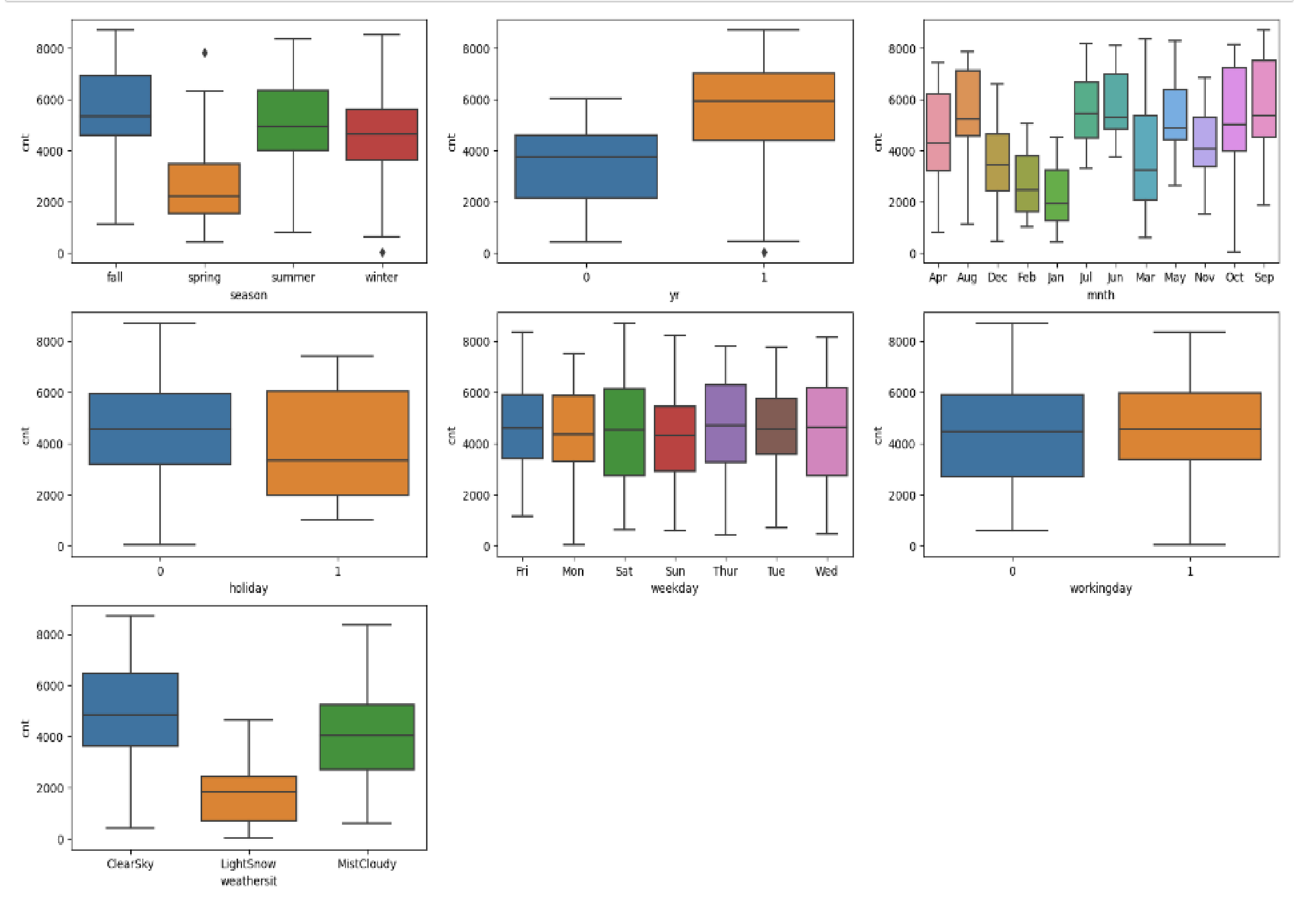
LINEAR REGRESSION SUBJECTIVE QUESTIONS

SAI BHARATH PARSAM

FROM YOUR ANALYSIS OF THE CATEGORICAL VARIABLES FROM THE DATASET, WHAT COULD YOU INFER ABOUT THEIR EFFECT ON THE DEPENDENT VARIABLE?

- Season - Spring season has lowest biking sharing as compared to another season
- Year (yr) - Bike sharing increased by year. Which mean 2019 has more rides compared to 2018.Indicates growth in business.
- Month - Same trend as Seasons. There is a decline in bike sharing count from Oct to Mar due to winter and spring season.
- Holiday - It has little dip on holidays.
- Weekday - It does not have a particular trend.
- Workingday - It does not have a particular trend
- Weather - "Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds" has least number of rides.

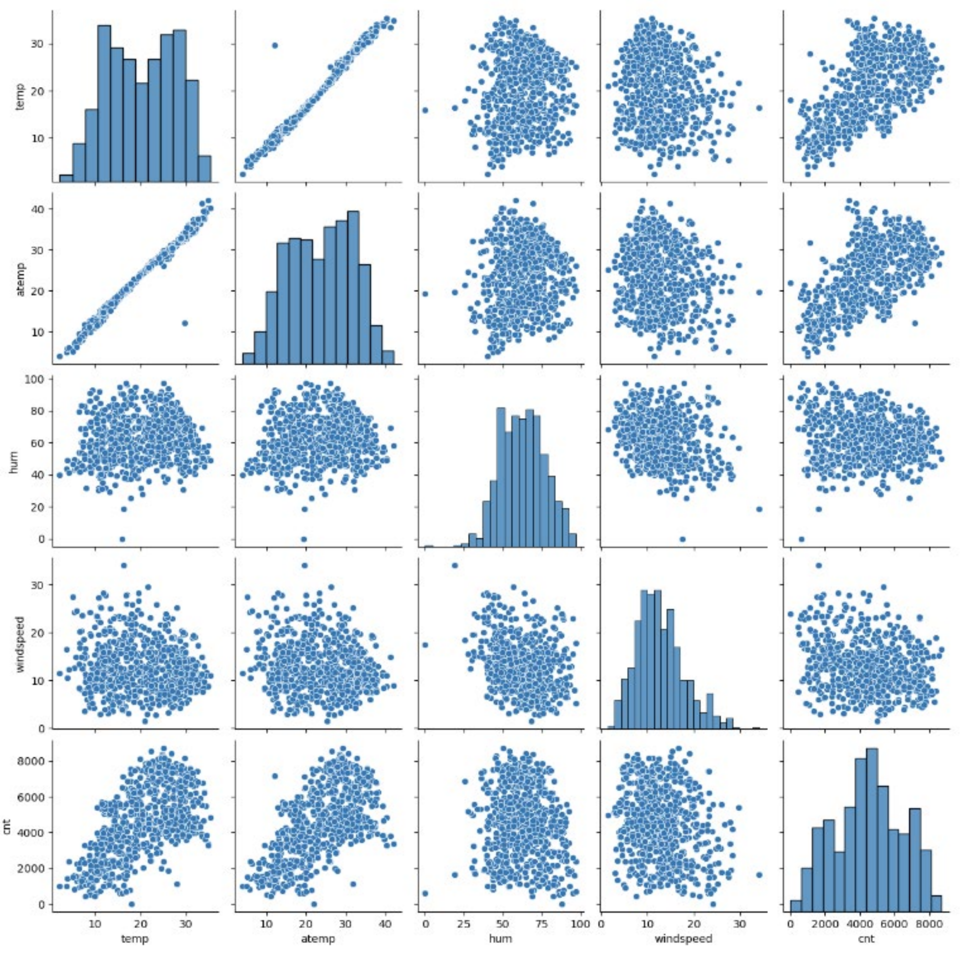


WHY IS IT IMPORTANT TO USE **DROP_FIRST=TRUE** DURING DUMMY VARIABLE CREATION?

- We use **drop_first=True** while creating the dummy variables is to reduce the number of dummy variables by 1 without loss of information. Its advantage is we reduce the number of variables that machine learning algorithm needs to learn. It also reduces the correlation between variables

LOOKING AT THE PAIR-PLOT AMONG THE NUMERICAL VARIABLES, WHICH ONE HAS THE HIGHEST CORRELATION WITH THE TARGET VARIABLE?

- Temp column has highest correlation



HOW DID YOU VALIDATE THE ASSUMPTIONS OF LINEAR REGRESSION AFTER BUILDING THE MODEL ON THE TRAINING SET?

- **1.Linear relationship**

  If we see the scatter plot the relation between temp and cnt appears to be linear. If we see other variables, they don't seem to be linear.

- **2.Multivariate normality**

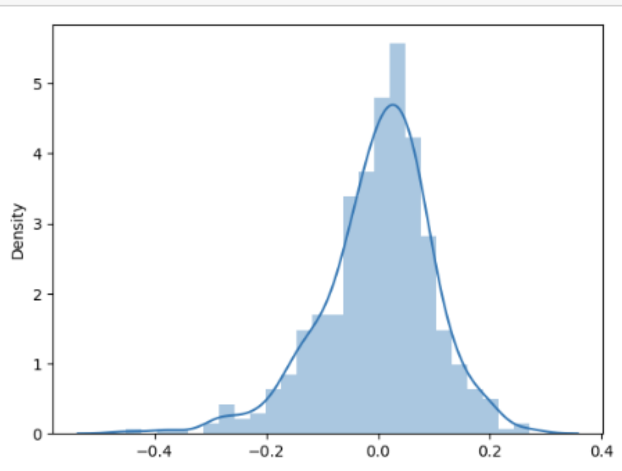
  This assumption requires that residuals are normally distributed. If we see the residual plot, it is normally distributed

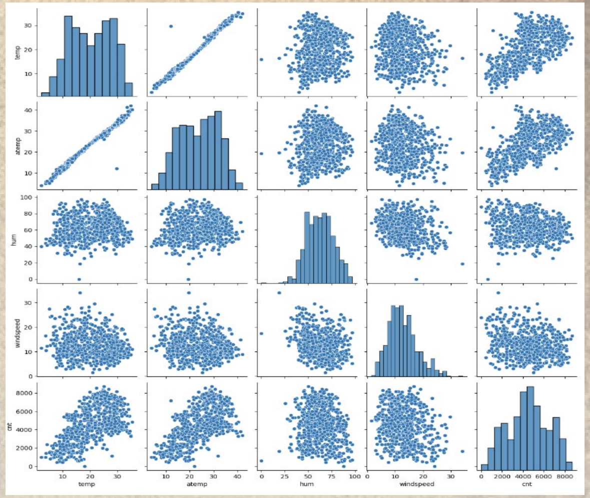
- **3. little multicollinearity**

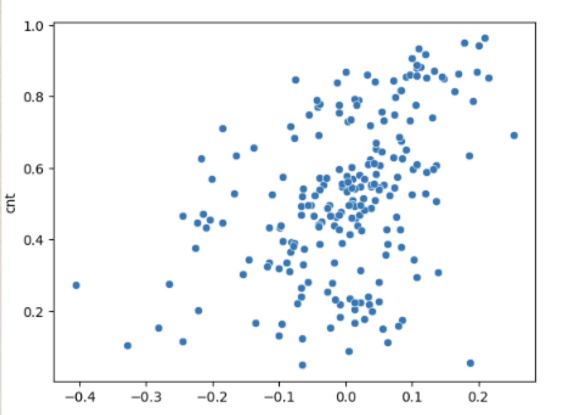
  This assumption is used to determine the relationship between independent variables. We need to ensure that there is less correlation between independent variables. We can use VIF or Correlation matrix to derive this. Scores less than 5 for VIF are desired.

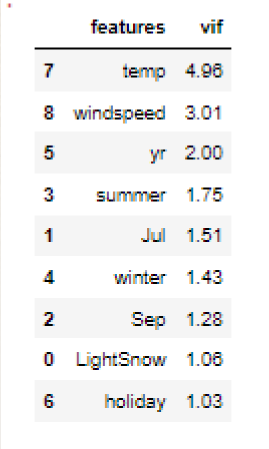
- **4.Homoscedasticity**

  This assumption need that error terms are having constant variance. It should not be increasing or decreasing like a cone but equally spread and concentrated near zero. As seen in below scatter plot the variance is evenly spread









	features	vif
7	temp	4.96
8	windspeed	3.01
5	yr	2.00
3	summer	1.75
1	Jul	1.51
4	winter	1.43
2	Sep	1.28
0	LightSnow	1.06
6	holiday	1.03

BASED ON THE FINAL MODEL, WHICH ARE THE TOP 3 FEATURES CONTRIBUTING SIGNIFICANTLY TOWARDS EXPLAINING THE DEMAND OF THE SHARED BIKES?

- Equation :

$$y = 0.084 - 0.2524\text{LightSnow} - 0.0313\text{Jul} + 0.0822\text{Sep} + 0.075\text{summer} + 0.1228\text{winter} + 0.2329\text{yr} - 0.0875\text{holiday} + 0.5854\text{temp} - 0.1459\text{windspeed}$$

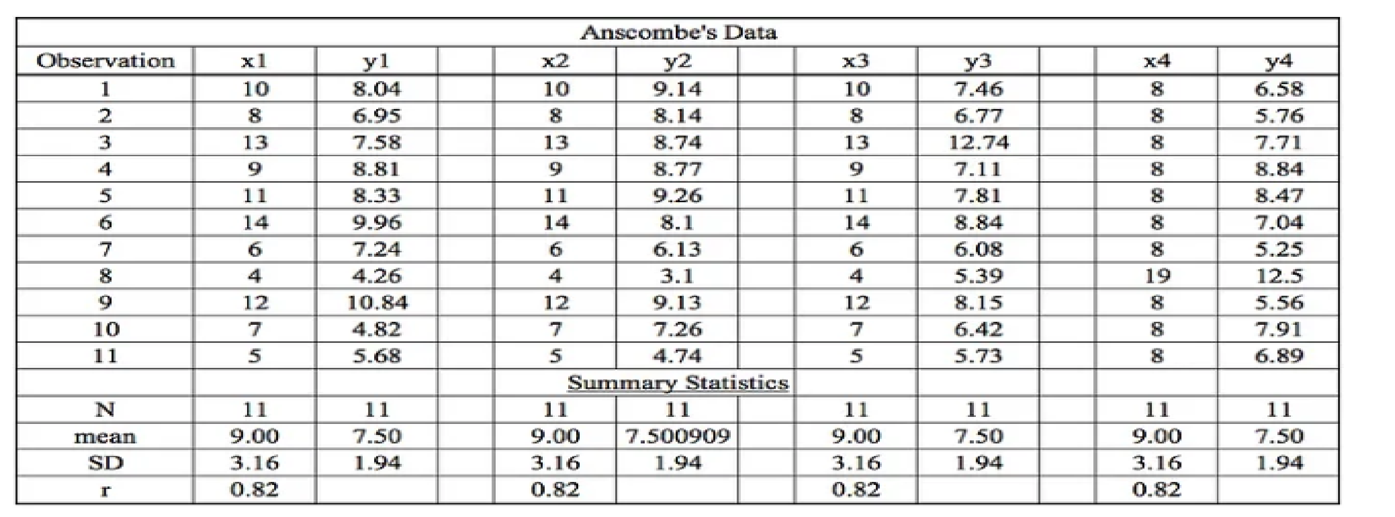
- temp , yr , LightSnow are the top 3 features contributing significantly towards explaining the demand of the shared bikes

EXPLAIN THE LINEAR REGRESSION ALGORITHM IN DETAIL.

- 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, newspaper on sales number. The basic idea over here is to identify an equation that helps to determine the predicted value here. sales with respect to to given input sales methods like tv or radio or newspaper etc also called independent variables.
- Equation for this plane or line is generally written by Y=B0+B1X1+B2X2+B3X3+….BnXn. Intent of regression is to identify B0 and B1…Bn. B0 is intercept
- 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
    - Analyzing the residuals and testing the model on test set
    - Validating the model

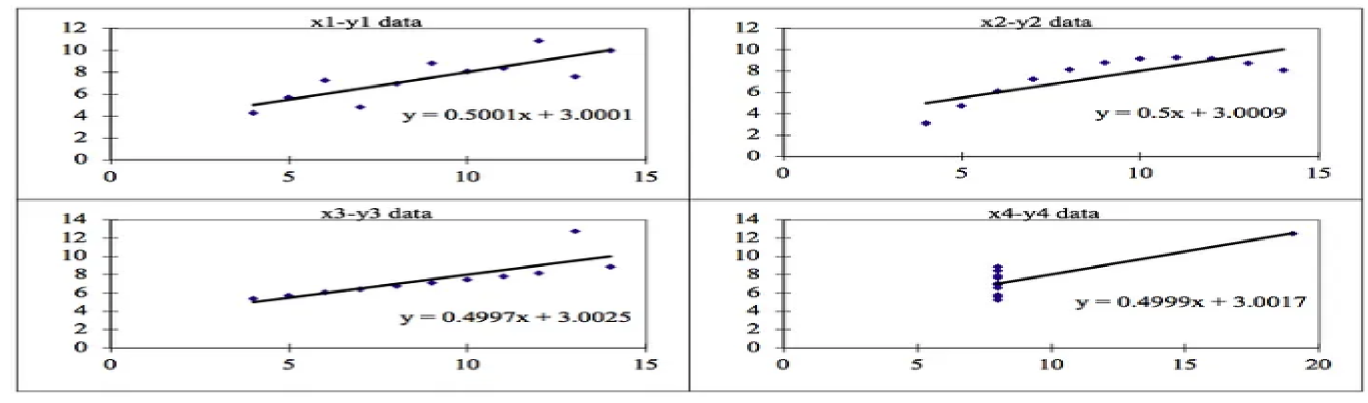
EXPLAIN THE ANSCOMBE'S QUARTET IN DETAIL.

- 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 they have very different distributions and appear very different when graphed."
- If we see 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 nonlinear 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



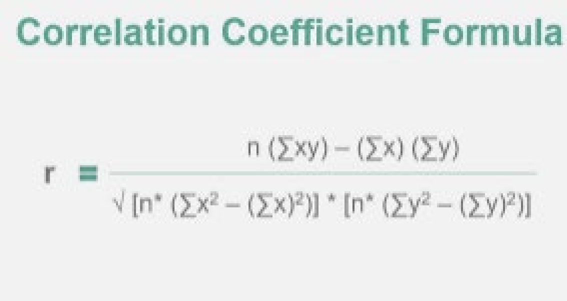
Anscombe's Data

Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
					Summary Statistics						
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	



WHAT IS PEARSON’S R?

- 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 another variable.
- The value of correlation is generally derived using corr function in python.



Correlation Coefficient Formula

$$r = \frac{n\,(\sum xy) - (\sum x)\,(\sum y)}{\sqrt{[n^*\,(\sum x^2 - (\sum x)^2)]\ ^*\ [n^*\,(\sum y^2 - (\sum y)^2)]}}$$

WHAT IS SCALING? WHY IS SCALING PERFORMED? WHAT IS THE DIFFERENCE BETWEEN NORMALIZED SCALING AND STANDARDIZED SCALING?

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-xmin)/(xmax-xmin)

whereas Standardized scaling scales value in such a way that mean lies at 0. It is computed by

(x-mean(x))/standard deviation(x)

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 As per formula shown below, if variables are highly corelated R2 becomes 1. This causes denominator to become 0 and hence infinite

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

WHAT IS A Q-Q PLOT? EXPLAIN THE USE AND IMPORTANCE OF A Q-Q PLOT IN LINEAR REGRESSION.

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