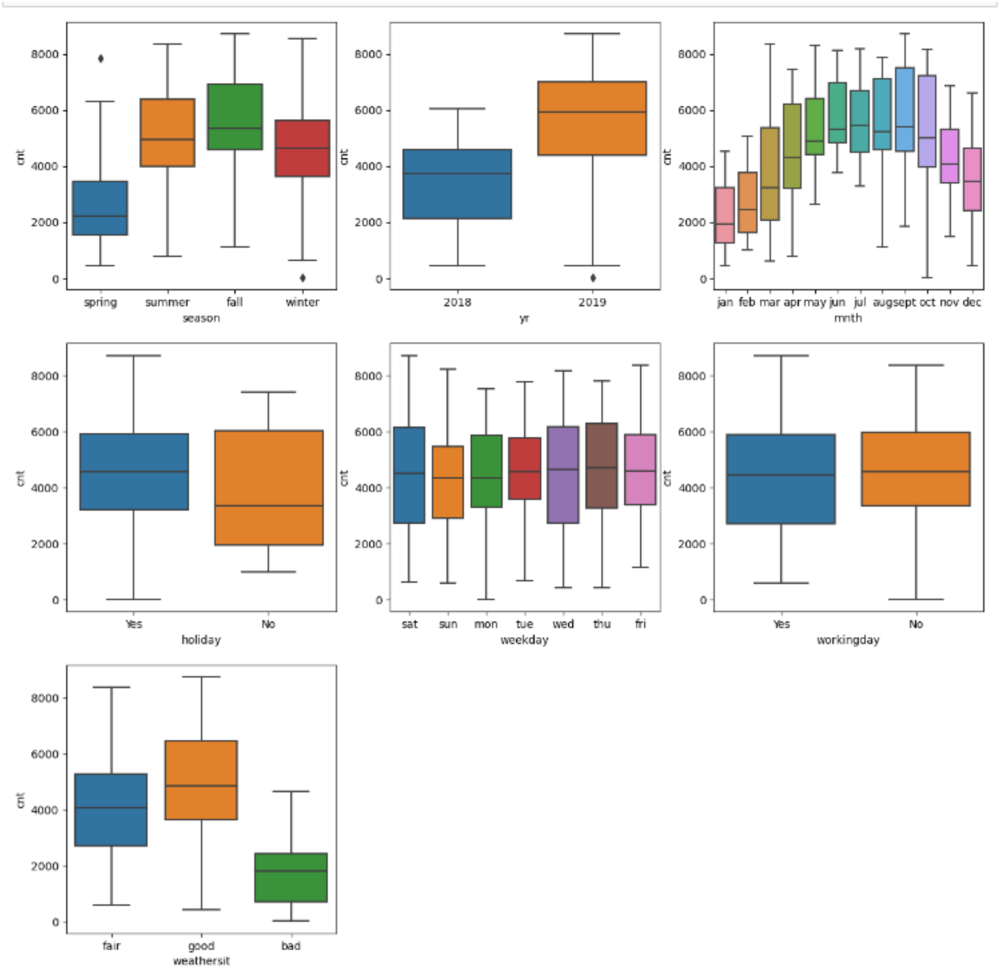
Assignment-based Subjective Questions

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)**



From the above analysis we can conclude that:

1. **Season "fall" has highest demand for rental bikes**
2. **Demand has increased in the year 2019 compared to 2018**
3. **Demand continuusly increases till July after that the demand decreases**
4. **When there is a holiday, demand has decreased.**
5. **The demand does not vary much with the day of the week or**
6. **The clear weathershit has highest demand**
7. **During September, bike sharing is more. During the year end and beginning, it is less, could be due to extereme weather conditions.**

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

Using drop_first=True during dummy variable creation is important to avoid the "dummy variable trap". Its important to drop first column because:

1. Multicollinearity: Without dropping a column, you introduce perfect multicollinearity.

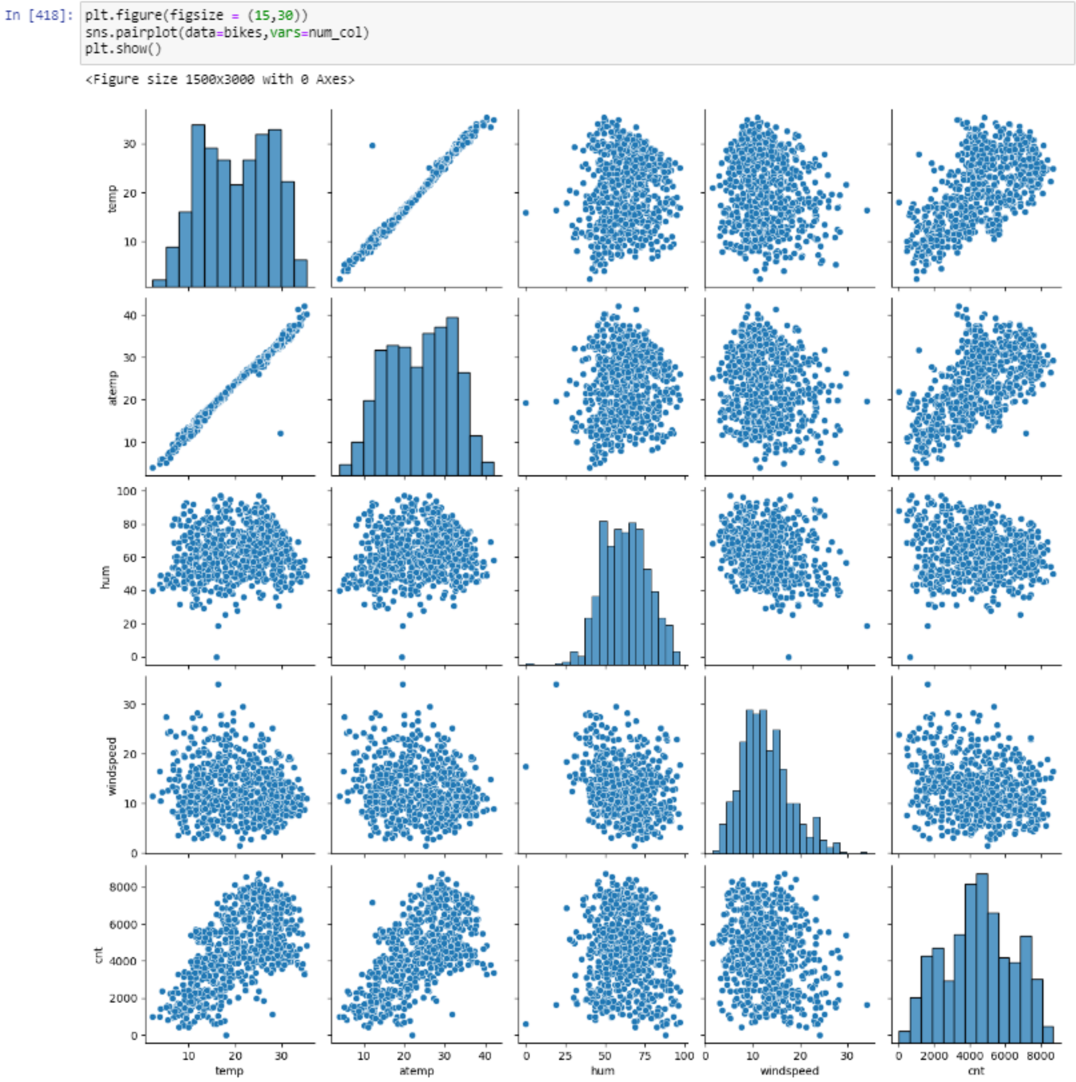
2. Redundancy: The dropped column's information is already represented by the others.

3. Model stability: Prevents singularity issues in matrix operations.

4. Degrees of freedom: Preserves the correct degrees of freedom in statistical tests.

5. Interpretation: Makes interpretation of coefficients more straightforward.

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



```
In [418]: plt.figure(figsize = (15,30))
          sns.pairplot(data=bikes,vars=num_col)
          plt.show()
```

```
<Figure size 1500x3000 with 0 Axes>
```

Looking at the pairplot we can say that the variable ‘temp’ has the highest correlation

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

To validate the assumptions of Linear Regression after building the model on the training set, I checked the following:

1. Residual terms are normally distributed around zero

2. Independence: The error terms are independent of each other

3. Homoscedasticity: Error terms have the same variance throughout

4. No multicollinearity: Calculate Variance Inflation Factor (VIF).

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

The top three features after :

1. temp - coefficient : 4390.8184

2. yr_2019 : 2085.8515

3. windspeed - coefficient : -1558.7046

General Subjective Questions

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

Linear regression is a statistical method used to model the relationship between a dependent variable (Y) and one or more independent variables (X). The algorithm aims to find the best-fitting linear equation of the form:

$$Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + ... + \beta_nX_n + \varepsilon$$

Where:

- Y is the dependent variable

- $X_1, X_2, ..., X_n$ are independent variables

- β_0 is the y-intercept

- $\beta_1, \beta_2, ..., \beta_n$ are the coefficients

- ε is the error term

The algorithm works as follows:

a) Initialize coefficients (usually to 0 or random small values).

b) For each data point, calculate the predicted Y using current coefficients.

c) Calculate the error (difference between predicted and actual Y).

d) Update coefficients to minimize the error, typically using methods like:

- Ordinary Least Squares (OLS)

- Gradient Descent

e) Repeat steps b-d until convergence (error is minimized or change in error is very small).

Example:

For simple linear regression (one independent variable):

$Y = \beta_0 + \beta_1 X$

The OLS formulas for β_1 and β_0 are:

$\beta_1 = \Sigma((x - \bar{x})(y - \bar{y})) / \Sigma((x - \bar{x})^2)$

$\beta_0 = \bar{y} - \beta_1\bar{x}$

Where $\bar{x}$ and $\bar{y}$ are the means of X and Y respectively.

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

Anscombe's Quartet is a set of four datasets created by statistician Francis Anscombe in 1973. Each dataset consists of 11 (x,y) points and has nearly identical simple statistical properties, but they appear very different when graphed.

Key points:

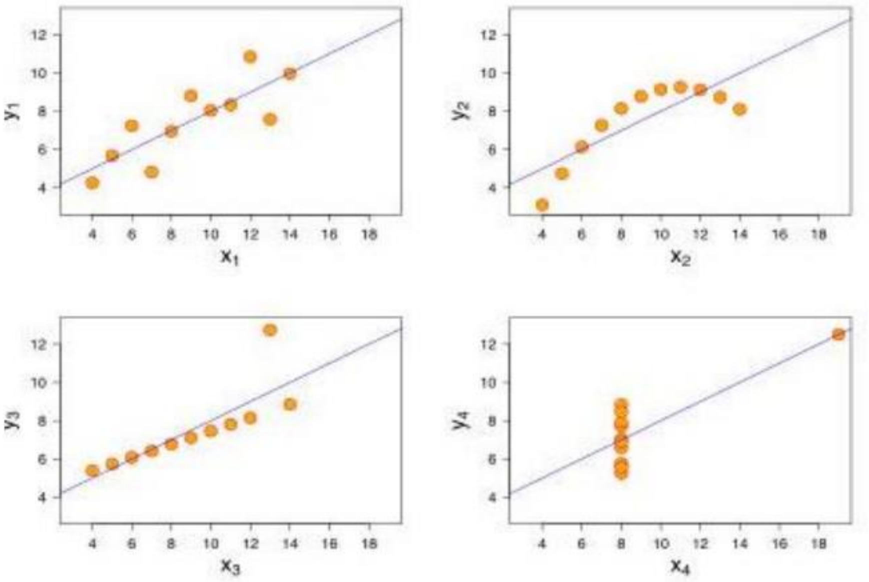
- All four datasets have the same mean of x (9.0) and y (7.5).

- They have the same variance of x (11.0) and y (4.1).

- They all have the same correlation coefficient (0.816).

- They all have the same linear regression line (y = 3 + 0.5x).

However, when plotted:



- Dataset 1: Shows a typical linear relationship.

- Dataset 2: Shows a clear non-linear relationship.

- Dataset 3: Shows a perfect linear relationship except for one outlier.

- Dataset 4: Shows a case where one outlier determines the regression line.

The importance of Anscombe's Quartet:

- It demonstrates the importance of visualizing data before analysis.

- It shows that summary statistics alone can be misleading.

- It emphasizes the need to check assumptions in regression analysis.

3. What is Pearson’s R? (3 marks)

Pearson's R, also known as the Pearson correlation coefficient, is a measure of linear correlation between two variables. It ranges from -1 to +1, where:

- +1 indicates a perfect positive linear correlation

- 0 indicates no linear correlation

- -1 indicates a perfect negative linear correlation

Formula:

$R = \Sigma((x - \bar{x})(y - \bar{y})) / \sqrt{(\Sigma(x - \bar{x})^2 * \Sigma(y - \bar{y})^2)}$

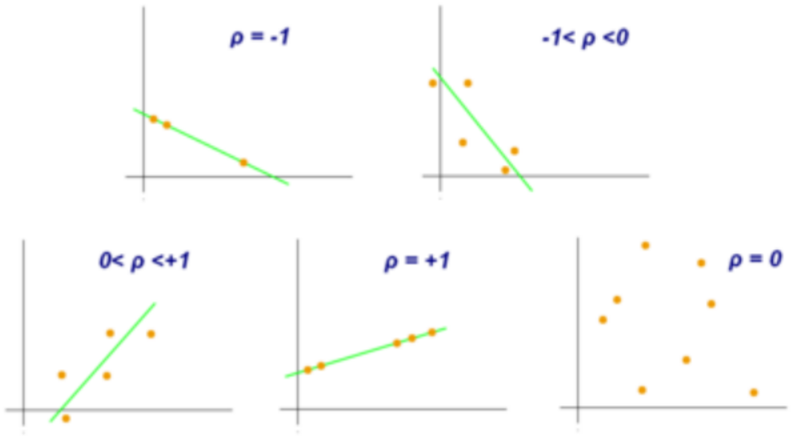
Where $\bar{x}$ and $\bar{y}$ are the means of X and Y respectively.

Interpretation:

- $|R| < 0.3$: Weak correlation

- $0.3 \leq |R| < 0.7$: Moderate correlation

- $|R| \geq 0.7$: Strong correlation



Pearson's R is used to assess the strength and direction of the linear relationship between two continuous variables.

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

Scaling is the process of transforming data to fit within a specific range. It's performed to:

- Ensure all features contribute equally to the model

- Improve convergence speed for gradient-based algorithms

- Prevent features with larger magnitudes from dominating the model

Two common types of scaling:

a) Normalized Scaling (Min-Max Scaling):

Scales data to a fixed range, typically [0, 1].

Formula: X_scaled = (X - X_min) / (X_max - X_min)

b) Standardized Scaling (Z-score Normalization):

Transforms data to have a mean of 0 and standard deviation of 1.

Formula: X_scaled = (X - μ) / σ

Where μ is the mean and σ is the standard deviation.

Differences:

- Normalized scaling bounds the data, while standardized scaling doesn't.

- Standardized scaling is less affected by outliers.

- Normalized scaling preserves zero values in sparse data.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

VIF (Variance Inflation Factor) becomes infinite when there's perfect multicollinearity between independent variables. This happens when:

- One variable is an exact linear combination of others.

- There's a perfect correlation between two or more variables.

Example:

Consider variables X_1, X_2, and X_3, where $X_3 = 2X_1 + 3X_2$

The VIF for X_3 would be infinite because it's perfectly predicted by X_1 and X_2.

Implications:

- The model becomes unstable and unreliable.

- Coefficient estimates become highly sensitive to small changes in the model.

- Standard errors of the coefficients inflate, making it difficult to interpret their significance.

To resolve, remove or combine the perfectly correlated variables.

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 follows a particular theoretical distribution, often the normal distribution.

How it works:

- The quantiles of the observed data are plotted against the quantiles of the theoretical distribution.

- If the points roughly follow a straight line, it suggests the data follows the theoretical distribution.

Use in Linear Regression:

- To check the normality assumption of residuals.

- Plot the standardized residuals against theoretical quantiles of a normal distribution.

Importance:

- Helps identify deviations from normality, such as skewness or heavy tails.

- Can reveal outliers or heteroscedasticity.

- Assists in validating the assumptions of linear regression, ensuring the model's reliability and the validity of statistical inferences.

Example interpretation:

- Straight line: Residuals are normally distributed.

- S-shaped curve: Residuals are skewed.

- Curve at the ends: Residuals have heavy tails.