Assignment-based Subjective Questions

Question 1: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

We have identified few categorical variables like 'season', 'weatherise' in this dataset and created dummy variables to represent each variable to build the model.

The categorical variables 'season', 'weathersit' are directly & indirectly influencing the user to utilise the bike sharing platform. We could infer the following observations.

- In season like 'winter' and 'summer' has positive coefficient and directly proportional to the number of users utilising the bike sharing platform.

- When the weather situation like mist & cloudy, light snow, thunderstorm and scattered cloudy affects the users to utilise bike sharing platform less i.e. not as much.

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

- When creating dummy variables, to represent 'n' levels, we need 'm - 1' dummy variables to represent all the levels.

- This helps to minimum number of variables and build simple model to explain the interpretability of the model.

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

From the pair plot, the variables 'temp' or 'feeltemp' has linear relationship with the target variable 'count'. When the temperature increases, there is linear increase in numbers of users using bike sharing platform.

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

We validate the assumption of linear regression as follows

1. The residuals (errors) are independent each other (there is no co-relation between the errors of series of time).
2. There is no multi-collinearity, the independent variables is not correlated by checking the VIF is within acceptable range (< 5).
3. The residuals have constant value at every level of x, that is Homoscedasticity.
4. The error terms has to follow normal distribution.

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

From the final model, we got the variables ‘temperature’, ‘year’ and ‘winter’ season contributes significantly to explain the demands of the shared bikes.

General Subjective Questions

Question 1: Explain the linear regression algorithm in detail.

Linear regression is which explains the relationship between dependent variable and independent variable using straight line (while plotting).

In the below graph, it explains the linear relation between marketing budget and sales, we plotted the straight line using data points.

The standard equation of regression line is as follows

y = B0 + B1 * X1 + B2 * X2 + B3 * X3 ….. + BnXn. (B0 -> intercept, Bn -> slope)

The best straight fit line is plotted by minimising the residual sum of squares (RSS) which is equal to sum of squares of residual for each data point in the plot.

RSS is calculated using ordinary least square method i.e.

RSS = E (Yi - B0 - B1Xi - B2Xi ……. - BnXi) where i : 1 to n

We can asses the strength of linear regression using following metrics

* R^2 (coeff of determination)
* Residual standard error (RSE)

R^2 (coefficient of determination) is a number, which explains the what portion of data is varied in the given developed model and it’s calculated here.

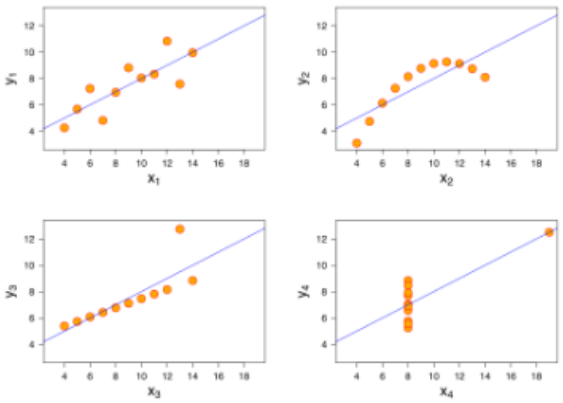
R^2 = 1 - (RSS / TSS). RSS : Residual sum of square, TSS: Sum of errors from mean

Question 2: Explain the Anscombe's quartet in detail?

The Anscombe's quartet explains the linear relationship about data,

* The data has any outlier
* It models only in linear relationship
* Few assumption requires to make the inference.

This chart explains the detail clear.



* In first plot, the data is distributed random & has linear relationship. In second plot, though data is linear relationship but it cannot handle any type of data.
* Third and fourth plot has linear relationship with data but it's sensitive to outliers.

Question 3: What is Pearson's R ?

Pearson's R is correlation coefficient used in linear regression. It measure the strength of correlation between two variables of data.

Pearson R can range from -1 to 1 i.e., when R = -1, perfect negative linear relation between two variables, 0 means no linear relationship, 1 indicates positive relationship between two variables.

When compared to p-value, p-value measures how likely you would observe a correlation of this strength under the null hypothesis.

Question 4: What is scaling? Why is scaling performed? What is the difference between normalised scaling and standardised scaling?

Scaling is method used to normalise the values of independent variables in the data. When we build the model, the independent variables whose values are in different scale. Due to this, we may get the weird coefficient value, it becomes is difficult to interpret the value.

So we do Normalised scaling and Standardised scaling.

* Normalised scaling is performed in a way the values are lie between 0 and 1 using maximum and minimum values in data.

    x = x - min(x) / max(x) - min(x)

* Standardised scaling is performed in a way the mean is 0 and standard deviation is 1.

x = x - mean(x) / std(x)

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

When VIF is infinite, it means there is perfect correlation between two independent variables i.e. R^2 = 1. The R^2 is used to calculate in VIF = 1/ (1 - R^2) and becomes infinite.

To solve this issue, we need to drop a variable which is causing this perfect multicollinearity.

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

Q-Q plot is known as Quantile-Quantile plot, we plot the quantiles of sample distribution against normal distribution. From this, we can determine the dataset follows any type of distribution like normal, exponential or uniform.

ML models works best under some distribution assumptions. Knowing which distribution we are working with can helps us to select the best model.

The power of Q-Q plot lies in ability to summarise any distribution visually. It's useful to find

* Two population are from same distribution
* Residuals follow normal distribution and have a normal error term assumption in regression
* Skewness of distribution.