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)

From the boxplots plotted in the notebok

1. Summer and Fall seasons show very high demand for rental bikes. Mostly because in these seasons heat is high and bike rides make the wind blow over you which is pleasant.
2. There is a very significant increase in the demand from 2018 to 2019. Mostly because of more awareness of this option.
3. Monthly demand shows increase from Jan to Sep with Sep being the highest. As holidays come by, the demand decreases.
4. Median demand is less on holidays. Probably because these bikes are mostly used for commute.
5. Weekday and Working day don't show much trend.
6. Clear weather has highest demand, for obvious reasons. Rainy day has least demand.

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

This is done because a categorical variable with N values can be fully represented as a numerical variable with N - 1 variables. When all N - 1 variables are 0, they represent Nth value. This is done to reduce multi-colinearity in the model.

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

‘Temp’ or ‘atemp’ has the highest correlation since the scatter rises at an almost 60 degree angle.

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

The following things were checked:

1. Normality of error terms - Check Residual analysis section, the errors terms are normally distributed with mean at 0.
2. Non Multi-colinearity between variables - We have removed variables with high multicolinearity inorder to have independent variables drive the target variable.
3. Homoscedasticity - There is no visible pattern in residuals of the model
4. Residual independent of variables - There is no driving of residuals by any variables showing any pattern in variance.
5. Linearity - We considered this model only when we so linear correlation between some variables with target.

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

Based on the coefficients:

1. Temperature - Positively affecting
2. Year - Positively affecting
3. Rainy Weather - Negatively affecting

General Subjective Questions

1. Explain the linear regression algorithm in detail.

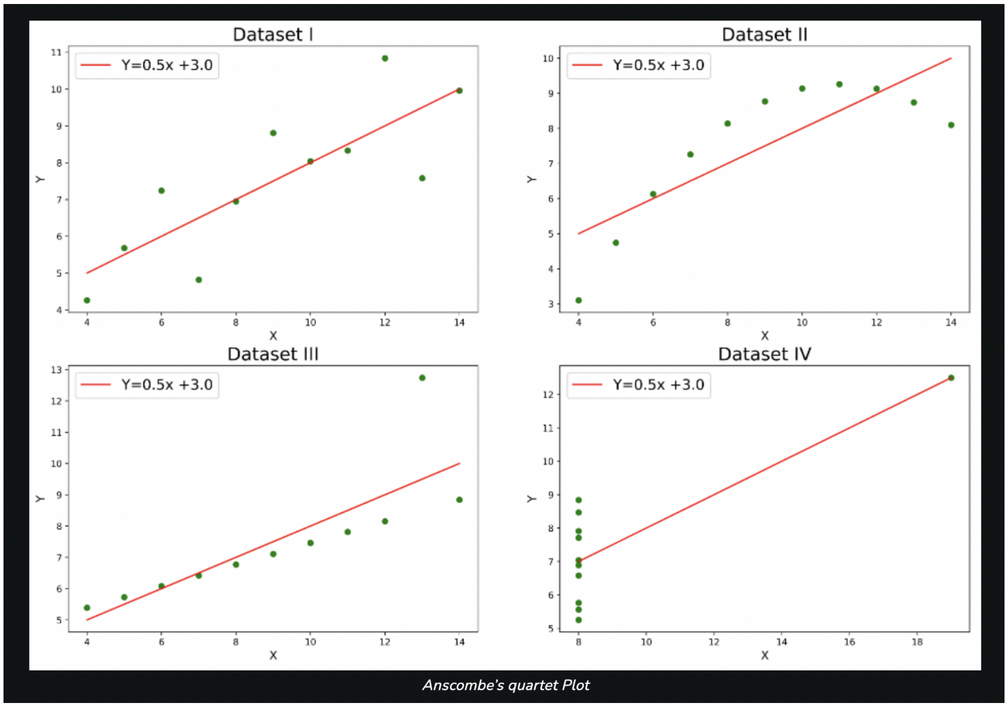
Linear Regression is a Supervised ML method to predict a target or dependent variable from a given set of independent variables by predicting a best fitting linear equation of Y = m1x1 + m2x2 + … + c.

There can be Simple (with one independent variable) or Multi LR models with many independent variables all contributing positively or negatively in a linear fashion to the target variable.

The best fit line is calculated by a cost function.

2. Explain the Anscombe's quartet in detail.

These consist of four datasets with identical summaries like variance, R-sq, Linear Regression line, means etc.
But these datasets have different representations when plotted using scatterplot.



Anscombe's quartet Plot

All these plots have same regression line and similar statistics, but visually they are quite different.
This is used to show the importance of EDA and not just rely on stats.

Explanation of this output:

1. In the first one(top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
2. In the second one(top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.
3. In the third one(bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
4. The fourth one(bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

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?

It is a statistical unit that measures the direction and strength of linear correlation between two variables and is represented by ‘r’. The value ranges from -1 to 1.

This is how it is interpreted:

1. r = 0: There is no linear association.
2. r > 0 < 5: There is a weak association.
3. r > 5 < 8: There is a moderate association.
4. r > 8: There is a strong association.

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

Scaling is either shifting the whole data such that the mean is 0, Or compressing the data to have the same representation but between 0 and 1.

It is performed so that multiple variables are with different range of values affect the model in the same way. If some values are far higher than others, the model will tend to ignore the smaller values and we will have incorrect model.

Nomalized Scaling - compresses all the values between 0 and 1 by calculating min/max. (X - Xmin)/(Xmax - Xmin)
Standardized Scaling - It uses mean and SD to shift the values such that mean is always 0 and SD always between -1 to 1. (X - Xmean)/Xsd

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

This can happen when there is a perfect linear correlation between two independent variables. In such a case Rsq becomes 1 and so VIF = 1/(1 - Rsq) = 1/0 = Infinite.

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

It is called a quantile-quantile plot, and is used to graphically compare two probability distributions and assess if they come from the same population.

It is created by plotting the quantiles of one distribution against the other. The quantiles of a distribution are the values below which a certain percentage of the data points fall. For example, the 25th percentile is the value below which 25% of the data points fall.

If the two distributions are identical, then the points on the Q-Q plot will fall on a straight line. If the two distributions are different, then the points on the Q-Q plot will deviate from the line. The amount of deviation can be used to measure the difference between the two distributions.

In LR they are used to assess if the residuals are normally distributed or not. It can also be used to plot distributions of test and train datasets and determine if the test data is coming from the same population as the train dataset.