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)

   - Season: Bike rental demand exhibited a clear seasonal pattern, with the highest median observed during Fall (Season 3), indicating peak usage during this season. In contrast, demand was lowest during Spring (Season 1).
   - Year (Yr): There was an increase in bike rentals from 2018 to 2019, with higher user counts recorded in the latter year. This trend suggests a growing popularity or increased accessibility of bike rentals over time.
   - Holiday: Bike rentals showed a decrease during holidays, indicating that fewer people chose to rent bikes
   - Weekday: The demand for bike rentals remained relatively stable throughout the week, indicating consistent usage patterns regardless of the weekday.
   - Workingday: There was no significant difference in bike rental counts between working days and non-working days.
   - Weather situation (Weathersit): Bike rentals were significantly lower during heavy rain or snow, which are adverse weather conditions. In contrast, the highest rental counts were observed during clear or partly cloudy weather conditions.
- Month (Mnth): Bike rentals peaked in September. Conversely, rentals tended to be lower in December.

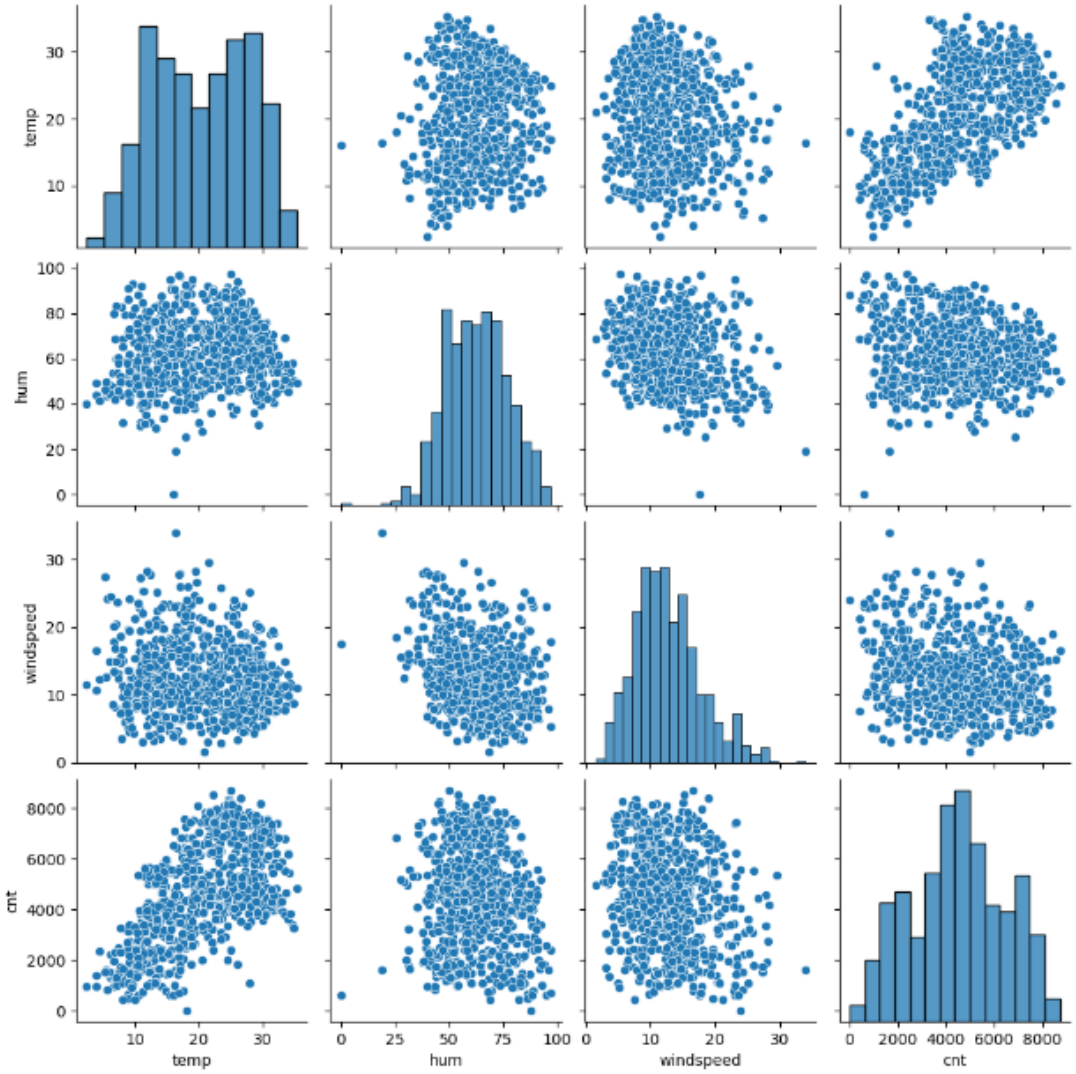
2. **Why is it important to use drop_first=True during dummy variable creation?**
   (2mark)

drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.
if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

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

The variable 'temp' exhibits the strongest correla>on 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.



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

   - Linear Relationship between Variables: We assessed the relationship between independent and dependent variables using a pairplot visualization of numeric variables. This helped us determine if the variables exhibit linear relationships, which is a key assumption for linear regression.

   - Normality of Residuals: To ensure the validity of residuals, we examined their distribution using a distplot. This allowed us to verify if the residuals follow a normal

distribution and are centered around zero (mean = 0). This is crucial as linear regression assumes that the errors (residuals) are normally distributed.

- Multicollinearity Check: We evaluated multicollinearity among independent variables using the Variance Inflation Factor (VIF). VIF quantifies how strongly the feature variables in our model are correlated with each other. High VIF values indicate strong multicollinearity, which can affect the reliability of coefficient estimates in linear regression.

By validating these assumptions—linear relationship, normality of residuals, and absence of multicollinearity—we ensure that the linear regression model is appropriate and reliable for making predictions based on our dataset.

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

cnt = 0.165227 + 0.487171 x{temp} + 0.239914 x{yr} + 0.068579 x{season_Winter} + 0.065961 x{mnth_September} + 0.051029 x{weekday_Sunday} + 0.047744 x{season_Summer} + 0.039219 x{workingday} - 0.034183 x{mnth_January} - 0.048140 x{holiday} - 0.051145 x{mnth_July} - 0.053503 x{season_Spring} - 0.067509 x{weathersit_Moderate} - 0.185137 x{windspeed}

The top 3 significant features are:

- Temperature (temp)
- Winter season (season_Winter)
- Calendar year (yr)

General Subjective Questions

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

Linear regression is a machine learning algorithm based on supervised learning. It performs a regression task, which means it predicts a continuous output variable (y) based on one or more input variables (x). It is mostly used for finding out the linear relationship between variables and forecasting.
The basic idea of linear regression is to find a line that best fits the data points, such that the distance between the line and the data points is minimized. The line can be represented by an equation of the form:

$y = \theta 0 + \theta 1x$
where θ0 is the intercept (the value of y when x is zero) and θ1 is the slope (the change in y for a unit change in x). These are called the parameters or coefficients of the linear model.

To find the best values of θ0 and θ1, we need to define a cost function that measures how well the line fits the data. A common choice is the mean squared error (MSE), which is the average of the squared differences between the actual y values and the predicted y values:

$MSE = (1/n) * \sum(y - y')^2$

where n is the number of data points, y is the actual value, and y' is the predicted value.

The goal is to minimize the MSE by adjusting θ0 and θ1. There are different methods to do this, such as gradient descent, normal equation, or using libraries like scikit-learn.

Linear regression can also be extended to multiple input variables (x1, x2, ..., xn), in which case the equation becomes:

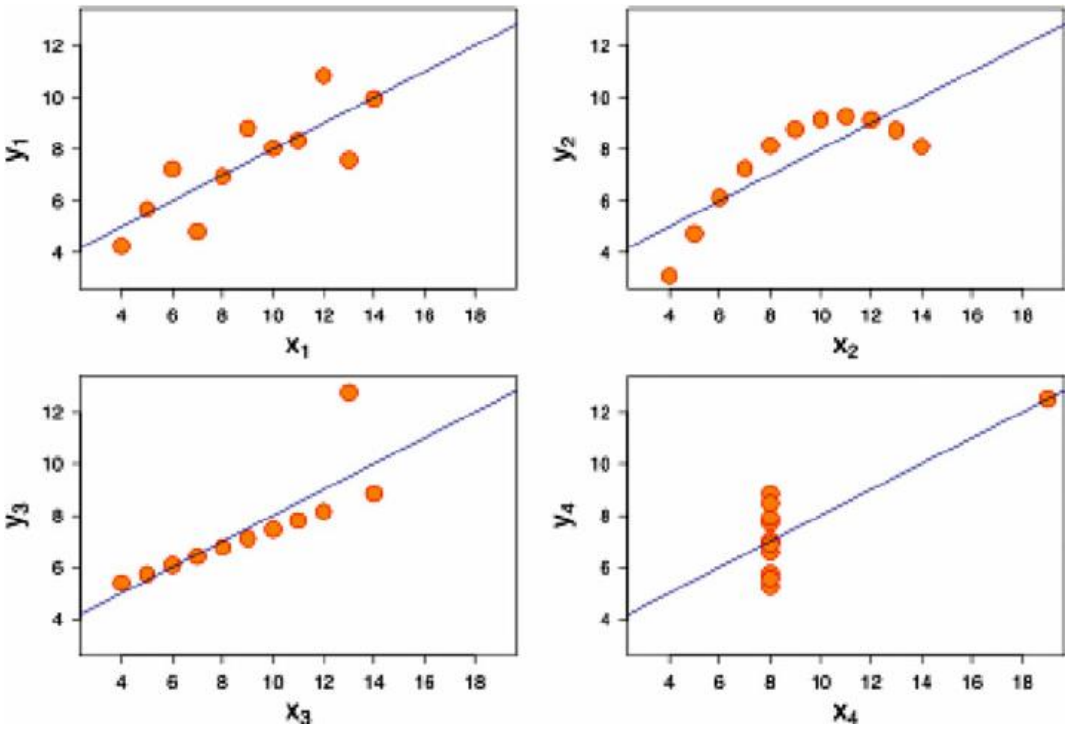
$y = \theta 0 + \theta 1x1 + \theta 2x2 + ... + \theta nxn$

Limitations are: it assumes a linear relationship between the input variables and the output variable, which may not always be the case. Another limitation is that it may be sensitive to outliers or multicollinearity.

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

Anscombe's Quartet was developed by statistician Francis Anscombe. It includes four data sets that have almost identical statistical features, but they have a very different distribution and look totally different when plotted on a graph. It was developed to emphasize both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties.

- The first scatter plot (top left) appears to be a simple linear relationship.
- The second graph (top right) is not distributed normally; while there is a relation between them is not linear.
- In the third graph (bottom left), the distribution is linear, but should have a different regression line the calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.



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

The Pearson's R (also known as Pearson's correlation coefficients) measures the strength between the different variables and the relation with each other. The Pearon's R returns values between -1 and 1. The interpretation of the coefficients are:

- • -1 coefficient indicates strong inversely proportional relationship.
- • 0 coefficient indicates no relationship.
- • 1 coefficient indicates strong proportional relationship.

$$r = \frac{n(\Sigma x * y) - (\Sigma x) * (\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2] * [n\Sigma y^2 - (\Sigma y)^2]}}$$

Where:
N = the number of pairs of scores
Σxy = the sum of the products of paired scores
Σx = the sum of x scores
Σy = the sum of y scores
Σx2 = the sum of squared x scores
Σy2 = the sum of squared y scores

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

(3 marks)

Feature scaling is a method used to normalize or standardize the range of independent variables or features of data. It is performed during the data pre-processing stage to deal with varying values in the dataset. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, irrespective of the units of the values.

● Normalization is generally used when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbours and Neural Networks.

● Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

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

(3 marks)

The VIF (Variance Inflation Factor) gives how much the variance of the coefficient estimate is being inflated by collinearity. If there is perfect correlation, then VIF = infinity. It gives a basic quantitative idea about how much the feature variables are correlated with each other. It is an extremely important parameter to test our linear model.

$$VIF = \frac{1}{1 - R^2}$$

Where R-1 is the R-square value of that independent variable which we want to check how well this independent variable is explained well by other independent variables. If that independent variable can be explained perfectly by other independent variables, then it will have perfect correlation and it's R-squared value will be equal to 1. So, VIF = 1/(1-1) which gives VIF = 1/0 which results in "infinity" The numerical value for VIF tells you (in decimal form) what percentage the variance (i.e. the standard error squared) is inflated for each coefficient. For example, a VIF of 1.9 tells you that the variance of a particular coefficient is 90% bigger than what you would expect if there was no multicollinearity — if there was no correlation with other predictors.

A rule of thumb for interpreting the variance inflation factor:

- 1 = not correlated.

- Between 1 and 5 = moderately correlated.
    - Greater than 5 = highly correlated.

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

(3 marks)

A Q-Q (Quantile-Quantile) plot is a graphical tool used to assess whether a dataset conforms to a specific theoretical distribution, such as the normal distribution. It compares the quantiles of the observed data against the quantiles of the expected theoretical distribution. If the points on the Q-Q plot approximately align along a straight line, it suggests that the data fits well with the chosen theoretical distribution.

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

1. **Normality Assessment**:
    - **Use**: In linear regression, it is often assumed that the residuals (the differences between observed and predicted values) follow a normal distribution. Q-Q plots are essential for verifying this assumption.
    - **Importance**: Deviations of residuals from normality can impact the reliability of statistical inferences derived from the regression model.
2. **Identifying Outliers**:
    - **Use**: Outliers within the residuals can be detected by observing points that deviate noticeably from the expected straight line on the Q-Q plot.
    - **Importance**: Outliers influence the estimation of model parameters and may indicate data points that are not well-captured by the regression model.
3. **Model Fit Assessment**:
    - **Use**: Q-Q plots visually assess how well the residuals conform to a normal distribution.
    - **Importance**: A good fit of the model is crucial for accurate predictions. Departures from normality in residuals may indicate deficiencies in the regression model.
4. **Validity of Statistical Tests**:
    - **Use**: Normality assumptions of residuals are important when conducting hypothesis tests or constructing confidence intervals.
    - **Importance**: Violations of normality assumptions can lead to inaccurate p-values and confidence intervals, compromising the validity of statistical inferences.

Interpretation of Q-Q Plots:

- A Q-Q plot where points closely follow a straight line suggests that residuals are approximately normally distributed.
- Deviations from the straight line indicate departures from normality, which may require further investigation or model refinement.

Q-Q plots serve as powerful diagnostic tools in linear regression for assessing the normality of residuals, detecting outliers, and ensuring the validity of statistical inferences. They offer an intuitive and visual approach to validating the assumptions underlying regression models.