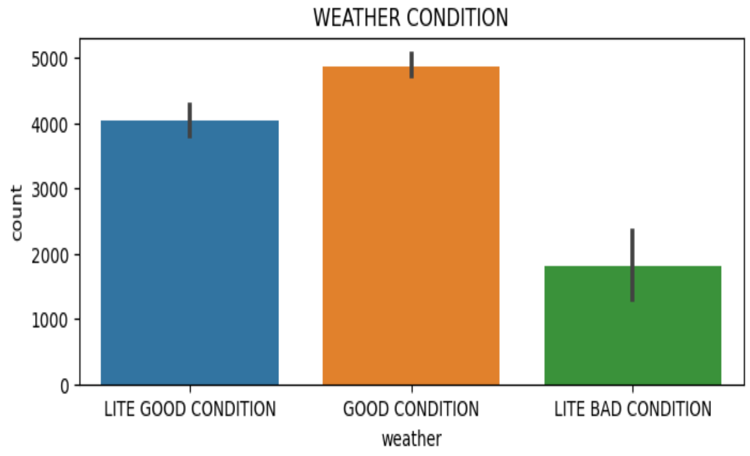
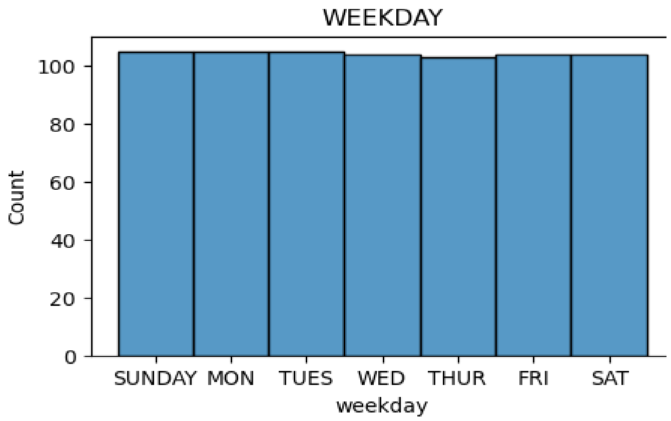
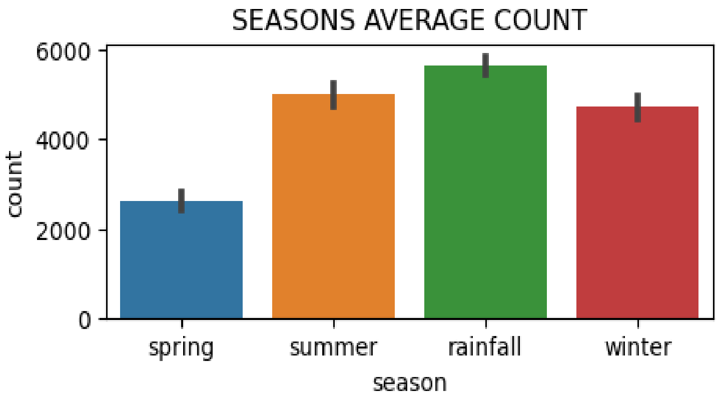
-: Assignment-based Subjective Questions :-

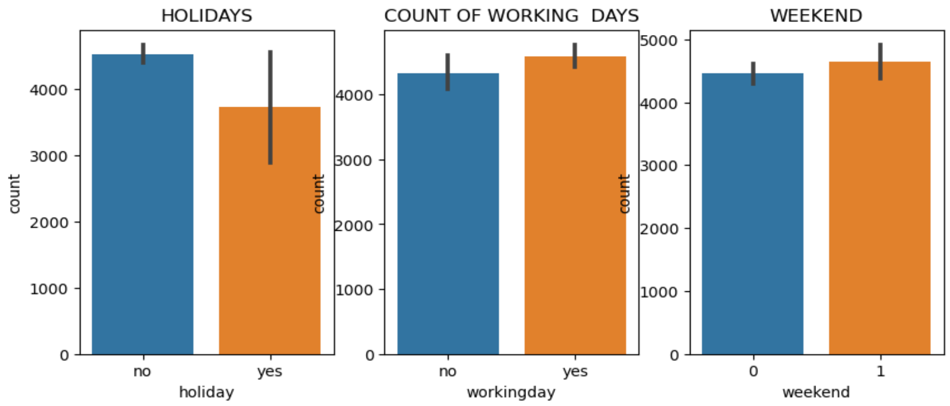
1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

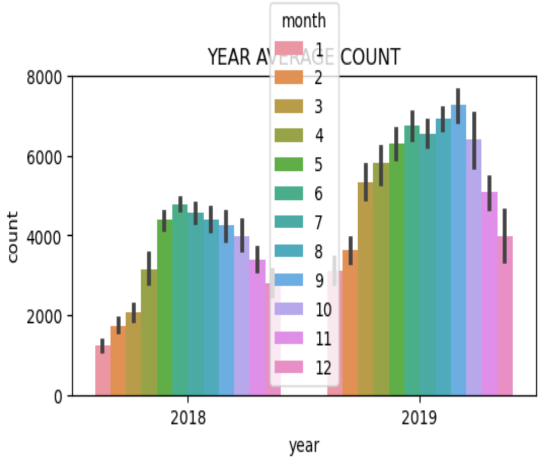
ANS : FROM MY ANALYSIS ON THE CATEGORICAL VARIABLES FROM THE DATA SET I OBSERVED THAT , IN OUR DATA SET WE HAVE CATEGORICAL VARIBLES THEY ARE WEATHER,SEASON, YEAR, MONTH,HOLIDAY,WEEKDAY AND WORKING DAY .ON THE DEPENDENT VARIBLE THERE IS LESS DEMAND OF BIKE IN WEATHER CONDITION AND WHEN THERE IS A WEEKEND THERE IS A DEMAND OF BIKE LET ME SHOW SOME GRAPH WHICH SHOWS THE BEHAVIOUR OF THE CATEGO RICAL VARIBLE ON THE DEPENDENT VARIABLE.

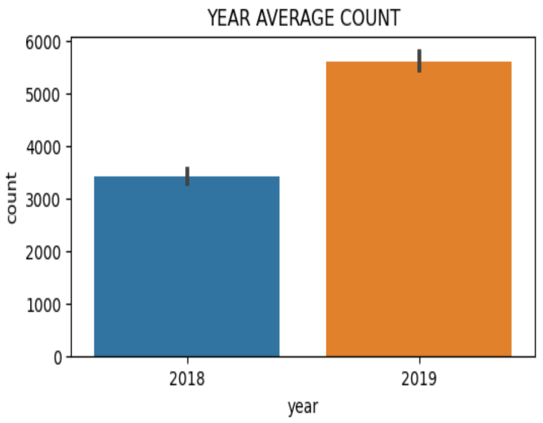












FROM THE GRAPH WE HAVE OBSERVED THAT

1. WHEN THERE IS HOLIDAY THE DEMAND IN BIKE IS LESS

2.IN THE SPRING SEASON THE BIKE DEMAND IS VERY VERY LESS

3.AND ALSO THE BIKE DEMAND IS INCRESED VERY HIGH IN THE YEAR 2019 AS COMAPRE TO 2018

4. WHEN THERE IS WEEKEND THERE IS HIGH DEMAND IN BIKE

5. WHEN THE WEATHER IS IN GOOD CONDITION THE BIKE DEMAND IS HIGH

6. IN WEEKDAY ALL DAYS ARE EQUAL BUT THERE IS SLIGHT INCREASE IN SUN,MON,TUE,FRI & SAT

7. WHEN IT IS RAIN FALL THE DEMAND OF BIKE SHARE IS VERY HIGH WHEN WE COMPARE TO ANOTHER SEASON

---******--------------------------------****-----------------------------------

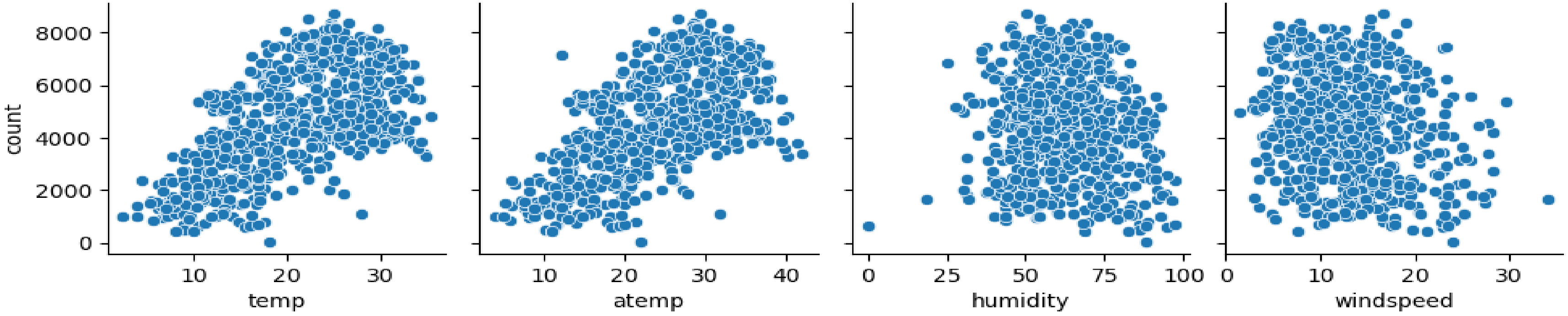
2. Why Is It Important To Use Drop_First=True During Dummy Variable Creation ?

ANSWER : WHEN WE CREATE A DUMMY VARIABLE FOR A CATEGORICAL COLUMNS . WHEN THE COLUMNS ARE CREATED WE HAVE TO DROP THE FIRST ONE COLUMN EXAMPLE IF WE HAVE **N COLUMNS, THEN WE HAVE TO N-1 COUMNS** .GENERALLY WE DROP THE FIRST COLUMN **DUE TO MULTICOLLINEARITY**. BY DROPPING A DUMMY VARIABLE COLUMN, IN GENERAL, IF WE HAVE NUMBER OF CATEGORIES, WE WILL USE DUMMY VARIABLES. **DROPPING ONE DUMMY VARIABLE TO PROTECT FROM THE DUMMY VARIABLE TRAP.**

ONE DUMMY VARIABLE IS HIGHLY CORRELATED WITH OTHER DUMMY VARIABLES. USING ALL DUMMY VARIABLES FOR REGRESSION MODELS LEADS TO A ***DUMMY VARIABLE TRAP***. SO, THE REGRESSION MODELS SHOULD BE DESIGNED **TO EXCLUDE ONE DUMMY VARIABLE**.

---******--------------------------------------****---

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



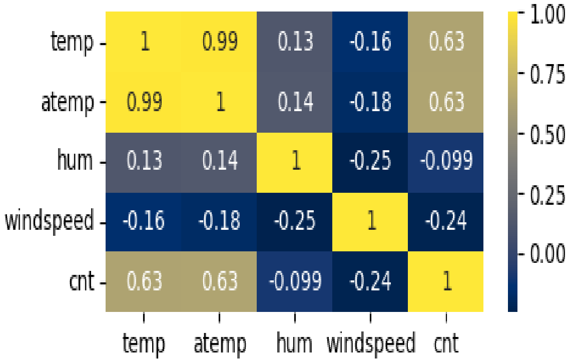
ANSWER : FROM THE ABOVE PAIR PLOT WE OBSERVED THAT

a) THERE IS HIGH POSITIVE CORRELATION BETWEEN THE TEMPERATURE AND COUNT (0.63)

b) HIGHLY POSITIVE CORRELATION BETWEEN THE ATEMP AND COUNT (0.63)

c) FROM THE HEATMAP AND PAIRPLOT WE OBSERED THAT WINDSPEED AND CNT ARE (-0.24) NEGATIVELY CORRELATED WITH EACHOTHER

d) HUM AND CNT ARE -VE CORRELATED (-0.099)

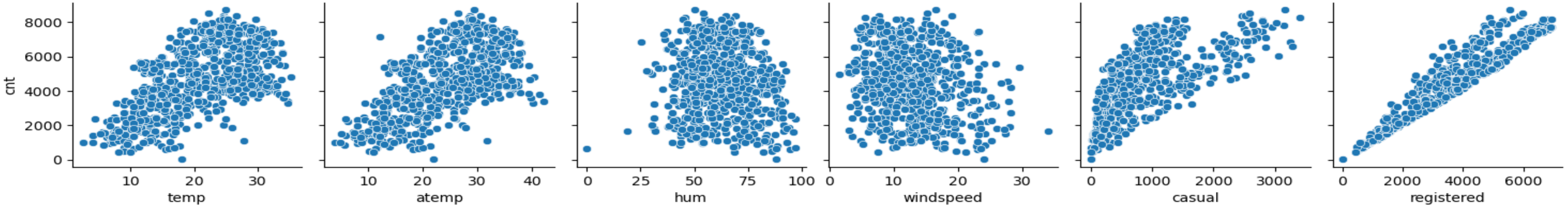


4 How did you validate the assumptions of linear regression after building the model on the training set?

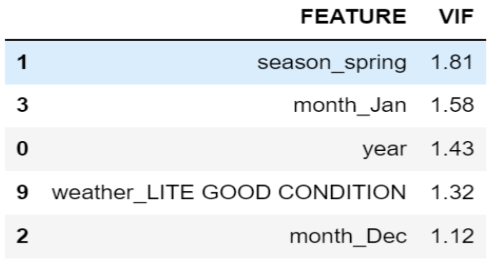
ANS: ASSUMPTION OF LINEAR REGRESSION:

- ✓ LINEAR RELATION SHIP BETWEEN THE INPUT AND TARGET VARIABLE OR X AND Y
- ✓ NO MULTICOLLINEARITY
- ✓ NORMALITY OF RESIDUALS
- ✓ HOMOSCEDASTICITY
- ✓ NO AUTOCORRELATION OF ERRORS

1 LINEAR RELATION SHIP BETWEEN THE INPUT AND TARGET VARIABLE OR X AND Y : THERE SHOULD BE A LINEAR REALTION BETWEEN X AND Y . FROM THE DATA SET WE HAVE OBSERVED THAT THERE IS LINEAR RELATION SHIP FROM THE GRAPH TEMP ATMEP CASUAL REGISTERD ARE LINEAR AND WHERE AS HUM AND WINDSPEED NEGATIVE LINEAR

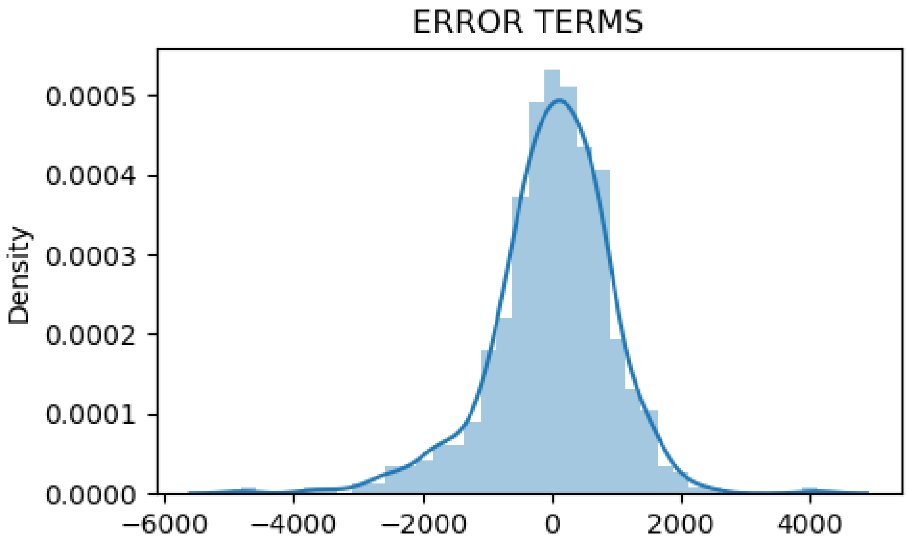


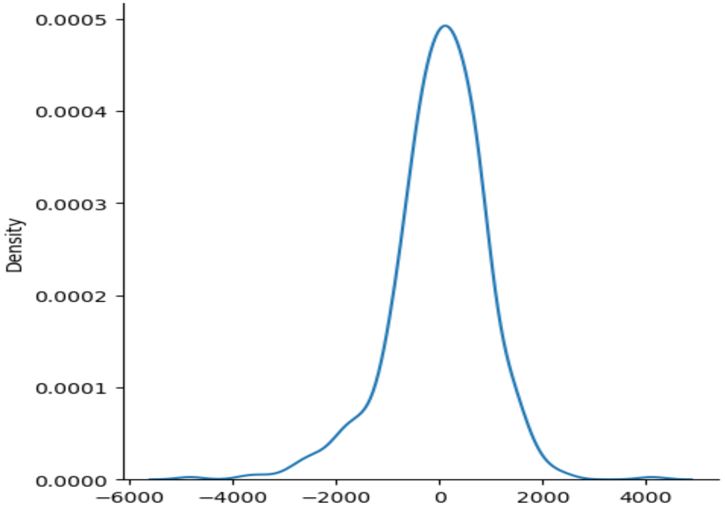
2 NO MULTICOLLINEARITY : THERE SHOULD NOT BE MULTICOLLINEARITY IT INDICATES THE INPUT COLUMN SHOULD BE INDEPENDENT SHOULD NOT DEPENDENT ON INPUT COLUMN FOR DETECTING THE MULTICOLLINEARITY WE HAVE TO USE VIF (VARIENCE INFLATION FACTOR) IF THE VIF IS LESS THEN <5 THEN WE CAN CONCLUD THAT THERE IS NO MULTICOLLINEARITY IN OUR MODEL



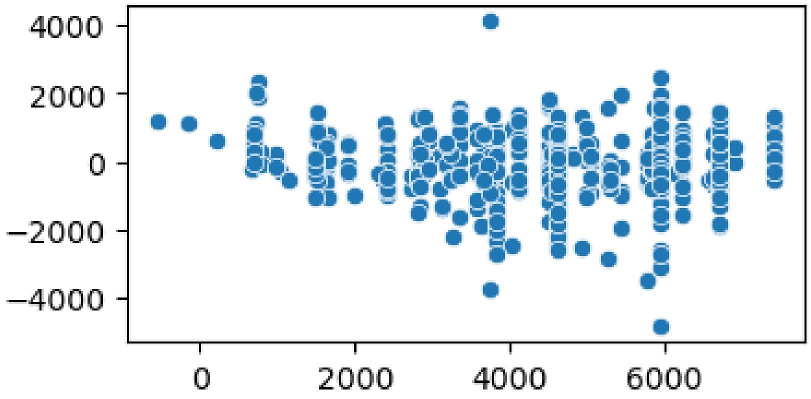
	FEATURE	VIF
1	season_spring	1.81
3	month_Jan	1.58
0	year	1.43
9	weather_LITE GOOD CONDITION	1.32
2	month_Dec	1.12

3 NORMALITY OF RESIDUALS : RESIDUAL MEANS(THE DISTANCE BETWEEN THE ACTUAL VALUE AND PREDICTED VALUE IS KNOW AS RESIDUAL) WHEN WE DO PREDICTIONS WE WILL GET RESIDUAL IF WE PLOT THE RESIDUAL IT SHOULD BE NORMAL .IT MEANS THAT MEAN SHOLUD BE CLOSE TO ZERO . NORMALITY OF RESIDUAL MEANS THE ERRORS SHOULD BE NORMALLY DISTRIBUTED.FROM THE GRAPH WE CAN CONCLUDE THAT RESIDUALS ARE NORMMALLY DISTRIBUTED





4 HOMOSCEDASTICITY : IN REGRESSION ANALYSIS, HOMOSCEDASTICITY MEANS THE VARIANCE OF THE DEPENDENT VARIABLE IS THE SAME FOR ALL THE DATA. SO, IN HOMOSCEDASTICITY, THE RESIDUAL TERM IS CONSTANT ACROSS OBSERVATIONS. SIMPLY, AS THE VALUE OF THE DEPENDENT VARIABLE CHANGES, THE ERROR TERM DOES NOT VARY MUCH. WHEN WE PLOT .WHEN WE PLOT A RESIDUAL ON THE SCATTER PLOT ,THE SPREAD OF THE DOTS SHOULD BE EQUAL



5 NO AUTOCORRELATION OF ERRORS (*RESIDUAL ERRORS SHOULD BE INDEPENDENCE (WITHOUT AUTOCORRELATION)):*

There is no autocorrelation of errors. Linear regression model assumes that error terms are independent. This means that the error term of one observation is not influenced by the error term of another observation. In case it is not so, it is termed as autocorrelation.

THESE ARE THE VALIDATE THE ASSUMPTIONS OF LINEAR REGRESSION AFTER BUILDING THE MODEL ON THE TRAINING SET

---********-----------------------------------********--

5 Based On The Final Model, Which Are The Top 3 Features Contributing Significantly Towards Explaining The Demand Of The Shared Bikes ?

ANS : BASED ON THE FINAL MODEL, WHICH ARE THE TOP 3 FEATURES CONTRIBUTING SIGNIFICANTLY TOWARDS EXPLAINING THE DEMAND OF THE SHARED BIKES I.E TEMP WHICH HAS HIGH DEMAND , YEAR 2019 SAYS THAT THERE WILL BE A HIGH DEMAND OF BIKE AND SEASON WINTER ALSO EXPLAIN THE HIGH DEMAND OF THE BIKE AND ALSO WEEKDAY_SUNDAY IS NEXT VARIBLE WHICH EXPLAIN HIGH DEMAND OF BIKE

General Subjective Questions :

1 . EXPLAIN THE LINEAR REGRESSION ALGORITHM IN DETAIL.

ANS : REGRESSION IS NOTHING BUT THE OUT PUT VARIABLE TOBE PREDICTED IS A CONTINUOUS VARIABLE.

LINEAR REGRESSION: IT IS MACHINE LEARNING TECHNIQUE WHICH PREDICT THE VARIABLE BASED ON THE CONTINUOUS OUTCOMES. LINEAR REGRESSION COMES UNDER SUPERVISED MACHINE LEARNING.

SUPERVISED AND UNSUPERVISED LEARNING

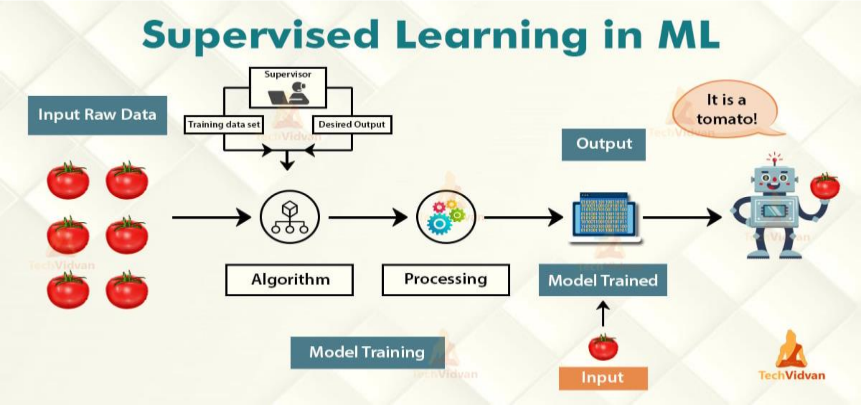
SUPERVISED MACHINE LEARNING: IT IS COMPLET LABELED DATA BASED ON THE PREVIOUS DATA THE MACHINE LEARNING MODELS ARE FORMED

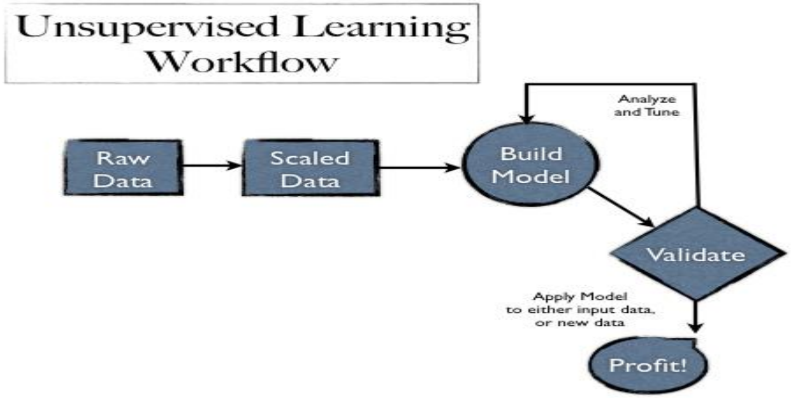
- LINEAR REGRESSION : IT USE LABEL DATA AND IT SHOULD BE CONTINUOUS VALUE
- CLASSIFICATION : CLASSIFICATION REFERS TO A PREDICTIVE MODELING PROBLEM WHERE A CLASS LABEL IS PREDICTED FOR A GIVEN EXAMPLE OF INPUT DATA.
    - ♦ DETECTING THE MAIL WHICH IS SPAM OR HAM

UNSUPERVISED MACHINE LEARNING : THE DATA IS UNLABELED DATA THERE IS NO PREVIOUS DATA TO FIND ANY THING

CLUSTRING : CLUSTRING COMES UNDER UNSUPERVISE MACHINE LEARNING MODEL .

-: FLOW CHART OF SUPERVISED AND UNSUPERVISED MACHINE LEARNING :-



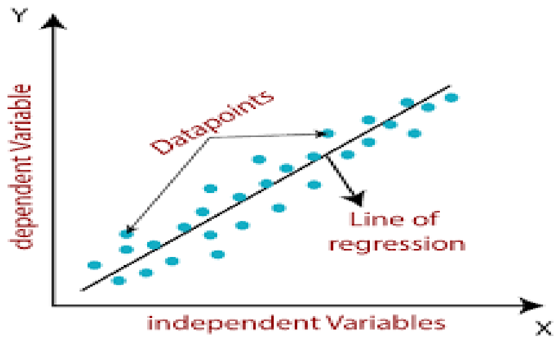

Unsupervised Learning Workflow
Raw Data
Scaled Data
Build Model
Analyze and Tune
Validate
Apply Model to either input data, or new data
Profit!

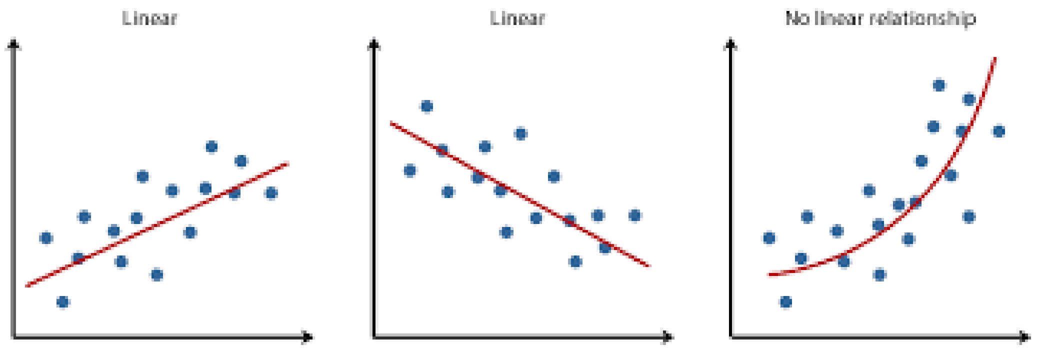
LINEAR REGRESSION IS CLASSIFIED IN TO TWO TYPES

- SIMPLE LINEAR REGRESSION
- MULTIPLE LINEAR REGRESSION

SIMPLE LINEAR REGRESSION : IT IS MATHEMATICIAL STRAIGHT LINE EQUATION $Y=MX+C$ OR $Y=B_0+ B_1X$ (WHERE X IS A FEATURE VARIABLES AND M IS SLOPE AND C IS INTERCEPT) WHICH EXPLAINS THE RELATION SHIP BETWEEN THE ONE INDEPENDENT VARIABLE AND ONE DEPENDENT VARIABLE USING A STRAIGHT LINE AND THE STRAIGHT LINE IS PLOT ON THE SCATTER PLOT. IN SIMPLE LINEAR REGRESSION THERE WILL BE LINEAR AND NON LINEAR

MULTIPLE LINNEAR REGRESION : IT IS MATHEMATICIAL STRAIGHT LINE EQUATION $Y=MX_1MX_2+MX_3+.....MX_n +C$ OR $Y=B_0+ BX_1BX_2+BX_3+.....BX_n$ (WHERE X IS A FEATURE VARIABLES AND M IS SLOPE AND C IS INTERCEPT) WHICH EXPLAINS THE RELATION SHIP BETWEEN THE MULTIPLE INDEPENDENT VARIABLE AND ONE DEPENDENT VARIABLE USING A STRAIGHT LINE AND THE STRAIGHT LINE IS PLOT ON THE SCATTER PLOT. IN MULTIPLE LINEAR REGRESSIO THERE WILL BE LINEAR AND NON LINEAR



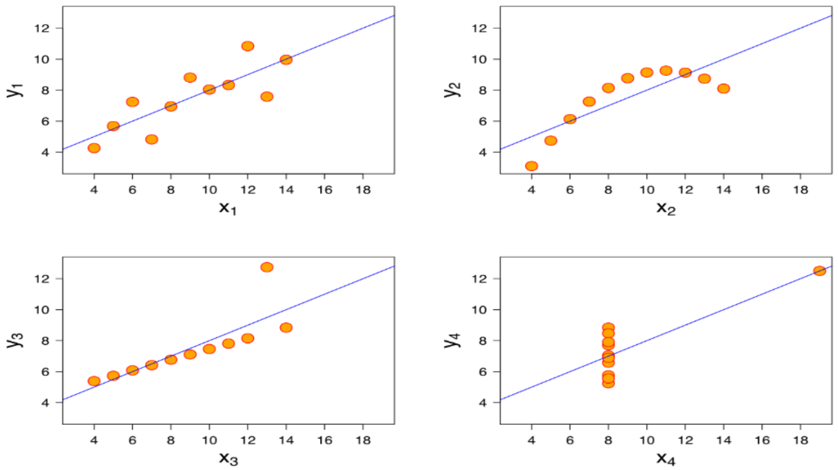


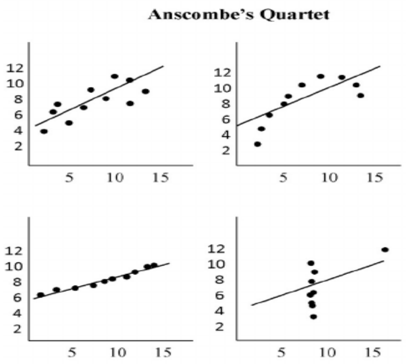
Copyright 2014. Laerd Statistics.

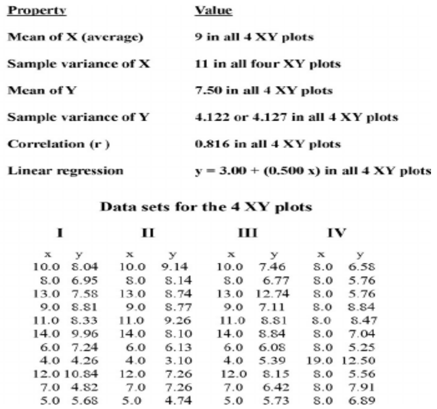
---********------------------------------------********--

2 . Explain the Anscombe's quartet in detail.?

ANS : THE STATISTICIAN FRANCIS ANSCOMBE TO DEMONSTRATE BOTH THE IMPORTANCE OF GRAPHING DATA WHEN ANALYZING IT, AND THE EFFECT OF OUTLIERS AND OTHER INFLUENTIAL OBSERVATIONS ON STATISTICAL PROPERTIES. ANSCOMBE'S QUARTET TELLS US ABOUT THE IMPORTANCE OF VISUALIZING DATA BEFORE APPLYING VARIOUS ALGORITHMS TO BUILD MODELS. ANSCOMBE'S QUARTET HIGHLIGHTS THE IMPORTANCE OF PLOTTING DATA TO CONFIRM THE VALIDITY OF THE MODEL FIT. IN EACH PANEL, THE PEARSON CORRELATION BETWEEN THE X AND Y VALUES IS THE SAME, R = . 816. IN FACT, THE FOUR DIFFERENT DATA SETS ARE ALSO EQUAL IN TERMS OF THE MEAN AND VARIANCE OF THE X AND Y VALUES . IT IS USED FOR OUTLIER DETECTION







Property	Value
Mean of X (average)	9 in all 4 XY plots
Sample variance of X	11 in all four XY plots
Mean of Y	7.50 in all 4 XY plots
Sample variance of Y	4.122 or 4.127 in all 4 XY plots
Correlation (r)	0.816 in all 4 XY plots
Linear regression	y = 3.00 + (0.500 x) in all 4 XY plots

Data sets for the 4 XY plots

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	5.76
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	8.81	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	7.26	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

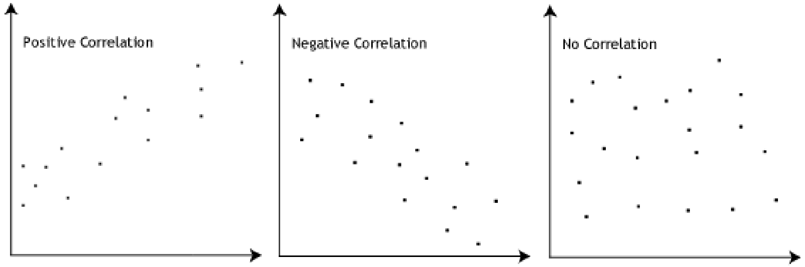
---********-------------------------------------********---------------------------------------

3 What is Pearson's R?

ANS: IN STATISTICS, THE PEARSON CORRELATION COEFFICIENT (PCC), ALSO REFERRED TO AS PEARSON'S R, THE PEARSON PRODUCT-MOMENT CORRELATION COEFFICIENT (PPMCC), OR THE BIVARIATE CORRELATION, IS A MEASURE OF LINEAR CORRELATION BETWEEN TWO SETS OF DATA. IT IS THE COVARIANCE OF TWO VARIABLES, DIVIDED BY 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.

THE PEARSON'S CORRELATION COEFFICIENT VARIES BETWEEN -1 AND +1 WHERE:

- R = 1 MEANS THE DATA IS PERFECTLY LINEAR WITH A POSITIVE SLOPE (I.E., BOTH VARIABLES TEND TO CHANGE IN THE SAME DIRECTION)
- R = -1 MEANS THE DATA IS PERFECTLY LINEAR WITH A NEGATIVE SLOPE (I.E., BOTH VARIABLES TEND TO CHANGE IN DIFFERENT DIRECTIONS)
- R = 0 MEANS THERE IS NO LINEAR ASSOCIATION
- R > 0 < 5 MEANS THERE IS A WEAK ASSOCIATION
- R > 5 < 8 MEANS THERE IS A MODERATE ASSOCIATION
- R > 8 MEANS THERE IS A STRONG ASSOCIATION



PEARSON R FORMULA

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

HERE

- ✓ **R=CORRELATION COEFFICENT**
- ✓ **X_I = VALUES OF THE X-VARIABLE IN A SAMPLE**
- ✓ **$\overline{X}$ = MEAN OF THE VARIABLE OF THE X VARIABLE**
- ✓ **Y_I = VALUES OF THE Y VARIABLE IN A SAMPLE**
- ✓ **$\overline{Y}$ = MEAN OF THE VARIABLE OF THE Y VARIABLE**

---******---********--**

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

ANS FEATURE SCALING : FEATURE SCALING IS A TECHNIQUE TO STANDARDIZE THE INDEPENDENT FEATURES PRESENT IN THE DATA IN A FIXED RANGE. IT IS PREFORMED DURING THE DATA PREPROCESSING TO HANDLE HIGHLY VARYING MAGNITUDE OR VALUES OR UNIT. IF FEATURE SCALING IS NOT DONE, THEN A MACHINE LEARNING ALGORITHM TENDS TO WEIGH GREATER VALUES, HIGHER AND CONSIDER SMALLER VALUES AS THE LOWER VALUES, REGARDLESS OF THE UNIT OF THE VALUES. IT BRING THE VALUE IN A FIXED RANGE

EXAMPLE : IF WE HAVE DATA LIKE IQ AND SALARY THOSE WHO HAVE HIGH IQ HAS HIGH SALARY IF WE CALCULATE THERE DISTANCE BETWEEN SALRAY AND IQ WE WILL GET A FAR POINT OF SALARY TO MAKE THEM IN A FIXED RANGE OR TO MAKE THEM CLOSE WE USE FEATURE SCALING

FEATURE SCALING HAS TWO TECHNIQUES NAMED

- ✓ NORMALIZATION ALSO CALLED HAS MIN MAX SCALING
- ✓ STANDARDIZATION

NORMALIZATION

MIN-MAX NORMALIZATION: **THIS TECHNIQUE RE-SCALES A FEATURE OR OBSERVATION VALUE WITH DISTRIBUTION VALUE BETWEEN 0 AND 1.**

FORMULA FOR MIN MAX SCALING :

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

STANDARDIZATION: IT IS A VERY EFFECTIVE TECHNIQUE WHICH RE-SCALES A FEATURE VALUE SO THAT IT HAS DISTRIBUTION WITH 0 MEAN VALUE AND VARIANCE EQUALS TO 1.

FORMULA FOR STANDARDIZATION :

$$x_{scaled} = \frac{x - mean}{sd}$$

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

ANS: **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).

IF ALL THE INDEPENDENT VARIABLES ARE ORTHOGONAL TO EACH OTHER, THEN VIF = 1.0. IF THERE IS PERFECT CORRELATION, THEN VIF = INFINITY. A LARGE VALUE OF VIF INDICATES THAT THERE IS A CORRELATION BETWEEN THE VARIABLE

AN INFINITE VALUE OF VIF FOR A GIVEN INDEPENDENT VARIABLE INDICATES THAT IT CAN BE PERFECTLY PREDICTED BY OTHER VARIABLES IN THE MODEL.LOOKING AT THE EQUATION ABOVE, THIS HAPPENS WHEN R^2 APPROACHES 1.

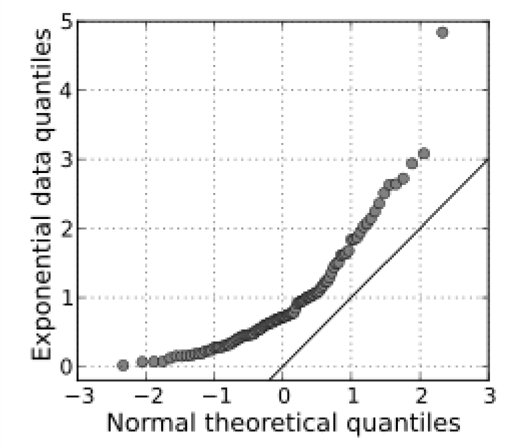
---********___________________________________********___________________________

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

ANS: **WHEN THE QUANTILES OF TWO VARIABLES ARE PLOTTED AGAINST EACH OTHER, THEN THE PLOT OBTAINED IS KNOWN AS QUANTILE – QUANTILE PLOT OR QQPLOT. THIS PLOT PROVIDES A SUMMARY OF WHETHER THE DISTRIBUTIONS OF TWO VARIABLES ARE SIMILAR OR NOT WITH RESPECT TO THE LOCATIONS. PLOTS ARE ALSO KNOWN AS QUANTILE-QUANTILE PLOTS. AS THE NAME SUGGESTS, THEY PLOT THE QUANTILES OF A SAMPLE DISTRIBUTION AGAINST QUANTILES OF A THEORETICAL DISTRIBUTION. DOING THIS HELPS US DETERMINE IF A DATASET FOLLOWS ANY PARTICULAR TYPE OF PROBABILITY DISTRIBUTION LIKE NORMAL, UNIFORM, EXPONENTIAL.**

Q-Q PLOTS (QUANTILE-QUANTILE PLOTS) ARE PLOTS OF TWO QUANTILES AGAINST EACH OTHER. A QUANTILE IS A FRACTION WHERE CERTAIN VALUES FALL BELOW THAT QUANTILE. FOR EXAMPLE, THE MEDIAN IS A QUANTILE WHERE 50% OF THE DATA FALL BELOW THAT POINT AND 50% LIE ABOVE IT. THE PURPOSE OF Q Q PLOTS IS TO FIND OUT IF TWO SETS OF DATA COME FROM THE SAME DISTRIBUTION. A 45 DEGREE ANGLE IS PLOTTED ON THE Q Q PLOT; IF THE TWO DATA SETS COME FROM A COMMON DISTRIBUTION, THE POINTS WILL FALL ON THAT REFERENCE LINE.

A Q Q PLOT SHOWING THE 45 DEGREE REFERENCE LINE:



IF THE TWO DISTRIBUTIONS BEING COMPARED ARE SIMILAR, THE POINTS IN THE Q–Q PLOT WILL APPROXIMATELY LIE ON THE LINE Y = X. IF THE DISTRIBUTIONS ARE LINEARLY RELATED, THE POINTS IN THE Q–Q PLOT WILL APPROXIMATELY LIE ON A LINE, BUT NOT NECESSARILY ON THE LINE Y = X. Q–Q PLOTS CAN ALSO BE USED AS A GRAPHICAL MEANS OF ESTIMATING PARAMETERS IN A LOCATION-SCALE FAMILY OF DISTRIBUTIONS.

A Q–Q PLOT IS USED TO 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.