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?

Answer >> Observation on Categorical variable is following:

a. Summer and Fall are the best season for Bike Service

b. Months from May to Oct are good for Bike Service

c. People like to use Bike more on Weekdays.

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

Answer >> Using "drop_first=True" ensures that optimal number of dummy variables are created which is equal to "n-1" where "n" is number of unique elements in categorical feature. Its helps to reduce the number of columns and thus helps in performance of model.

Example: Say a categorical column has 3 unique values A, B, C, these can be converted individually to three different columns A, B, C or can be converted to 2 columns say B,C and having zero in B,C indicates presence of A.

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

Answer >> "atemp" feature has highest co-relation with target variable "cnt"

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

Answer >> By predicting on the train set and then comparing the values with actual values to get difference between errors and then plotting so see if error distribution is centred around zero or not.

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

Answer >> Top three features are following:

1. Yr (Year)
2. Holiday
3. temp

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer >> Linear regression is a machine learning algorithm which falls into supervised category (labels are present in data), in which we try to model and find best fit linear line between independent and dependent variables.

Linear regression tries to fit line of the following form:

$Y = b0 + b1.X1 + b2.X2 \ldots\ldots bn.Xn$

Where X1,X2…Xn are independent variables while Y is the dependent variable which will be predicted by the final model.

In simplistic term (simple linear regression) with one feature, equation is of standard line equation:

Y = mX + C

Where “m” is slope and “C” is the intercept at Y-axis.

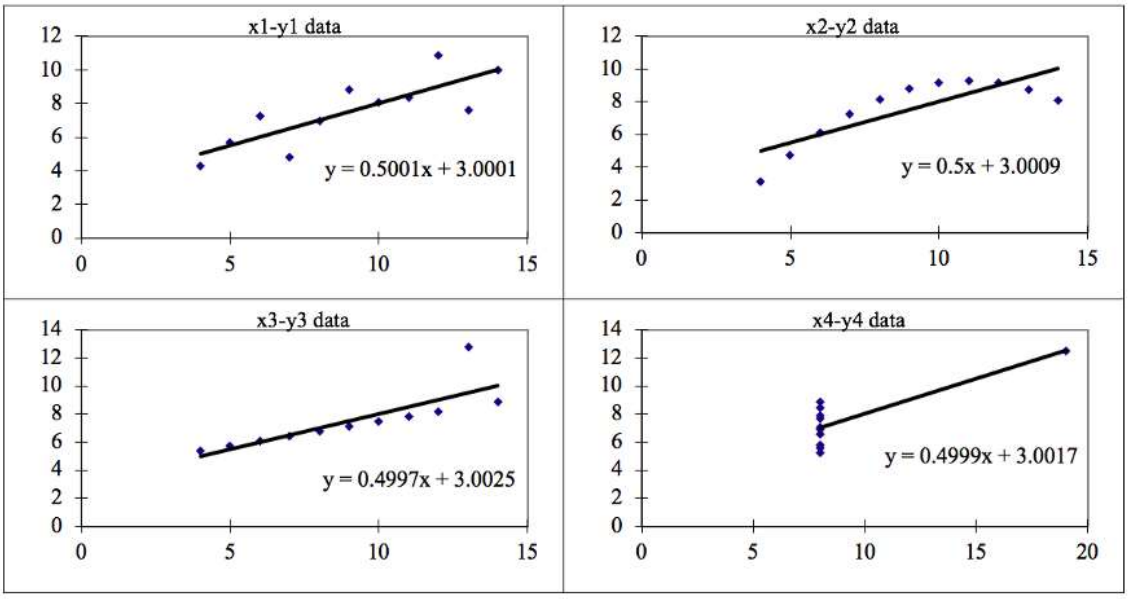
Linear regression is based on certain assumptions, which are as following:

a. Linear Relationship between X and Y
b. Homoscedasticity: Error (Residual) is constant when plotted against value of Y
c. Independence: Observations are independent of each other.
d. multi-collinearity is not present, due to presence of multi-collinearity model will not be able to tell due to which data (Corelated one) is responsible for change in dependent variable.
e. Residual errors are normally distributed, which can shown using plotting.

Linear regression guarantees interpolation of data and not extrapolation of data, meaning: prediction works well only when independent variables are in similar range on which model was trained.

2. Explain the Anscombe’s quartet in detail.

Answer >> Anscombe’s quartet tells us about the importance of visualization before starting with model preparation. It consists of set of four data set which are nearly identical in simple descriptive statistics (like number of samples, Mean, Standard deviation) but modelling of linear regression on this data set fails as distribution of data is very different.



3. What is Pearson’s R?

Answer >> Pearson’s R or Pearson’s Coefficient is used to find how strong a relationship is between two variables, this is very used in statistics and used commonly while linear model generation. This relationship is helpful to find how one variable will behave on any change on other variable. Pearson’s R can have values between -1 and 1 where:

- 1 indicates strong positive relationship
- 0 indicates no relationship
- -1 indicates strong negative relationship

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

Answer >> Scaling is a technique to fit data (independent features) into fixed range. Scaling is useful when dealing more features/columns to generate model like in multiple linear regression. In such cases, it's not necessary that all columns will have data in similar range for example if a data set has Temperature, Humidity, Windspeed as features then these will have data at very different scale, in such cases scaling is used to bring all data into Fixed range. Normalized and standardized are two scaling techniques which are used to perform scaling.

In case of standardized scaling, scaled data is centred around zero while in Normalized scaling, scaled data lies between 0-1.

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

Answer >> If we have perfect corelation in that case VIF will be infinite and R-Square will be 1, this also suggest that this variable can be expressed by combination of other variables. In such case, it better to drop that column from model generation.

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

Answer >> Q-Q plot or Quantile-Quantile plot is a graphical tool which plots two quantile against each other and helps to find out if two sets of data come from the same distribution (Normal, Exponential). A quantile is a fraction where certain values fall below that quantile.

If given distributions are similar or linearly related, then points in the plot will approximately lie on the line, for similar distribution line would be y = x.

In Linear Regression, Q-Q plot is used in Residual Analysis, to check if Errors maps to Normal distribution or not.