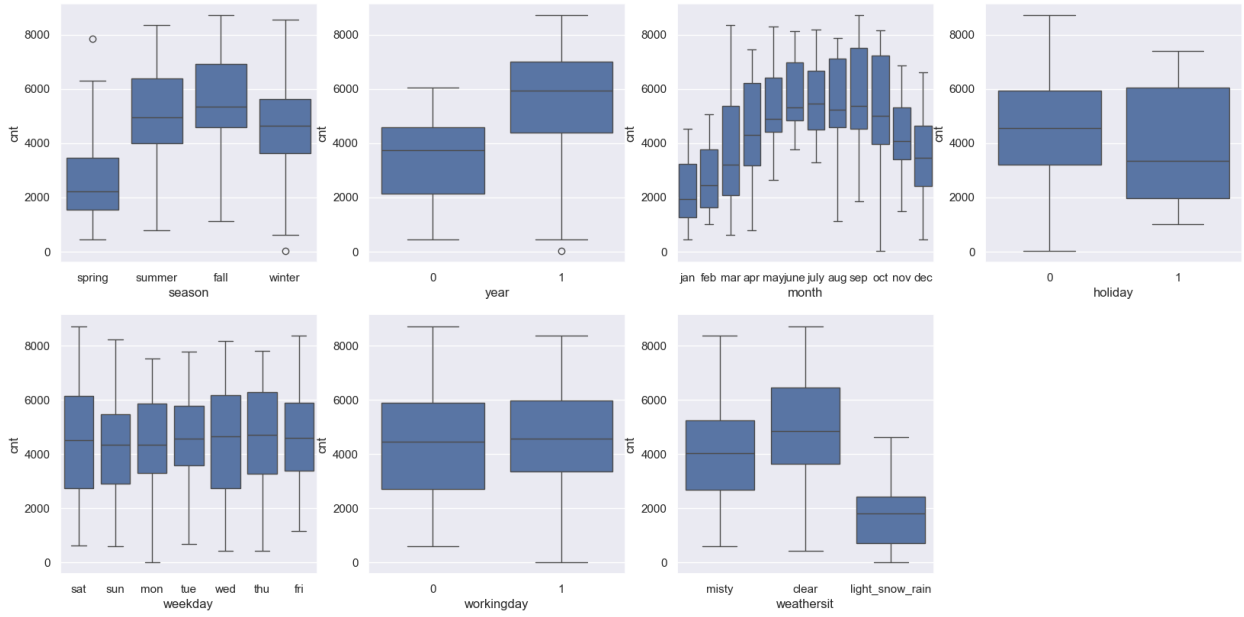
Assignment-based Subjective Questions

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. =>

There are 7 categorical variables (**season, month, year, weekday, holiday, workingday and weathersit**) in given data set for bike sharing.

- season 3 (fall) has the highest demand of the bikes
- demand for rental bikes in 2019 has increased as compared to 2018.
- On Monthly basis demand of rental bikes shows continues increase till the month of June, while in month of July and August the is slight fall in demand of rental bikes. September is having highest demand.
- Post September month, rental bikes demand has gradually decreased.
- On holidays, demand of rental bike is lesser.
- Clear weather situation: (Clear, few clouds, Partly cloudy, Partly cloudy) has highest demand as compared to other weather conditions.
- 6th day (Friday) of the week is having more demand for rental bikes.



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

Ans. => With creation of dummy variables from categorical data, it is always suggested to drop or use **first= True** to avoid the problem of Multicollinearity in regression model.

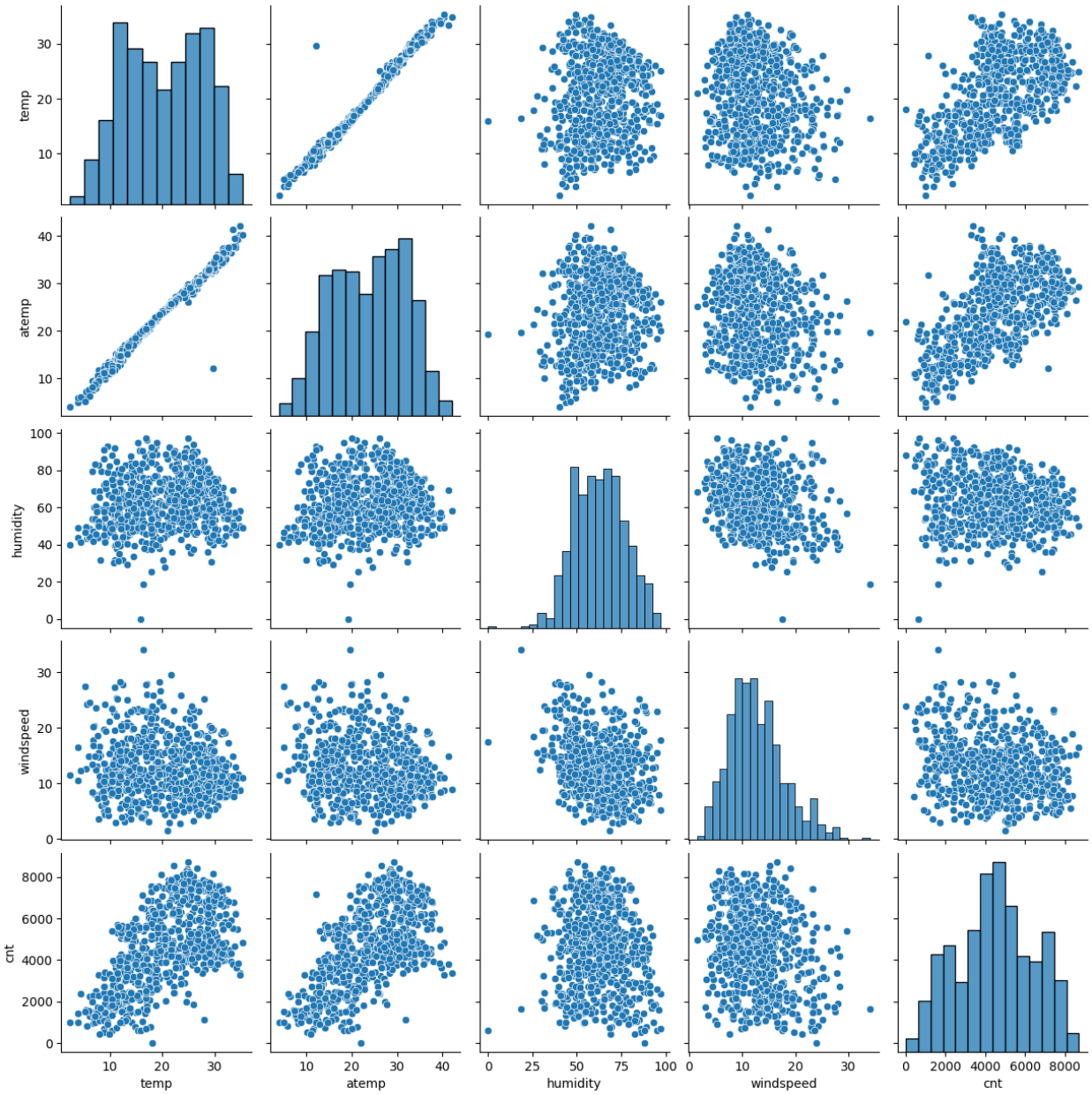
By setting **drop_first=True**, we only create n-1 dummy variables instead of n. This eliminates one of the dummy variables, which effectively serves as a reference or baseline category. The

remaining n-1 dummy variables can then be used to compare the effect of each category relative to this baseline, without introducing collinearity issues.

Multicollinearity is a condition with occurs when independent variables in regression model is highly correlated with each other. It can cause problem to calculate coefficients accurately and as result model will be less reliable.

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

Ans. => `**temp**` and `**atemp**` has highest correlation.



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

Ans. => Validation the assumptions of Linear Regression after building the model on the training set is very crucial for ensuring the model's result is reliable and the interpretations are valid. To perform the validation, we perform the following checks:

- Linearity check: Draw a scatter plot to predict values against actual values or residuals to see if there is a linear relationship.
- Normality of Residuals: Plotting histogram of residuals should ideally resemble a normal distribution with center at 0.
- No Multicollinearity: check for VIF and p-value, High VIF value greater than 10 is problematic and below 5 is best case. Also, P-value should be less than 0.05 and there should not be high value for either of them.

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 temp, year, and winter.

General Subjective Questions

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

Ans. => Linear regression is fundamental and widely used algorithm in the field of statistics and Machine Learning. The main goal of linear regression is to predict the value of the dependent variable (y) based on the values of independent variables (x). In linear regression, it is assumed that there is linear relationship between the dependent variable (y) and the independent variables (x).

It can be of 2 types as mentioned below:

Simple Linear Regression: It involves only one independent variable and the relationship is modelled with a straight line.

$y=\beta0+\beta1x+\epsilon$

where:

y is the dependent variable.

x is the independent variable.

β0 is the intercept of the line (the value of y when x = 0).

β1 is the slope of the line (the change in y for a one-unit change in x).

ϵ is the error term or residual (the difference between the observed and predicted values of y).

Multiple Linear Regression: It is dependent on multiple variables and the equation of model is given as $y=\beta0+\beta1x1+\beta2x2+\cdots+\beta nxn+\epsilon$

Where:

X1,x2,…,xn are the independent variables

β1,β2,…,βn are the coefficients for each independent variable.

A Linear Regression can have positive or negative linear relationship. The goal of linear regression is to determine best-fit line in case of simple linear regression and hyperplane in case of multiple variables to minimize the discrepancy between the predicted values and the actual values. This is typically achieved using the **Least Square** method. To evaluate the performance of the linear regression R-squared (R^2) , Mean squared error (MSE), Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) can be used.

Assumptions:

Linearity: The relationship between the dependent and independent variables is linear.

Independence: Observations are independent of each other.

Homoscedasticity: The residuals have constant variance.

Normality of Errors: The residuals are normally distributed (primarily for hypothesis testing and confidence intervals).

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

Ans. => Anscombe's Quartet is a powerful illustration of why visual exploration of data is essential in statistical analysis. It underscores the importance of looking beyond summary statistics to understand the true nature of the data and make informed decisions based on accurate insights. Anscombe's quartet comprises of a set of 4 datasets, having identical descriptive statistical properties in terms of means, variance, R-Squared, correlations, and the linear regression lines but having different representations when we scatter plots on a graph.

The four datasets of Anscombe's quartet.

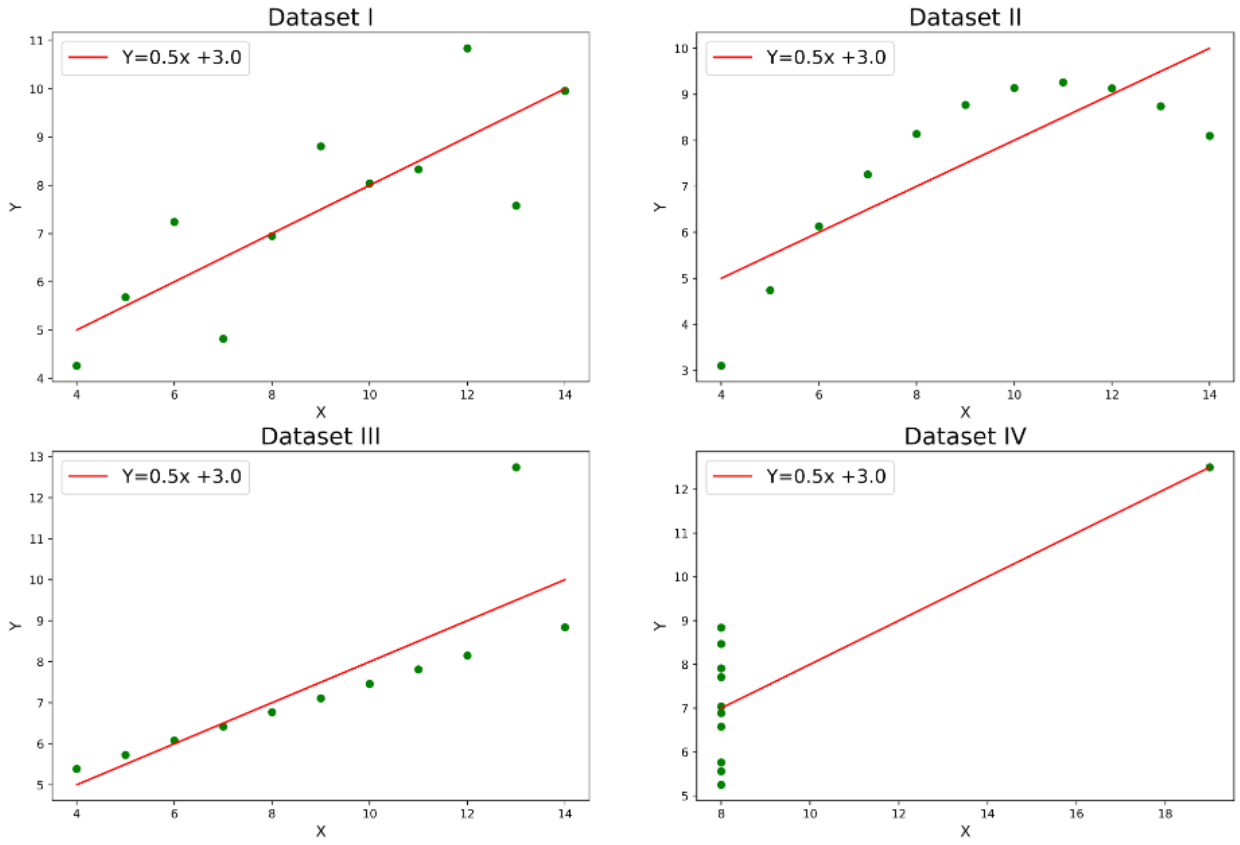
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	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

Output:

	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

Clearly, there is identical descriptive statistics summary which lead to believe that the datasets are essentially the same.

However, when examining the scatter plots of these datasets, we'll observe the inherent differences.

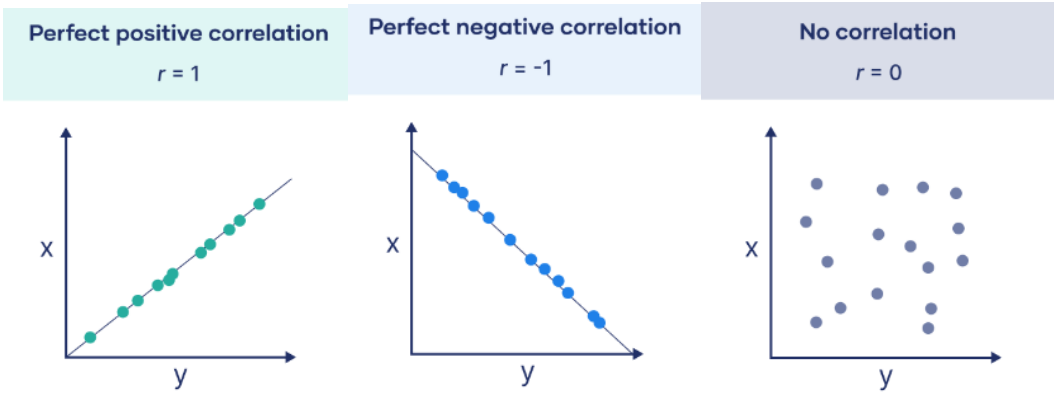


3. What is Pearson's R? (3 marks)

Ans. => Pearson's R (Pearson correlation coefficient) is a measure of the linear relationship between two variables. It quantifies the strength and direction of the linear relationship and is widely used in statistics and data analysis.

Pearson's R ranges from -1 to +1,

- +1 is perfect positive linear relationship
- -1 is perfect negative linear relationship
- And 0 is no linear relationship



Pearson's R can be highly sensitive to outliers. An outlier can significantly affect the correlation coefficient.

Strength of Pearson's R can be described in following ways:

- 0.1 to 0.3: Weak correlation
- 0.3 to 0.5: Moderate correlation
- 0.5 to 1.0: Strong correlation

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

Ans. => Scaling is a preprocessing step in data analysis and machine learning where independent features or variables are transformed to a common scale. This process is crucial for several reasons and can be performed using different techniques.

Feature with larger ranges, values or weight can disproportionately influence the model. Also, some algorithm assumes that the features are on similar scale. For instance, gradient descent-base methods like linear regression and logistic regression converge faster and more effectively with the scaled features. Thus, it is important to perform scaling.

We have 2 types of scaling. They are:

- Normalized scaling: Normalization (min-max scaling) transforms features to a fixed range usually between 0 and 1.
- Standardized scaling: Standardization (z-score normalization) transforms features to have a mean of 0 ad a standard deviation of 1.

S.No.	Normalized Scaling	Standardized Scaling
1	Scales between 0 and 1.	It is not bound to any range.
2	Compresses extreme values.	Centers data around 0 and scales by standard deviation.
3	Sensitive to outliers, and can compress data if outliers are present.	Less sensitive to outliers and outliers may still be visible but they are scaled.
4	It is used when features are on different scale or needs to be scaled to a specific range.	It is used when data should be centered and scaled for algorithms or to ensure zero mean and unit standard deviation.
5	Scikit-Learn provides a transformer known as MinMaxScaler for Normalization.	SciKit-Lear provides a transformer called StandardScaler for standardization.

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

Ans. => VIF stands for Variance Inflation Factor. It is a measure used to detect multicollinearity in regression models. Multicollinearity occurs when independent variables in a regression model are highly correlated, which can make it really difficult to determine the individual effects of each predictor on the dependent variable.

VIF might be infinite when there are specific conditions like perfect multicollinearity, Linear dependence among predictors and singular matrix.

Generally, it due to perfect multicollinearity which occurs when one predictor variable is a perfect linear combination of one or more other predictor variables. For example, we have two predictors, $X1$ and X2, where $X2=2X1$.The correlation between $X1$ and X2 is perfect (1 or -1), which causes 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. => A Q-Q stands for Quantile-Quantile plot. A Q-Q plot is a valuable diagnostic tool in linear regression for assessing whether the residuals (errors) follow a normal distribution. It helps in validating the assumptions of normality and identifying potential issues with the model fit. By visually comparing the quantiles of the observed data against the quantiles of a theoretical distribution, a Q-Q plot provides insight into the appropriateness of the linear regression model and the reliability of its predictions.

Its is used to check normality of residuals, model diagnostics and access the fit quality in linear regression model.

- Check Normality Residuals – residuals (errors) are normally distributed. This assumption is important for conducting reliable hypothesis tests and constructing confidence intervals.
- Model Diagnostics- It helps tin diagnosing issues related to the residuals such as skewness or heavy tails.
- Accessing fit Quality: For well-fitting linear regression model with normally distributed residual, the Q-Q plot show points that lie close to a straight line. If the plot shows systematic deviations from the line, suggests that the model may not be appropriate or the residuals might not be normally distributed.

Q-Q plot is important to get visual insight as it provides a straightforward visual method to assess normality and diagnose model fit. Also, it helps in validating the assumptions of linear regression and ensuring the robustness of the models predictions.