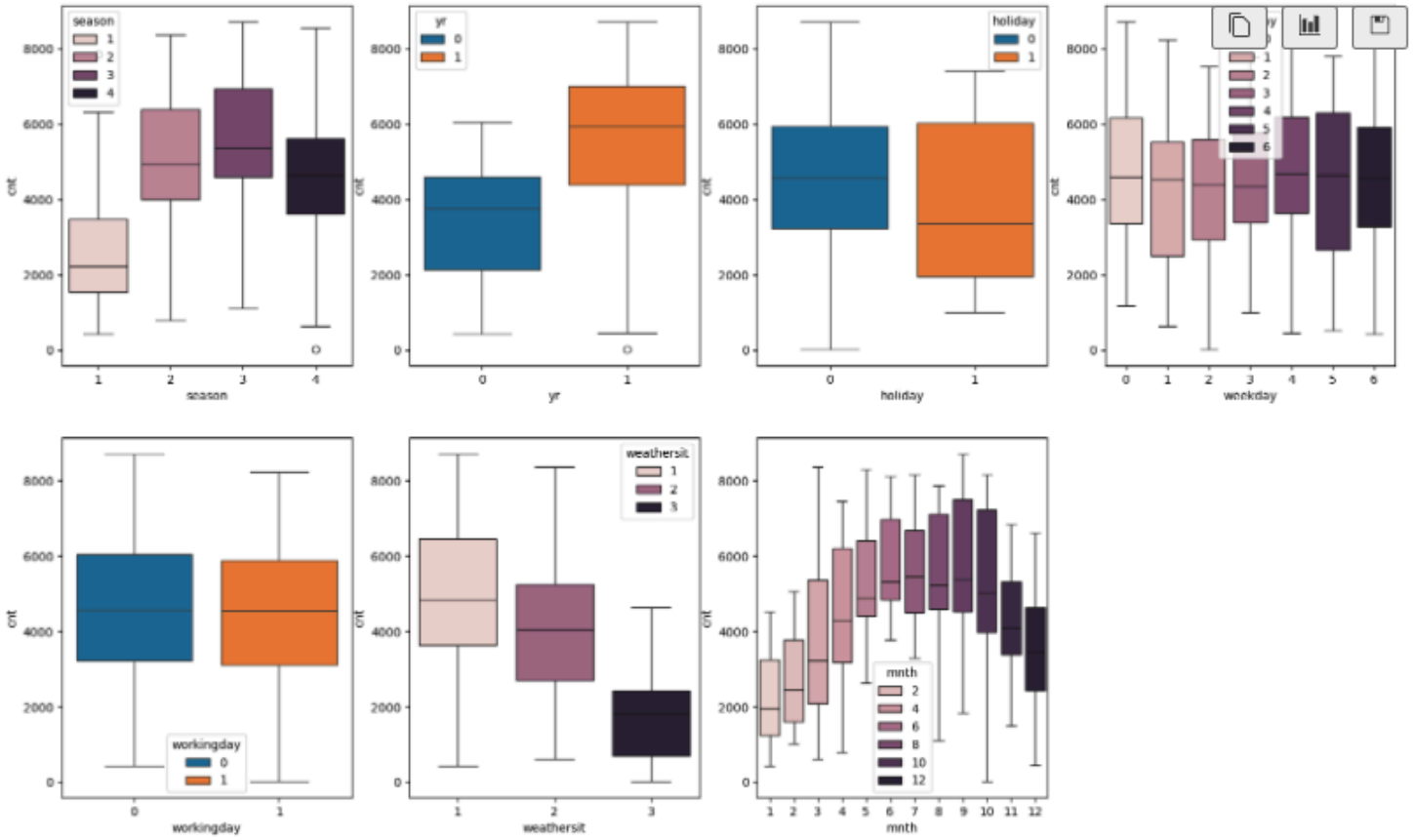
Assignment submission by RITWIK RAJ

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 :

Below is the analysis of categorical variables vs the target variable ‘cnt’ (number of bikes issued) from the dataset done with the help of boxplots .



The inference that can be drawn about their effect on the dependent variable is :

1. **Season** :
   The summer, fall and winter seasons show a higher affiliation for bikes being rented .

   - Summer and Fall generally offer favourable weather for outdoor activities . Comfortable temperatures, longer daylight hours, and pleasant conditions make biking more enjoyable.
   - These seasons often coincide with vacation periods, leading to an increase in tourism and leisure activities, including biking.
   - Seasonal Events and Festivals : Becomes a mode of transport for transportation or recreation .

2. **Month :**
   The trend seems to be at its peak during the summer / fall months , gradually starting to decline during the winter months . Follows the same trend as the “Season” feature .

3. **Holidays :**
   The demand increases when it's a holiday . Considering that people rent bikes for leisure , it might be the reason behind the increase in bike rentals .

4. **Weekday :**
   Nothing definitive can be deciphered with the distribution from the weekdays chart .

5. **Weather Situation :**
   As can be seen from the chart , people prefer to ride in clear/cloudy weather conditions , and refrain from doing the same during turbulent / rainy conditions .

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

Answer :

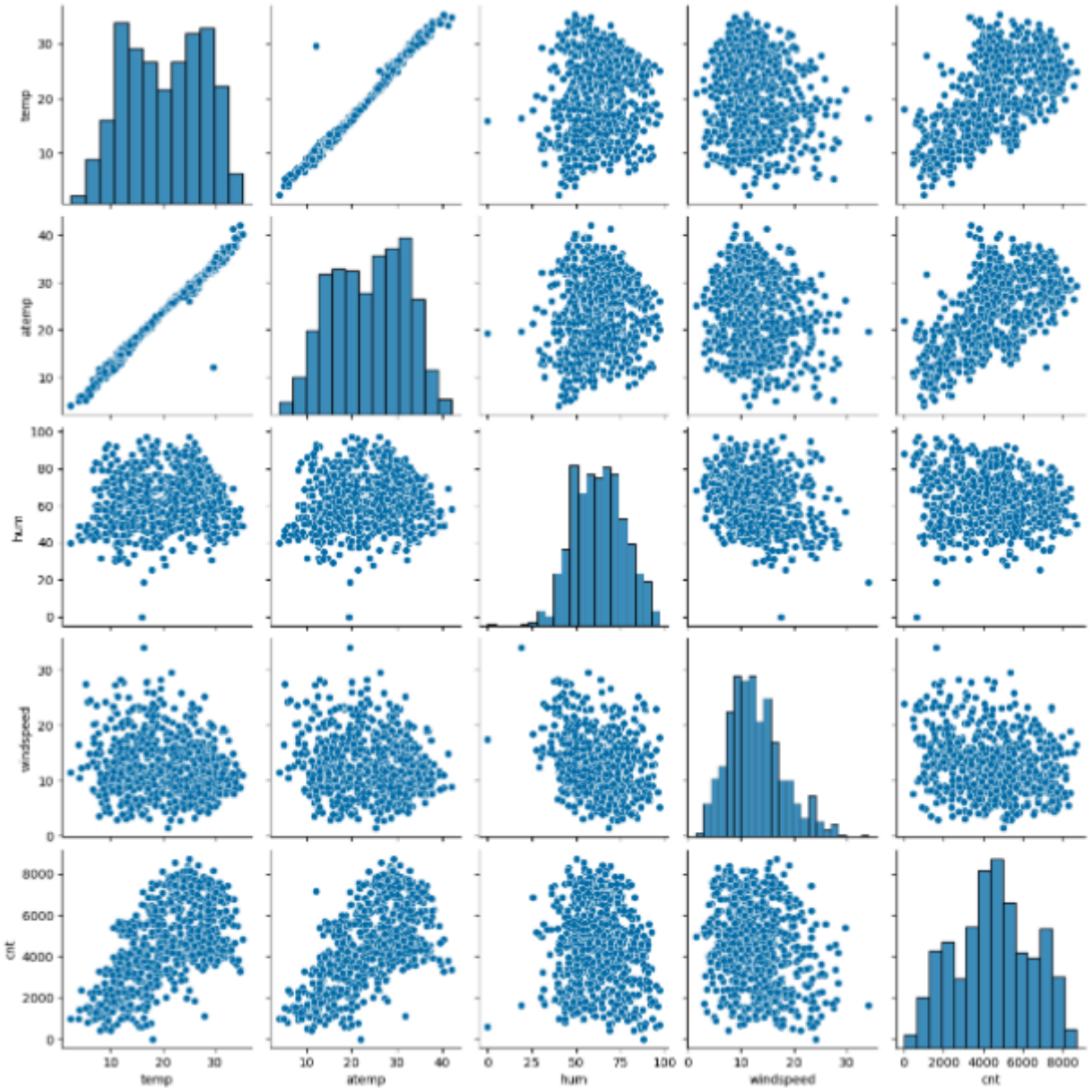
Using `drop_first=True` during dummy variable creation is important because it helps prevent multicollinearity in regression models, specifically the issue known as the "**dummy variable trap**."

- **Dummy Variable Trap** : When you create dummy variables for a categorical feature, you generate a binary variable (0 or 1) for each category. If you include all these dummy variables in your regression model, one of them will be perfectly predictable from the others, leading to multicollinearity. This makes the model parameters difficult to interpret and can mislead us from the statistical significance of the features .

- **Dropping the First Category** : By setting `drop_first=True`, you drop one of the dummy variables (usually the first one). This approach avoids multicollinearity while still allowing the model to capture the influence of the categorical feature.

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

Answer :




```
df[['temp','atemp','hum','windspeed','cnt']].corr()['cnt'].sort_values(ascending=False)
```

✓ 0.0s

```
cnt          1.000000
atemp        0.630685
temp         0.627044
hum         -0.098543
windspeed   -0.235132
Name: cnt, dtype: float64
```

'temp'(temperature)(~0.63) and 'atemp'(adjusted temperature)(~0.63) show the highest correlation with the target variable .

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

Answer :

To validate the assumptions of Linear Regression after building the model on the training set, we follow these steps:

1. **Linearity**: The residuals should show no discernible patterns, indicating a linear relationship between the independent and dependent variables.
2. **Independence of residuals:** Residuals should be independent, with no significant autocorrelation.
3. **Homoscedasticity**: The residuals should display constant variance across all levels of predicted values.
4. **Normality of Residuals**: The residuals should be approximately normally distributed
5. **Multicollinearity** : VIF values should generally be below 5, indicating low multicollinearity. High VIF values point to potential multicollinearity, which can distort the estimation of coefficients.

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

Answer :

The top 3 features :

- Temperature : The temperature plays a big role in the demand for rental bikes.
- Year : Although the dataset contained data from only 2 years , it exhibited a distinctive pattern .
- Season : Weather situation plays a big role as well , with people's preference being distinctively clear .

GENERAL SUBJECTIVE QUESTIONS

Question 1 : Explain the linear regression algorithm in detail.

Answer :

Linear regression is a supervised machine learning technique that models the relationship between a dependent variable and one or more independent features by fitting a linear equation to observed data.

- **Simple Linear Regression** is used when there is a single independent feature. It models the relationship with a straight line.

  The equation for simple linear regression is:
  $y=\beta 0 + \beta 1X$
  Where:

- Y is the dependent variable
- X is the independent variable
- β0 is the intercept
- β1 is the slope

- **Multiple Linear Regression** involves multiple independent features and models the relationship with a hyperplane in multidimensional space.

  The equation for multiple linear regression is:

  y=β0+β1X1+β2X2+.....βnXn

  Where :
  - Y is the dependent variable
  - X1, X2, ..., Xn are the independent variables
  - β0 is the intercept
  - β1, β2, ..., βn are the slopes

- Additionally :

  - **Univariate Linear Regression** refers to the case where there is one dependent variable and one or more independent variables.

  - **Multivariate Regression** involves multiple dependent variables being predicted from one or more independent variables.

Question 2 : Explain the Anscombe's quartet in detail.

Answer :

Anscombe's Quartet is a famous set of four datasets that were constructed by the statistician Francis Anscombe in 1973 to demonstrate the importance of graphical analysis in statistics.

The quartet illustrates how datasets with identical statistical properties can have very different distributions and relationships when visualised.

This underscores the idea that descriptive statistics alone may not fully capture the characteristics of the data.

Anscombe's Quartet consists of four datasets, each with the following properties:

- Mean of x values
- Mean of y values
- Variance of x values
- Variance of y values
- Correlation coefficient between x and y
- Regression line (least squares fit)

For each dataset in the quartet:

- Mean of x values: 9.0
- Mean of y values: 7.5
- Variance of x values: 11.0
- Variance of y values: 4.12
- Correlation coefficient between x and y: 0.816
- Regression line equation: y=3+0.5x

Despite having the same statistical summary, the datasets exhibit different patterns when plotted.

Datasets in Anscombe's Quartet

1. Dataset I (Linear Relationship):

- **Description**: This dataset shows a linear relationship between x and y. The data points form a straight line when plotted.
- **Graph**: A scatter plot with a clear linear trend.

2. Dataset II (Nonlinear Relationship):

- **Description**: This dataset also has the same statistical properties as the first but shows a curved relationship. The points follow a parabolic pattern.
- **Graph**: A scatter plot where data points follow a curve rather than a straight line.

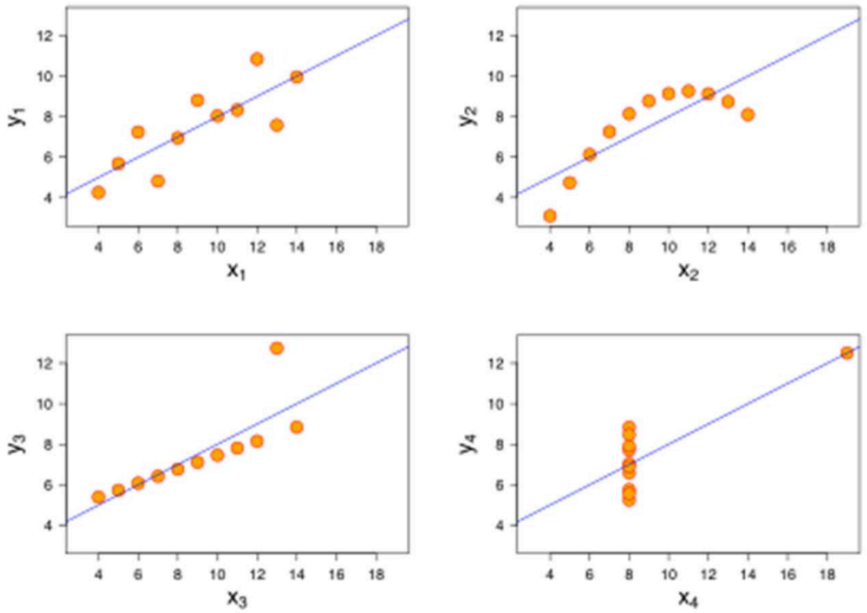
3. Dataset III (Outliers):

- **Description**: This dataset contains a single outlier that significantly affects the distribution. All other data points form a linear relationship, but the outlier skews the results.
- **Graph**: A scatter plot with a prominent outlier that disrupts the linear trend.

4. Dataset IV (Horizontal Line with Outlier):

- **Description**: This dataset has a strong vertical line with a single outlier. The outlier affects the slope of the regression line significantly, while most points are clustered horizontally.
- **Graph**: A scatter plot where most data points lie on a horizontal line with one significant outlier affecting the regression.

The four datasets compose Anscombe's quartet. All four sets have identical statistical parameters, but the graphs show them to be considerably different



The four datasets compose Anscombe's quartet. All four sets have identical statistical parameters, but the graphs show them to be considerably different

Question 3 : What is Pearson's R?

Answer :

Pearson's R, also known as the **Pearson correlation coefficient,** is a measure of the linear relationship between two continuous variables. It quantifies the degree to which the variables are related and provides insight into the strength and direction of their relationship.

Mathematically, it is given by:

$$r = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

Where:

- cov(X,Y) is the covariance between the variables X and Y.
- σX is the standard deviation of X.
- σY is the standard deviation of Y.

The formula for Pearson's R is :

$$r = \frac{n\sum(X_i Y_i) - \sum X_i \sum Y_i}{\sqrt{[n\sum(X_i^2) - (\sum X_i)^2][n\sum(Y_i^2) - (\sum Y_i)^2]}}$$

Where:

- n is the number of data points.
- Xi and Yi are individual data points of variables X and Y respectively.

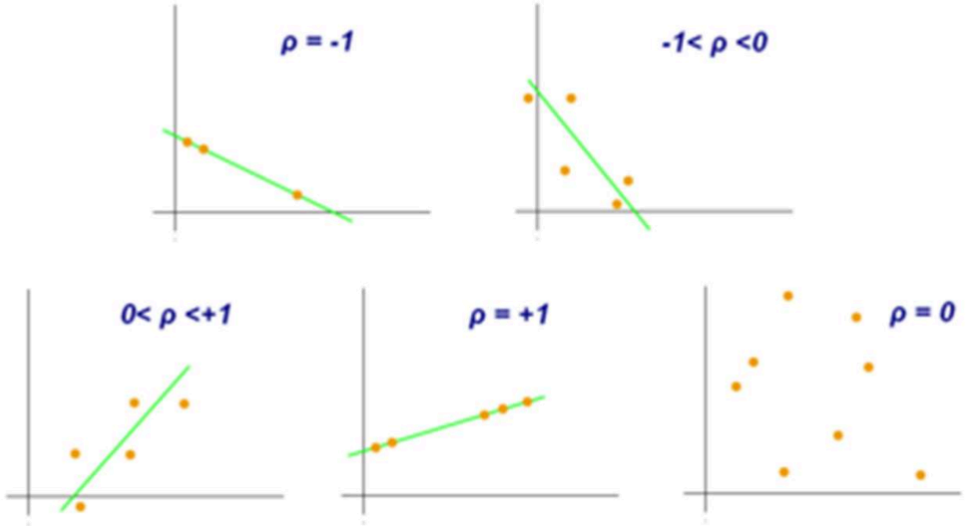
Pearson's R ranges from -1 to 1:

- r = 1: Perfect positive linear relationship.

- r = -1: Perfect negative linear relationship.
- r = 0: No linear relationship.

Interpretation:

- **Positive Correlation**: A positive Pearson's R value indicates that as one variable increases, the other variable also tends to increase.
- **Negative Correlation**: A negative Pearson's R value indicates that as one variable increases, the other variable tends to decrease.
- **Magnitude**: The closer the absolute value of Pearson's R is to 1, the stronger the linear relationship between the variables.



Examples of scatter diagrams with different values of correlation coefficient (ρ)

Question 4 : What is scaling? Why is scaling performed? What is the difference between normalised scaling and standardised scaling?

Answer :

Scaling refers to the process of adjusting the range and distribution of features (variables) in a dataset. This is typically done to ensure that different features contribute equally to the analysis or model, especially when they are on different scales or units.

Scaling is performed for several reasons:

1. Improve Model Performance:
    - Many machine learning algorithms, especially those that rely on distance metrics (e.g., k- nearest neighbours, support vector machines) or gradient-based optimization (e.g., linear regression, logistic regression), perform better when features are on similar scales.
    - Algorithms may converge faster and perform more accurately when features are scaled properly.

2 . Ensure Equal Weight:
- Features with larger ranges or different units can disproportionately influence the model's performance. Scaling ensures that each feature contributes equally to the model.

3 . Normalise Data Distribution:
- Scaling can make the data distribution more uniform, which helps in meeting the assumptions of some statistical methods and models.

4 . Handle Different Units:
- When features are measured in different units (e.g., height in cm and weight in kg), scaling ensures that the model interprets these features on a comparable scale.

Difference between normalised scaling and standardised scaling

Normalised scaling	Standardised scaling
Scales data to a specific range, typically [0, 1]. The result is that the transformed features have minimum and maximum values defined by this range.	Transforms data to have a mean of 0 and a standard deviation of 1. There are no fixed bounds; the transformed values can be any real number.
Does not alter the distribution shape; it simply rescales the data within a specific range.	Centres the data around the mean and scales it according to the standard deviation, making the data follow a standard normal distribution if the original data was normally distributed.
Useful when you need bounded data or when features have different units and need to be scaled to a common range.	Useful for algorithms that assume normality or when you want to compare features on a common scale regardless of their original distribution
Sensitive to outliers because outliers can skew the minimum and maximum values.	More robust to outliers compared to normalisation because it is based on mean and standard deviation.

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

Answer :

A Variance Inflation Factor value can become infinite when there is perfect multicollinearity in your data.

Perfect multicollinearity occurs when one predictor variable in a regression model is an exact linear combination of one or more of the other predictor variables. This means that the variable can be perfectly predicted by the others.

How does it happen ?
- VIF measures how much the variance of a regression coefficient is inflated due to multicollinearity. It's calculated as

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

where R^2 is the coefficient of determination obtained when a predictor is regressed against all other predictors in the model.

- If there is perfect multicollinearity, R^2 for the regression of one predictor on the others will be 1. When $R^2 = 1$, the denominator of the VIF formula $(1–R^2)$ becomes zero, making the VIF value **infinite**.

The consequences for this is that an infinite VIF indicates that the regression model has perfect multicollinearity, meaning one variable is completely redundant as it doesn't provide any new information beyond what is already captured by the other variables.

To address this issue, you typically need to remove one of the perfectly collinear variables or combine them into a single predictor to reduce the multicollinearity and obtain a finite VIF value.

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

Answer :

A Q-Q plot is a scatter plot where the quantiles of a dataset are plotted against the quantiles of a theoretical distribution. For a normal Q-Q plot, the quantiles of the data are plotted against the quantiles of a standard normal distribution.

In the context of linear regression, a Q-Q plot is crucial for validating assumptions, particularly the normality of residuals. Here's why it is important:

1. **Assess Normality of Residuals**:
   - Assumption Check: Linear regression assumes that the residuals (errors) of the model are normally distributed. This assumption is important for making valid inferences and constructing confidence intervals for predictions.

2. **Detect Deviations from Normality**:
   - Straight Line: If the residuals follow a normal distribution, the points on the Q-Q plot will lie approximately along the 45-degree reference line.

3. **Identify Outliers and Influential Points**:
   - Outliers: Points that deviate significantly from the line can highlight outliers or influential data points that might affect the regression model.

4. **Enhance Model Interpretation**:
   - Residuals Distribution: Understanding the distribution of residuals helps in assessing the quality of the model fit and in interpreting the results more accurately.