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?

Ans) I have done analysis on categorical columns using the boxplot. Below are the few points we can infer about their effect on the dependent variable:

- The maximum number of bike bookings are made during the fall season.
- The number of bookings has increased in 2019.
- Most of the bookings has been done during the month of May, June, July, Aug, Sep and Oct.
- The number of bookings increases during the holidays.
- Clear weather attracted more booking which seems obvious.
- The number of bookings seems to be equal in both working days and non-working days.
- Thu, Fir, Sat and Sun have more number of bookings.

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

Ans) It is important to use drop_first=True, because it drops the first column and reduces the extra column which was created during the dummy variable creation. Hence it reduces the correlations created among dummy variables.

Let's take an example, we have 3 types of values in categorical column and we want to create dummy variable for that column. If one variable is not A and B, then It is obvious C.

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

Ans) The 'temp' variable has the highest correlation with the target variable.

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

Ans) The followings are the assumptions of Linear Regression, I validate after building the model on the training set,

- There should be a linear relation ship between x and y.
- There is no correlation between the independent variables.
- The error terms should be normally distributed.
- The error terms should have constant variance.

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

Ans) The top 3 features contributing significantly towards explaining the demand of the shared bikes are,

- temp
- winter
- Sep

General Subjective Questions

1) Explain the linear regression algorithm in detail.

Ans) Linear regression which explains the relationship between a dependent variable and one independent variable using a straight line. The straight line is plotted on the scatter plot of these two points.

The standard equation of the regression line is given by the following expression,

$$Y = mX+c$$

Where, Y = Dependent variable

X = Independent variable

m = Coefficient or Slope

c = Intercept

The linear relationship can be two types,

- Positive linear relationship
- Negative linear relationship

The linear regression is two types,

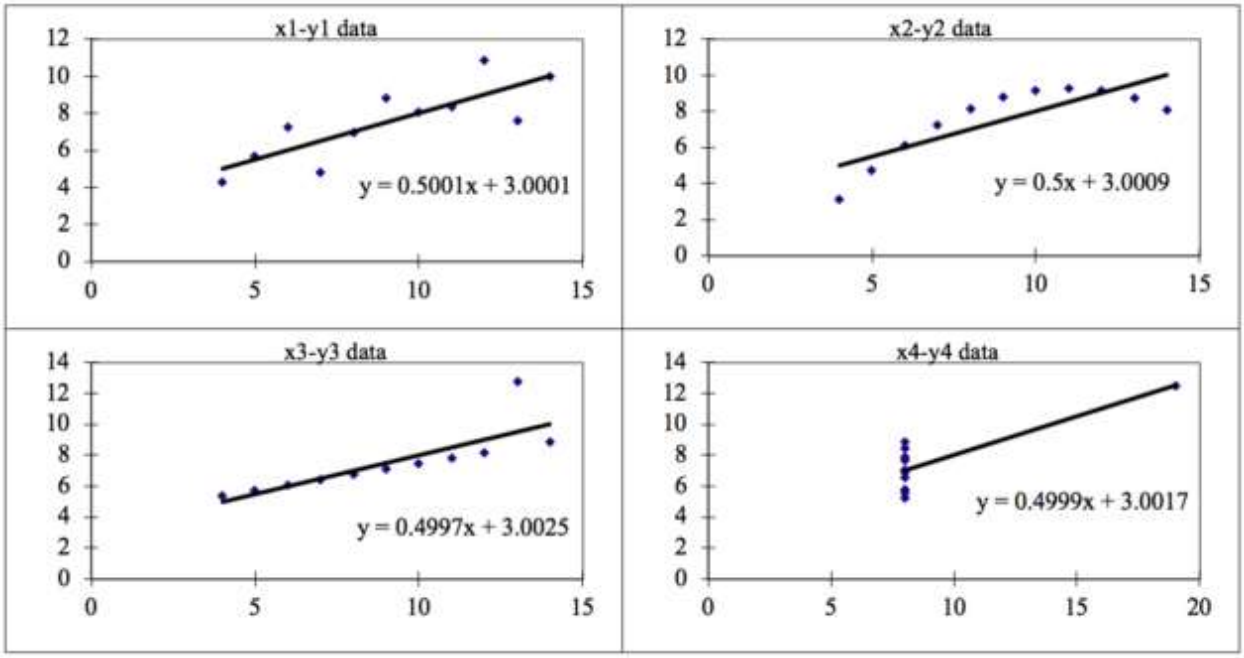
- Simple linear regression
- Multiple linear regression

The following are some assumptions about dataset that is made by Linear Regression model:

- There should be linear relation ship between dependent and independent variable.
- There is no correlation between independent variables.
- The error terms should be normally distributed.
- The error terms should have constant variance.

2) Explain the Anscombe's quartet in detail.

Ans) It was developed by statistician Francis Anscombe. It is use to show the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties. There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x, y points in all four datasets.



The four datasets can be described as:

1. **Dataset 1:** this fits the linear regression model pretty well.

2. **Dataset 2:** this could not fit linear regression model on the data quite well as the data is non-linear.

3. **Dataset 3:** shows the outliers involved in the dataset which cannot be handled by linear regression model

4. **Dataset 4:** shows the outliers involved in the dataset which cannot be handled by linear regression model

3) What is Pearson's R?

Ans) Pearson's R is use to define the strength of the linear relationship between the variables. If the variables go up and down together, the correlation coefficient will be positive. If the variables go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative. The Pearson correlation coefficient can take a range of values from +1 to -1. A value of 0 indicates that there is no relationship between the two variables. A value greater than 0 indicates a positive relationship; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative relationship; that is, as the value of one variable increases, the value of the other variable decreases.

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

Ans) Scaling is a process in which we scale the independent variables in to a fixed range.

Scaling is performed for ease of interpretation and faster convergence for gradient descent methods. Scaling just affects the coefficients and none of the other parameters, such as t-statistic, F-statistic, p-values and R-squared.

The difference between normalized scaling and standardized scaling are,

<u>normalized scaling:</u>

- Minimum and maximum value of features are used for scaling.
- Scales values between (0, 1).
- SKLearn provides a transformer called MinMaxScaler() for Normalization.
- It is used when features are of different scales.
- It is really affected by outliers.

<u>Standardized scaling:</u>

- Mean and standard deviation is used for scaling.
- It is not bounded to a certain range. The mean value lies at the center, so some values are come in negative range.
- SKLearn provides a transformer called StandardScaler() for standardization.
- It is used when we want to ensure zero mean and unit standard deviation.
- It is much less affected by outliers.

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

Ans) If there is perfect correlation between two independent variables, then the value of VIF is infinity. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity. In the case of perfect correlation, we get R-squared (R2) value is equal to 1, which lead to 1/ (1-R2) infinity. To solve this, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

Ans) The Q-Q plot is a graphical technique for determining if two data sets come from populations with a common distribution. A Q-Q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 30% quantile is the point at which 30% of the data fall below and 70% fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions. When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.