## 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?

wheathersit, season, weekday, mnth are the categorical variables in the dataset , following are the observations related to categorical variables that directly affect the dependent variable

- '**Summer**' season, a unit increase in the season w.r.t Spring season(reference categorical value) increases the bike hire count by .099 and '**Winter**' season , a unit increase in the season w.r.t Spring season increases the bike hire count by .16, as per EDA '**Fall**' season seems to have attracted more booking. And, in each season the booking count has increased drastically from 2018 to 2019
- another categorical variable impacting the dependent variable is mnth, especially '**September**' month a unit increase in the value w.r.t January month increase the bike hire count by .126
- wheathersit is another categorical data directly impacting the count, a unit increase in the value of '**Cloudy/Mist**' wheathersit w.r.t '**Clear**' wheathersit decreases the bike hire count .089 and a unit increase in the value of '**Snow/Rain**' wheathersit w.r.t '**Clear**' wheathersit decreases the bike hire count .089 units

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

- **Avoiding Multicollinearity**: When creating dummy variables for a categorical feature, each category is represented by a separate binary (0 or 1) variable. If all categories are included as dummy variables, one of the variables can be perfectly predicted from the others (i.e., they are linearly dependent), leading to multicollinearity. By dropping the first dummy variable, this dependency is removed, and the model avoids multicollinearity
- **Reference Category**: Dropping the first dummy variable establishes a baseline or reference category. The coefficients of the remaining dummy variables can be interpreted relative to this reference category. This helps in understanding the impact of different categories compared to the baseline.
- **Simplifying the Model**: By dropping one dummy variable, the model is simplified, reducing the number of variables without losing any significant information. This

makes the model easier to interpret and can sometimes improve computational efficiency

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

temp & atemp showing highest correlation with ‘cnt’ of .63

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

- **Linearity**: The relationship between the independent variables and the dependent variable should be linear, it is checked by plotting scatter chart with residuals and predicted values the it should not show any patterns
- **Independence**: The residuals (errors) should be independent of each other, it is checked by plotting the residuals against the order in the list, the chart was not showing any visible pattern
- **Homoscedasticity**: The residuals should have constant variance at every level of the independent variables, the residual vs predicted value plot showig not sign of a visible pattern
- **Normality of Residuals**: The residuals should be approximately normally distributed, the residual distribution plot was showing normal distribution of residual values
- **Multicollinearity check** : The VIF matrix showing values less than 5

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

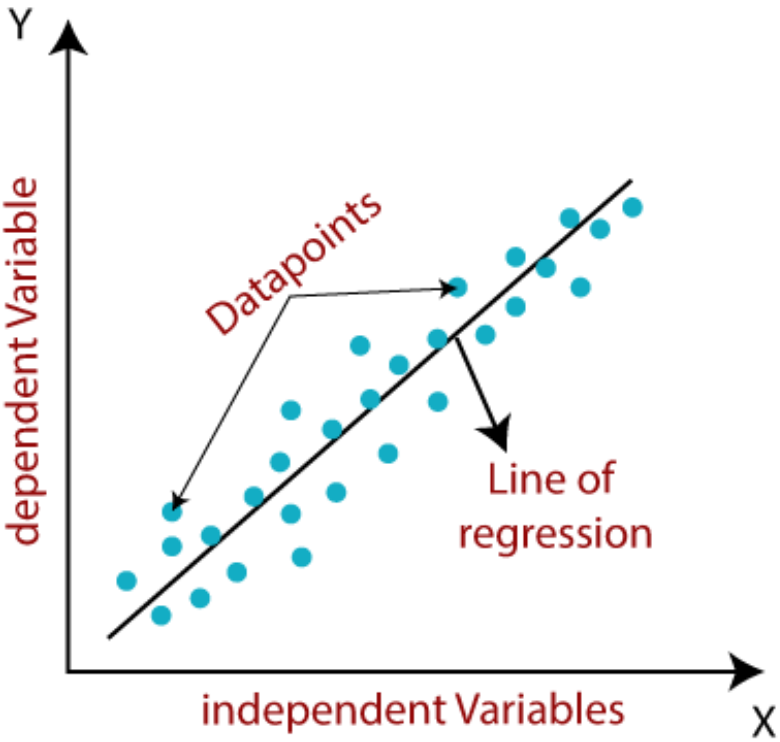
- yr
- temp
- season ( Winter value 4 in dataset)

## General Subjective Questions

### 1. Explain the linear regression algorithm in detail.

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. The simplest form is simple linear regression, which involves a single independent variable. In this model, the relationship is described using a linear equation: y=mx+c, where y is the dependent variable, x is the independent variable, m is the slope of the line (representing the rate of change in y for a unit change in x), and c is the y-intercept (the value of y when x is zero). The goal of linear regression is to find the values of m and c that minimize the sum of the squared differences between the observed values and the values predicted by the linear model, a method known as least squares estimation.

Fig below shows example of simple linear regression .



Multiple linear regression extends this concept to multiple independent variables, allowing for the modeling of more complex relationships. The equation in this case is y=b0+b1x1+b2x2+...+bnxn, where y is the dependent variable, x1,x2,...,xn are the independent variables, and b0,b1,b2,...,bn are the coefficients representing the contributions of each independent variable to the dependent variable. The fitting process involves finding the set of coefficients that minimizes the sum of the squared residuals, which represent the differences between the observed and predicted values. Linear regression assumes a linear relationship, independence of errors, homoscedasticity (constant variance of errors), and normality of error distribution. When these assumptions are met, linear regression provides a powerful and interpretable tool for understanding and predicting relationships between variables

**Best fit line :** When working with linear regression, our main goal is to find the best fit line that means the error between predicted values and actual values should be minimized. The best fit line will have the least error.

**Cost function** : The different values for weights or coefficient of lines ives the different line of regression, and the cost function is used to estimate the values of the coefficient for the best fit line.

Cost function optimizes the regression coefficients or weights. It measures how a linear regression model is performing

Linear Regression, we use the Mean Squared Error (MSE) cost function, which is the average of squared error occurred between the predicted values and actual values

**Residuals**: The distance between the actual value and predicted values is called residual. If the observed points are far from the regression line, then the residual will be high, and so cost function will high. If the scatter points are close to the regression line, then the residual will be small and hence the cost function

**Gradient Descent**

Gradient descent is used to minimize the MSE by calculating the gradient of the cost function.

**2. Explain the Anscombe’s quartet in detail.**

Anscombe's quartet consists of four datasets that have nearly identical simple descriptive statistics, such as mean, variance, correlation, and linear regression line, yet appear very different when graphed. Created by Francis Anscombe in 1973, this quartet demonstrates the importance of graphing data before analyzing it and highlights how relying solely on statistical properties can be misleading. Each dataset in the quartet has the same linear regression line, the same mean of x and y, the same variance, and the same correlation coefficient, yet they tell very different stories when visualized, emphasizing the significance of visual data analysis

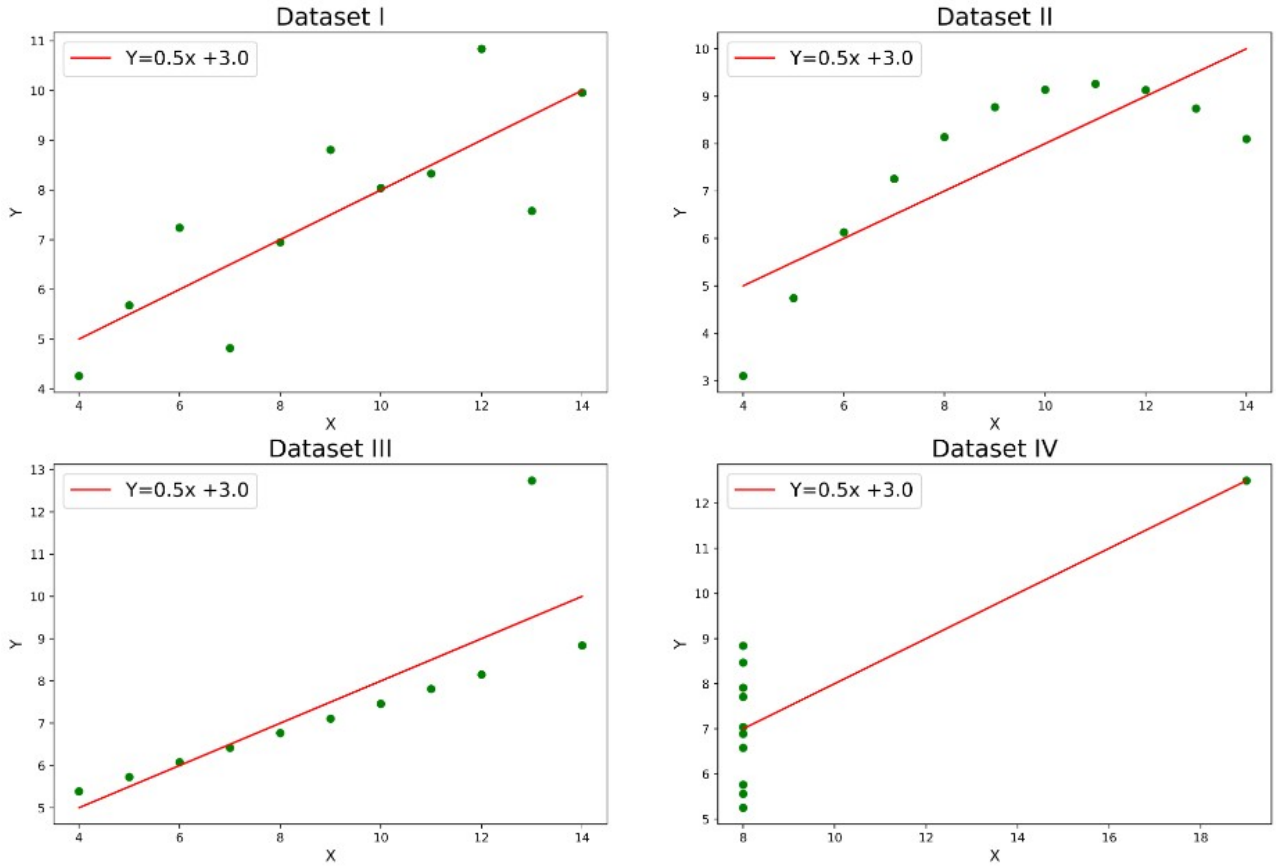
see the following data set, with x & y values in each data set

| I | | II | | III | | IV | |
|---|---|---|---|---|---|---|---|
| x | y | x | y | x | y | x | y |
| 10.0 | 8.04 | 10.0 | 9.14 | 10.0 | 7.46 | 8.0 | 6.58 |
| 8.0 | 6.95 | 8.0 | 8.14 | 8.0 | 6.77 | 8.0 | 5.76 |
| 13.0 | 7.58 | 13.0 | 8.74 | 13.0 | 12.74 | 8.0 | 7.71 |
| 9.0 | 8.81 | 9.0 | 8.77 | 9.0 | 7.11 | 8.0 | 8.84 |
| 11.0 | 8.33 | 11.0 | 9.26 | 11.0 | 7.81 | 8.0 | 8.47 |
| 14.0 | 9.96 | 14.0 | 8.10 | 14.0 | 8.84 | 8.0 | 7.04 |
| 6.0 | 7.24 | 6.0 | 6.13 | 6.0 | 6.08 | 8.0 | 5.25 |
| 4.0 | 4.26 | 4.0 | 3.10 | 4.0 | 5.39 | 19.0 | 12.50 |
| 12.0 | 10.84 | 12.0 | 9.13 | 12.0 | 8.15 | 8.0 | 5.56 |
| 7.0 | 4.82 | 7.0 | 7.26 | 7.0 | 6.42 | 8.0 | 7.91 |
| 5.0 | 5.68 | 5.0 | 4.74 | 5.0 | 5.73 | 8.0 | 6.89 |

the stastistical summary of all data set is almost equal as given below

| | I | II | III | IV |
|---|---|---|---|---|
| Mean_x | 9.000000 | 9.000000 | 9.000000 | 9.000000 |
| Variance_x | 11.000000 | 11.000000 | 11.000000 | 11.000000 |
| Mean_y | 7.500909 | 7.500909 | 7.500000 | 7.500909 |
| Variance_y | 4.127269 | 4.127629 | 4.122620 | 4.123249 |
| Correlation | 0.816421 | 0.816237 | 0.816287 | 0.816521 |
| Linear Regression slope | 0.500091 | 0.500000 | 0.499727 | 0.499909 |
| Linear Regression intercept | 3.000091 | 3.000909 | 3.002455 | 3.001727 |

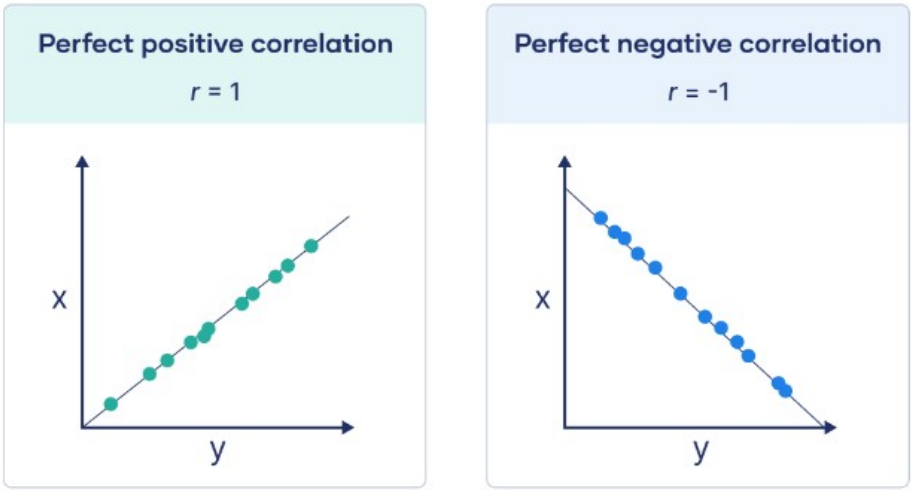
but when analyse the data by plotting it in graph show different behaviour, as shown below




**3. What is Pearson's R?**

Pearson's R, also known as the Pearson correlation coefficient, is a measure of the linear relationship between two variables. It quantifies the degree to which the variables move together, with values ranging from -1 to 1. A Pearson's R of 1 indicates a perfect positive linear correlation, meaning that as one variable increases, the other also increases proportionally. A value of -1 indicates a perfect negative linear correlation, meaning that as one variable increases, the other decreases proportionally. A value of 0 indicates no linear correlation between the variables. Pearson's R is widely used in statistics to understand and quantify the strength and direction of the relationship between two continuous variables

*figure below shows the the different correlation between X & Y data set*




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

Scaling is a data preprocessing technique used to adjust the range and distribution of features in a dataset. It is performed to ensure that different features contribute equally to the analysis and to improve the performance of machine learning algorithms, which can be sensitive to the scale of input data. Scaling helps algorithms converge faster and achieve better accuracy

Normalized scaling, or min-max scaling, transforms data to fit within a specific range, typically [0, 1] or [-1, 1]

Standardized scaling, or Z-score normalization, transforms data to have a mean of 0 and a standard deviation of 1

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

The value of Variance Inflation Factor (VIF) can become infinite when there is perfect multicollinearity among the predictors in a regression model. This means that one predictor variable is an exact linear combination of one or more other predictor variables.

In mathematical terms, the VIF for a given predictor Xi is calculated as:

VIF(Xi)=1/ (1−Ri^2)

where Ri^2 is the coefficient of determination of the regression of Xi on all the other predictors. When Xi is perfectly collinear with the other predictors, Ri^2 =1. Substituting this into the VIF formula results in a denominator of zero, making the VIF value infinite

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

A Q-Q (quantile-quantile) plot is a graphical tool used to assess if a dataset follows a particular theoretical distribution, typically the normal distribution. It plots the quantiles of the data against the quantiles of the chosen theoretical distribution. If the data points lie approximately along a straight line, it indicates that the data distribution matches the theoretical distribution

In the context of linear regression, a Q-Q plot is primarily used to check the normality of residuals . The assumptions of linear regression include that the residuals are normally distributed, which is critical for making valid inferences about the model coefficients and predictions.

*Figure below shows a q-q plot*

Here, as the data points approximately follow a straight line in the Q-Q plot, it suggests that the dataset is consistent with the assumed theoretical distribution, which in this case we assumed to be the normal distribution