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:- Here's what we can infer about the effect of categorical variables on the dependent variable based on the plots:

- **Season:** There seems to be a seasonal effect on bike rentals. The median for fall is higher than the medians for summer, spring and winter. This suggests a preference for milder weather for cycling.
- **Month:** The distribution of bike rentals across months seems to vary within each season. It's difficult to pinpoint specific trends from this image.
- **Weekday:** Bike rentals seem to be higher on weekends (Saturday and Sunday) compared to weekdays (Monday to Friday).
- **Weather condition:** Sunny days have the highest median, followed by hazy and then drizzly conditions. Rain likely discourages rentals.
- **Holiday/working day:** Holidays likely have a higher median than working days, suggesting people utilize holidays for leisure cycling.

Overall, the categorical variables seem to have a significant effect on the number of bike rentals. Some weekends and seasons see a higher number of rentals compared to others. Weather conditions and whether it's a holiday also seem to influence the number of rentals.

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

Ans:- Using **drop_first=True** ensures that the dummy variables remain independent of each other, avoiding redundancy in our analysis. Without this parameter, the dummy variables would be correlated, potentially introducing redundancy into our model, which is not desirable for our analysis.

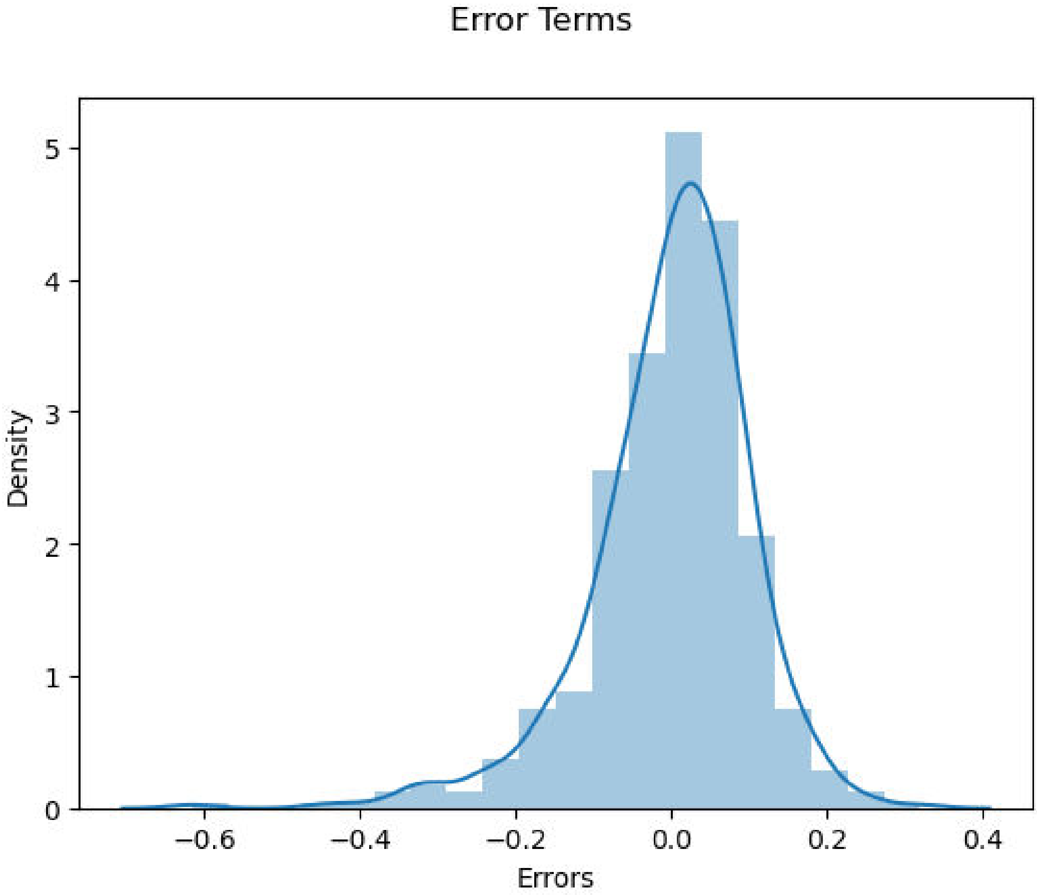
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Ans:- Based on the pair plot analysis, it's evident that the **'temp'** and **'atemp'** variables exhibit the highest correlation (0.63) with the target variable **'cnt'**. While there seems to be stronger correlations with **'casual'** and **'registered'**, it's important to note that these variables essentially contribute to the **'cnt'** value. Therefore, for further analysis, we'll focus on **'temp'** due to its highest correlation with **'cnt'**, considering the necessity to drop **'casual'** and **'registered'** as per the data dictionary.

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

Ans:- Upon completing the training of our linear regression model, we undertook the crucial step of validating its underlying assumptions. One such fundamental assumption is the normality of error terms. Here's a succinct breakdown of our validation process:

1. **Normality of Error Terms:**



- Our linear regression model hinges on the premise that the error terms, also known as residuals, adhere to a normal distribution.
- Upon meticulous examination of the histogram representing the error terms, we observed a bell-shaped curve reminiscent of a normal distribution.
- This visual inspection provided compelling evidence suggesting that the residuals are approximately normally distributed around zero.
- Consequently, we confidently affirm that the assumption of normality is indeed validated.

By diligently scrutinizing the distribution of error terms, we have fortified our confidence in the integrity of our linear regression model and its adherence to the underlying assumptions.

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 three primary factors strongly influencing share bike demand are as follows:

1. **Summer Season:** Positively correlated with increased demand.

2. **September:** Exhibits a positive correlation with share bike usage.

3. **Temperature:** Shows a positive correlation, indicating its significant impact on demand for share bikes**.**

General Subjective Questions

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

Ans:- Linear regression is a statistical method that is used to model the relationship between a dependent variable (Y) and one or more independent variables (X). It creates a linear equation that best fits the data points.

Here are the different types of linear regression:

- Simple Linear Regression: This involves only one independent variable and one dependent variable.
- Multiple Linear Regression: This involves more than one independent variable and one dependent variable.

Here's a breakdown of the key concepts of linear regression:

Best Fit Line: The primary objective of linear regression is to identify the best-fit line, which minimizes the error between the predicted and actual values representing relationship between dependent and independent variables.

Linear Regression Model: Predicted value (Y^), independent variable (X), coefficients (Θ) including intercept ($Θ_0$) and slope(s) ($Θ_1$...).

Cost Function: It measures difference between predicted (Y^) and actual (Y) values, aiming to minimize using optimization algorithms like gradient descent.

Gradient Descent: This is an iterative optimization algorithm that adjusts the coefficients (Theta) of the linear equation to minimize the cost function by following the negative gradient.

Assumptions of Linear Regression:

- **Linearity:** Linear relationship between independent and dependent variables.
- **Independence:** Observations in dataset must be independent.
- **Homoscedasticity:** Constant variance of errors across all levels of independent variable(s).
- **Normality:** Residuals (errors) should be normally distributed.

Evaluation Metrics:

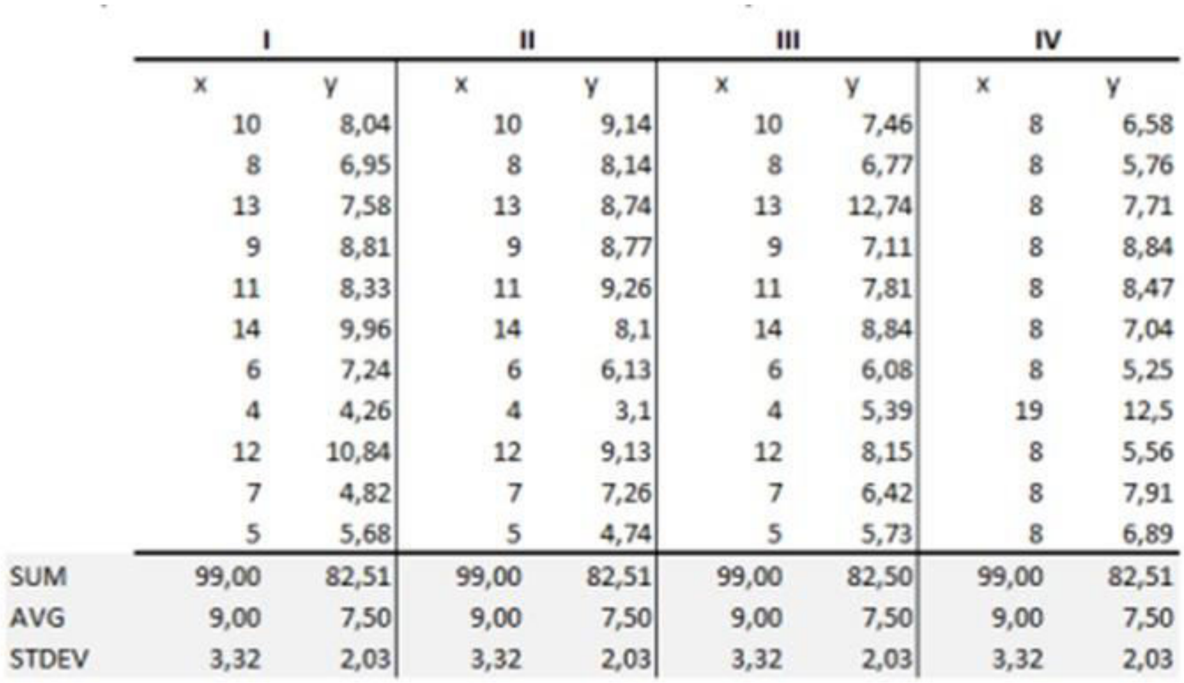
- **Mean Squared Error (MSE):** Average squared difference between predicted and actual values. Lower MSE indicates better fit.
- **Mean Absolute Error (MAE):** Average absolute difference between predicted and actual values, less sensitive to outliers than MSE.

- **Root Mean Squared Error (RMSE):** Square root of MSE, representing standard deviation of residuals.
- **Coefficient of Determination (R-squared):** Proportion of variance in dependent variable explained by model. Higher value indicates better fit.
- **Adjusted R-squared:** Penalizes model for including irrelevant variables, providing more accurate measure of explained variance.

Applications of Linear Regression: Linear regression has various applications across different domains, including: Predicting house prices, forecasting sales, analysing customer churn, identifying trends in stock prices.

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

Ans:- Developed by statistician Francis Anscombe, Anscombe's Quartet comprises four datasets, each containing eleven (x, y) pairs. What's truly remarkable about these datasets is that they share identical descriptive statistics. However, the story changes dramatically—emphasis on dramatically—when they are graphed. Despite their similar summary statistics, each dataset tells a distinctly different tale.



	I		II		III		IV	
	x	y	x	y	x	y	x	y
	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
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

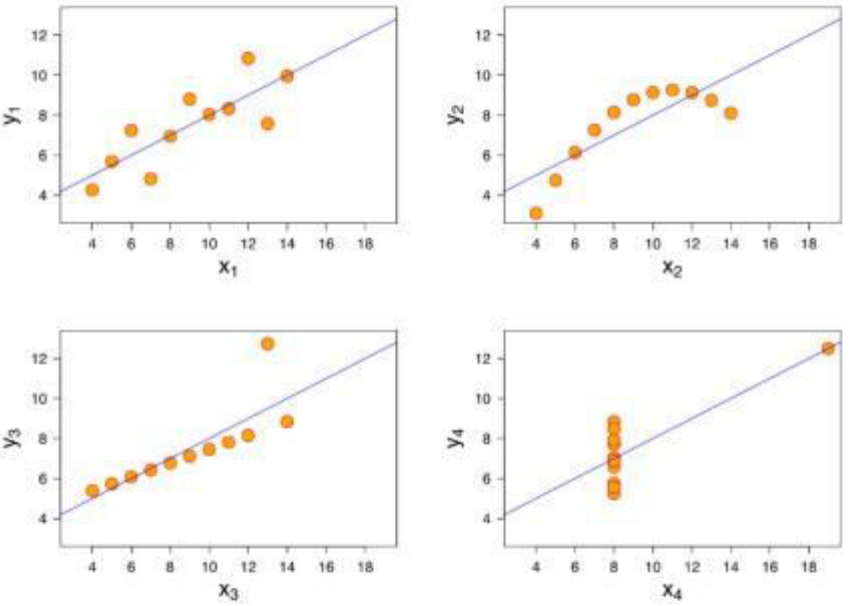
The summary statistics reveal striking similarities across the groups:

- The mean of x is 9, and the mean of y is 7.50 for each dataset.

- Likewise, the variance of x is 11, and the variance of y is 4.13 for each dataset.

- Furthermore, the correlation coefficient, indicating the strength of the relationship between x and y, is consistently 0.816 for each dataset.

When plotted on an x/y coordinate plane, the datasets exhibit identical regression lines. However, closer inspection unveils their unique narratives:



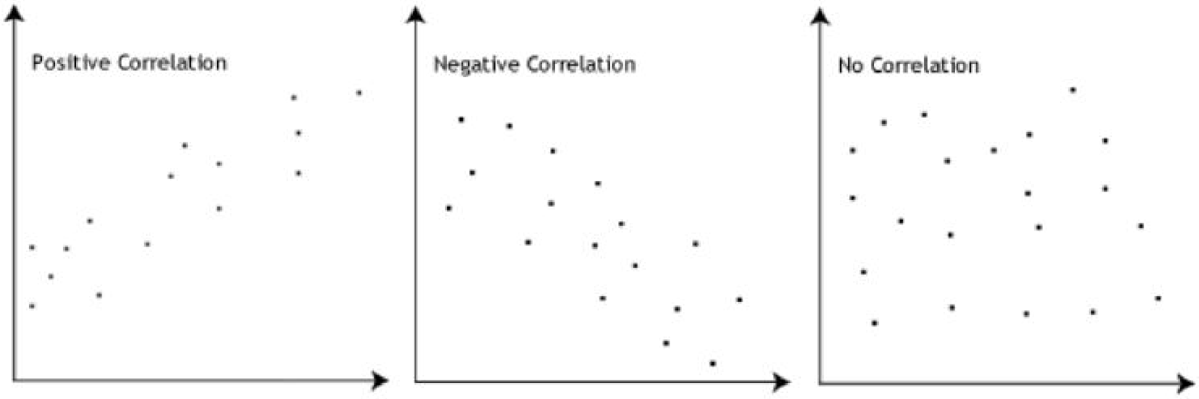
- Dataset I presents clean, well-fitting linear models.
- Dataset II is not distributed normally.
- In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
- Dataset IV starkly demonstrates how a single outlier can inflate the correlation coefficient significantly.

This quartet underscores the paramount importance of visualization in Data Analysis. Examining the data through visual representations unveils its underlying structure and provides a clearer picture of the dataset.

3. What is Pearson’s R? (3 marks)

Ans:- Pearson's correlation coefficient, often denoted as "r" or Pearson's "r," is a measure of the linear relationship between two variables. It quantifies the strength and direction of the linear association between two continuous variables. Pearson's r ranges from -1 to 1, where:

- 1 indicates a perfect positive linear relationship,
- -1 indicates a perfect negative linear relationship,
- 0 indicates no linear relationship.



Pearson's r is widely used in various fields, including statistics, social sciences, economics, and many others, to assess the strength and direction of relationships between variables. However, it's important to note that Pearson's correlation coefficient only measures linear relationships and may not capture non-linear associations. Additionally, correlation does not imply causation, so even if two variables are strongly correlated, it does not necessarily mean that changes in one variable cause changes in the other.

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

Ans:- Scaling is a crucial preprocessing step in data analysis and machine learning, ensuring that variables are on a comparable scale. It involves adjusting the range or distribution of variables so that they have similar magnitudes, which can facilitate comparison and improve the performance of certain algorithms. Scaling is particularly important when dealing with features that have different units of measurement or widely different ranges. It's vital for:

1. **Improving algorithm performance:** Scaling helps algorithms like SVM, KNN, and gradient descent-based models converge faster and perform better by ensuring features contribute equally.
2. **Facilitating interpretation:** It makes it easier to interpret feature coefficients in linear models, avoiding dominance by features with larger magnitudes.
3. **Enhancing visualization:** Scaling enables accurate visualizations, preventing overshadowing of features and facilitating clearer understanding of relationships.

There are two common types of scaling:

- **Normalized scaling:** Transforms data to a range typically between 0 and 1, preserving relative relationships between points.
- **Standardized scaling:** Gives data a mean of 0 and a standard deviation of 1, centering it around the mean.

In summary, normalized scaling adjusts the range, while standardized scaling adjusts the distribution. The choice depends on specific analysis or algorithm requirements.

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

(3 marks)

Ans:- The Variance Inflation Factor (VIF) can become infinite in multiple regression analysis due to perfect multicollinearity between a variable and others in the model. Multicollinearity occurs when independent variables are highly correlated. When perfect multicollinearity exists, one variable can be expressed exactly as a linear combination of others. In such cases, the VIF formula denominator becomes zero, resulting in an undefined or infinite VIF. Perfect multicollinearity affects parameter estimation and interpretation by making it impossible to estimate unique coefficients and inflating standard errors, leading to imprecise estimates and potentially misleading conclusions about variable relationships.

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 (quantile-quantile) plot is a graphical tool used to assess whether a set of data follows a particular distribution, such as the normal distribution. It compares the quantiles of the dataset against the quantiles of a theoretical distribution, typically the normal distribution, by plotting them against each other.

Here's how a Q-Q plot is constructed:

1. First, the data is sorted in ascending order.
2. Next, the quantiles of the dataset are calculated. These quantiles represent the percentage points in the dataset (e.g., the 25th percentile, 50th percentile, 75th percentile, etc.).
3. Similarly, quantiles for the theoretical distribution (e.g., the normal distribution) are calculated.
4. The quantiles of the dataset are plotted against the quantiles of the theoretical distribution. If the dataset follows the theoretical distribution, the points on the Q-Q plot will fall along a straight line.

The use and importance of a Q-Q plot in linear regression are as follows:

1. **Assumption checking**: It helps assess if residuals follow a normal distribution. Deviations from a straight line indicate departures from normality.
2. **Identifying outliers**: Outliers, appearing as points deviating from the expected line, can be spotted.
3. **Model assessment**: Q-Q plots visually evaluate the model's adequacy. Non-normal residuals suggest the model might not capture data relationships adequately, requiring adjustments or transformations.

In summary, Q-Q plots are valuable tools for assessing the assumptions of linear regression models, identifying outliers, and evaluating model adequacy. They provide insights into the distributional properties of the residuals and help ensure the validity and reliability of regression analysis results.