{s}$$

Where:
xi is a data point (x1, x2…xn).
$\bar{x}$ is the sample mean.
s is the sample standard deviation.

The normalization is carried out using the formula:

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Where,
x is a data point
x_{min} is minimum value out of all data points
x_{max} is maximum value out of all data points


input scale
Standardized [$N(\mu=0,\ \sigma=1)$]
min-max scaled [min=0, max=1]
7
6
5
4
3
2
1
0
−1
−2
−5
0
5
10
15
20

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

Ans:- The variance inflation factor (VIF) quantifies the extent of correlation between one predictor and the other predictors in a model. It is used for diagnosing collinearity/multicollinearity. Higher values signify that it is difficult to assess accurately the contribution of predictors to a model.

The extent to which a predictor is correlated with the other predictor variables in a linear regression can be quantified as the R-squared statistic of the regression where the predictor of interest is predicted by all the other predictor variables. The variance inflation for a variable is then computed as:

$$VIF = \frac{1}{1 - R^2}$$

Ideally, the VIF values should be below 10 for a predictor to significantly contribute in the model. But as a thumb rule, we also investigate the predictors who have VIF > 5.

From the above formula, for VIF to be infinite, R^2 should be equal to 1. In simple words, an infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables

Remedies for high VIF:

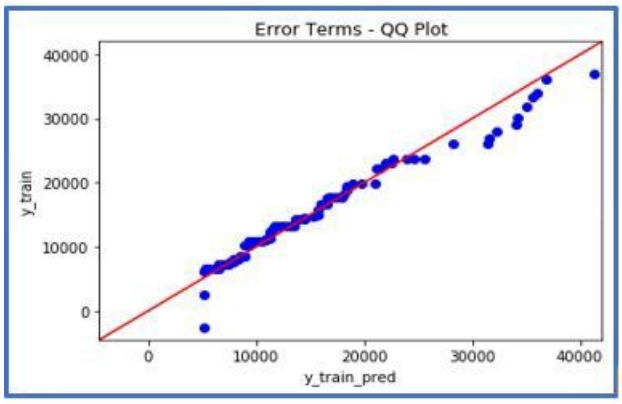
1. You can choose to drop the predictor with high VIF and re-build the model
2. You can use some derived variable from the predictor with high VIF and re-build the model to access its significance.

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

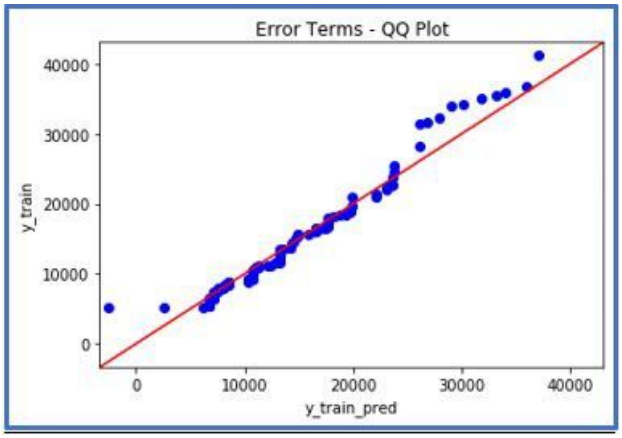
Ans:- Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.
This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using the Q-Q plot that both the data sets are from populations with the same distributions.

Few advantages:
a) It can be used with sample sizes also
b) 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.
It is used to check following scenarios:
If two data sets —
i. come from populations with a common distribution
ii. have common location and scale
iii. have similar distributional shapes
iv. have similar tail behavior

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

In python, statsmodels.api provide qqplot and qqplot_2samples to plot Q-Q graph for single and two different data sets respectively.