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?**

   Answer: Here are some of the inferences I made from my analysis of categorical variables from the dataset on the dependent variable (Count).

   - Fall has the highest median, which is expected as weather conditions are most optimal to ride bike followed by summer.
   - Median bike rents are increasing year on as year 2019 has a higher median then 2018, it might be due the fact that bike rentals are getting popular and people are becoming more aware about environment.
   - Overall spread in the month plot is reflection of season plot as fall months have higher median.
   - People rent more on non - holidays compared to holidays, so reason might be they prefer to spend time with family and use personal vehicle instead of bike-rentals.
   - Overall median across all days is same but spread for Thursday and Friday is bigger may be evident that those who have plans for Saturday might not rent bikes as it a non-working day.
   - People rented more bikes on working day rather than non-working day, so reason might be the weather is more clear on working day and they also prefer to spend time with family and use personal vehicle instead of bike-rentals.

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

   Answer: When creating dummy variables from categorical data, using drop_first=True is important for the following reasons:

   - Avoiding Multicollinearity (Dummy Variable Trap): Multicollinearity occurs when independent variables are highly correlated, causing issues in regression models. Including all categories as dummy variables creates perfect multicollinearity, known as the Dummy Variable Trap.

     Solution: drop_first=True drops one category, preventing this issue by removing the redundant variable.

   - Interpretability of Coefficients: Dropping the first category makes the coefficients of the remaining dummy variables interpretable as changes relative to the baseline or reference category. This relative interpretation often provides more meaningful insights.

   - Reduced Complexity: Dropping one category reduces the number of variables in the model, leading to a more parsimonious and interpretable model without losing any information.

Example:

For a categorical variable Color with categories Red, Blue, and Green:

- Without Dropping: All three dummies (Color_Red, Color_Blue, Color_Green) lead to multicollinearity.

	Blue	Green	Red
0	0	0	1
1	1	0	0
2	0	1	0
3	1	0	0
4	0	0	1

- With Dropping: Drop Red, leaving Color_Blue and Color_Green as dummies. The baseline is now Red.

	Blue	Green
0	0	0
1	1	0
2	0	1
3	1	0
4	0	0

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

Answer: There is linear relationship between temp and atemp as we observed from pairplot. 'temp' 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: By plotting the residuals distribution we came out a normal distribution by mean value is 0.

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

Answer: Based on the final model the top 3 features contributing significantly towards the demand of the shared bikes are:

a. Year : 0.2340
b. Temp : 0.4782
c. Winter : 0.0969

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

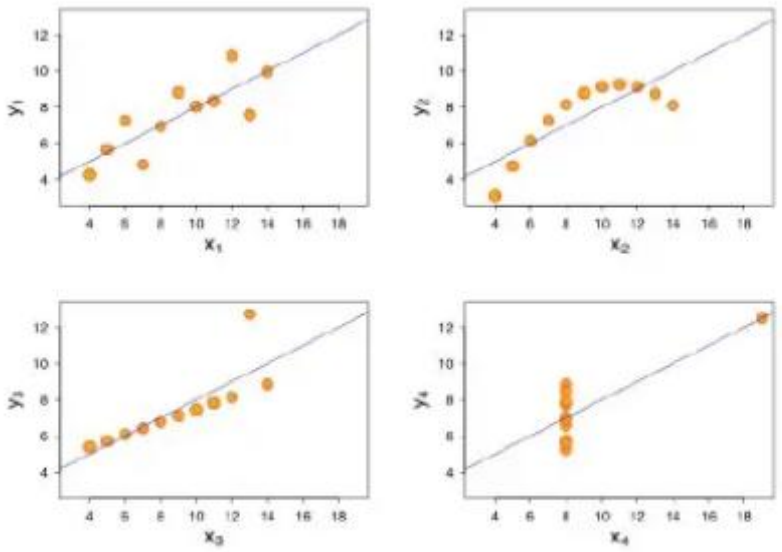
Answer: A linear regression algorithm tries to explain the relationship between independent and dependent variable using a straight line. It is applicable to numerical variables only. Following steps are performed while doing linear regression:

- The dataset is divided into test and training data
- Train data is divided into features (independent) and target (dependent) datasets
- A linear model is fitted using the training dataset. Internally the api's from python uses gradient descent algorithm to find the coefficients of the best fit line. The gradient descent algorithm works by minimising the cost function. A typical example of cost function is residual sum of squares.
- In case of multiple features, the predicted variable is a hyperplane instead of line.
- The predicted variable takes the following form:
  $Y = \beta0+\beta1x1+\beta2x2+\beta3x3++ \beta nxn$

a. The predicted variable is than compared with test data and assumptions are checked.

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

Answer: Anscombe's quartet comprises of four data sets that have nearly identical simple descriptive statistics but have quite different distribution when visualized graphically. The simple statistics consist of mean, sample variance of x and y, correlation coefficient, linear regression line and R-Square value. Anscombe's Quartet shows that multiple data sets with many similar statistical properties can still be vastly different from one another when graphed.



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

Answer: Pearson's R, also known as Pearson's correlation coefficient, is a statistical measure used to assess the strength and direction of the linear relationship between two continuous variables. It is one of the most common correlation coefficients used in statistics and data analysis. Key features of Pearson's R:

a. Value Range:
- The value of Pearson's R ranges from -1 to 1.
- +1 indicates a perfect positive linear correlation.
- -1 indicates a perfect negative linear correlation.
- 0 indicates no linear correlation between the variables.

b. Interpretation:
- Positive R (>0): As one variable increases, the other variable also tends to increase.
- Negative R (<0): As one variable increases, the other variable tends to decrease.
- Magnitude:
  - 0.1 to 0.3: Small or weak correlation.
  - 0.3 to 0.5: Medium or moderate correlation.
  - 0.5 to 1.0: Large or strong correlation.

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

   Answer: An infinite Variance Inflation Factor (VIF) typically arises from perfect multicollinearity, where one independent variable is an exact linear combination of others. This situation often occurs with duplicate variables, the dummy variable trap (when all categories are included without dropping one), or other forms of linear dependence. When the correlation between variables is perfect, the R2R^2R2 value for the variable becomes 1, causing the VIF to approach infinity. To resolve this, you can remove redundant variables, drop one dummy variable to avoid the trap, or use techniques like regularization or Principal Component Analysis (PCA) to mitigate multicollinearity.

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

   An infinite Variance Inflation Factor (VIF) occurs due to perfect multicollinearity among independent variables. This happens when one variable is an exact linear combination of others, making the model's design matrix singular. As a result, the R^2 value for the variable becomes 1, causing the VIF to approach infinity. This indicates redundancy among variables and can severely impact the stability and interpretability of regression coefficients. To address this, remove redundant variables or use regularization techniques to mitigate multicollinearity.

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 with a theoretical distribution, typically the normal distribution. It plots the quantiles of the dataset against the quantiles of the theoretical distribution.

   Use and Importance in Linear Regression
   Assess Normality: A Q-Q plot helps verify if the residuals (errors) of a linear regression model are normally distributed, which is a key assumption for valid hypothesis tests and confidence intervals.

   Model Diagnostics: Deviations from the diagonal line in the plot indicate departures from normality. This can signal problems such as skewness or kurtosis in the residuals, suggesting that the model may not be appropriate.