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**

Answer: Inferences from box plot of categorical variable with dependent variable:

1: More people use bikes when sky is clear or Few clouds or Partly cloudy

2:People share more bikes in fall and summer

3:Median of bike sharing is most in the month of June.

4.Very few number of people prefer bikes in winters specially in January

5.Median of bike sharing is most on Monday ,Thursday and Friday

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

Answer : To eliminate redundant information in feature.

Lets consider following example

row	Is Odd number	Is even number
0	0	1
1	1	0

row	Is Odd number
0	0
1	1

Redundant info leads to multicollinearity which is not good for our model.

In above example,both the tables provide same info so we don't really need dummy for both, even and odd features

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

Answer: "temp" has highest correlation with target variable.

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

Answer : 1.By Checking if there is multicollinearity.Visualised numerical columns using heatmap and checked whether any high correlation exist between variables.Also checked VIF which is calculated as below

VIF=1/(1-R2)

2. Checked linear relationship between predictors and dependent variables
3. Plotted distplot to check if errors are normally distributed which was found to be normally distributed

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

Answer:

1. for every increase in temp by 0.4480 there is increase in count of bike sharing

2. when there is decrease in humidity by .24 unit there is unit increase in count

3. with decrease in windspeed by .21 unit there there is unit increase in count

General Subjective Questions

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

Answer : Linear regression is model used in predictive analysis.It attempts to find relation between a dependent variables and one or more independent variables.There are set of predictor variables which independent and they are used to predict dependent variables.

A simple linear regression is expressed as y= c +b*X.

Where c= constant , b= coefficient

Least Square Regression :

It is on of most common method for fitting the line .In this method we calculate best fitting line by minimizing the sum of squares of the difference between actual and predicted point.

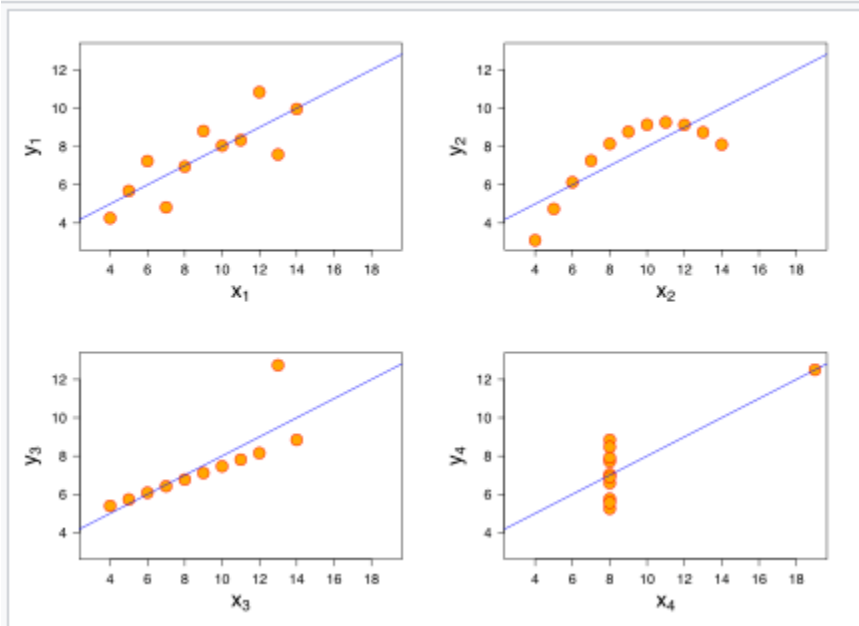
Residuals:

Residuals are difference between fitted line and observed values.

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

Answer:

Anscmbe;s quarlet comprises four datasets that have identical statisctics still they have very different distribution and they appeared completed different when plotted .



Source of picture : Wikipedia

1st picture appeared to have linear relationship.Two variable looked correlated with normal distribution

2nd picture appeared to have some relationship between two variables but it wasn't normally distributed.

3rd picture appeared to have linear distribution but its line is different

4 th picture showed that when there is high value point is in distribution it can misdirect because it can produce high correlation coefficient.

3. What is Pearson's R?

Answer: Pearson coefficient if covariance of two variables say X and Y divided by the product of their standard deviation.

Pearson coefficient is denoted by a greek letter called rho which is ρ

ρ = (cov(X,Y))/ std deviation of X * standard deviation of Y

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

Answer : It is a way which is applied to independent variables to normalize data within a range.It helps us in calculations.

Sometime when we collect data,some features are contains values which vary in high magnitude,units and Range.By Scaling we normalize the highly varying data with others so that calculations can be easier.

Normalized Scaling is also called min-max scaler:

Min Max Scaler: X = (X – Min (X))/(max(X)-min(X))

Standardization scaling :

Standardization normalises the values by replacing values of their Z score

X= (X-mean(X)) / sd(X)

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

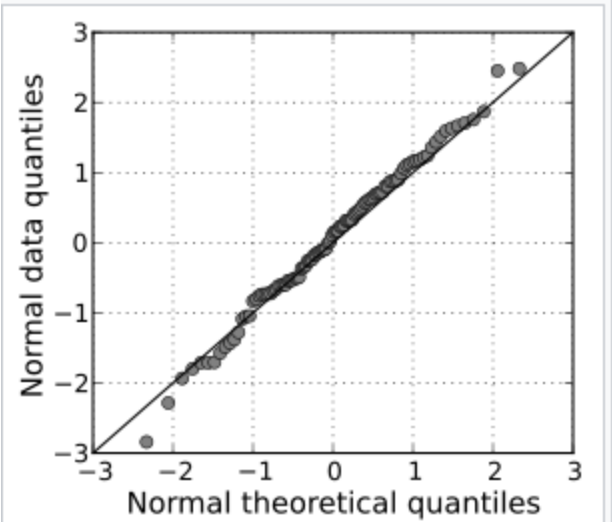
Answer: When there is perfect correlation between predicting variables the VIF = infinity

An infinite value indicates that 1 unit increase in 1 variable will lead to increase by exactly 1 unit in other variable.

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

Answer: Scatter plot created by two set of quantiles against each other is called Q- Q plot .

If they come from same set of distribution they would roughly form a line.


Normal data quantiles
Normal theoretical quantiles
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3