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 :- Have done analysis on categorical columns using the boxplot and bar plot. Below are the few points we can infer from the visualization –

a. Fall season seems to have attracted more booking. And, in each season the booking count has increased drastically from 2018 to 2019

b. Most of the bookings has been done during the month of may, june, july, aug, sep and oct. Trend increased starting of the year till mid of the year and then it started decreasing as we approached the end of year.

c. Clear weather attracted more booking which seems obvious

d. 2019 attracted a greater number of booking from the previous year, which shows good progress in terms of business.

e. Thu, Fir, Sat and Sun have a greater number of bookings as compared to the start of the week.

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

Ans :- drop_first = True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

Let's say we have 3 types of values in Categorical column, and we want to create dummy variable for that column. If one variable is not A and B, then It is obvious C. So, we do not need 3rd variable to identify the C.

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

Ans:- “temp” variable has the highest correlation with the target variable.

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

Ans :- Have validated the assumption of Linear Regression Model based on below 5 assumptions –

- ✓ **Normality of error terms -** Error terms should be normally distributed
- ✓ **Multicollinearity check** - There should be insignificant multicollinearity among variables.
- ✓ **Linear relationship validation** - Linearity should be visible among variables
- ✓ **Homoscedasticity -** There should be no visible pattern in residual values.
- ✓ **Independence of residuals** - No auto-correlation

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

Ans:- Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes –

i. temp

ii. winter

iii. sep

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans:-

Linear regression is a type of supervised learning algorithm. It is a predictive modelling technique which gives us the relationship between a target (dependent) variable and one or more predictors (independent variables). This technique is used when the target variable is continuous in nature e.g exam scores of students, temperate across days etc

There are two types of linear regressions -

1. Simple linear regression &

2. Multiple linear regression

- Simple linear regression is applied when there is only one predictor. Multiple linear regression is used when there are more than one predictors.
- The model works based on finding the best fit line that represents the dependent variable. Best fit line is found by minimizing the residual error. Residual error is the difference between actual value and predicted value.

For a simple linear regression, predicted values are calculated using the formulae -

Ypred = B0 + B1X

For a multiple linear regression, predicted values are calculated using the formulae -

Ypred = B0 + B1X1+ B2X2….BnXn

Ypred represents the predicted value of the target variable B0 is the intercept. It is value of Y when X is zero

Bn represents the slope or coefficient. It represents the degree of impact the independent variable has on the target variable. A positive value indicates positive correlation whilst a negative value represents a negative coefficient

A linear regression model works out the B0 & Bn in such a way that residual errors are minimised. A gradient descent algorithm is used to identify this.

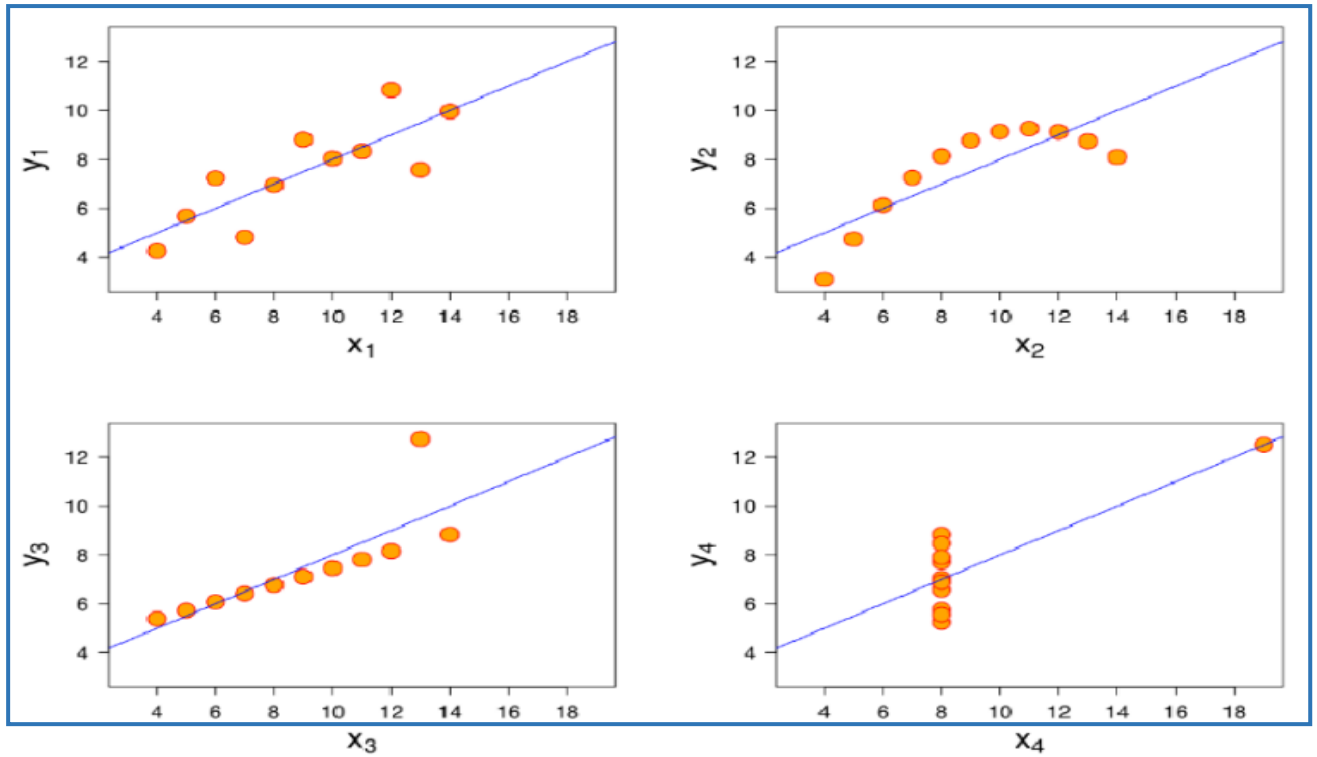
R2 is used as a measure of percentage of variance explained by the model. A higher value of R2 indicates a better fit. R2 = 1- RSS/TSS where RSS is residual sum of squares and TSS is total sum of squares

2. Explain the Anscombe’s quartet in detail.

Ans:-

Anscombe’s Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties. There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.

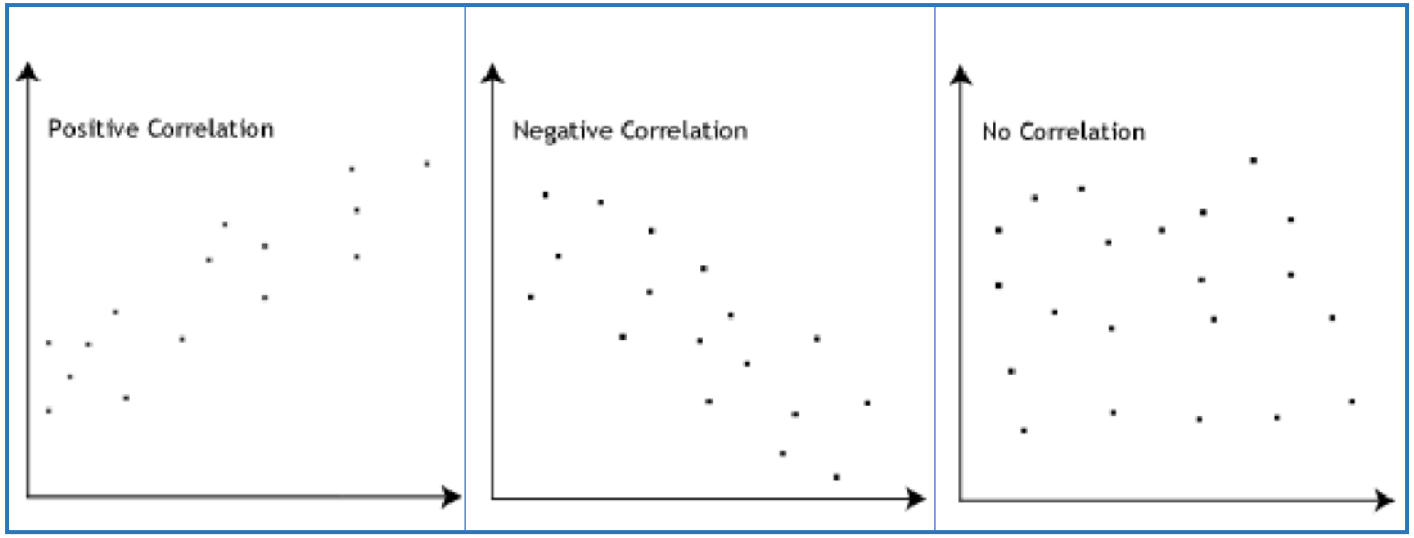


3. What is Pearson's R?

Ans:-

Pearson's R is a correlation coefficient that quantifies the linear relationship between two variables. It takes values from -1 to +1. Variables are strongly correlated when the coefficient values are closer to -1 and +1. A lower value (closer to 0) indicates a weaker correlation.

Two variables have a positive correlation when coefficient lies between 0 & 1. On the other hand, they are negatively correlated when coefficient lies between -1 & 0. A positive correlation indicates that the value of a variable increases with an increase in value of the other variable. A negative correlation indicates that the value of a variable decreases with an increase in value of the other variable.



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

Ans:-

Scaling is a way of making values of all features within the same range.

In a multiple regression model with multiple independent features (predictors), each variable/feature can have different scales. This will result in models having very high/low coefficients which can be difficult to interpret. For e.g categorical variables will take values of 0 or 1. However, continuous variables can take values across the spectrum.

- Scaling is performed for two main reasons -

1. Ease of interpretation &

2. Faster convergence for gradient descent methods

Scaling only affects the coefficients and does not impact the other parameters. There are two popular methods of scaling

Standardization: Features are scaled in such a way that mean of the feature is zero and standard deviation is 1

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

Normalization: feature is scaled in such a way that all values lie between zero and one.This is referred to as mix/max scaling as it uses the minimum and maximum values of the feature

Xs = (Xi- min(X)) / (max(X) - min(X))

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

Ans:-

In a multiple linear regression model, VIF is a measure that helps identify multicollinearity and deal with it. Multicollinearity occurs when one or more independent features in a linear regression are correlated. VIF is measured using the following formula - VIF = 1/ (1-R^2). Where R2 is the square of residual value.

When R^2 is equal to 1, then the denominator becomes zero. As a result, VIF becomes infinite. A single or a combination of independent features can completely explain this independent feature being evaluated. Since these features can explain the independent feature, it is recommended to drop the feature with infinite VIF value.

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

Ans:-

Q-Q plots stands for Quantile-Quantile plots. It is used to identify the type of probability distribution (normal, uniform or exponential) for a given data. Quantiles of the data are plotted against theoretical quantiles of above-mentioned distribution types to identify best fit.

zero. When residual errors are not normally distributed, the model is not acceptable. A Q-Q plot can be used to identify whether residual errors are normally distributed. Residual errors are sorted from small to large and plotted against z-score of points dividing a normally distributed curve with ‘N+1’ segments where ‘N’ is the number of data points of the residual errors. Residual errors are normally distributed when all data points align closely to the 45-degree line.