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] The categorical variables present in the data set are

a) Year b) Month c) Season 4) Working Day 5) Holiday 6) Weekday

Following are the inference about their effect on dependent variable “Count” or cnt

1. Average of Count is more in the year 2019 from year 2018. This shows bike sharing system is gaining popularity over the year
2. Average count is less for holiday. This shows there is less demand for bike sharing during holidays
3. Average count gradually increases from January to May, then remains stable and start decreasing from October.
4. Average count is more in Fall season and less in Spring season
5. Working day and Weekday have less impact to count.

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

[Ans] During dummy variable creation each value is encoded as multiple columns where the value 1 in the column represent actual value and 0 represent other. So when one value can be represented as all other having 0. So when first is set as True , one value can represented as all Zero. Thus it helps in reducing 1 extra column and thereby reducing collinearity among columns

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

[Ans] Temperature has highest correlation with target variable

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

[Ans] The assumption are validate by following below steps

1. Normality: Plotting residual and checking for mean approximately equal to zero. Plotting distplot of residual to check if is follows normal disbution with mean around zero
2. Homodescasticity: This is validated by plotting residual with respected to fitted variable(y_train) in scatter plot and checking if the point are around horizontal line in the middle
3. Independence: By plotting the correlation heatmap among dependent variables and check is any cell has very high value

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

[Ans] The top 3 features which contribute significantly explaining demand of bikes are

a) Temperature : The demand is positively related to temperature
b) Year: The demand is increasing year of year
c) Weather Situation: If weather is bad (light rainy or snowy) the demand decreases

General Subjective Questions

1. Explain the linear regression algorithm in detail.

[Ans] Linear Regression is a machine learning algorithm which is used find the linear relationship between a target variable and set of feature variables. The linear relationship can be positive or negative. When relation is positive the target variable increase with increase in feature variable and is the relation is negative target variable decrease with increase in feature variable.

Linear Regression is categorized into two types

a) Simple Linear regression: When there is only one feature variable
b) Multiple Linear regression: When the target variable is dependent on multiple feature variables

Linear regression can be described as below mathematical equation

$$Y=\beta_0 + \beta_1*X_1+ \beta_2*X_2+ \beta_3*X_3 \ldots\ldots+ \beta_n*X_n$$

Where Y = Target Variable

$X_1, X_2 \ldots$ Xn are n independent variables

β_0 = Constant or intercept

$\beta_1, \beta_2, \beta_n$ are called co-efficient associated with respective independent variables

Gradient Descent method can be utilized to find the values of these coefficients.

Linear regression can only be applied when the target variable of continuous in nature. However the feature variables can still be categorical. But to apply regression , categorical variables are encoded as continuous variables. One such encoding is one hot encoding.

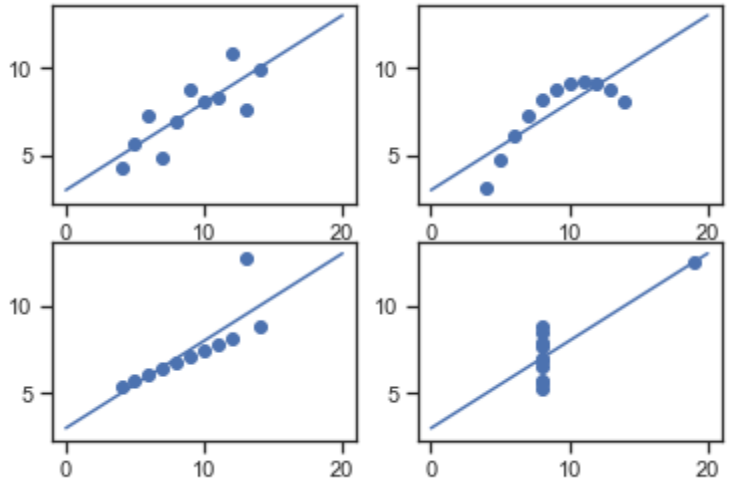
Some of examples where Linear regression is used are

1. To find how much marketing investment is needed to increase sales
2. Effect weather condition on cricket scrore
3. Effect of a drug on health parameters

2. Explain the Anscombe's quartet in detail.

[Ans] Anscombe's quartet is an example proposed by statistician Francis Anscombe to emphasize importance of data visualization . The example consists of 4 data sets consisting of 11 x-y data points which have same statistical parameters (mean,variance, correlation , regression line etc).

However then x y point are plotted , they look very different from each other



This shows that the statistical properties can be misleading and data visualization is important to assert if the relationship between the variables are valid .

3. What is Pearson's R?

[Ans] Pearson's R describes strength and direction of relationship between two quantitative/numerical variables. The value of Pearson R can be between -1 and 1.

The strength of R value shows how much the two variables are related linearly and sign shows if it is positive or negative relationship.

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

r= Pearson correlation coefficient
x=Values in the first set of data

y= Values in the second set of data

n= Total number of values

Pearson R between 2 variables can be calculated when the two variables are numerical and follows normal distribution.

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

[Ans] Scaling is method to normalize the feature variables.

When the target variable is dependent multiple features, the variables can have different ranges. So machine learning models like multiple linear regression which depend of gradient descent algorithm does not perform well when it is subject different range of values for different features. That's before applying such algorithms feature scaling is applied. Scaling only affect coefficient , the other statical parameters such as R square values , p values are not affected via scaling

Normalized Scaling is technique where the values of a random variable are shfted and rescaled so that they lie in between 0 and 1

Xnorm= (X-Xmin)/(Xmax-Xmin)

X = Value of a variable at a data point

Xmin= Minimum value of variable X

Xmax= Maximun value of variable X

Xnorm= normalized value

Standardization scaling is a technique where the values of variable are centered arrounf mean with unit standard deviation

Xnorm= (X-Xmean)/sigma

Xmean= mean of random variable X

Sigma= standard deviation of X

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

[Ans] $VIFi = 1/ 1-Ri^2$

So When Ri^2 is 1 then VIF can become infinity

Ri^2 represents the unadjusted coefficient of determination for regressing the ith independent variable on the remaining ones.

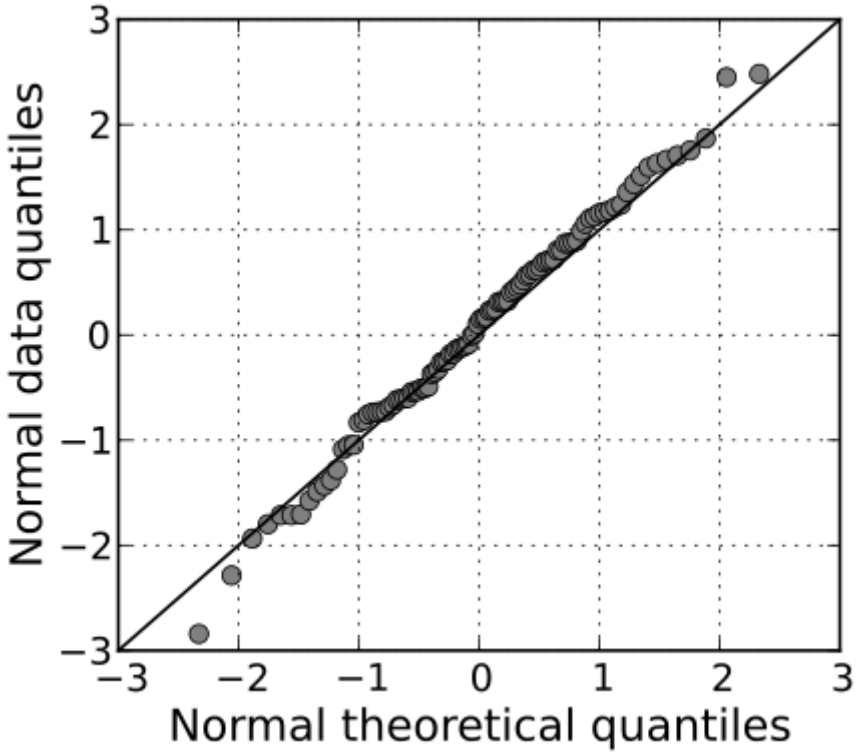
When an independent variable can be exactly explained by other variables Ri^2 will be 1 causing VIF to be infinite. For example, if there is variable x and its square x^2 are considered feature variables it will result VIF infinite for x and x^2 as they are exactly relating and Ri2 becomes 1

Another example if there is dummy variables and two columns are exactly related. For example when we use 2 columns for gender (M/F) they are exactly opposite . They will be exactly correlated and will result in VIF inifinity

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

[Ans] Q-Q Plot or quantile plot is used to determine if two datasets come from a common distribution such as normal, exponential. Most of the cases such as in Linear regression it is used to validate normality property

It is plotted with the theoretical quantile(0.95,0.997,0.01etc) of basic normal distribution with mean 0 and standard deviation 1 on X axis and the variable under considerations quantile values on Y axis.



To determine if the residuals are following a normal distribution Q-Q plot can be used. A perfect straight line at 45 angle shows data is normally distributed