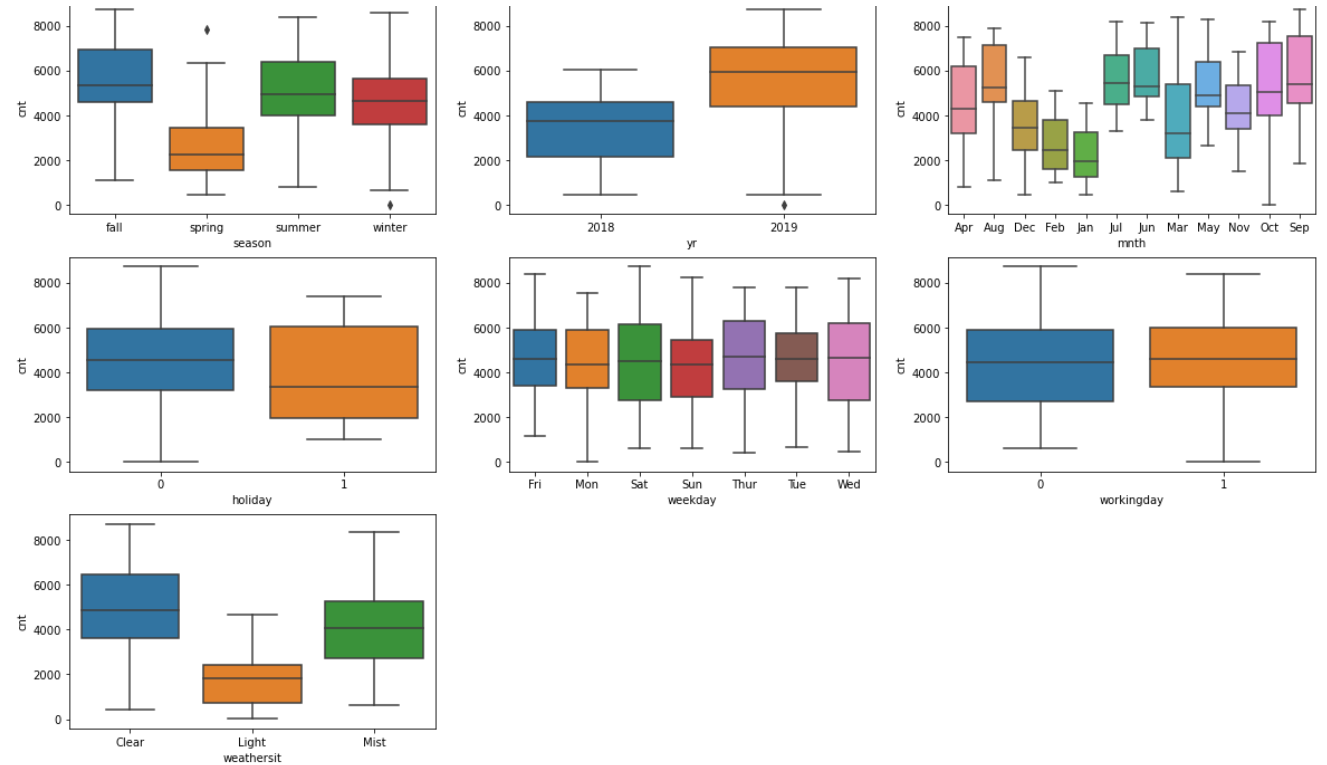


1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer –

- **Season** - Spring season has lowest biking sharing as compared to other season
- - **Year** - Bike sharing increases by year(2019 > 2018)
- - **Month** - There is a decline in bike sharing count from OCT to march due winter and spring season
- - **Weekday** - It does not have a particular trend
- **Weather** - Clear, Few clouds, Partly cloudy, Partly cloudy has higher number of bike sharing as compared to other weather
- **Holiday** – Holiday has a dip in bike sharing count

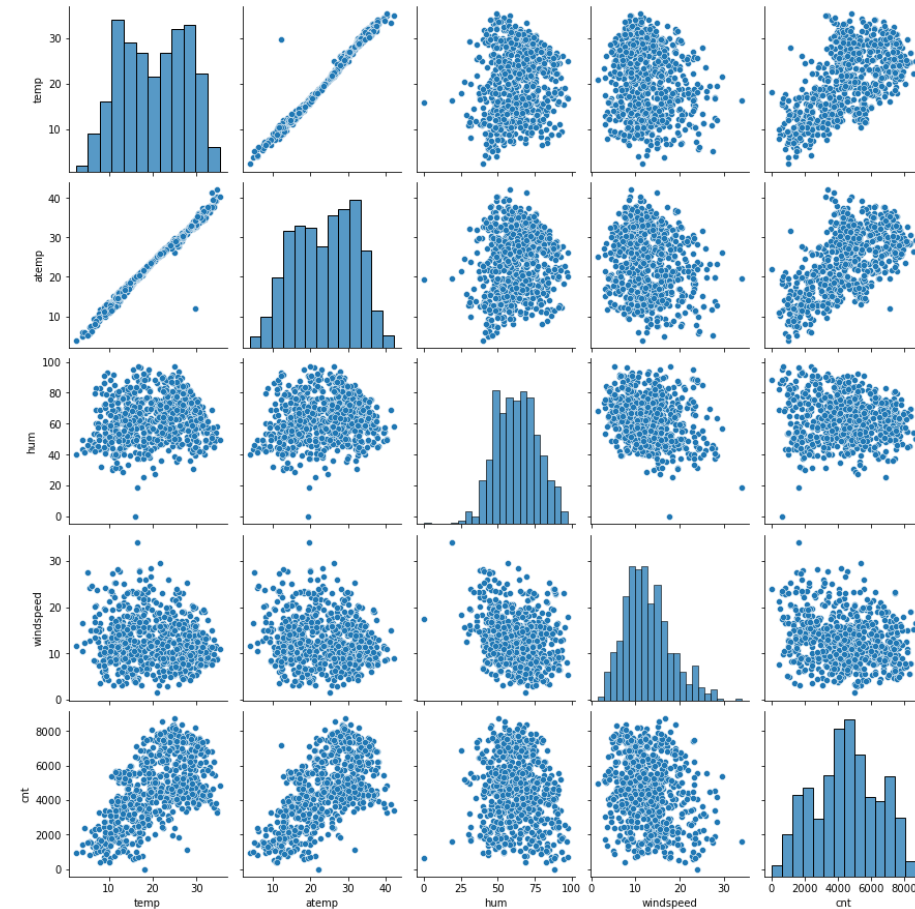


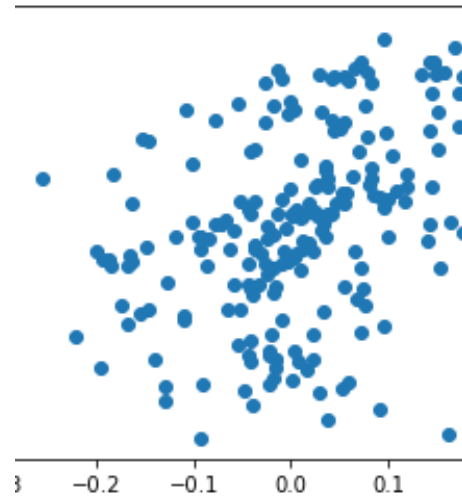
2. Why is it important to use `drop_first=True` during dummy variable creation?

- Answer
 - `Drop_first= True` or dummy encoding basically reduces the number of dummy variables by 1 without losing any information. It also reduces correlation created between variables. Another advantage is we reduce the number of variables that machine learning algorithm needs to learn.

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

- Answer :
- Temp has the highest correlation with cnt variable





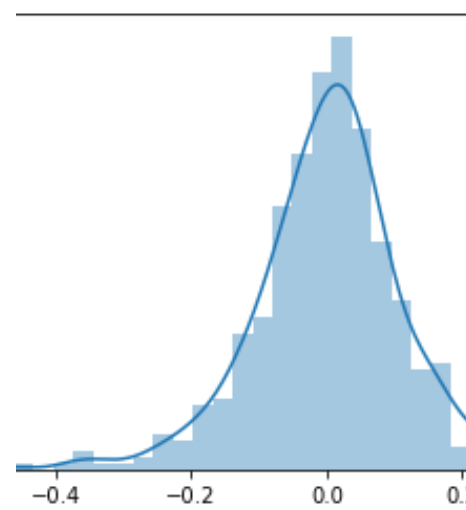
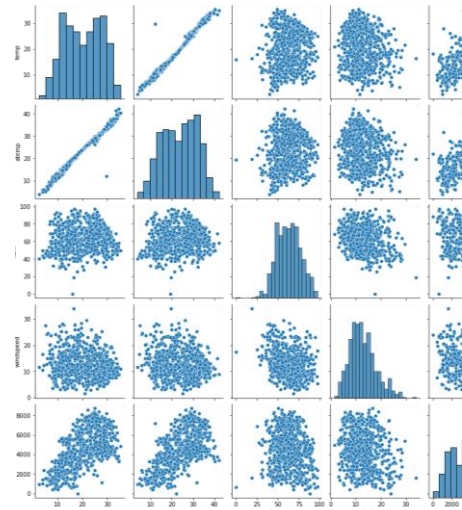
```

>> vif = cal_vif(x_train_rfe)
vif

```

37]:

	features	vif
6	hum	11.45
5	temp	7.01
4	workingday	4.53
7	windspeed	3.60
0	2019	2.02
2	Sat	1.78
3	winter	1.48
1	Light	1.08



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

- **Answer:** There are 4 basic assumptions of linear regression . Using the below steps we can validate. And also r2_score will help the accuracy of model.
1. Linear relationship
 - If we see the scatter plot the relation between temp and cnt appears to be linear. If we see other variables the don't seem to be linear.
 2. Multivariate normality
 - Multivariate normality. This assumption requires that residuals are normally distributed. If we see the residual plot it is normally distributed
 3. No or little multicollinearity.
 - This assumption is used to determine the relationship between independent variables. We need to ensure that there is less correlation between independent variables. We can use VIF or Corelation matrix to derive this. Scores less than 5 for VIF are desired.
 4. Homoscedasticity
 - This assumption need that error terms are having constant variance. It should not be increasing or decreasing like a cone but equally spread. As seen in below scatter plot the variance is evenly spread

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

- Answer:
- Temp, Year and winter are the top 3 features contributing significantly towards explaining the demand of the shared bikes.
- Equation is $0.0537 + 0.2330yr - 0.2541Light + 0.0616Sat + 0.1036winter + 0.0514workingday + 0.6020temp - 0.1387windspeed$

General Subjective Questions

1. Explain the linear regression algorithm in detail?

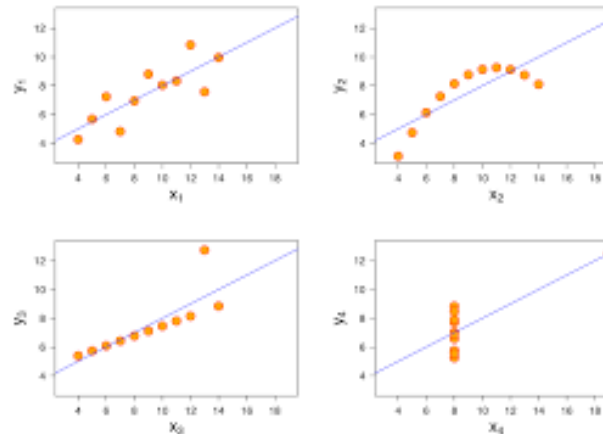
Answer:

- Linear regression is one of the supervised learning algorithms that works by fitting a line or a plane given a set of attributes or predictors. To illustrate this let's take an example where a team wants to determine influence of marketing methods like tv, radio on sales number. The basic idea over here is to identify an equation that helps to determine the predicted value here. sales wrt to given input sales methods like tv/ radio etc also called independent variables.
- Equation for this plane or line is generally written by $Y=B_0+B_1X_1+B_2X_2+B_3X_3+....B_nX_n$. Intent of regression is to identify B_0 and $B_1...B_n$
- We typically use techniques like differentiation or gradient descent to identify coefficients. In addition to this we strive to achieve 4 principles of regression Homoscedasticity, no multicollinearity, multivariate normality.
- The effectiveness of algorithm is usually derived by methods like RSS. We try to get least RSS.
- For any successful linear regression, we perform following steps
 - Understanding data
 - Performing analysis and plotting
 - Preparing data by scaling or dummy variables
 - Splitting data into train and test set(70:30 or 80:20)
 - Building model by leveraging VIF and p-Values either through top-down or bottom up approach on train set
 - Analysing the residuals and Testing the model on test set
 - . Validating the model

2. Explain the Anscombe's quartet in detail

Answer:

- Anscombe's quartet shows that a dataset with similar statistical properties can still be different when graphed. As per wikipedia "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."
- If we see above data we observe the summary stats are nearly identical for all datasets but their plots are varying. There are some observations of this
 - Data can be non linear or linear.
 - There can be outliers which can or cannot be handled by linear regression model.
 - This brings an important conclusion that plot of data is needed before right model is picked for a given dataset




	I		II		III		IV	
	x	y	x	y	x	y	x	y
0	10	8.04	10	9.14	10	7.46	8	6.58
1	8	6.95	8	8.14	8	6.77	8	5.76
2	13	7.58	13	8.74	13	12.74	8	7.71
3	9	8.81	9	8.77	9	7.11	8	8.84
4	11	8.33	11	9.26	11	7.81	8	8.47
5	14	9.96	14	8.10	14	8.84	8	7.04
6	6	7.24	6	6.13	6	6.08	8	5.25
7	4	4.26	4	3.10	4	5.39	19	12.50
8	12	10.84	12	9.13	12	8.15	8	5.56
9	7	4.82	7	7.26	7	6.42	8	7.91
10	5	5.68	5	4.74	5	5.73	8	6.89

3. What is Pearson's R?

Answer

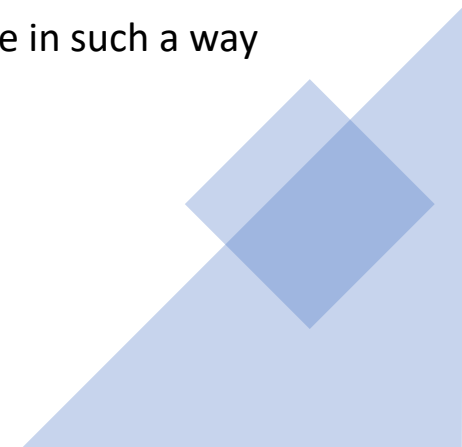
- Pearson coefficient is the measure of linear correlation between two sets of data. The value of this typically lies between -1 and 1. A value of 0 means no correlation where any thing greater than 0 means a positive tendency of increase with increase on other variable. A negative value means tendency of decrease with increase in value of other variable.
- The value of corelation is generally derived using corr function in python.

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



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

Answer:

- Scaling is the process of bringing all that variables in uniform or comparable measuring scale. This is needed specially to avoid coefficients swinging on extreme end, that leads to difficulty in interpretation of model.
 - This help in speeding up Beta derivation using gradient descent.
 - Normalized scaling or Min Max scaling tries to fit data in [0 and 1] scale by doing
$$(x - x_{\min}) / (x_{\max} - x_{\min})$$
whereas Standardized scaling scales value in such a way that mean lies at 0. It is computed by
$$(x - \text{mean}(x)) / \text{standard deviation}(x)$$
- 

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

- An infinite VIF means that the corresponding variable can be expressed linearly by other variables. Take this formula, if variables are highly correlated R^2 becomes 1. This causes denominator to become 0 and hence infinite

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

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

- Answer
 - QQ plot or the quantiles plot is a scatter plot created by plotting 2 quantiles against each other. It helps us in identifying the normality of a distribution. If a distribution is normal then it follows a straight line. This is especially significant to validate the assumption that residual follow normal distribution. Also, it also can be used to confirm that data comes from same distribution