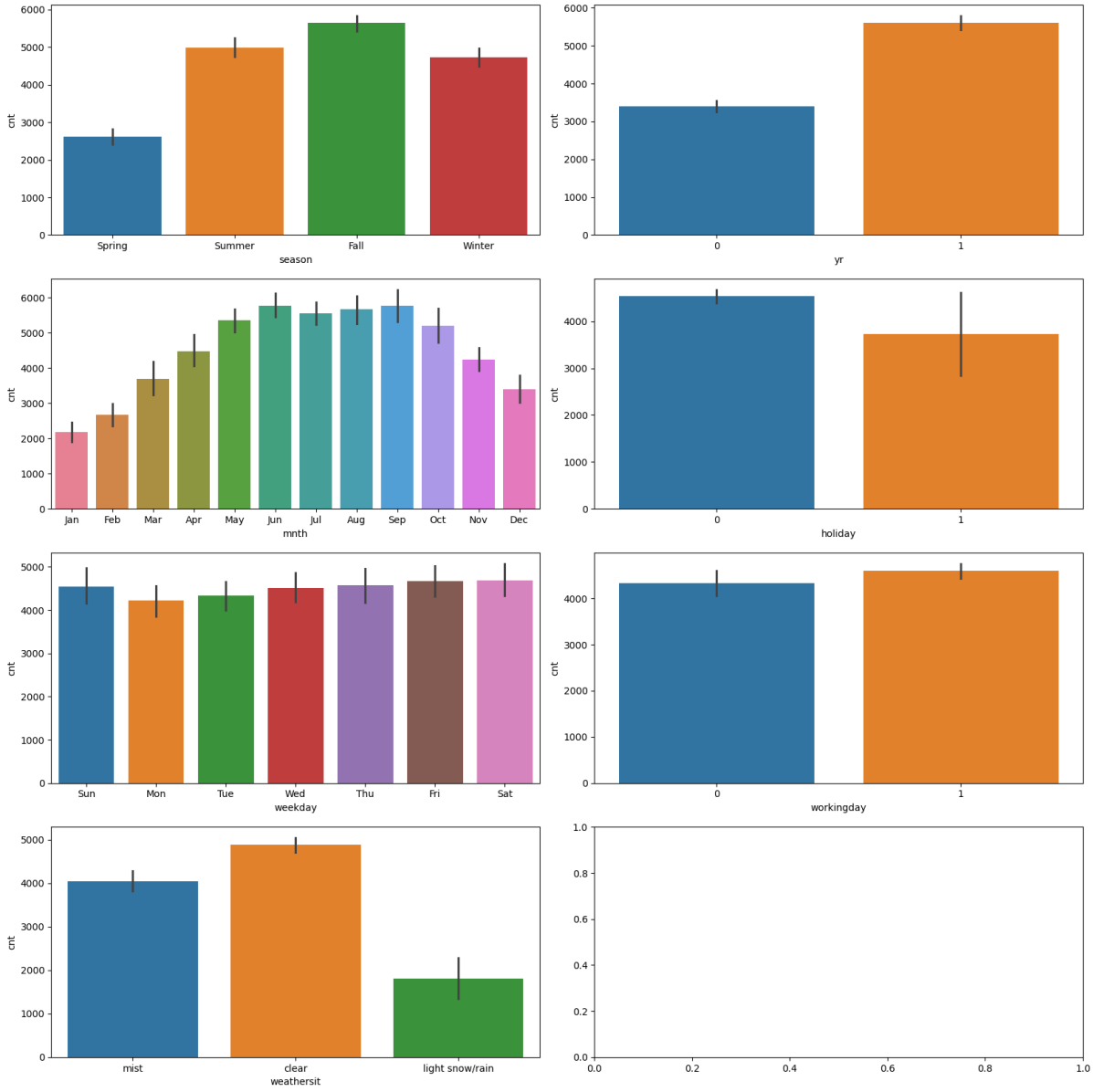
Assignment-based Subjective Questions

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



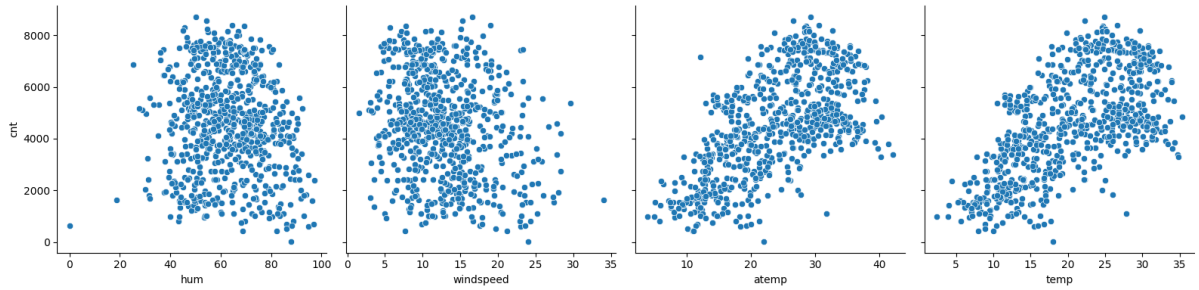
Based on the analysis of categorical variables from the dataset, it can be inferred that the independent variable weekday, mnth, season, year, weathersit have a significant impact on the dependent variable cnt. The analysis shows that there is a clear pattern of increase in the total number of bikes rented as the day of the week progresses to weekends, June and September, Fall season, and the year 2019 . This suggests that there may be a positive correlation between these variables and the total number of rented bikes.

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

Drop_first=True means that one level of each categorical variable is dropped, which reduces the number of dummy variables by one and ensures that they are linearly independent. This makes the

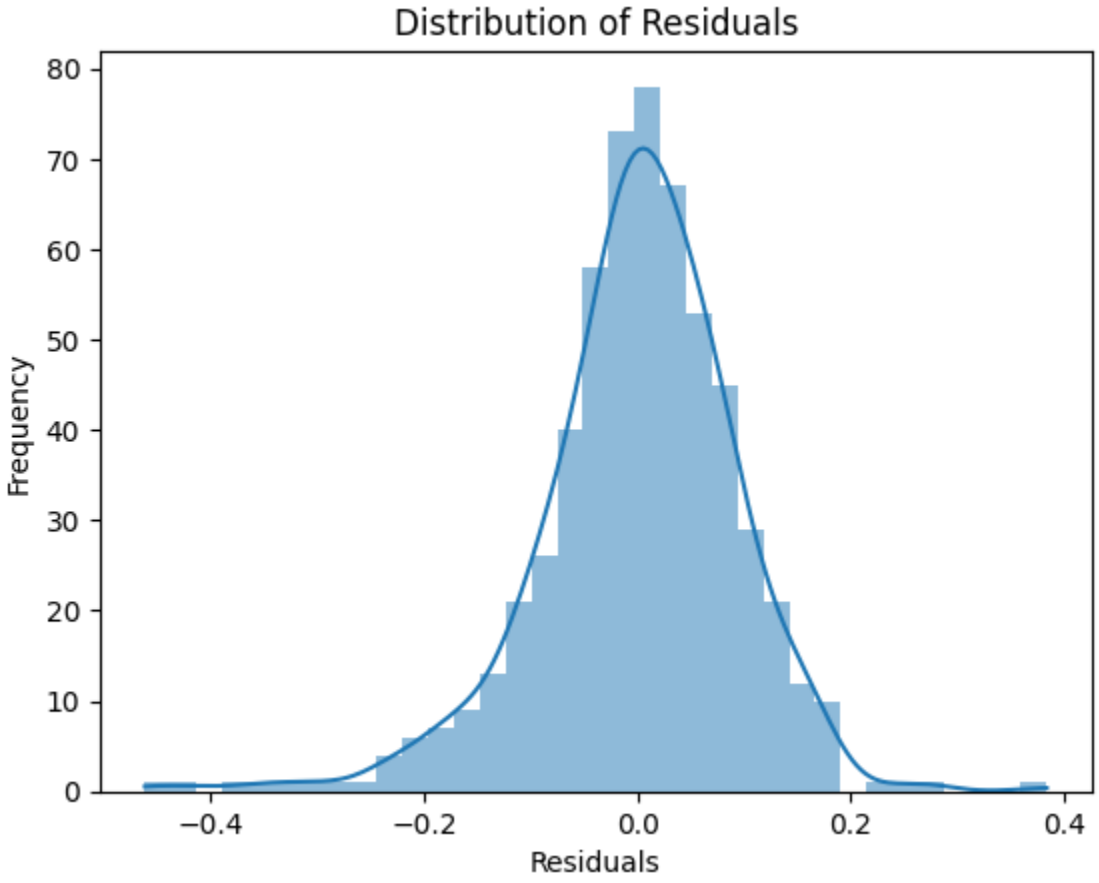
dummy variables more interpretable, prevent multicollinearity and ensure efficient model performance.

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



The one that has the highest correlation with the target variable is the 'temp' variable

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



	feature	VIF
0	temp	4.973062
2	workingday	4.190392
1	yr	2.053449
4	season_Winter	1.933493
5	weekday_Sun	1.696790
7	weathersit_mist	1.536788
3	season_Spring	1.482773
9	mnth_Nov	1.481072
8	mnth_Jul	1.306706
6	weathersit_light snow/rain	1.060059

I validate the assumptions of Linear Regression such as normality of Residuals by plotting the Residuals, linearity by examining the 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)

Top 3 features are:

1. Temp: the temperature has the most impact on shared bikes rental. As the temperature increase, the demand for shared bikes also increase
2. Yr: has a positive impact on the demand for shared bikes, the demand for shared bikes increases in 2019
3. Weathersit: rhe weather condition has a significant impact on the demand for shared bikes. The demand for shared bikes is higher on clear days than on rainy or cloudy days.

General Subjective Questions

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

Linear regression is a fundamental statistical method used for modeling the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the predictors and the target variable.

The equation for simple linear regression is: $y = \beta0 + \beta1x1 + \beta2x2 + \ldots + \beta nxn$

where y is the dependent variable, x1, x2, …, xn are the independent variables, and β1, β2, …, βn are the coefficients of the linear equation, β0 is the intercept (where the regression line crosses the y-axis).

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

Anscombe's quartet is a collection of four datasets that share summary statistics such as mean, variance, correlation, and linear regression line yet appear significantly differently when displayed on a graph.

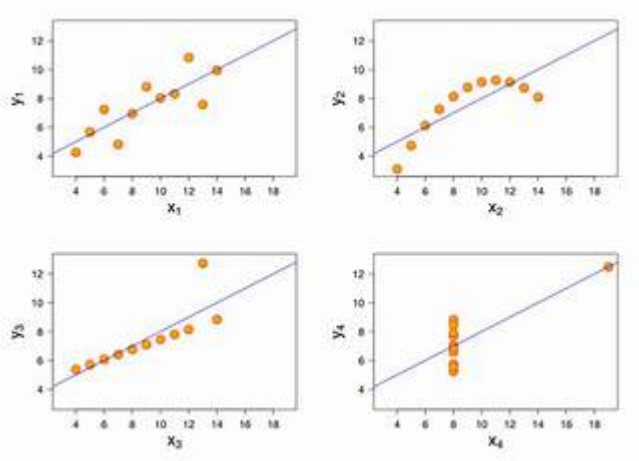


Figure 1: From Wikipedia, the free encyclopedia

Anscombe's quartet is commonly used to show how crucial it is to visually explore data before diving into analyzing it based on a specific relationship. It highlights that basic statistics might not be enough to fully describe real-world datasets.

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

'Pearson's R is a measure of the linear correlation between two variables. It can vary from -1 to 1, with -1 being a perfect negative linear relationship, 0 representing no linear relationship, and 1 representing a perfect positive linear relationship.'

The formula for Pearson's R is:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 (y_i - \bar{y})^2}}$$

- X_I and y_I are the two variables
- $\bar{x}$ and $\bar{y}$ are the means of x and y

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
    - Scaling is a data preprocessing technique that converts variable values to a similar scale.
    - It is performed to make sure that no variable has an excessive effect on a model's learning algorithm because of bigger scales or different units.
    - Difference between normalized scaling and standardized scaling: Normalized scaling modifies data to fit within a given range, such as 0-1 or 0-100, whereas standardized scaling transforms data to have a mean of zero and a standard deviation of one.
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

The value of VIF became infinite thanks to the dataset's perfect multicollinearity, which suggests that one independent variable can be completely predicted by another independent variable or a linear combination of other 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 plot compares two probability distributions by putting their quantiles against one other. It compares the quantiles of the dataset's distribution to the quantiles of a theoretical distribution, to determine whether a dataset follows a specific probability distribution, usually the normal distribution.

Q-Q plots are critical in linear regression for evaluating the normality assumption of residuals.