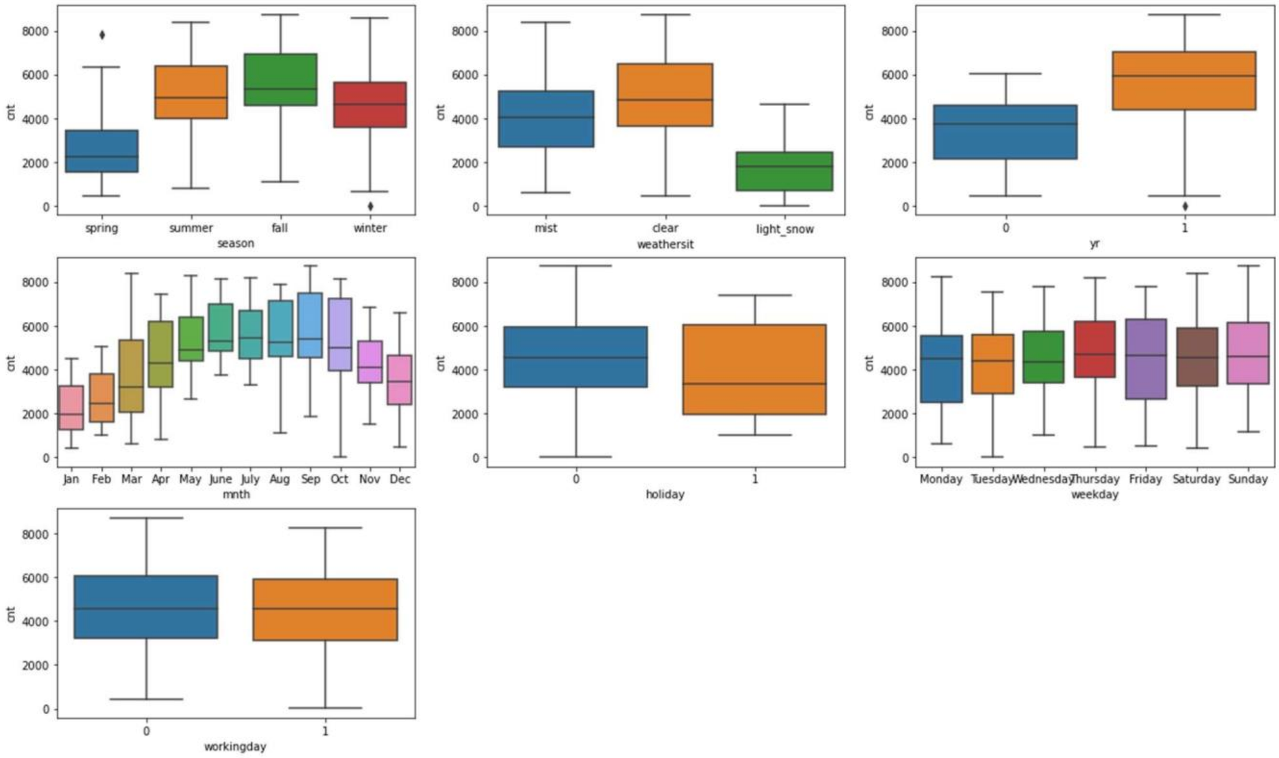
Assignment-based Subjective Questions

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 : In the Boom-bike assignment dataset, the categorical variables include Season, Weather situation, Months, Weekday, Workingday, Holiday, and Year.



1. The demand for shared bikes is highest during the fall season and lowest in the spring.
2. Spring, spanning from December to March, sees minimal demand for shared bikes.
3. September experiences the peak demand for shared bikes.
4. January records the lowest demand for shared bikes.
5. Demand for shared bikes increases when the weather is clear, while it decreases during snowfall.
6. Demand for shared bikes drops during holidays.
7. December and January, marking the start of spring, still experience winter effects and snowfall, leading to reduced bike demand. Demand begins to rise again from February.
8. Shared bike demand is heavily influenced by weather conditions; it increases with clear weather and decreases with snowfall.
9. Demand for shared bikes grew in 2019 compared to 2018, indicating a potential increase in demand as conditions normalize and restrictions ease.
10. Working days and weekdays have minimal impact on bike demand.

11. July is an exception, showing a slight dip in demand despite pleasant weather.

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

Answer: Using drop_first=True when creating dummy variables is crucial because it helps eliminate the redundant column that would otherwise be created. This reduces the correlation among the dummy variables.

For example, if the Season variable has four categories—Fall, Spring, Summer, and Winter—creating dummy variables without drop_first=True would result in:

- Fall = 1000
- Spring = 0100
- Summer = 0010
- Winter = 0001

Instead, by using drop_first=True, we can reduce the number of dummy variables to three:

- Spring = 100
- Summer = 010
- Winter = 001
- Fall = 000 (implied as the reference category if all other variables are 0)

This approach effectively drops the first category (Fall) and uses it as the baseline, avoiding unnecessary multicollinearity.

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

Answer: The independent variable temperature (temp) is strongly correlated with the target variable, demand for shared bikes (cnt).

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

The assumptions of linear regression are:

1. **Linearity**: There should be a linear relationship between the independent and dependent variables. To check this, plot a scatterplot of `X_train_index` versus residuals. The assumption holds if the residuals are evenly spread around a horizontal line without distinct patterns, indicating a linear relationship.

2. **Assumptions about Residuals**:
   - **Normality**: Create a distribution plot of the residuals (`sns.distplot(y_train - y_train_pred, bins=20)`). If the residuals are normally distributed, the assumption of normality is satisfied.
   - **Zero Mean**: The residuals should be normally distributed with a mean of zero. Check this in the distribution plot.
   - **Constant Variance (Homoscedasticity)**: Examine a scatterplot of residuals versus `X_train_index`. The variance of the residuals should be consistent across all x-axis values. If the plot shows a pattern, the assumption of constant variance is violated.
   - **Independence**: Residuals should not exhibit any patterns when plotted against `X_train_index`. There should be no significant pairwise correlation between residuals.
3. **Assumptions about the Estimator**:
   - **Linearly Independent Predictors**: Independent variables should be linearly independent (no multicollinearity). Check the Variance Inflation Factor (VIF) for each predictor. If VIF < 4 for all predictors, multicollinearity is not a concern.
   - **Error-Free Measurement**: Independent variables should be measured accurately. If the p-value for each predictor is zero or less than 0.05, the predictor is considered significant.

Understanding these assumptions is key to ensuring the validity of the linear regression model.

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

Answer: The top three features significantly influencing the demand for shared bikes are temperature (temp), year (yr), and the Spring season (season_spring):

1. Demand for shared bikes increases with higher temperatures and clear weather, while it decreases during snowfall or low temperatures.

2. There is expected to be higher demand for shared bikes in the coming years. As conditions return to normal after the COVID-19 lockdowns, Boom-Bikes could see substantial profits in the shared bike market.

3. Demand for shared bikes tends to decrease during the spring season. To boost demand, Boom-Bikes might consider offering promotions during this period and expanding their business operations for the spring season.

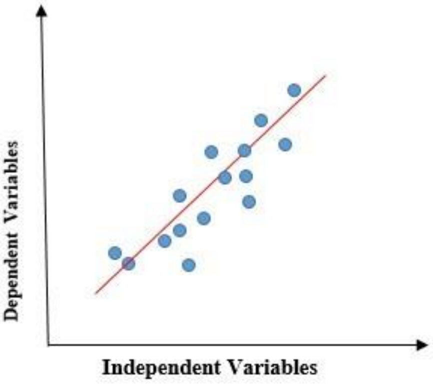
General Subjective Questions

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

Answer: 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:

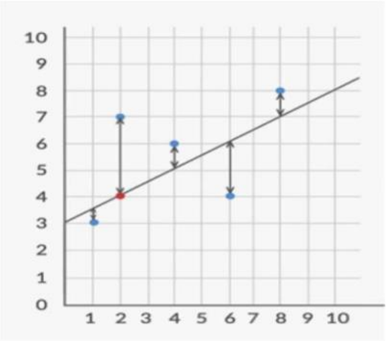
The best-fit line is determined by minimizing the Residual Sum of Squares (RSS), which is the sum of the squared residuals for each data point. The residual for a given data point is calculated as the difference between the actual value of the dependent variable and its predicted value (ei=yi−y^ie_i = y_i - \hat{y}_iei=yi−y^i).

Linear Regression uses the slope-intercept form to compute the best-fit line: y=a0+a1xy = a_0 + a_1 xy=a0+a1x where:

- yyy is the dependent variable
- xxx is the independent variable
- a0a_0a0 is the intercept
- a1a_1a1 is the slope of the line

Here, a0a_0a0 and a1a_1a1 are the coefficients of the line. The Linear Regression algorithm finds these coefficients using the gradient descent method, an iterative process that minimizes the Sum of Squared Errors to achieve the best linear fit.

Residuals:



Residuals are calculated as: ei=yi−y^ie_i = y_i - \hat{y}_iei=yi−y^i

Ordinary Least Squares Method:

The Residual Sum of Squares (RSS) is given by: RSS=e12+e22+e32+⋯+en2\text{RSS} = e_1^2 + e_2^2 + e_3^2 + \cdots + e_n^2RSS=e12+e22+e32+⋯+en2 RSS=(y1−(a0+a1x1))2+(y2−(a0+a1x2))2+⋯+(yn−(a0+a1xn))2\text{RSS} = (y_1 - (a_0 + a_1 x_1))^2 + (y_2 - (a_0 + a_1 x_2))^2 + \cdots + (y_n - (a_0 + a_1 x_n))^2RSS=(y1−(a0+a1x1))2+(y2−(a0+a1x2))2+⋯+(yn−(a0+a1xn))2 RSS=∑i=1n(yi−(a0+a1xi))2\text{RSS} = \sum_{i=1}^{n} (y_i - (a_0 + a_1 x_i))^2RSS=∑i=1n(yi−(a0+a1xi))2

Cost Function:

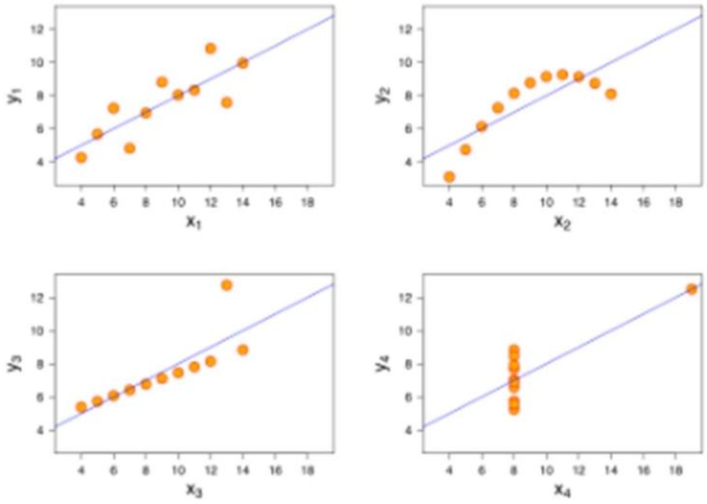
The cost function optimizes the regression coefficients and measures the performance of the linear regression model. It evaluates how well the mapping function (also known as the Hypothesis function) maps the input variable to the output variable.

In Linear Regression, the Mean Squared Error (MSE) cost function is used, which is the average of the squared differences between predicted and actual values: MSE=1N∑i=1n(yi−(a0+a1xi))2\text{MSE} = \frac{1}{N} \sum_{i=1}^{n} (y_i - (a_0 + a_1 x_i))^2MSE=N1∑i=1n(yi−(a0+a1xi))2

The goal is to adjust the values of a0a_0a0 and a1a_1a1 so that the MSE is minimized. These parameters can be optimized using the gradient descent method to achieve the minimum cost function value.

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

Answer : 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 thir 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)

Answer: 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)

Answer: 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. Formula used : x = $\frac{\text{x-min (x)}}{\text{max(x) -min (x)}}$ 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 = $\frac{\text{x-mean(x)}}{\text{sd(x)}}$ 5. It is less affected by outliers. 6. Value is not bound to certain range

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

Answer: 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 help 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 line is plotted on the Q-Q plot:

- If the two datasets originate from the same distribution, the points will align with this reference line.
- If the points of the quantiles deviate significantly from the 45-degree line, it suggests that the datasets come from different distributions.