Assignment-based Subjective Question

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**
   Answer:
   1. Booking count has significantly increased for each season in 2019 compared to 2018 and Fall season seems to have attracted more booking in both the years
   2. There is a good increase in bookings during the months of may, june, july, aug, sep and oct compared other months of the year. We can see a trend in increasing starting of the year till mid of the year and then it started decreasing as we approached the end of year. Number of bookings for each month have increased in 2019 as compared to 2018.
   3. Clear weather has more booking and is obvious. Comparing 2018 and 2019, booking has been increased for each weather conditions in 2019
   4. Thursday, Friday, Saturday and Sunday have a greater number of bookings as compared to the start of the week for both 2018 and 2019
   5. Holidays show a smaller number of bookings and quite obvious that people would want to stay at home and spent time with well beings
   6. Not much difference between a working and non-working day for both 2018 and 2019
   7. Certainly 2019 got a greater number of bookings compared 2018 in every category and conditions

2. **Why is it important to use drop_first=True during dummy variable creation?**
   Answer:
   When creating dummy variables for categorical data in regression analysis, it is important to use drop_first=True in order to avoid multicollinearity problems known as the "dummy variable trap". The dummy variable trap occurs when the model has perfect multicollinearity, which means that one of the dummy variables can be perfectly predicted by the others. This happens because the dummy variables are not independent; one of them can be expressed as a linear combination of the others.
   Example

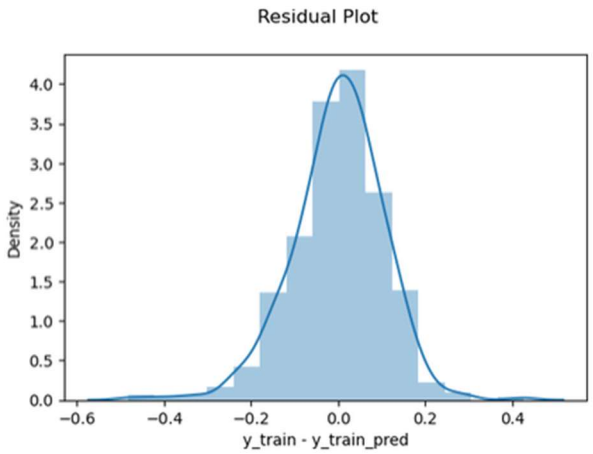
   Consider a categorical variable Colour with three categories: Red, Blue, and Green. Without dropping the first category, you would create three dummy variables:

   Color_Red
   Color_Blue
   Color_Green

   If you include all three dummy variables in a regression model, you introduce perfect multicollinearity. This is because if you know the values of Color_Red and Color_Blue, you can perfectly predict the value of Color_Green (i.e., Color_Green = 1 - Color_Red - Color_Blue).

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

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
   Answer:
   Please find below details for assumptions related to the LR Model
   - **Normality of error terms**
     Error terms should be normally distributed



The residual plot is following Normal distribution centered around zero

- **Multicollinearity check**

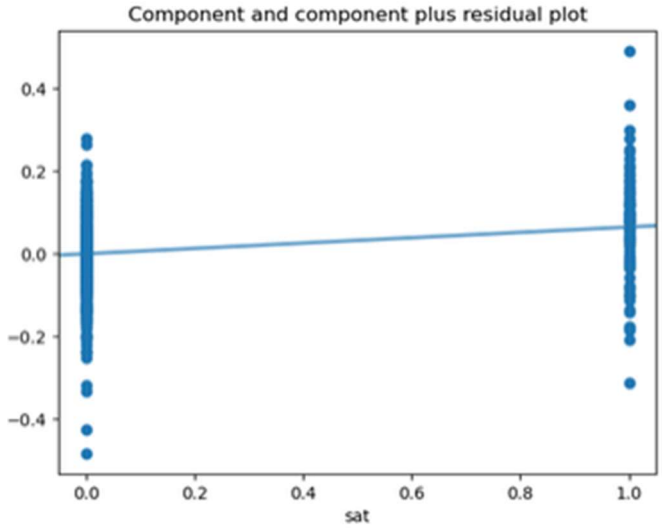
  There should be insignificant multicollinearity among variables. The details of VIF shows no multicollinearity

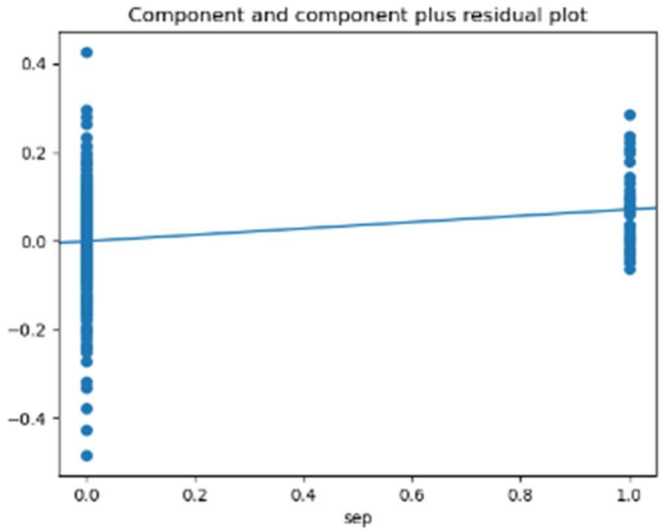
[73]:

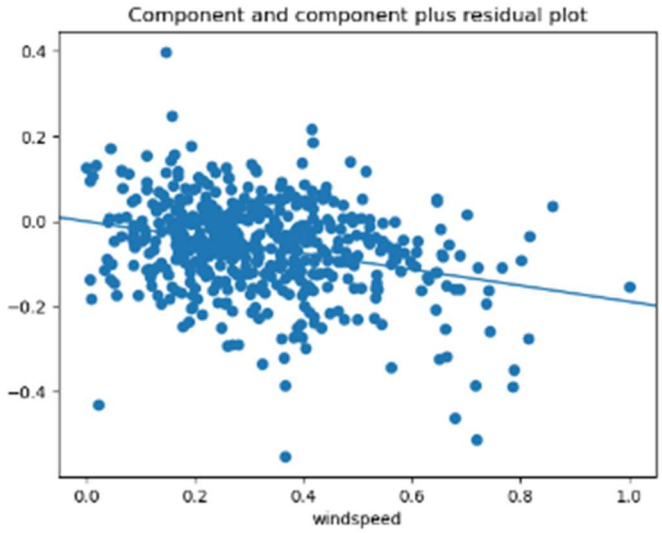
	Features	VIF
2	windspeed	4.04
1	workingday	3.29
8	spring	2.65
9	summer	2.00
0	year	1.88
10	winter	1.73
3	jan	1.60
7	Misty	1.57
5	sat	1.56
4	sep	1.18
6	Light_snowrain	1.08

- **Linear relationship validation**

  Linearity should be visible among variables

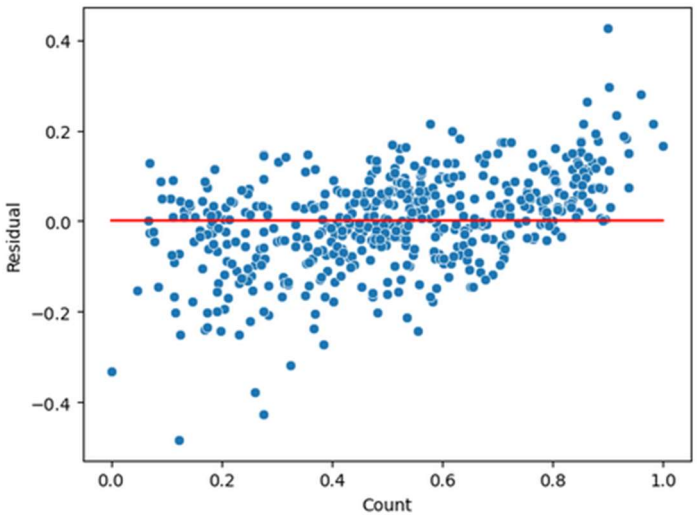

Component and component plus residual plot
0.4
0.2
0.0
−0.2
−0.4
0.0
0.2
0.4
0.6
0.8
1.0
sat


Component and component plus residual plot
0.4
0.2
0.0
−0.2
−0.4
0.0
0.2
0.4
0.6
0.8
1.0
sep


Component and component plus residual plot
0.4
0.2
0.0
−0.2
−0.4
0.0
0.2
0.4
0.6
0.8
1.0
windspeed

- **Homoscedasticity**

  There should be no visible pattern in residual values.



No visible patterns observed from above plot for residuals.

- **Independence of residuals**

  No auto-correlation

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

Answer:

a) sep
b) sat
c) workingday

General Subjective Questions

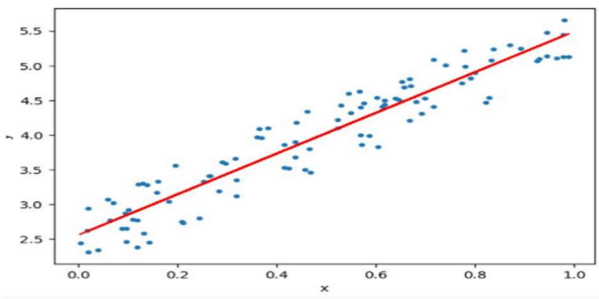
1. **Explain the linear regression algorithm in detail.**

Answer:

A statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables.
Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).

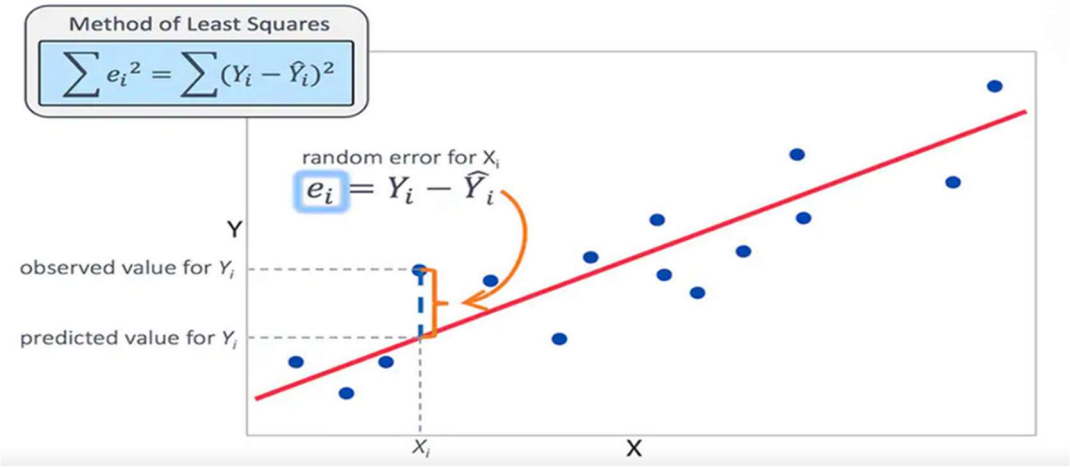
Mathematically the relationship can be represented with the help of following equation –
Y = mX + c
Here, Y is the dependent variable we are trying to predict.
X is the independent variable we are using to make predictions.
m is the slope of the regression line which represents the effect X has on Y
c is a constant, known as the Y-intercept. If X = 0, Y would be equal to c.

In the figure below, we can see that we have two variables x and y which have some scattered. Through the scatter plot, there is a line passing through. This called a **Regression line** or the **best fit line**.



Objective

The objective of linear regression is to minimize the sum of the squared differences between the observed values and the predicted values (residuals). This method is known as Ordinary Least Squares (OLS).



Steps in Linear Regression

Data Collection and Preparation:
Gather and clean the data.
Split the data into training and testing sets.

Exploratory Data Analysis (EDA):
Visualize the data to understand relationships between variables.
Check for linearity, correlation, and outliers.

Feature Selection and Engineering:
Select relevant features that influence the dependent variable.
Create new features if necessary.

Model Training:
Use the training data to fit the linear regression model.
Calculate the coefficients (β0,β1,…,βnβ0,β1,…,βn).

Model Evaluation:
Evaluate the model using the testing data.
Common metrics: R-squared (R2R2), Mean Squared Error (MSE), Root Mean Squared Error (RMSE).

Prediction:
Use the trained model to make predictions on new data.

Assumptions of Linear Regression

Linearity: The relationship between the independent and dependent variables is linear.
Independence: The residuals (errors) are independent.
Homoscedasticity: The residuals have constant variance.
Normality: The residuals are normally distributed.
No Multicollinearity: The independent variables are not highly correlated.

2. Explain the Anscombe's quartet in detail.

Answer:

Anscombe's Quartet is the modal example to demonstrate the importance of data visualization which was developed by the statistician Francis Anscombe in 1973 to signify both the importance of plotting data before analysing it with statistical properties. It comprises of four data-set and each data-set consists of eleven (x,y) points. The basic thing to analyse 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 behaviour irrespective of statistical analysis.

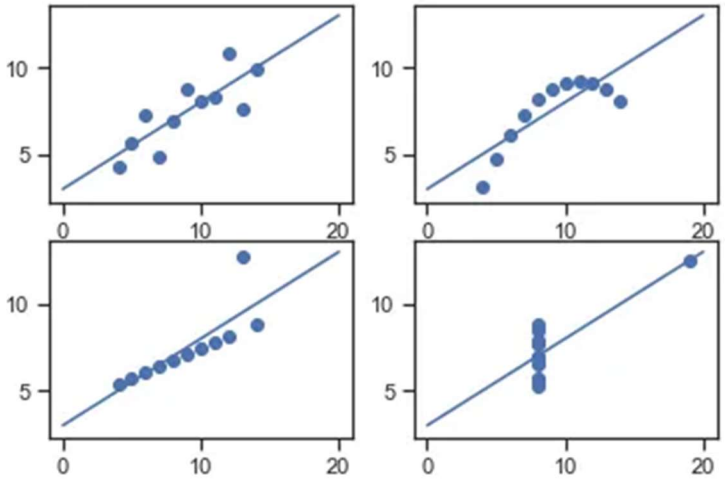
x1	y1	x2	y2	x3	y3	x4	y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Four Data-sets

Apply the statistical formula on the above data-set,

Average Value of x = 9
Average Value of y = 7.50
Variance of x = 11
Variance of y =4.12
Correlation Coefficient = 0.816
Linear Regression Equation : y = 0.5 x + 3

However, the statistical analysis of these four data-sets is pretty much similar. But when we plot these four data-sets across the x & y coordinate plane, we get the following results & each pictorial view represent the different behaviour.



Graphical Representation of Anscombe's Quartet

Data-set I — consists of a set of (x,y) points that represent a linear relationship with some variance.
Data-set II — shows a curve shape but doesn't show a linear relationship (might be quadratic?).
Data-set III — looks like a tight linear relationship between x and y, except for one large outlier.
Data-set IV — looks like the value of x remains constant, except for one outlier as well.

```
_____Mean of x_____
x1 : 9.000      x2 : 9.000      x3 : 9.000      x4 : 9.000

_____Mean of y_____
y1 : 7.501      y2 : 7.501      y3 : 7.500      y4 : 7.501

_____Variance of x_____
x1 : 11.000     x2 : 11.000     x3 : 11.000     x4 : 11.000

_____Variance of y_____
y1 : 4.127      y2 : 4.128      y3 : 4.123      y4 : 4.123

_____Correlation of x & y_____
x1/y1 : 0.816   x2/y2 : 0.816   x3/y3 : 0.816   x4/y4 : 0.817
```

Data-sets which are identical over a number of statistical properties, yet produce dissimilar graphs, are frequently used to illustrate the importance of graphical representations when exploring data. This isn't to say that summary statistics are useless. They're just misleading on their own. It's important to use these as just one tool in a larger data analysis process. Visualizing our data allows us to revisit our summary statistics and re-contextualize them as needed.

Importance of Anscombe's Quartet

Anscombe's quartet demonstrates several important points:

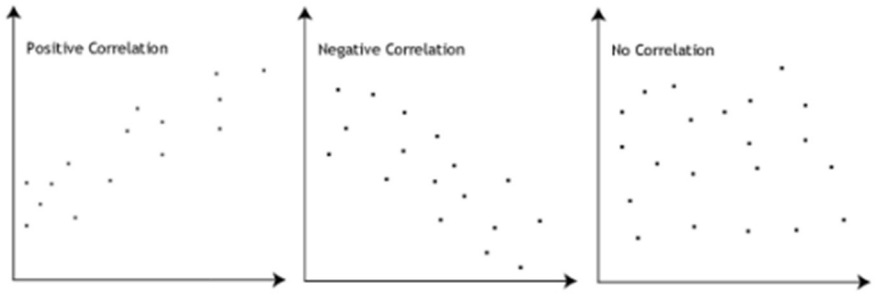
Graphical Analysis: Summary statistics alone can be misleading. Visualizing data is crucial for identifying patterns, relationships, and anomalies that statistics might miss.
Outliers Impact: Outliers can significantly affect regression models and correlations, highlighting the need to carefully inspect and handle them.
Model Fit: The appropriateness of a model cannot be determined by summary statistics alone. It is essential to visualize data to ensure the chosen model fits well.

By understanding and applying the lessons from Anscombe's quartet, analysts can avoid common pitfalls in data analysis and ensure more accurate and insightful conclusions.

3. What is Pearson's R?

Answer:
Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

The Pearson correlation coefficient, r, can take a range of values from +1 to -1.
A value of 0 indicates that there is no association between the two variables.
A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable.
A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the
diagram below:



The formula for Pearson's R is:

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

Where:

- x_i and y_i are the individual sample points.
- $\bar{x}$ and $\bar{y}$ are the means of the x and y variables, respectively.

Calculation Steps

Compute the mean of both variables x and y.
Subtract the mean from each individual observation to obtain the deviations.
Multiply the deviations of x and y for each observation.
Sum the products of the deviations.
Square the deviations of x and y separately, then sum these squares.
Divide the sum of the products of deviations by the square root of the product of the sums of the squared deviations.

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

Answer:

Scaling is the process of adjusting the range of feature values in a dataset so that they fit within a specific range, often to improve the performance and training stability of machine learning algorithms.

Why Scaling is Performed

Improves Model Performance:

Many machine learning algorithms perform better when features are on a similar scale. For example, gradient descent converges faster when features are scaled.

Enhances Model Interpretability:

Scaling can help make the features more interpretable and comparable, especially when they are on different scales.

Prevents Bias Towards Features:

Algorithms that compute distances between data points (e.g., k-NN, SVM) or that involve regularization (e.g., linear regression with L2 regularization) can be biased by the scale of features.

Key Differences between Normalized Scaling Vs Standardized Scaling

Aspect	Normalized Scaling	Standardized Scaling
Formula	$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$	$x' = \frac{x - \mu}{\sigma}$
Range	Typically [0, 1] or [-1, 1]	Mean = 0, Standard Deviation = 1
Sensitivity to Outliers	High	Moderate
Use Case	Non-Gaussian distributions, bounded range	Gaussian distributions, many ML algorithms
Effect	Rescales data to a fixed range	Centers data around mean with unit variance

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

Answer:

The Variance Inflation Factor (VIF) measures how much the variance of a regression coefficient is inflated due to multicollinearity with other predictor variables. When the VIF value is infinite, it indicates perfect multicollinearity.

Why VIF Becomes Infinite

Perfect Multicollinearity: This occurs when one predictor variable in a regression model is an exact linear combination of one or more of the other predictor variables. This means that there is a perfect correlation (either positive or negative) between some predictors.

Mathematical Explanation

The VIF for a predictor Xi is calculated as follows:

$$\text{VIF}(X_i) = \frac{1}{1-R_i^2}$$

Where R_i^2 is the coefficient of determination of the regression of X_i on all other predictors. If $R_i^2 = 1$ (indicating perfect multicollinearity), then:

$$\text{VIF}(X_i) = \frac{1}{1-1} = \frac{1}{0} = \infty$$

Causes of Perfect Multicollinearity

Duplicate Variables: Including the same variable more than once in the regression model.
Linear Dependence: One variable is a perfect linear function of another, such as:
Sum of variables (e.g., X3=X1+X2X3=X1+X2).
Constant multiples (e.g., X2=2X1X2=2X1).
Dummy Variable Trap: Including all dummy variables for a categorical feature without dropping one (when using one-hot encoding), causing perfect multicollinearity.

Handling Infinite VIF

Remove One of the Collinear Variables: Identify and remove one of the perfectly collinear variables.
Combine Collinear Variables: Combine collinear variables into a single feature if they convey the same information.

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

Answer:

A Q-Q (quantile-quantile) plot is a graphical tool used to compare the distribution of a dataset to a theoretical distribution, typically the normal distribution. It plots the quantiles of the dataset against the quantiles of the theoretical distribution.

How to Read a Q-Q Plot

Straight Line: If the points lie approximately along a straight line, the data follows the theoretical distribution.
Deviations: Systematic deviations from the line indicate departures from the theoretical distribution.

Steps to Create a Q-Q Plot

Sort the Data: Order the sample data from smallest to largest.
Calculate Quantiles: Calculate the quantiles for the ordered data.
Determine Theoretical Quantiles: Calculate the corresponding quantiles from the theoretical distribution.
Plot: Plot the sample quantiles against the theoretical quantiles.

Use and Importance of a Q-Q Plot in Linear Regression

In the context of linear regression, a Q-Q plot is crucial for diagnosing the normality assumption of the residuals.
Linear regression relies on several assumptions, one of which is that the residuals (errors) are normally distributed. This assumption affects the validity of hypothesis tests and confidence intervals for the regression coefficients.

Steps for Using Q-Q Plot in Linear Regression

Fit the Regression Model: Fit the linear regression model to the data.
Obtain Residuals: Calculate the residuals (differences between observed and predicted values).
Create Q-Q Plot of Residuals: Generate a Q-Q plot of the residuals against a normal distribution.