Name : Shubhashree P

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? (3 marks)

a)From Boxplot observation weekday may not be the dominant feature for the analysis of dependent variable –median value almost same for most of the weekdays.

b)There is less demand for 2018 more demand in 2019

c)Season is having partial impact - indicates that more bikes are rent during fall season.

d)The no holiday day(working day) and holiday box plots indicate that more bikes are rent during no holiday(normal working days) than on weekends or holidays.

e)Partial impact on month-June ,July, Aug, Sept ,Oct months have good demand and Sept month have high demand

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

It will help in reducing the extra column created during (categorical) dummy variable creation, It reduces the correlation created among the dummy variable.

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

Column Temp and atemp have highest correlation each other.

*they have positive correlation with demand(cnt)

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

- By building linear Regression model by considering all features ,we try to estimate mean square error-(one and above)
- R2 values are reduced from each iteration (by observing high VIF and low P value)
- Linearity, mean of residual Analysis etc.
- Validated by Scatter plot – Actual v/s Predicted

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

The 3 features are

1) year -2019-(Positive Correlation)
2) temp (positive correlation)

3) weathersit(negative Correlation)

General Subjective Questions

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

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables.

Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

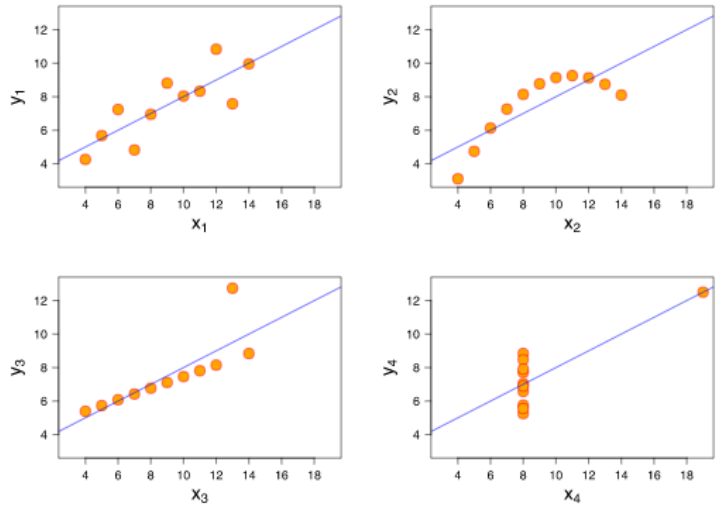
Mathematically, we can write a linear regression equation as:
$y=a+bx$ where a = intercept and b = slope

Here, x and y are two variables on the regression line.
b = Slope of the line
a = y-intercept of the line
x = Independent variable from dataset
y = Dependent variable from dataset
In Linear Regression calculate Root Mean Square Error(RMSE) to predict the next weight value.

- Dependent variable should be numeric and the response variable is continuous to value, It is based on the least square estimation.

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

Anscombe to illustrate the importance of plotting data before you analyze it and build your model. 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.



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

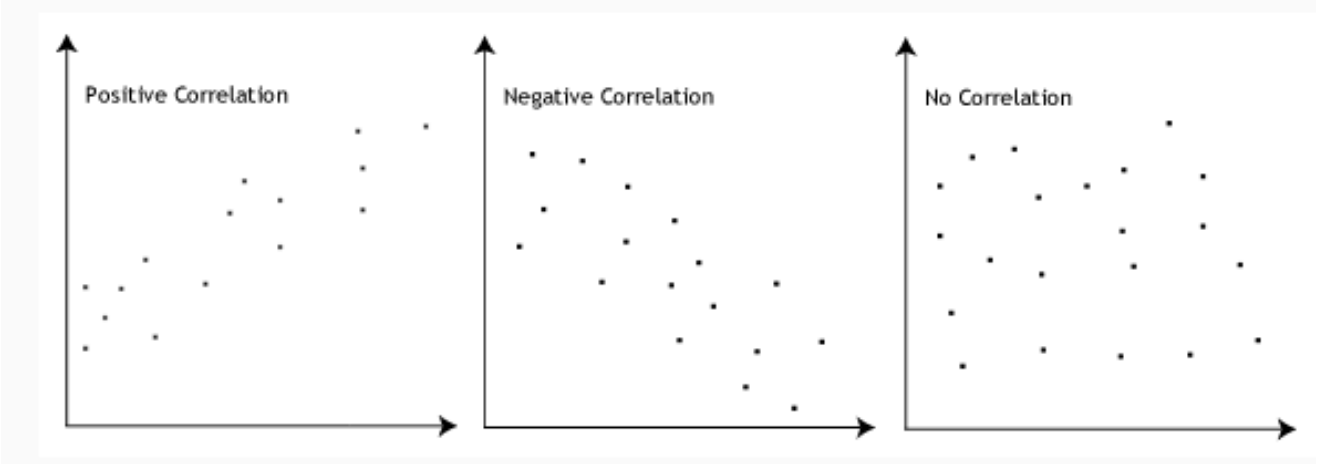
In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1.

The Pearson's correlation coefficient varies between -1 and +1 where:

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



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

Normalization typically means rescales the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1

S.NO.	Normalisation	Standardisation
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.

S.NO.	Normalisation	Standardisation
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called Standard Scaler for standardization.
6.	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
7.	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8.	It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

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

If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables

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

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is

plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

A Q Q plot showing the 45 degree reference line: