Assignment-based Subjective Questions- Vikas Bhartiya DS-43

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)**

Ans: I have plotted the box plot for categorical column Vs Target variable and can infer the following:

Season- Season no.2 (Summer) and Season no.3(Fall) are when highest number of the bikes are rented.

Mnth:- June,July and Aug are the month when highest number of bikes are rented and in Jan lowest number of bikes are rented.

Weathersit:- Weathersit no.1 (clear or partly cloudy) highest number of bikes are rented almost 3500 to 6500 in numbers.

Holiday: We can see from the box plot when its holiday median is higher when its not the holiday.

Weekday:- thu fri and sat are more booking compare to the other day

Working Day:- Almost equal number of booking whether its working day or non working day.

Year: Year 2019 has more number of booking comparing to the last year 2018.

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

Ans:- When we create the dummy variable all categories convert into 0 and 1. That means if have N category then n column will be created and corresponding to that in every row that category will be 1 and rest all will be zero.

We do drop_first because if all other are zero that means it is dropped category.

In other word we can conclude all N features from N-1 features.

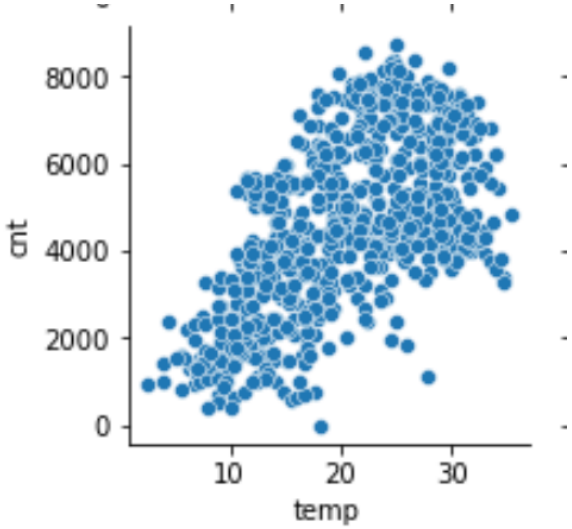
For example from the assignment we can see from the first row that weathersit_2 and weathersit_3 are zero that means it is weathersit=1

weathersit_2	weathersit_3
0	0
0	0
0	0
1	0
0	0

So keeping weathersit_1 as a feature is redundant and we do not want include redundant feature in our model.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Ans:- from the pair plot we can see temp has highest correlation with the target variable



4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans:- In my assignment section -9 , I have validated assumption

- Error terms are normally distributed- Using distribution plot
- Error terms are independent- Using reg plot
- Multicorrelation- using VIF and heatmap
- Homoscedasticity- using reg plot

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

Ans:- Top 3 features are-
1. Temperature: coff is 0.5435
2.Weathersit_3: coff is -0.2756
-3.Year: coff is 0.23 means.

General Subjective Questions

1.Explain the linear regression algorithm in detail. (4 marks)

Ans:- Linear Regression is machine learning algorithm which used to predict the dependent variable based on independent variable having linear relationship between them.

There are two type of Linear Regression

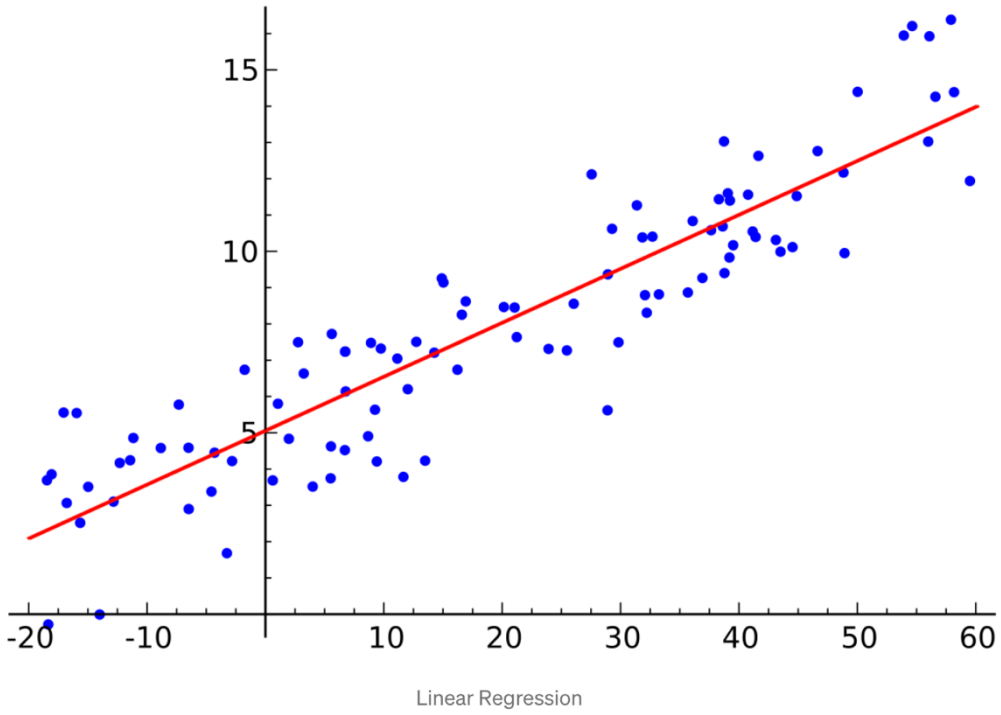
1. Simple Linear Regression
2. Multiple Linear regression.

Simple Linear Regression: In this type of regression there is only one independent variable.

Eqaution for this is Y=mx+c

Where y in dependent variable and x is independent variable

m is the slope and c is the constant.



Linear Regression

Assumption of Simple Linear Regression:

1. Linear relationship between X and Y
2. Error terms are normally distributed(no X and Y)
3. Error terms are independent to each other.
4. Error terms has the constant variance(homoscedasticity)

   Parameter to access the model:

   1. t-statistic:- used to determine he p value and hence, used to determine whether the model is significant or not.
   2. F Satistic: Used to assess whether the overall model fit is significant or not. Generally the higher the value of F statistic the more significant a model turn out to be.
   3. R -squared: After it has been concluded that model fit is significant the r-squared value tells the extent of fit, i.e. how well the straight line describes the variance in data. Its value ranges from 0 to 1, with the value 1 being the best and 0 is worst.

Multiple Linear Regression:- In this there are several independent variable . We have to find the best fit equation that can predict the dependent variable Y for the different independent variable.

Eqaution: Y=c+m1x1+m2x2+…. +mn xn+ e

Where x1,x2,…xn are the independent variable

Y is dependent variable

c is constant

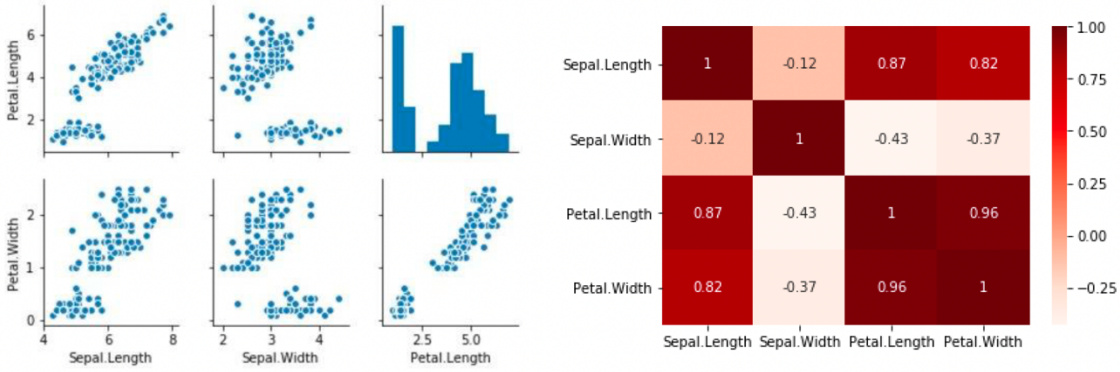
m1,m2…mn are the coefficient.

e is the error term.

So we can see that Multiple Linear regression is the combination of the many Simple Linear Regression.

Following to be considered when using the Multiple Linear Regression:

1. Adding more variable helps to add information to predict Y variable.
2. Model is now a hyperplane instead of the simple line.
3. Coefficient are obtained by minimising the sum of squared error same as the simple Linear regression.
4. All the assumption of the simple Linear regression still hold true in Multiple Linear Regression.
5. Adding more variable always is not good as it may overfit and becoming too complex.
6. Multicollinearity associated between the independent variable. Which can be seen using the heatmap and corelation matrix and calculated by VIF
7. VIF >5 should never be ignored.



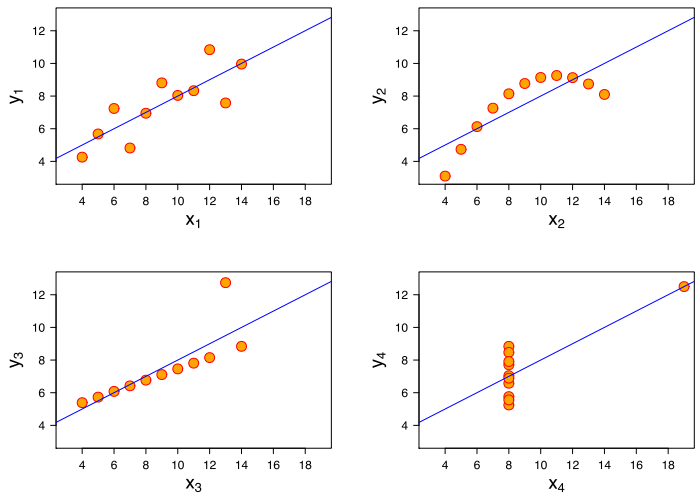
8. Feature selection is very important to use the significant variable in the model.
9. Where there are many independent variable there might possible that these variables are on different scale fro that we have to Feature scaling of these variable either by using MinMax scaler or the Standardizing.
10. Feature scalling affect only the coefficient but not other parameter like t-statistic, F-statistic,p-value, R-square etc.
11. We have to change the categorical variable into dummies.

2.Explain the Anscombe's quartet in detail.(3 marks)

Ans: Anscombe's quartet was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting data before you analyze it and build your model. These four data sets have nearly the same statistical observations, which provide the same information (involving variance and mean) for each x and y point in all four data sets. However, when you plot these data sets, they look very different from one another.

Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

We can define these four plots as follows:



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

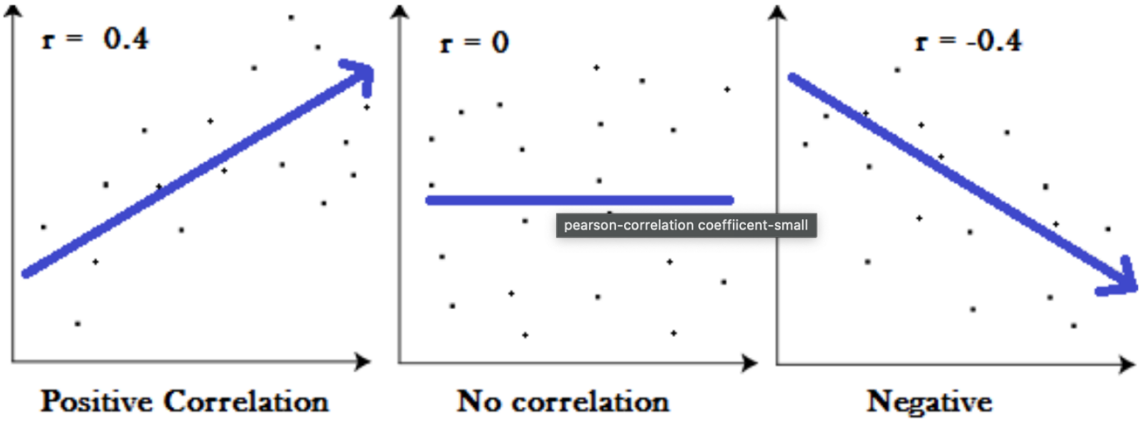
- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.
- The second graph (top right); while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third graph (bottom left), the modelled relationship is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets

3.What is Pearson's R? (3 marks)

Ans:- Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

- Value of the Pearson R lies between -1 and 1.
- 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



Most commonly use formula for Pearson R is

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

> For calculating in Python we can import pearson R from module scipy.stats

4.What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Ans:- Scaling comes under parts of the data pre-processing while applying the Machine Learning algorithm. We also know the machine learning algorithm train model according to the data present to them.

So if the value of the features are closer to each other Machine Learning model trained better rather than high difference in the dataset.

So in case there is large difference between the dataset, scalling is the feature via which we convert all dataset to same scale to learn the machine faster and increase the accuracy.

There are two types of scalling

1. **Min Max scaler(Normalization):-** In this method we fit the data in scale range from 0 to 1. So that data points can come closer to each other.

   In this method min value of the data is 0 and max value of the data is 1.

   We can represent the normalization as follows

$$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Where x is any value from the feature x and min(X) is the minimum value from the feature and max(x) is the maximum value of the feature.

From sklearn library we can scalled the dataframe like below

```
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler() scaler.fit(df)
 scaled_features = scaler.transform(df)
print(scaled_features)
```

Standardization:- Like normalization, standardization is also required in some forms of machine learning when the input data points are scaled in different scales. Standardization can be a common scale for these data points.

Basic concept of the standardization is means is zero and standard deviation becomes 1.

Mathematically we can represent it as follow

$$x_{\text{stand}} = \frac{x - \text{mean}(x)}{\text{standard deviation } (x)}$$

From sklearn we can convert data (df) to standard scale as below.

```
from sklearn.preprocessing import StandardScaler
sc_X = StandardScaler()
sc_X = sc_X.fit_transform(df)
print(sc_X)
```

Standard scaler is bounded by the scale as we have seen in the min max scaler, also it can be used if the data is following the Gaussian distribution.

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

Ans:- VIF represent the multicollinearity between the features.
When the value of the vif is infinity means that the variable are perfectly co related.
In this case r2=1 means 1-1/r2 gives the infinite value.
To rectify with this kind of the issue we have to drop one variable.

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

Ans:-

In statistics, a Q–Q plot (quantile-quantile plot) is a probability plot, a graphical method for comparing two probability distributions by plotting their quantiles against each other

QQ plot can also be used to determine whether or not two distributions are similar or not. If they are quite similar you can expect the QQ plot to be more linear. The linearity assumption can best be tested with scatter plots. Secondly, the linear regression analysis requires all variables to be multivariate normal. This assumption can best be checked with a histogram or a Q-Q-Plot.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

Few advantages:
a) It can be used with sample sizes also
b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

As we build your machine learning model, ensure we check the distribution of the error terms or prediction error using a Q-Q plot. If there is a significant deviation from the mean, we might want to check the distribution of your feature variable and consider transforming them into a normal shape.