Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer: The categorical variables in the dataset were "season", "year", "month", "holiday", "weekday", "workingday" and "weathersit". These were visualized using a boxplot. These variables had the following effect on our dependent variable.

- Season:
  - Fall: Bike demand is highest in Fall.
  - Summer and Winter: Summer and winter has intermediate value of count with summer having greater count among the two.
  - Spring: The demand for bikes is lowest in spring, possibly due to less favourable weather.
- Year:
  - 2019 vs. 2018: There is a clear increase in bike demand from the year 2018 to 2019. This trend suggests that the bike- sharing program gained popularity over this period.
- Month:
  - High-Demand Months: June, July, August and September are the months with the highest bike demand. Out of all these months, September has seen the highest no of rentals. This could be due to the warm summer weather, which is ideal for biking.
  - Low-Demand Months: December has seen the lowest no of rentals. January, February, and December see the less bike demand, likely due to colder winter weather, which discourages biking.
  - Holiday:
  - Holidays vs Non-holidays: Bike demand is higher on holidays compared to non-holidays. This increase can be attributed to people having more leisure time and choosing to bike for recreation or errands on holidays.
  - Weekday:
  - Even Distribution: Bike demand is relatively evenly distributed across all weekdays, indicating consistent usage throughout the week.
  - Slightly Higher on Fridays and Saturdays: There is a slight increase in bike usage on Fridays and Saturdays, though the difference is not very pronounced.
  - Weathersit:
  - Clear Weather: The highest bike demand occurs during clear weather conditions, due to favorable weather.
  - Adverse Weather: Bike demand decreases significantly during misty conditions, light snow/rain, and heavy snow/rain. There are no users when there is heavy snow/rain. The least demand is observed during light snow/rain, as adverse weather conditions make biking less appealing and potentially hazardous.

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

Answer: Using 'drop_first=True' during dummy variable creation is important for the following reasons:

- Preventing Multicollinearity: Including all dummy variables for a categorical feature can lead to multicollinearity, which occurs when predictor variables are highly correlated. This can make it difficult to determine the individual effect of each variable on the target variable. By dropping the first dummy variable, we avoid this issue and ensure that the remaining

dummies can provide the necessary information without redundancy.

- Reducing Redundancy: By dropping the first category, the total number of dummy variables is reduced, which simplifies the model and improves efficiency.
- Model Interpretability: This practice ensures that the model remains interpretable and free from redundant variables, making it easier to understand and analyze the effects of other predictors.

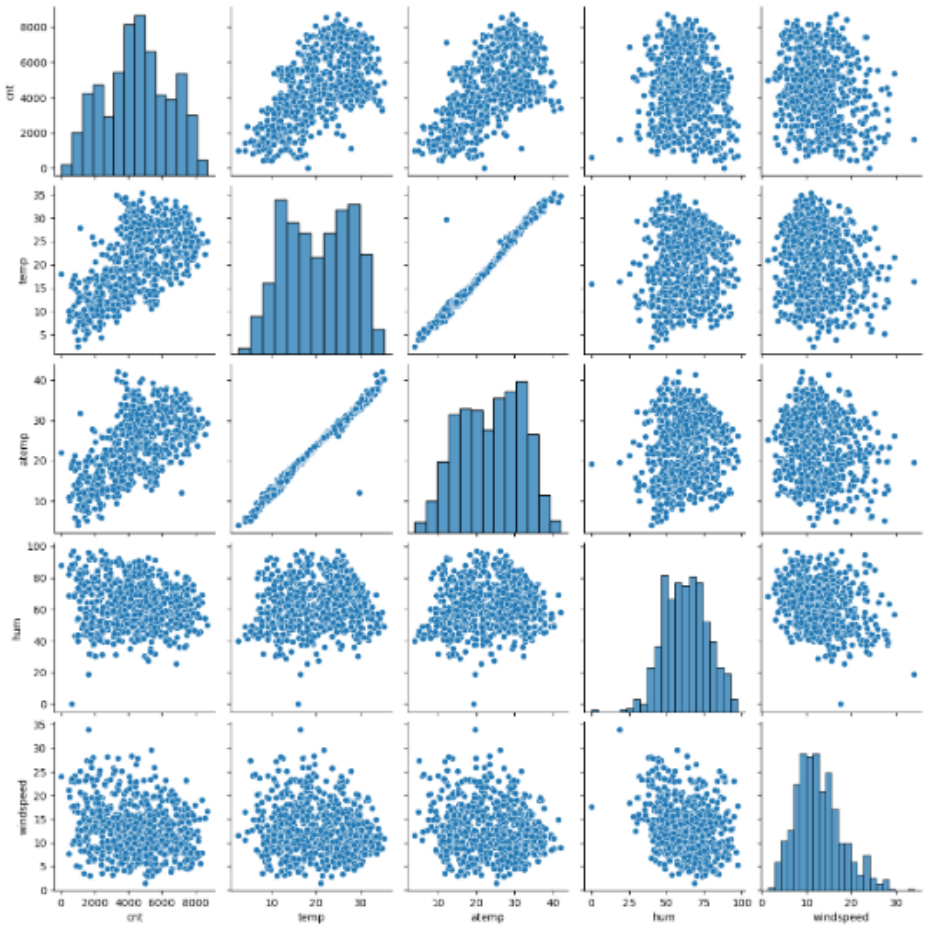
Syntax - drop_first: bool, default False, which implies whether to get n-1 dummies out of n categorical levels by removing the first level.

Example - Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not B and C, then it is obvious A. So, we do not need any extra variable or column to identify the A. Hence, we can drop first column A as it is redundant. So, for two dummy columns B and C the combination will be

A will be denoted by 00, B will be denoted by 10 and C will be represented by 01.

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

Answer: In the pair-plot analysis, the two temperature variables, "temp" and "atemp", show the highest correlation with the target variable "count" or "cnt". This strong positive correlation indicates that higher temperatures are associated with an increase in bike bookings.



Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer: Validated the assumption of Linear Regression Model based on below 5 assumptions

- Normality of Error Terms: If the residuals follow a normal distribution, the assumption is met.
  Histogram: Plotted a histogram of the residuals. If the residuals are normally distributed, the histogram should resemble a bell curve.
  Q-Q Plot: Plotted a Q-Q plot of the residuals. If the residuals are normally distributed, the points should lie along the 45-degreeline.
- Multicollinearity Check:
  Variance Inflation Factor (VIF): Calculated the VIF for each predictor variable. VIF values less than 10 indicate that multicollinearity is not a concern.
- Linear Relationship Validation:
  Residual Plot: Plotted residuals against the predicted values. If the residuals are randomly scattered around zero, it suggests that there is a linear relationship between the predictors and the response variable.
- Homoscedasticity:
  Residuals vs. Predicted Plot: Plotted residuals against the predicted values to check for constant variance. The absence of a clear pattern indicates homoscedasticity.
- Independence of residuals:
  Durbin-Watson Test: Have calculated and checked the Durbin-Watson statistic to detect autocorrelation in the residuals.

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

Answer: Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes.

1. Temperature - coefficient: (0.472207)
   - Higher temperatures are associated with increased bike usage. This positive coefficient suggests that as the temperature increases, the demand for bikes tends to rise.
2. weathersit - coefficient: Light Snow, Light Rain + Mist & Cloudy (-0.290800)
   - Adverse weather conditions, such as light snow or rain, discourage bike usage, negatively impacting the demand for shared bikes. The negative coefficient indicates that these weather conditions are likely to reduce bike usage.
3. year - coefficient: (0.234461)
   - The year 2019 seems to be a strong predictor, indicating an increasing trend in bike usage over time. This positive coefficient suggests that as the years progress, bike usage has shown an upward trend, especially in 2019.

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail.

Answer: 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:

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

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.

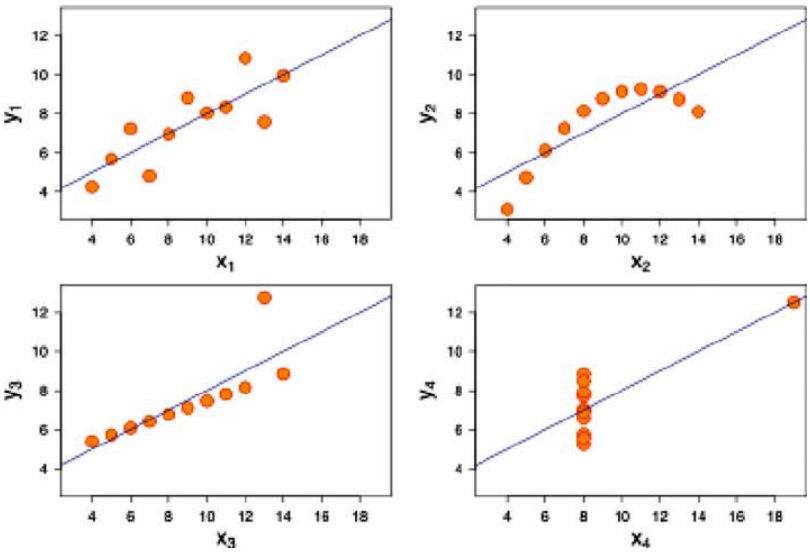
Question 7. Explain the Anscombe's quartet in detail.

Answer: 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

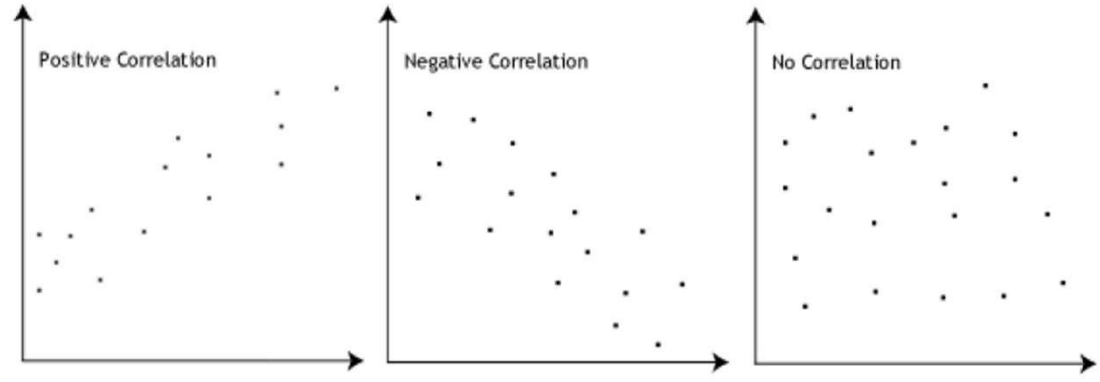
correlation coefficient from 1 to 0.816.

● 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.



Question 8. 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:



- Where, r = 1 means the data is perfectly linear with a positive slope
- r = -1 means the data is perfectly linear with a negative slope
- r = 0 means there is no linear association

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer: 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 Neighbors 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.

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Answer: 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 its 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.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Answer: The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

Use of Q-Q plot:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions. If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line $y = x$

If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line $y = x$. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.