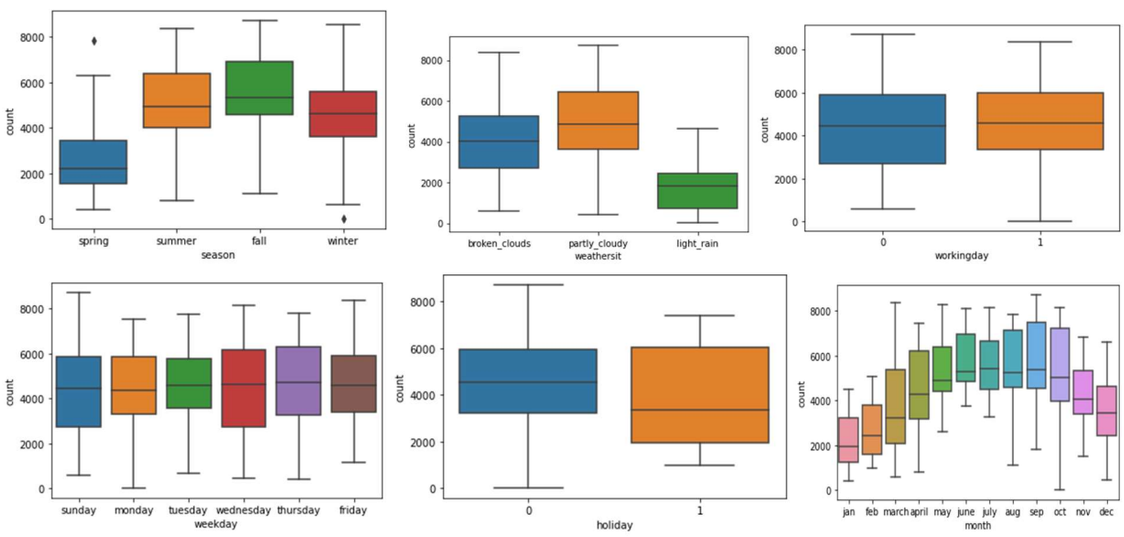
ASSIGNMENT BASED-SUBJECTIVE QUESTIONS

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



THE VARIOUS CATEGORICAL VARIABLES IN OUR DATASET ARE [“SEASON”,”WEATHERSIT”,”WORKINGDAY”,”WEEKDAY”,”HOLIDAY”,”MONTH”].THE DEPENDENT VARIABLE HERE IS THE COUNT VALUE WHICH IS ALSO CALLED TARGET VARIABLE SO THESE CATERGORICAL VARIABLES HAVE MAJOR EFFECT ON THE VALUE COUNT WHICH CAN BE SEEN IN THE ABOVE FIG.IT IS VISUALIZED USING BARPLOT GRAPH.

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

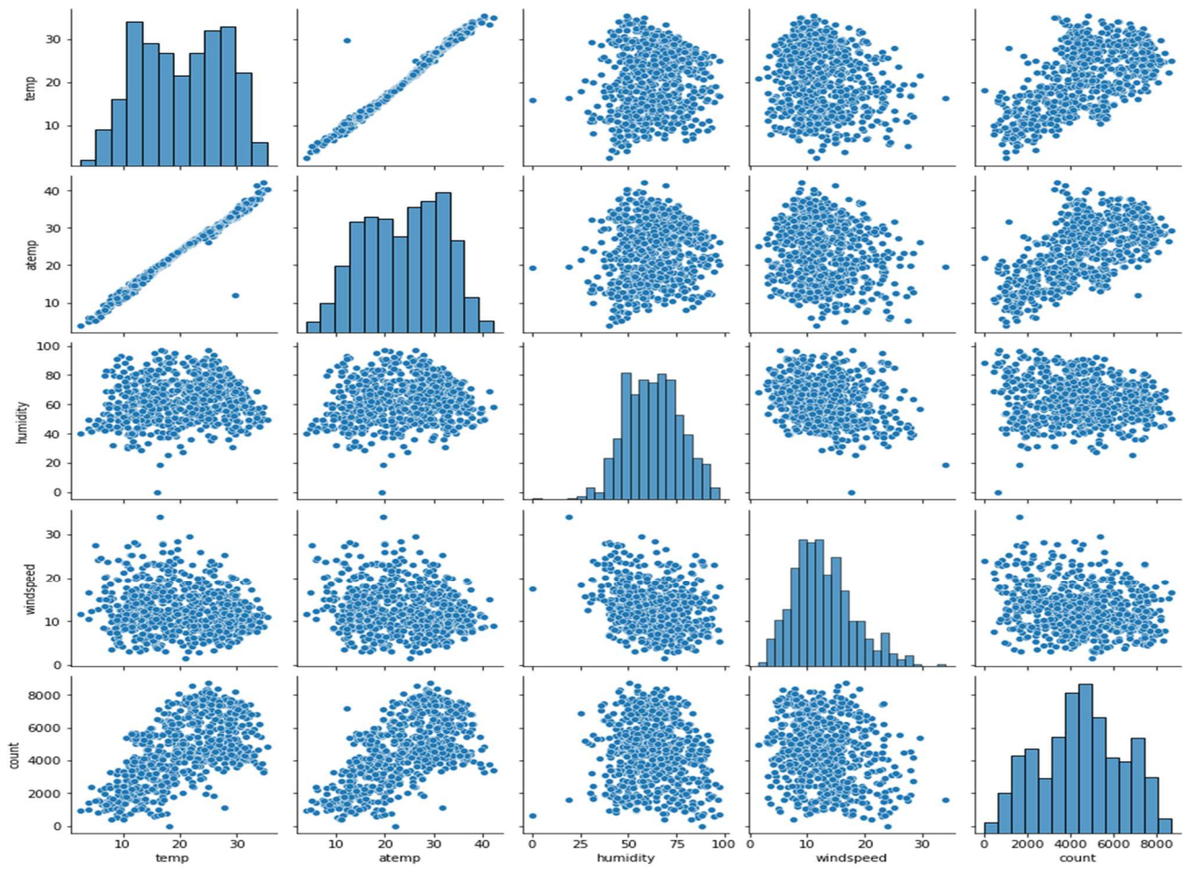
```
Period=pd.get_dummies(df['season'],drop_first=True)
Climate=pd.get_dummies(df['weathersit'],drop_first=True)
Day_of_week=pd.get_dummies(df['weekday'],drop_first=True)
Month=pd.get_dummies(df['month'],drop_first=True)
```

SO WHEN WE TAKE THE CASE OF SEASON WE ARE CREATING DUMMIES FOR IT SEASON HAS FOUR VALUES WHICH INCLUDE SPRING ,SUMMER, FALL, WINTER .DROP_FIRST ALLOWS US TO DROP THE FIRST VARIABLE AND IDENTIFY IT THROUGH ALL THE OTHER COLUMNS AS 0. SO IF

SUMMER,FALL,WINTER VALUES ARE 0 AND THE SPRING COLUMN BEING DROPED AS IT'S THE FIRST COLUMN .WHEN ALL THREE VALUES ARE 0 IT INDICATES AS SPRING .

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

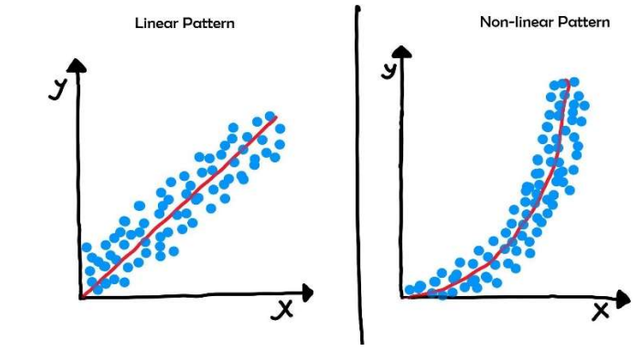
TEMP AND ATEMP VARIABLES HAVE HIGH CORRELATION IT LOOKS LIKE A LINEAR GRAPH.WE CAN CHECK IT FROM BELOW FIG.



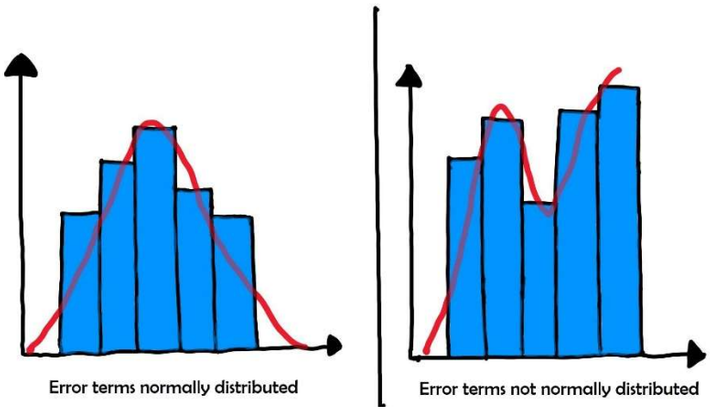
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

LINEAR REGRESSION IS VALIDATED BASED ON:

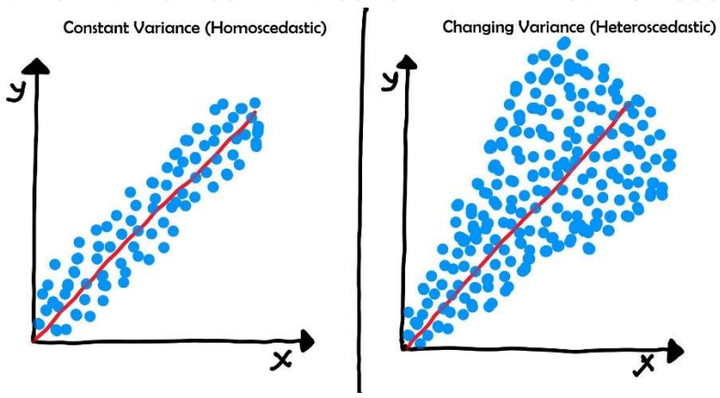
1.LINEAR RELATIONSHIP BETWEEN X AND Y



2.NORMAL DISTRIBUTION WITH MEAN=0



3.ERROR TERMS MUST HAVE CONSTANT VARIANCE: HOMOSCEDASTICITY



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

THE TOP THREE FEATURES THAT HAS A SIGNIFICANT IMPACT TOWARDS EXPLAINING THE DEMAND OF SHARED BIKES ARE YEAR,SPRING,LIGHT_RAIN.WE ARE CONSIDERING THEM BECAUSE THEY HAVE HIGH COEFFICENT VALUES.

GENERAL SUBJECTIVE QUESTIONS

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

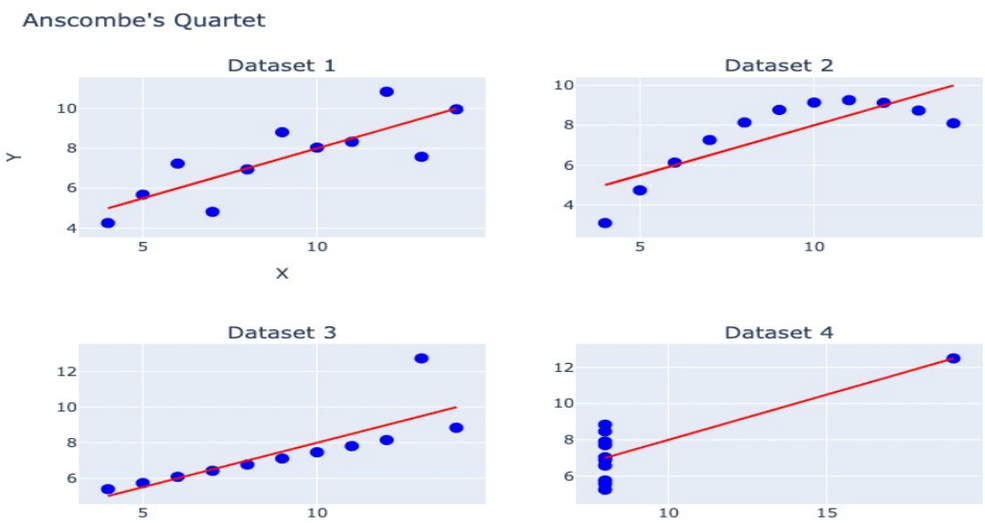
   A LINEAR REGRESSION MODEL EXPLAINS THE RELATIONSHIP BETWEEN INDEPENDENT AND DEPENDENT VARIABLES WITH THE HELP OF A STRAIGHT LINE.THE INDEPENDENT VARIABLE IS ALSO KNOWN AS PREDICTOR VARIABLE WHERE AS DEPENDENT VARIABLE IS KNOWN AS OUTPUT OR TARGET VARIABLE.THE STRAIGHT LINE IS PLOTTED ON THE SCATTER PLOT USING TWO POINTS.THE STANDARD EQUATION OF REGRESSION IS $Y = \beta_o + \beta_1 X$.DEPENDENT VARIABLE IS ALWAYS CONTINUOUS IN NATURE.IF A SINGLE VARIABLE IS GIVEN AS A INPUT THEN IT IS A SIMPLE LINEAR REGRESSION .IF THERE ARE MORE THAN ONE INDEPENDENT VARIABLES THEN IT IS CALLED MULTIPLE LINEAR REGRESSION.THE REGRESSION LINE CAN BE EITHER NEGETIVE OR POSITIVE RELATIONSHIP.WE NEED TO FIND BEST VALUES FOR β_o AND β_1 VALUES IN ORDER TO GET BEST FIT.SOME METHODS LIKE MEAN SQUARE ERROR,R-SQARED,ADJUSTED R-SQUARE COULD BE USED IN ORDER TO GET BEST FIT LINE.
   EX:

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

   ANSCOMBE QUARTET IS GROUP OF FOUR DATASETS (X,Y) WHICH HAVE SAME MEAN , SAME STANDARD DISTRIBUTION AND REGRESSION LINE EVEN THOUGH THEY HAVE VERY DIFFERENT DISTRIBUTIONS AND LOOK COMPLETELY DIFFERENT IN EACH

PLOT.BEFORE FITTING THE MODEL IT IS NECCESARY TO VISUALIZE THE DATA IN ORDER TO GET A BEST FIT OTHERWISE IT WILL BE EASY TO FOOL A REGRESSION ALGORITHM.

- **DATA SET 1:** THERE EXISTS LINEAR RELATIONSHIP BETWEEN THE POINTS.
- **DATA SET 2:** THE POINTS ARE NON-LINEAR HENCE NO LINEAR RELATIONSHIP.
- **DATA SET 3:** SHOWS AN OUTLIER WHICH IS NOT A GOOD FIT FOR LINEAR REGRESSION MODEL.
- **DATA SET 4:** THE OUTLINER GETS INVOLVED WITH THE DATASET WHICH MIGHT LEAD TO HIGH CORRELATION.



3. What is Pearson's R?

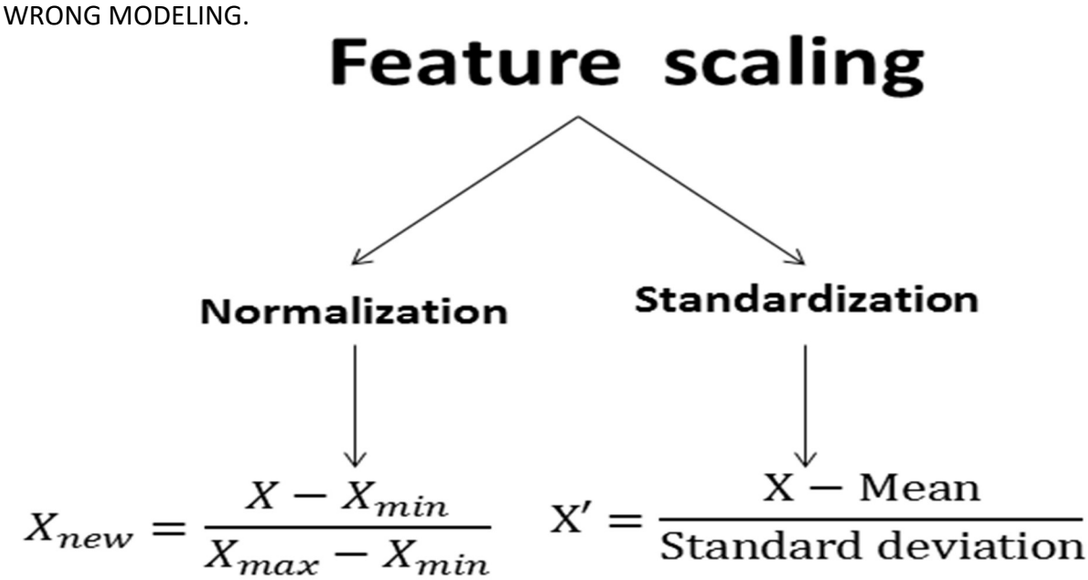
$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

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

PEARSONS R IS ALSO CALLED PEARSONS CORRELATION COEFFICIENT.IT IS THE COVARRIENCE OF THE TWO VARIABLES WHICH IS DIVIDED BY THE PRODUCT OF THEIR STANDARD DEVIATIONS.

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

SCALLING IS BASICALY TRANFORMING YOUR DATA SO THAT IT FITS WITH IN SPECIFIC SCALE.ITS LIKE PRE=PROCESSING STEP WHERE SCALING HELPS TO SPEED UP THE CALCULATIONS.
SCALING IS PERFORMED SO THAT IT DOESN'T WEIGH HIGH VALUES AND DOESN'T LEAD TO WRONG MODELING.



NORMALIZED SCALING VALUES LIE BETWEEN 0 AND 1.STANDARDIZED SCALING HAS A MEAN OF 0 AND STANDARD DEVIATION OF 1.NORMALIZATION SCALES THE MODEL USING MINIMUM AND MAXIMUM VALUES.STANDARDIZATION SCALING SCALES THE MODEL USING MEAN AND STANDARD DEVIATION.NORMALIZATION SCALING GETS AFFFECTED BY OUTLIERS WHERE AS STANDARD SCALING DOESN'T HAVE ANY EFFECT.

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

VIF-VARIANCE INFLATION FACTOR

IT HELPS IN EXPLAINING RELATIONSHIP BETWEEN INDEPENDENT VARIABLES.

IF VIF>10 IS DEFINETLY HIGH

VIF>5 SHOULD NOT BE IGNORED AND INSPECTED PROPERLY.

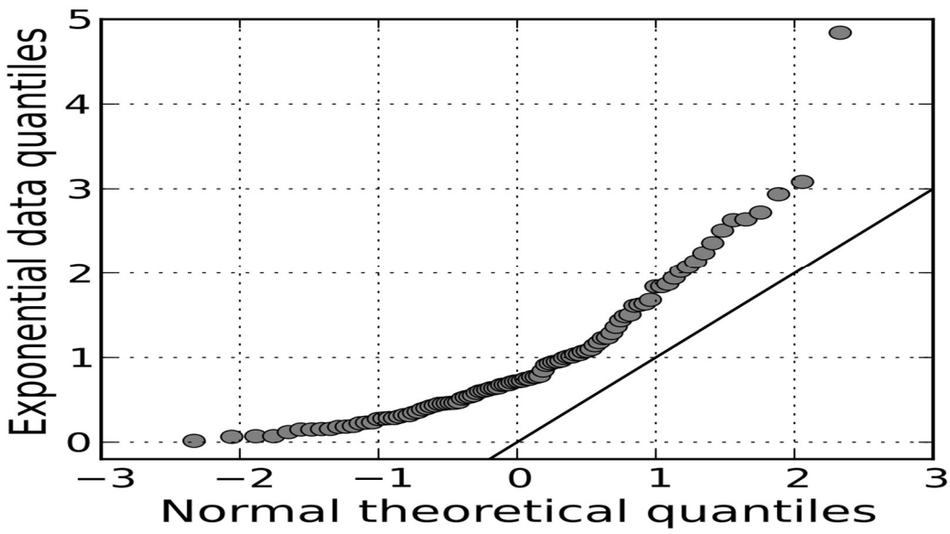
$$VIF_i = \frac{1}{1 - R_i^2}$$

WHEN R^2 VALUE IS 1 THEN VIF VALUE BECOMES INFINITE.THIS MIGHT LEAD TO MULTICOLLINEARITY

WHICH IS SOLVED BY USING DROPPING OF VARS, CREATE NEW VARIABLES USING THE OLDER ONES.

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

Q-Q PLOT IS BASICALLY IS A SCATTERED PLOT WHERE TWO DIFFERENT SET OF QUANTILES ARE PLOTTED AGAINST EACH OTHER. IT IS A TOOL WHICH HELPS US IN TELLING IF A PARTICULAR SET CAME FROM SPECIFIC PROBABILITY DISTRIBUTION.THEORETICAL POINTS ARE PLOTTED ON HORIZONTAL AXIS AND SAMPLE QUANTILES ARE PLOTTED ON VERTICAL AXIS. QUANTILE-QUANTILE PLOTS ARE MOSTLY USED FOR CHECKING THE NORMALITY OF DATA



THE IMPORTANCE OF Q-Q PLOTS IS TO VISUALLY CHECK THE DATA MEETS HOMOSCEDACITY AND NORMALITY ASSUMPTIONSOF LINEAR REGRESSION.WE HAVE TRAIN AND TEST DATASET BY USING Q-Q PLOT IT TELLS US WHETHER THEY BELONG TO SAME DISTRIBUTION OR NOT.