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)

Answer

* The box plots for categorical variables show that many categorical variables affect the demand.
* The season, month of year are major contributors to change in demand for bikes. Holidays also affect demand
* The categorical variables are also inter related. The seasons and months are correlated (with a factor of ) and hence only one may be considered.
* In case of weekdays, no major change is observed in demand.

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

Answer

* The dummy variable helps in encoding the various categories of a categorical variable.
* The dependant variable is then accounted for based on these categories which become independent variables (columns) after encoding.
* The line drop\_first = true ensures that the dependence between categories is reduced and one lesser variable is introduced after encoding

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

with the target variable? (1 mark)

* The ‘registered’ variable has the highest correlation with the demand

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

training set? (3 marks)

* The assumptions are validated by using the following
* -Coefficients
* -p value
* -VIF

5. Based on the final model, which are the top 3 features contributing significantly towards

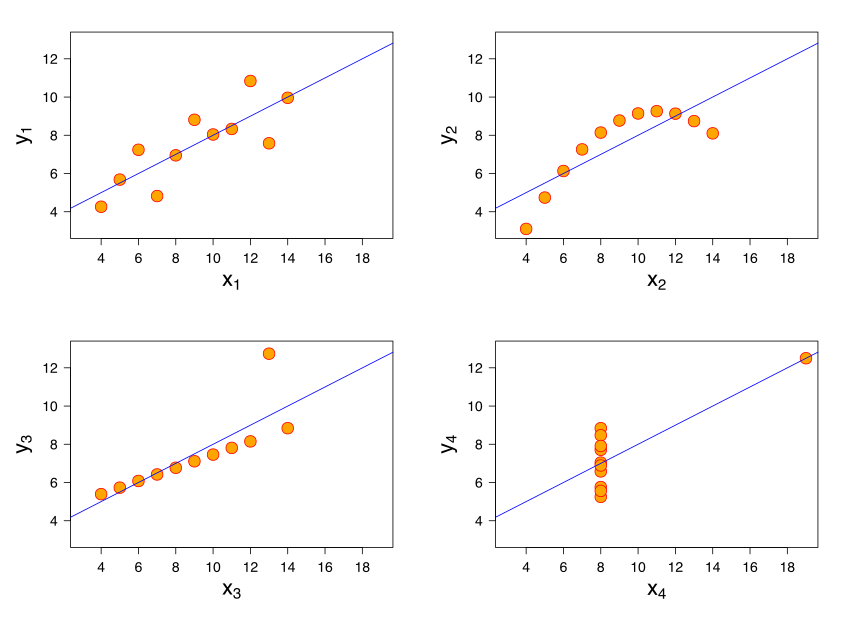
explaining the demand of the shared bikes?

* Whether a user is registered or not . Casual users are also important.
* Weather situation

General Subjective Questions

1. Explain the linear regression algorithm in detail.

2. Explain the Anscombe’s quartet in detail.

* Anscombe’s quartet basically highlights the importance of graphical methods to assess the suitability of fitted curves.
* Sometimes a curve fit is statistically described well for multiple datasets. However, visual appreciation elucidates the differences in data wrt linearity, outliers and coefficients.
* The Ascombe’s quartet is a set of four datasets for which one linear approximation fits well with identical statistical descriptives. 
* Thus it is very important that before fitting and finalising models, the data set should be visually appreciated to see if it is generally suitable for linear regression.

3. What is Pearson’s R?

* Pearson R is the measure of linear correlation between two sets of data.
* The pandas function of corr uses the Person method for determination of the correlation coefficient.
* It is defined as the ratio between the covariance of two variables and the product of their standard deviations.
* It has a value between -1 and 1 and indicates the strength and direction of linear relation between two datasets. It does not indicate any info on any other type of correlation.
* It is a good estimate to study multicollinearity between two variables during Linear Regression Modelling of a dataset.

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

and standardized scaling? (3 marks)

* Different variables may belong to different ranges/ magnitude relatively( x1 betn 20 and 30 ; x2 betn 2000 and 3000) .
* When a linear regression model is fitted on the data, the coefficients that are developed depend on the range and magnitudes of the variable.
* Thus a variable with a higher magnitude will have a small coefficient than that with a smaller magnitude. This could distort the appreciation of the coefficients that are evaluated by the model.
* Hence it is important o have scaling so that all coefficients of the model are comparable.
* Min Max Scaling/ Normalisation . This method uses the range of the variable to scale it. It results in the variable to be reduced to a value that is 0 - 1. This method normalises the outliers too.
* Standardisation - This method uses the mean of the data and it’s std deviation for scaling. This results in the mean of the data being centres around zero and all other data points shifting around it. Outliers remain unaffected in terms of their distance form the dataset in this method of data preparation.

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

(3 marks)

* the Variance Inflation Factor is a measure of correlation between different variables.
* In case the VIF is infinite it means that the Rsquared value for the variable is 1. This it means there is perfect correlation between the two variables.
* This happens when mukticolllinear variables are chosen for modelling and carrion in one variable directly maps to variation in the other.

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

(3 marks)

* A Q-Q plot is a graphical method, like a scatter plot, to assess the correlation between two probability distribution functions.
* In this the quantiles of the pdf are plotted against each other.
* The alignment of the data points is indicative of its normal distribution. A linear scatter of points indicates that the data is normally distributed.