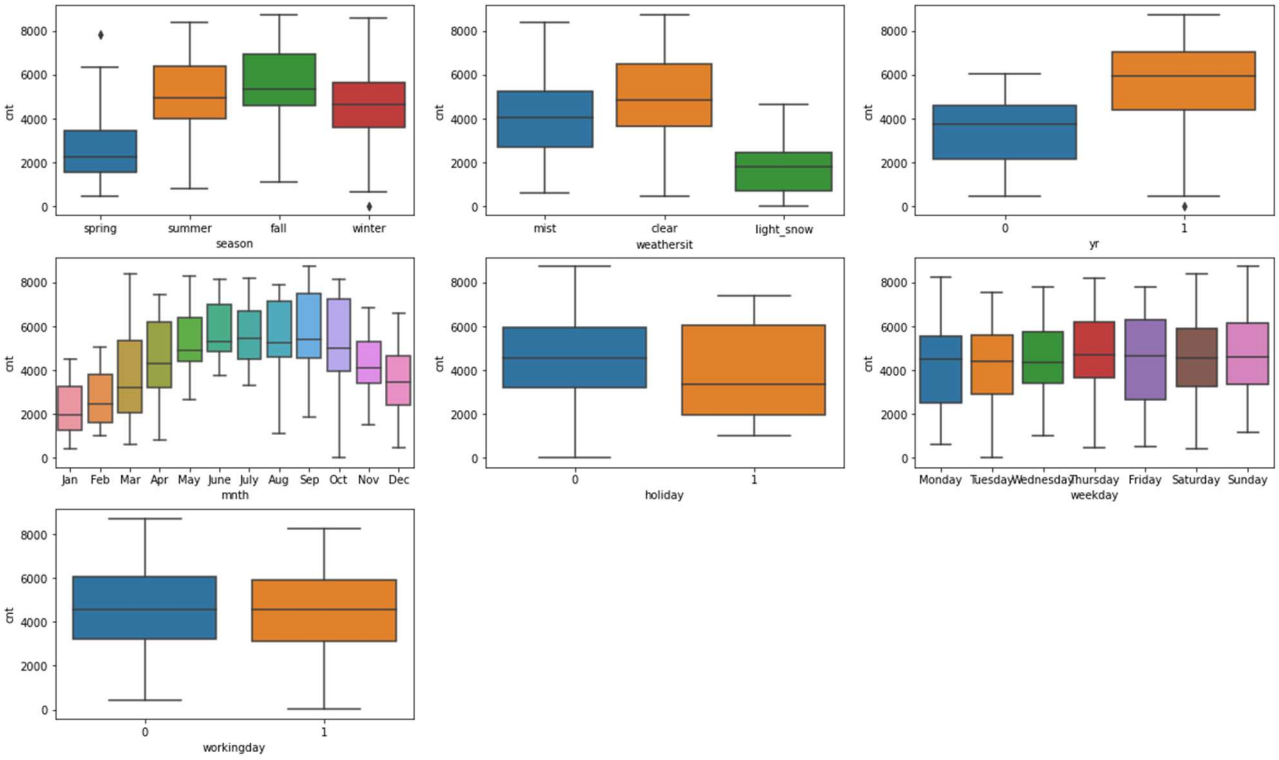
Assignment-based Subjective Questions

1.From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Ans: In the Boom-bike assignment dataset, the categorical variables are Season , Weather situation , months, weekday, workingday, holiday, year



1. In fall season the demand for shared bikes are more and in spring season shows least demand for shared bikes.
2. The spring season is from December to March, the demand for shared bike is very less in this season.
3. September month has highest demand for shared bikes.
4. January month has least demand for shared bikes.
5. The demand for shared bikes are more when the weather is clear and there is least demand when there is snow fall.
6. In holiday the demand for shared bike is reduces.
7. December and January is the beginning of Spring and still winter effect will be there, snow fall will be there. Therefore the demand for shared bikes is less. From Feb onwards again it pickups.
8. The demand for shared bikes more depends on weather, if the weather is clear then the demand for shared bikes is more and if there is a snow fall the demand for shared bikes reduces.
9. Demand for shared bikes is increased in 2019 compare to 2018.That means in coming year there is more demand for shared bikes when the situation gets normal and quarantines and lockdowns are over.
10. Working day and weekdays are not effecting more on the target variable.
11. July month is little special because there is a sudden little dip in demand for shared bikes even the weather is pleasant.

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

Ans: drop_first=True is important to use during dummy variable creation, because it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

Example : Season has 4 categories: Fall, Spring, Summer,and winter

If we create a dummy variable without drop_first = True.

Fall =1000

Spring =0100

Summer = 0010

Winter =0001

Instead of 4 dummy variables we can use only 3,

Spring =100

Summer =010

Winter=001

Fall =000 (Assumed that if all the above categories are 0 then it is the first category)

If we use drop_first =True during dummy variable creation it will drop the first category by considerings its value as 000.

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

Ans: The independent variable temperature (temp) feature is highly correlated with the target variable demand for shared bike(cnt).

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

Ans : The assumptions of linear regression are :

1. Linearity : There is a linear relationship between independent and dependent variable. Plot a Scatter plot for X_train_index vs residuals.

If there is a linearity in the plots then the assumption holds good. The equally spread residuals around a horizontal line without distinct patterns are a good indication of having the linear relationships.

2. Assumption about the residuals:
   a. Normality : Plot distribution plot for residuals (sns.distplot((y_train - y_train_pred), bins = 20))
      If the distribution of residuals is normally distributed then the assumption of Normality holds good.
   b. Zero mean : In the distribution plot if the error terms are normally distributed with mean equal to 0, the assumption holds good.
   c. Constant variance: Check this assumption by examining the scatterplot of "residuals versus X_train_index"; the variance of the residuals should be the same across all values of the *x*-axis. If the plot shows a pattern then variances are not consistent, and this assumption has not been met.
   d. Independent error : By examining the scatterplot of "residuals versus X_train_index"; There should not be any patterns.The pairwise correlation is zero.
3. Assumptions about the estimator:
   a. Independent variables are linearly independent of each other(no multicollinearity) Check the VIF value of each predictors. If the VIF < 4 for all the predictors, then there is no issue of multicollinearity.
   b. Independent variables are measured without errors: If the p-value of each predictors is equal to zero or less than 0.05 then we can say the variable is significant.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Ans: The top 3 features contributing significantly towards explaining the demand of the shared bikes are temperature(temp), year (yr), Spring season (season_spring)

1. The demand for shared bikes are more when the weather is clear, temp is high and there is least demand when there is snow fall or temp is low.
2. There is a high demand for the shared bikes in comming years. After corona lockdown, when the conditions become normal the 'Boom-Bikes' will get very good profit in shared bike business.
3. The demand for shared bike is reducing in spring season.The Boom bike company can give some **spring season offers** to increase the demand in this season. Expand the business in Spring season.

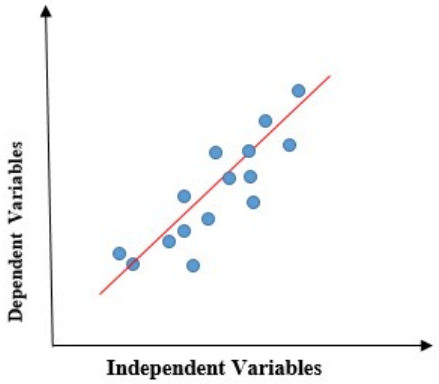
General Subjective Questions

1.Explain the linear regression algorithm in detail. (4 marks)

Ans: Linear Regression is a Supervised learning Algorithm. It is method of finding the best straight line fitting to the given data. i.e. finding the best linear relationship between the independent and dependent variable.

Linear Regression model used to predict the unseen dependent variable by using the independent variables. The Linear Regression Algorithm uses Least Sum of Residuals Squares to find the best linearly fitted model.

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output).



The above graph presents the linear relationship between the dependent variable and independent variables. When the value of x (**independent variable**) increases, the value of y (**dependent variable**) is likewise increasing.

Best-Fit Line:

The best-fit line is found by minimising the expression of RSS (Residual Sum of Squares) which is equal to the sum of squares of the residual for each data point in the plot. Residuals for any data point is found by subtracting predicted value of dependent variable from actual value of dependent variable ($e_i = y_i - y_{pred}$).

To calculate best-fit line linear regression uses a traditional slope-intercept form.

$$y = a_0 + a_1 x$$

Where , y = dependent variable

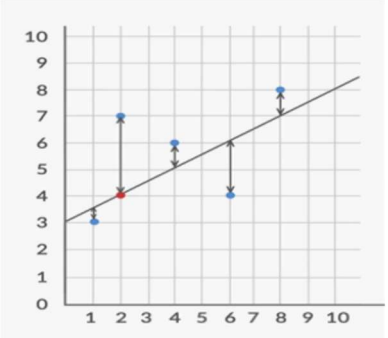
x= independent variable

a_0= intercept

a_1= slope of the line

where, a0 and a1 are the co-efficient of the line. The Linear Regression Algorithm will find these co-efficient by gradient decent method(iterative process) using Sum of Least Square Error to find the best linear fit model.

Residuals:



Residuals : $e_i = y_i - y_{pred}$

Ordinary Least Squares Method : RSS (Residual Sum of Squares)

$$RSS = e_1^2 + e_2^2 + e_3^2 + \cdots . + e_n^2$$

$$RSS = (y_1 - (a_0 + a_1 x_1))^2 + (y_2 - (a_0 + a_1 x_2))^2 \ldots . + (y_n - (a_0 + a_1 x_n))^2$$

$$RSS = \sum_{i=1}^{n} (y_i - (a_0 + a_1 x_i))^2$$

Cost function: Cost function optimizes the regression coefficients or weights and measures how a linear regression model is performing. The cost function is used to find the accuracy of the **mapping function** that maps the input variable to the output variable. This mapping function is also known as **the Hypothesis function**.

In Linear Regression, **Mean Squared Error (MSE)** cost function is used, which is the average of squared error that occurred between the predicted values and actual values.

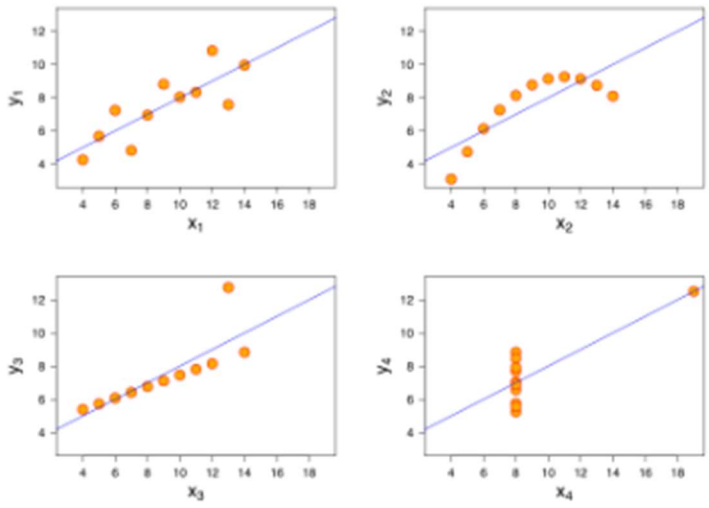
By simple linear equation y=mx+b we can calculate MSE as:

$$MSE = \frac{1}{N} \sum_{i=1}^{n} (y_i - (a_0 + a_1 x_i))^2$$

Using the MSE function, we will change the values of a_0 and a_1 such that the MSE value settles at the minima. Model parameters x_i, (a_0, a_1) can be manipulated to minimize the cost function. These parameters can be determined using the gradient descent method so that the cost function value is minimum.

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

Ans : Anscombe's Quartet can be defined as a group of four data sets which are **nearly identical in simple descriptive statistics**, but there are some peculiarities in the dataset that **fools the regression model** if built. They have very different distributions and **appear differently** when plotted on scatter plots.



In the above plot, the first one seems to be doing a decent job, the second one clearly shows that linear regression can only model linear relationships and is incapable of handling any other kind of data. The thirs and fourth images showcase the linear regression models sensitive to outliers. If outliers are not there, we could have got the great line through the data points. So, we should not run a regression without having a good look at our data.

Anscombe's Quartet illustrate the importance of plotting the graphs, visualizing the data that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. before analysing 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.

Also, the Linear Regression can be only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets.

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

Ans: Pearson's R or Pearson's correlation coefficient is defined in statistics as the measurement of the strength of the relationship between two variables and their association with each other. Pearson's R calculates the effect of change in one variable when the other variable changes. The Pearson's R tries to find out two things , the **strength** and the **direction** of the relationship from the given sample sizes.

Pearson correlation coefficient formula:

$$r = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

Where:
N = the number of pairs of scores
Σxy = the sum of the products of paired scores

Σx = the sum of x scores
Σy = the sum of y scores
Σx2 = the sum of squared x scores
Σy2 = the sum of squared y scores

The Pearson's R returns the values between -1 and 1.
Strength :The stronger the association between the two variables, the Pearson's R value incline towards 1 or -1. Attaining values of 1 or -1 signify that all the data points are plotted on the straight line of 'best fit.' It means that the change in factors of any variable does not weaken the correlation with the other variable. If the Pearson's R lies near 0, the more the variation in the variables.

Direction: The negative and positive sign of the Pearson's R tells the direction of the line.
The direction of the line indicates a positive linear or negative linear relationship between variables. If the line has an upward slope, the variables have a positive relationship. This means an increase in the value of one variable will lead to an increase in the value of the other variable. A negative correlation depicts a downward slope. This means an increase in the amount of one variable leads to a decrease in the value of another variable.

Pearson's R Correlation co-efficient is designed to find the correlation between the variables which shows linear relationship and it might not be a measure for if the relationship between the variables is non-linear.

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

Ans: Scaling is a step of data Pre-Processing of Machine Learning. Scaling is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm. Scaling reduced the iterative steps of Gradient Decent Algorithm to converge towards the best-fit Model.

Most of the times, the collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units it leads to incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

It is important to note that **scaling just affects the coefficients** and none of the other parameters like **t-statistic, F-statistic, p-values, R-squared**, etc. are affected by scaling.

Normalized Scaling	Standardized Scaling
1. It is also called Scaling Normalization. 2. Min-Max value of features are used for scaling. 3. Normalization brings all of the data in the range of 0 and 1. 4. Formaula used : $x = \frac{x-\min(x)}{\max(x)-\min(x)}$ 5. It is really affected by outliers 6. Scales value between (0,1) or (-1,1)	1. It is also called Z-Score Normalization 2. Mean and Standard deviation is used for scaling. 3. Standardization replaces the values by their Z -scores. It brings all of the data into standard normal distribution which has mean=0 and sd(standard deviation) =1 4. Formula used: $x = \frac{x-mean(x)}{sd(x)}$ 5. It is less affected by outliers. 6. Value is not bound to certain range

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

Ans : The value of VIF = infinity, shows that a perfect correlation between two independent variables. In case of perfect correlation, we get R-square =1, which leads to 1/(1-R-square) infinity. An infinity value of VIF indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

Before starts building a multiple linear regression we will do many assumptions, in that "issue of multicollinearity" is very important. We will assume that there is no multicollinearity, that means the selected independent variables are nor correlated with any of the other selected independent variables. But if there is perfect correlation between independent variables, the value of that particular variables VIF becomes infinity.

To solve this problem, 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. (3 marks)

Ans : Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q -Q plots is to find out if two sets of data come from the same distribution.

The use and importance of Q-Q Plot are ,

1. Q-Q plots helps in a scenario of 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.
2. Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

3. A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.

A 45 degree angle is plotted on the Q-Q plot;

- If the two data sets come from a common distribution, the points will fall on that reference line.
- If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis indicates the data sets come from different distribution