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)

1. Based on the weather conditions the booking count decrease
2. During the month of Aug, sept and Oct the demand is high
3. During summer and fall season the demand is high when compared with winter.
4. Demand increase from first to fourth month
5. On 2019 demand is high when compared to 2018
6. During working day the demand is high

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

Drop_first=True helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

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

with the target variable? (1 mark)

Highest correlation is casual and registered bike count with the total bike count.

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

training set? (3 marks)

Validate the assumptions of linear regression,

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

explaining the demand of the shared bikes? (2 marks)

1. During Holidays the booking count will decrease
2. As windspeed increase the bike sharing count will decrease
3. Less booking when the weather is bad.

General Subjective Questions:

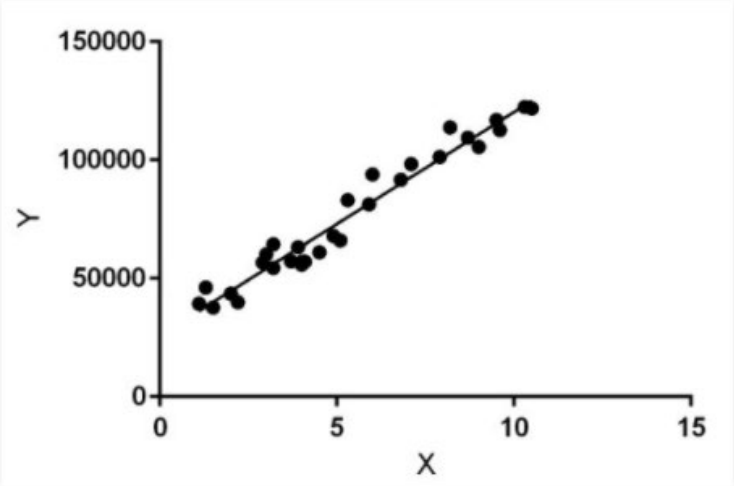
1. Explain the linear regression algorithm in detail.

Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour of your data based on some variables.

In linear regression as name suggests its a linear relationship between two variables which can be plotted on x-axis and y-axis.

Variable on X-axis is called independant variable and variable on Y-axis is called target. They should be lineraly correlated where X is increasing Y is also increasing or vice versa, which is either linear upward or downward relationship. In layman terms, as part regression analysis we try to calculate the best fit line which describes the relationship between the independant and dependent [target] variable.

This is represented with formula ==> Y = c + mX, where c is the intercept and m is the slope.



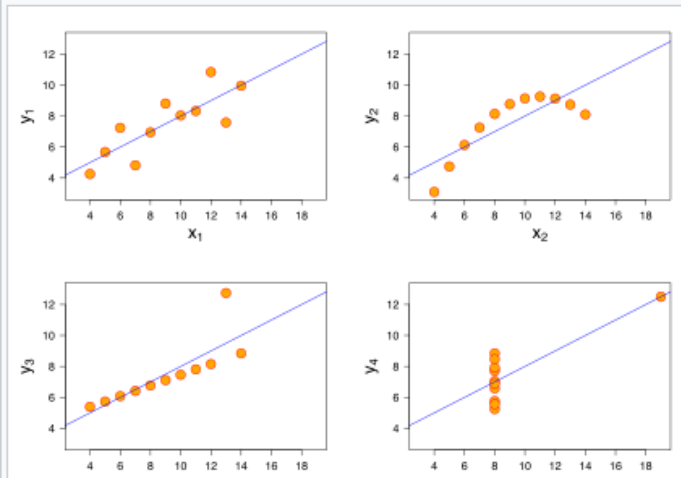
Some of the assumption in linear regression are :

1. A linear relationship between the independent and dependent variables .
2. The error residuals should follow a normal distribution like
3. The error is constant along the values of the dependent variable.
4. Strong correlations between each independent variable and the dependent variable, but no correlation between independent variables.
5. All independent variables are uncorrelated with the error term

2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. The basic thing to analyze about these data-sets is that they all share the same descriptive statistics(mean, variance, standard deviation etc) but different graphical representation.

Each graph plot shows the different behavior irrespective of statistical analysis.



3. What is Pearson's R? (3 marks)

Pearson's correlation (also called Pearson's R) is a correlation coefficient commonly used in linear regression. It is used to measure how strong a relationship is between two variables.

Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.

1. There are certain requirements for Pearson's Correlation Coefficient:
2. Scale of measurement should be interval or ratio
3. Variables should be approximately normally distributed
4. The association should be linear
5. There should be no outliers in the data

Limitation of Pearson's R is it cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.

Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

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

and standardized scaling? (3 marks)

It is a data Pre-Processing step which is applied to independent variables to standardize the data within a particular range. It speeds up the calculation in the model.

The collected data set contains features which may be in different units and range. If scaling is not done then algorithm only takes precedence of range over units and lead to incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level.

It is important to note that scaling just affects the coefficients and none of the other parameters

Differences are captured below:

Normalization	Standardization
Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.	Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation
Formula: $X' = \frac{X - X_{min}}{X_{max} - X_{min}}$	Formula : $X' = \frac{X - \mu}{\sigma}$
There is a bounding range	There is no bounding range

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

(3 marks)

Multicollinearity occurs when two or more independent variables are highly correlated with one another in a regression model.It can be a problem in a regression model because we would not be able to distinguish between the individual effects of the independent variables on the dependent variable.VIF determines the strength of the correlation between the independent variables. If there is perfect correlation between the variables, then VIF = infinity. During regression analysis we consider features VIF less than 5 for model training.

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

(3 marks)

The Q-Q [quantile-quantile plot] is a graphical method for determining whether two samples of data came from the same population or not. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value.

In linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.