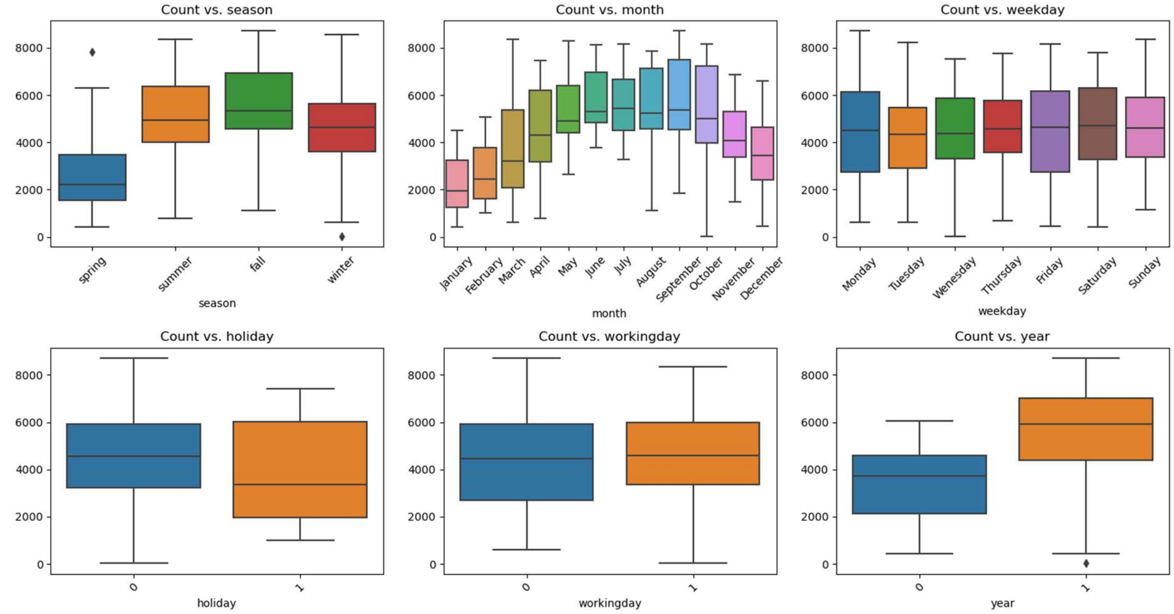
Assignment-based Subjective Questions

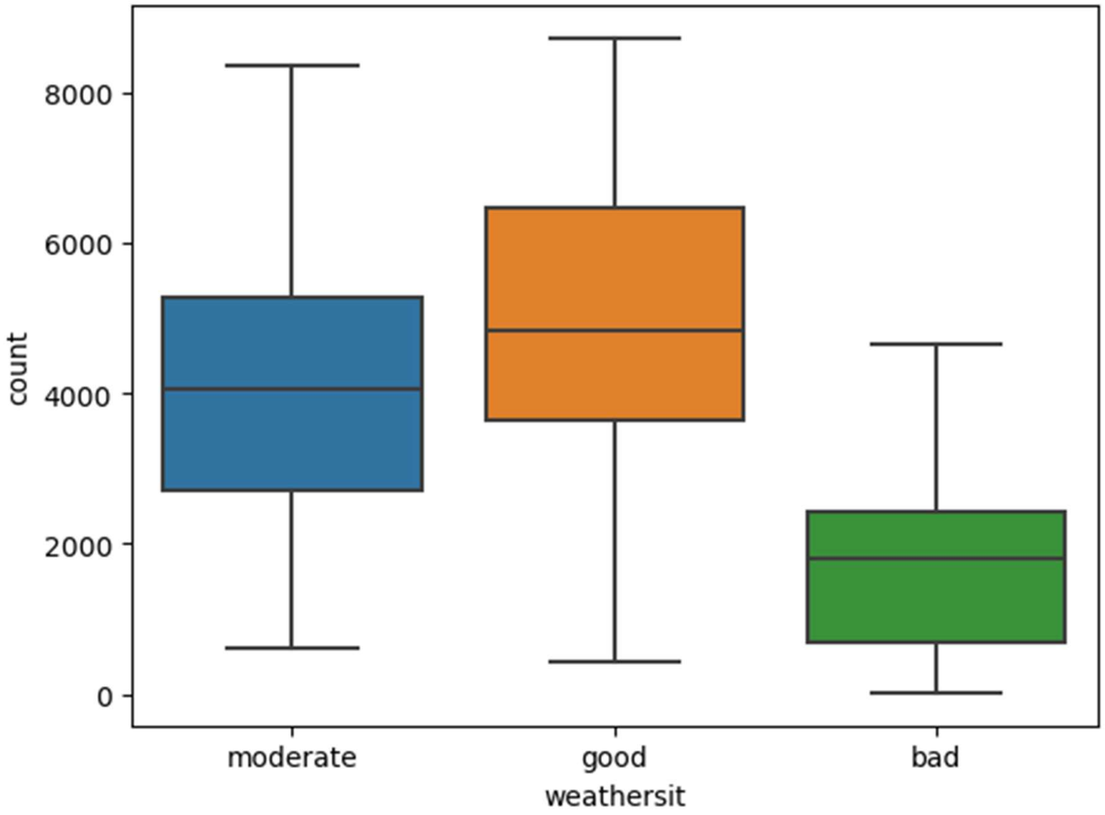
Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: Following are the analysis-

1. There is significant increase in bike demand from 2018 to 2019.
2. There is increase in bike demand in fall season and there is dip in spring season
3. Bike booking is high when weather is good and low when weather is bad.
4. Bike booking increase from May to October, this could be because of good weather condition and favorable season.
5. Booking is equally distributed between working and non-working day
6. On non-holidays, the booking count tends to be lower, which is reasonable as people may prefer spending time at home with family during holidays
7. Bookings were more prevalent on Friday, Saturday, and Sunday compared to the early days of the week.





Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: Dummy variable creation is a technique used in statistical modelling and machine learning to represent categorical variables It involves creating new binary (dummy) variables for each category of the original categorical variable. These dummy variables serve as indicators for the presence or absence of a specific category.

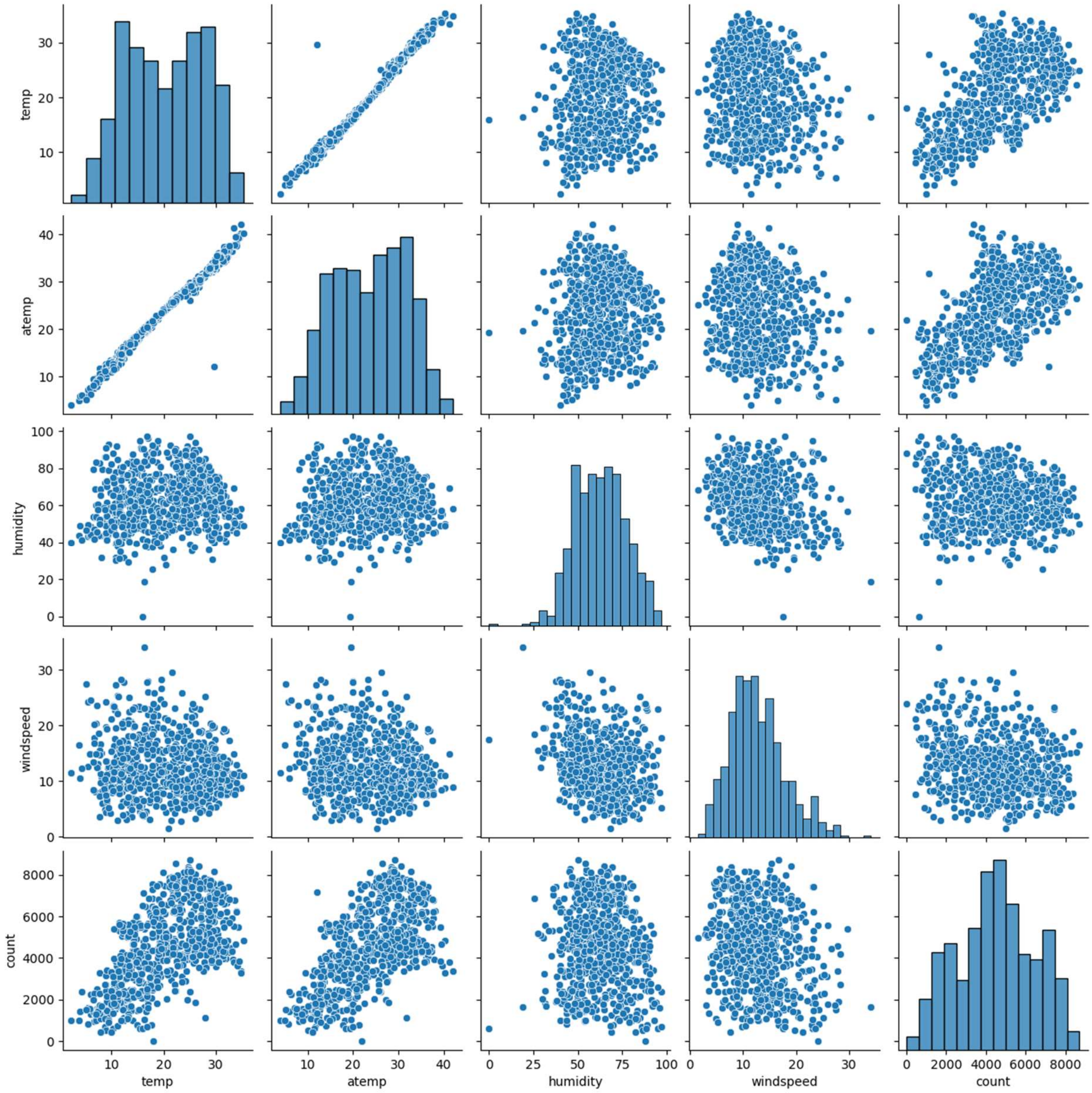
1. **Perfect Multicollinearity** - If we create dummy variable for every value of categorical variable, it will result in perfect multicollinearity. This means one of the dummy variables can be perfectly predicted from the others, leading in redundancy in the model. Dropping first column (or any1 column) will resolve this issue.
2. **Avoiding Redundancy-**The information about the omitted category is implicitly captured by the constant term in the model. Including all dummy variables would introduce redundancy.
3. **Enhancing Interpretability-** The coefficients of the dummy variables represent the change in the response variable compared to the omitted category. This makes the interpretation of the model more straight forward.

For example, if you have a variable "Color" with categories "Red," "Blue," and "Green," you would create two dummy variables, say "Is_Blue" and "Is_Green." The absence of both dummy variables implies that the color is "Red."

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

Total Marks: 1 mark (Do not edit)

Answer: The variable 'temp' and ‘atemp’ exhibits the strongest correlation with the target variable, as depicted in the graph below. Given that 'atemp' and 'temp' are redundant variables, only one of them is selected during the determination of the best fit line.

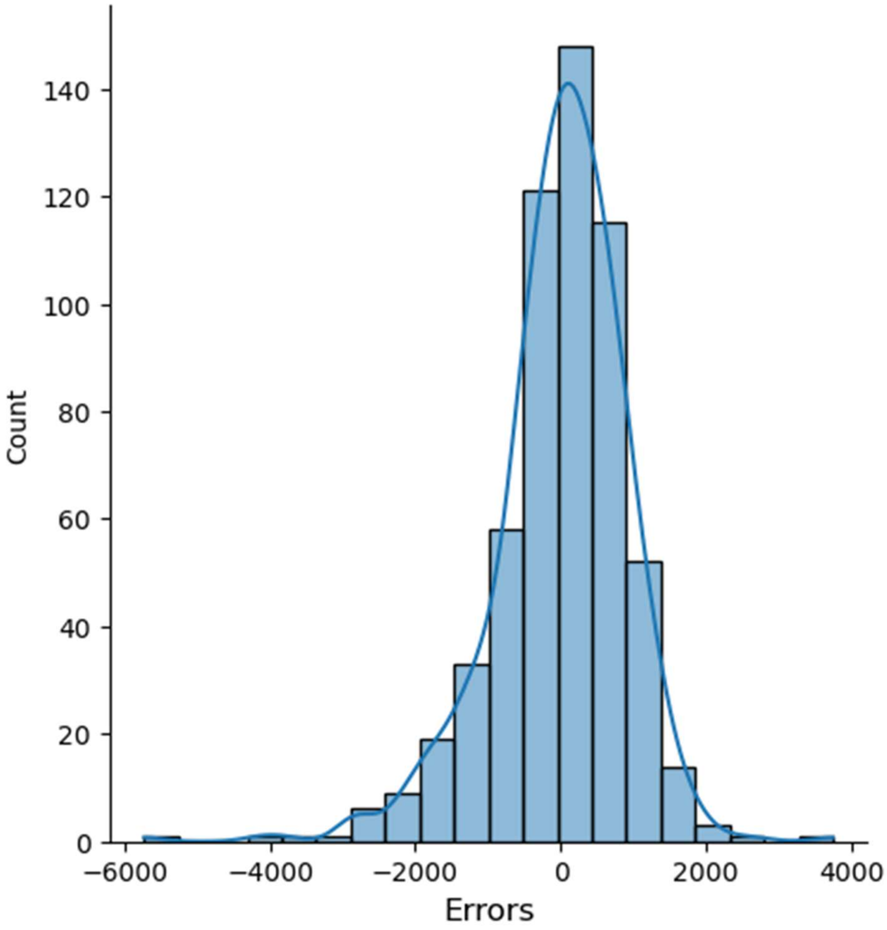


Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: Validating the assumptions of linear regression is a crucial step to ensure the reliability of the model. After building the model on the training set, here are the steps I followed to validate the assumptions:

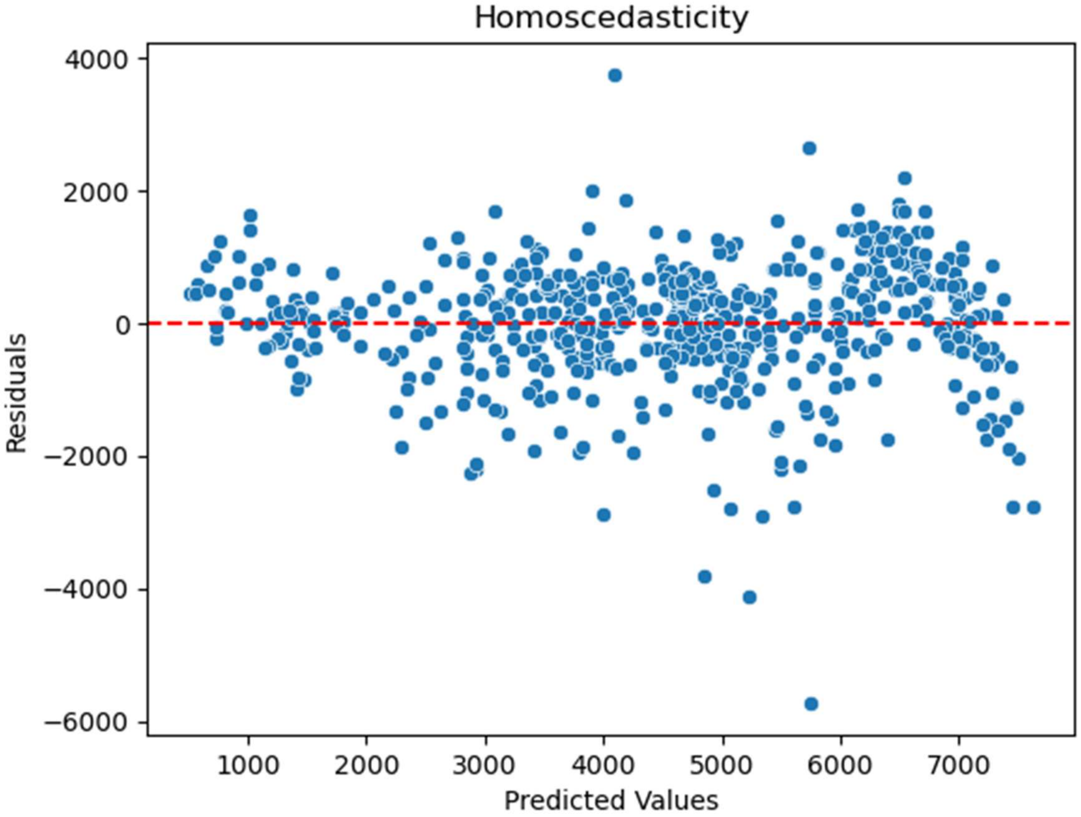
1. **Residual Analysis**
   We need examine the residuals (the differences between observed and predicted values) and confirm if Residuals are approximately normally distributed, and there should be no discernible patterns in the residual plot.



2. **Homoscedasticity (Constant Variance):**

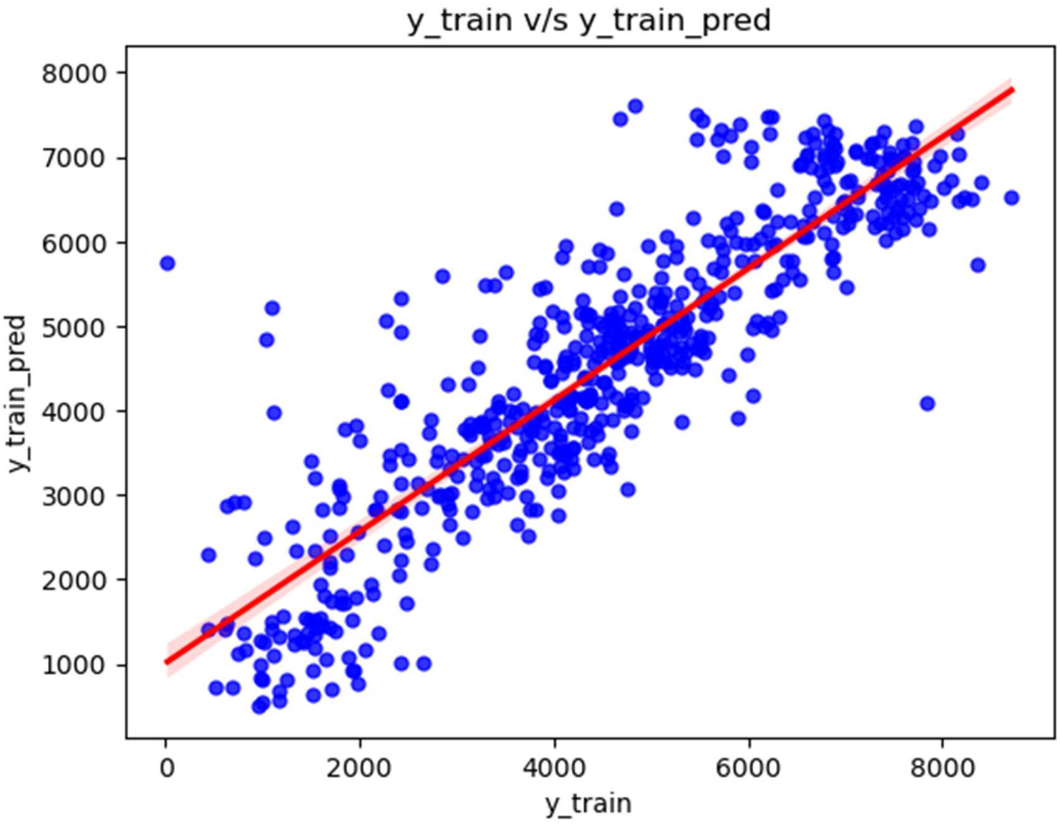
   We need to plot residuals against predicted values.

   Check: The spread of residuals should be roughly constant across all levels of the predicted values.



3. Linearity

We need to create a scatterplot of observed vs. predicted values.
Check: The points should fall approximately along a diagonal line, indica>ng a linear relationship.



4. **Independence of Residuals:**

   Examine residuals for autocorrelation.

   Check: There should be no discernible pattern in the residuals when plotted

5. **Multicollinearity**

   Process: Calculate Variance Inflation Factors (VIF) for predictor variables.

   Check: VIF values should be below a certain threshold (commonly 5 or 10) to ensure no problematic multicollinearity.

6. **Cross-Validation**

   Process: Validate the model on a test set or through cross-validation

   Check: Assess the model's performance on new data to ensure generalizability and consistency.

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: From the equation of best fit line,

count = 1314.37 x const + year x 2074.31 + 3456.38 x temp - 1261.71 x season_spring + 722.16 x season_winter - 624.52 x month_December - 758.49 x month_November + 776.01 x weathersit_good

Following feature significantly contribute to explaining demand of shared bike

1. Temperature
2. Year
3. Winter season

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Linear regression is a statistical method used for modelling the relationship between a dependent variable and one or more independent variables. It is widely used for predicting the value of the dependent variable based on the values of one or more independent variables. The basic idea is to find the best-fitting line (or hyperplane in the case of multiple independent variables) that minimizes the sum of the squared differences between the observed and predicted values of the dependent variable.

Linear regression algorithm follows following steps:

1. **Model Representation:**

Simple Linear Regression: In the case of a single independent variable, the model is represented as:

$$y = b_{!} + b_{"} \cdot x + \quad \varepsilon$$

where:

- y is the dependent variable,
- x is the independent variable,
- $b_{!}$ is the y-intercept (constant term), - $b_{"}$ is the slope of the line, and - ε represents the error term.

Multiple Linear Regression: When there are multiple independent variables, the model is extended to:

$$y = b_{!} + b_{"} \cdot x_{"} + b_{\#} \cdot x_{\#} + \cdots + b_{\$} \cdot x_{\$}$$

where:

- $(x_{"}, x_{\#}, \ldots, x_{\$})$ are the independent variables, and
- $(b_{!}, b_{"}, b_{\#}, \ldots, b_{\$})$ are the coefficients.

2. **Objective Function:**

The goal is to find the values of $(b_{!}, b_{"}, b_{\#}, \ldots, b_{\$})$ that minimize the sum of the squared differences between the observed and predicted values. This is expressed as the sum of squared errors (SSE) or mean squared error (MSE):

$$\mathrm{MSE} = \frac{1}{2m}\mathrm{V}^{'}(y_{\%} - y\mathrm{W}_{\&})^{\#}$$

%(" where (m) is the number of data points, ($y_{\%}$) is the observed value, and (yW$_{\&}$) is the predicted value.

3. Minimization:

To find the op>mal values of the coefficients, the algorithm uses optimization techniques such as gradient descent. The objective is to iteratively update the coefficients in the direction that minimizes the cost function.

4. Training the Model:

The model is trained on a dataset, where the algorithm learns the values of the coefficients that best fit the data. This involves feeding the algorithm input-output pairs and adjusting the coefficients un>l the model produces predictions close to the actual outcomes.

5. Prediction:

Once the model is trained, it can be used to make predictions on new, unseen data. The predicted values are obtained by plugging the new input values into the learned regression equation.

6. Evaluation:

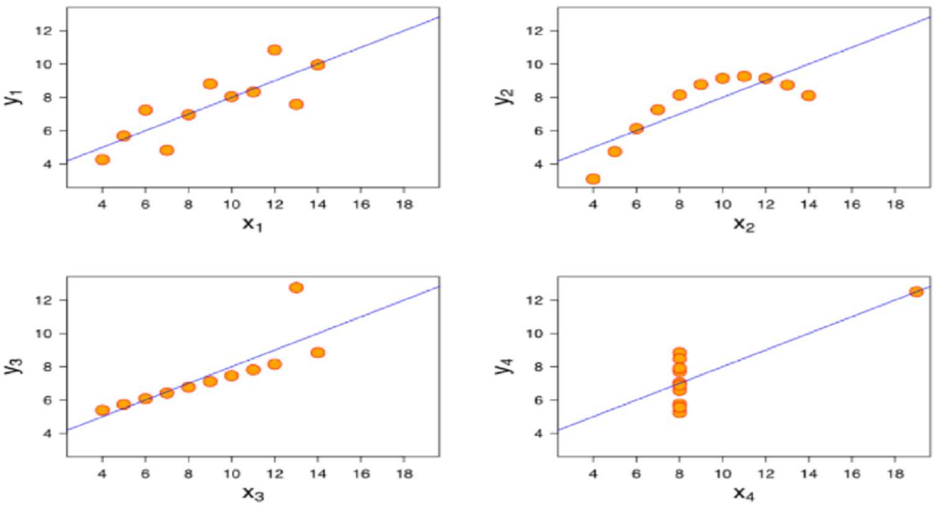
The model's performance is assessed using metrics such as ($R^{\#}$) (coefficient of determination), MSE, or other relevant metrics, depending on the context.

7. Assumptions:

Linear regression relies on the assumption of a linear relationship between independent and dependent variables, normally distributed errors, constant error variance (Homoscedasticity), and the absence of perfect multicollinearity, ensuring that there is no perfect linear relationship among the predictors.

Question 7. Explain the Anscombe’s quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

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. It was constructed to illustrate 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.



- 1st data set fit linear regression model as it seems to be linear relationship between X and y
- 2nd data set does not show a linear relationship between X and Y, which means it does not fit the linear regression model.
- 3rd data set shows some outliers present in the dataset which can’t be handled by a linear regression model.
- 4th data set has a high leverage point means it produces a high correlation coeff.

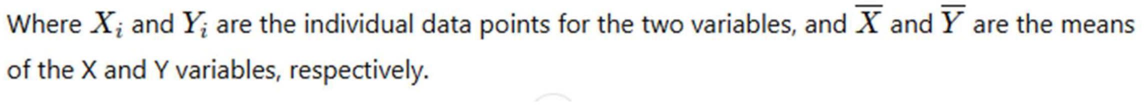
Its conclusion is that regression algorithms can be fooled so, it's important to data visualization before build machine learning model.

Question 8. What is Pearson’s R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

In Statistics, the Pearson's Correlation Coefficient is also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It is a statistic that measures the linear correlation between two variables.

Below is its formula for data X and Y

$$r = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum(X_i - \overline{X})^2 \sum(Y_i - \overline{Y})^2}}$$


Where X_i and Y_i are the individual data points for the two variables, and $\overline{X}$ and $\overline{Y}$ are the means of the X and Y variables, respectively.

Pearson's correlation coefficient is widely used in statistics to assess the strength and direction of the linear relationship between two variables. It's important to note that correlation does not imply causation, and a correlation coefficient close to zero does not necessarily mean the absence of a relationship; it only indicates the absence of a linear relationship.

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized

scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Scaling means you're transforming your data so that it fits within a specific scale. It is one type of data pre-processing step where we will fit data in specific scale and speed up the calculations in an algorithm. Collected data contains features varying in magnitudes, units and range. If scaling is not performed than algorithm tends to weigh high values magnitudes and ignore other parameters which will result in incorrect modeling. Difference between Normalizing Scaling and Standardize Scaling:

1. In normalized scaling minimum and maximum value of features being used whereas in Standardize scaling mean and standard deviation is used for scaling.
2. Normalized scaling is used when features are of different scales whereas standardized scaling is used to ensure zero mean and unit standard deviation.
3. Normalized scaling scales values between (0,1) or (-1,1) whereas standardized scaling is not having or is not bounded in a certain range.
4. Normalized scaling is affected by outliers whereas standardized scaling is not having any effect by outliers.
5. Normalized scaling is used when we don’t know about the distribution whereas standardized scaling is used when distribution is normal.
6. Normalized scaling is called as scaling normalization whereas standardized scaling is called as Z Score Normalization.

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

VIF (Variance Inflation Factor) is a measure of multicollinearity among predictor variables in a regression model. It quantifies how much the variance of a coefficient estimate is inflated due to correlations with other predictors.

Infinite VIF can happen when we have perfect multicollinearity. One predictor variable is an exact linear combination of other predictors. For instance, including both "temp" and "temp in celsius" in a model would result in perfect multicollinearity.

This can result in Model Instability, Unreliable Coefficient and Model Training Issue

---

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to compare two probability distributions by plotting their quantiles against each other. The plot helps to determine if the two datasets come from the same distribution. If the datasets are similar, the points on the Q-Q plot will form a straight line. A Q-Q plot can be used to assess whether a dataset follows a specific theoretical distribution. It is particularly useful for checking the normality assumption in linear regression, where the residuals are assumed to follow a normal distribution.

In the context of linear regression, a Q-Q plot is used to verify the assumption that the residuals (differences between observed and predicted values) are normally distributed. This assumption is critical because it affects the validity of hypothesis tests and confidence intervals in linear regression.

Key points for linear regression:

1. Checking Normality of Residuals: Linear regression assumes that the residuals are normally distributed. A Q-Q plot can help assess whether this assumption holds by comparing the quantiles of the residuals with the quantiles of a standard normal distribution.
2. Evaluating the Fit of a Model: If the Q-Q plot shows that the residuals deviate significantly from a straight line, this may indicate problems with the model, such as non-normality or outliers that may affect the results of the regression.

Advantages of Q-Q Plot:

- Sample Size Flexibility: The Q-Q plot can be used with any sample size, making it versatile for both small and large datasets.
- Detecting Distributional Characteristics: It helps identify shifts in location, scale, symmetry, and the presence of outliers in the data.
- Assessing Similarity Between Two Datasets: A Q-Q plot can also be used to compare two datasets to check if they come from populations with the same distribution, location, scale, or tail behavior.