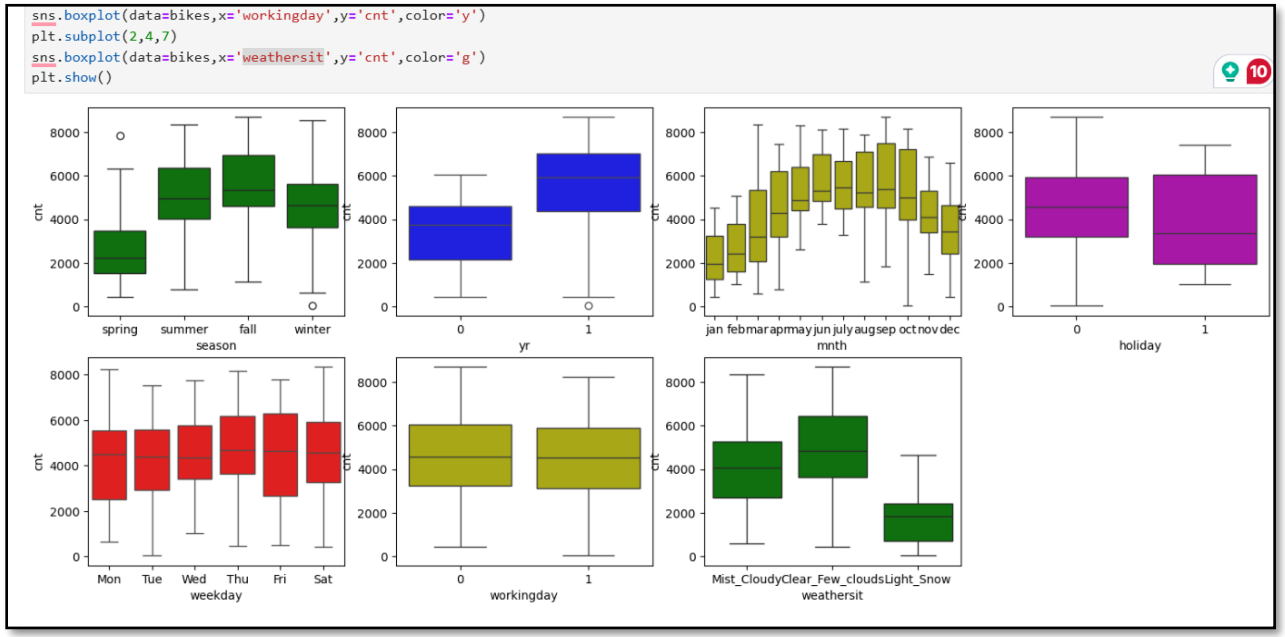
MLR Assignment

- Subjective Questions Of MLR

- By: Sachin Hase

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?**
   a. It has been clearly seen that from categorical var analysis that summer and fall season has positive relation with dependent var. i.e -cnt
   b. Year wise cnt value look like increasing
   c. Month wise analysis shows that starts from June to Sept has increased trend of cnt
   d. Weathersit shows that clear, few cloudy weathers have positive relation with no of people rent boom bikes.
   e. Similarly, non-holiday has positive relation with cnt.



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

a. When creating dummy variables using pandas.get_dummies, setting drop_first=True is crucial to avoid a "dummy variable trap" by removing one of the dummy columns, preventing multicollinearity between the newly created features, which can significantly impact the performance of your regression model.

b. Example we have column of Season which state Seasons like Spring,summer,fall & winter, there is no point to keep all column in dataset. It is obvious that if one is True then other Should be False. Hence, we just need to create (n-1) var col.

]:

	season	yr	mnth	holiday	weekday	
0	spring	0	jan	0	Mon	
1	spring	0	jan	0	Tue	
2	spring	0	jan	0	Wed	
3	spring	0	jan	0	Thu	
4	spring	0	jan	0	Fri	

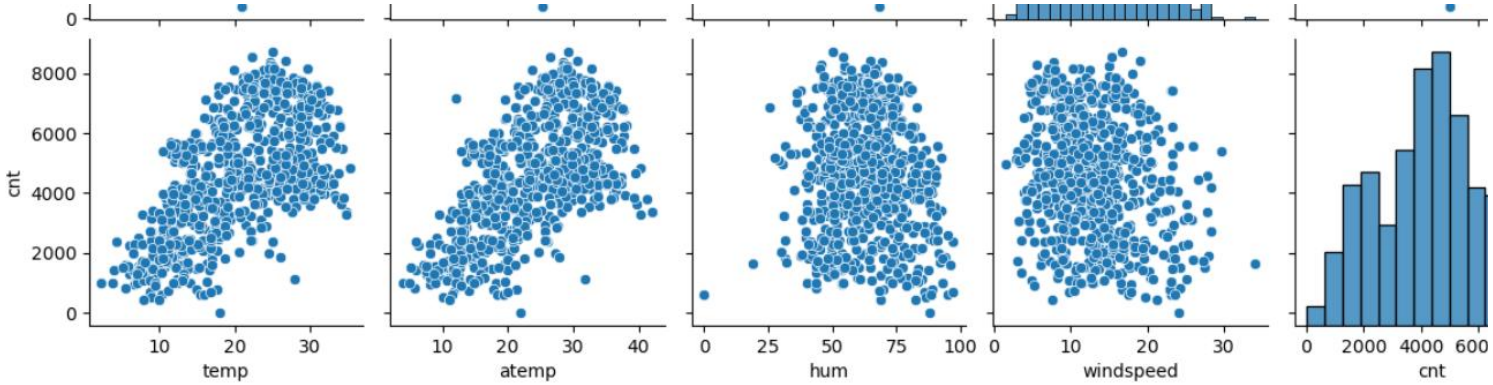
```
temp=pd.get_dummies(bikes[['season', 'mnth', '
temp.head()
```

	season_spring	season_summer	season_winter	r
0	1	0	0	
1	1	0	0	
2	1	0	0	
3	1	0	0	
4	1	0	0	

5 rows x 21 columns

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

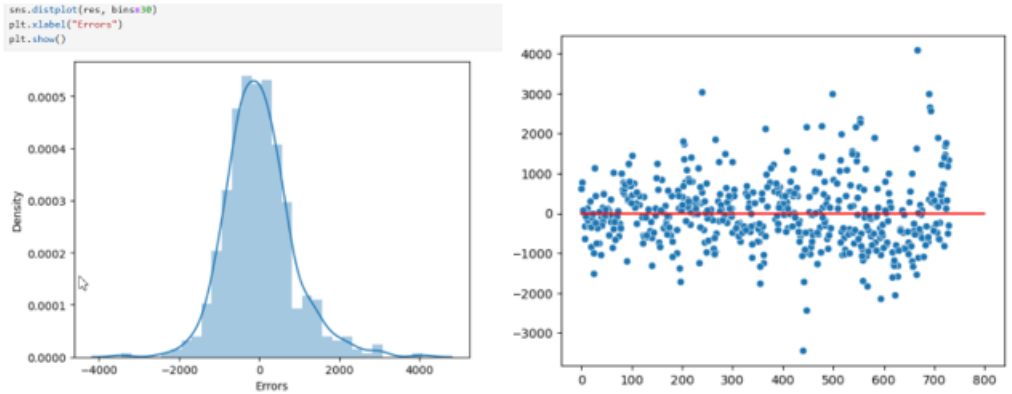
a. 'Temp' var has highest correlation with Target variable.



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

a. First, we check residual distribution- as per assumptions "Errors terms should be normally distributed with mean Zero."

b. Secondly, we can plot the scatter plot to check- Error terms are independent of each other.



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

a. Atemp: Feels like temperature
b. Weathersit: Clear, Few clouds, Partly cloudy, Partly cloudy
c. Yr: year

• General Subjective Questions

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

- Ans: Linear regression is a statistical method used in machine learning to model the relationship between a dependent variable (target) and one or more independent variables (features) by fitting a straight line through the data points, aiming to minimize the difference between predicted values and actual values; essentially, it predicts a continuous value based on a linear equation, where the goal is to find the "best fit" line that best represents the data points, allowing for predictions on new data based on this established relationship.
- Example: Consider the task of calculating blood pressure. In this case, height, weight, and amount of exercise can be considered independent variables. Here, we can use multiple linear regression to analyze the relationship between the three independent variables and one dependent variable, as all the variables considered are quantitative.

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

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

• The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables. When one variable changes, the other variable changes in the same direction.

• **Example:** Baby length & weight:

• The longer the baby, the heavier their weight.

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

What:

It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Why:

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

Example: As you can see that area values are in thousands and bedroom and bathrooms are very less. Because of these coefficients of bedroom and bathroom will be higher whereas area will be very less. eg Coef_area=0.001, coef_bedroom=754. So interpretation could be wrong, secondly scaling will minimize the cost function. Type of scaling

Min - Max scaling (Normalization): Between the 0 to 1 Adv- this method takes care of outliers - (x-xmin)/(xmax-xmin)

Standardization (mean-0,sigma-1): Will not take care of outliers. - (x-mu)/sigma

	price	area	bedrooms	bathrooms	stories	ma
0	13300000	7420	4	2	3	
1	12250000	8960	4	4	4	
2	12250000	9960	3	2	2	
3	12215000	7500	4	2	2	
4	11410000	7420	4	1	2	

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

- Ans: A VIF value is considered infinite when there is a perfect correlation between two or more independent variables in a regression model, essentially meaning one variable can be perfectly predicted by a linear combination of the others, leading to a situation called "perfect multicollinearity" where the formula for VIF results in division by zero, causing the infinite value.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

- Ans: A quantile-quantile (Q-Q) plot is a scatterplot that compares the quantiles of two data sets to determine if they come from the same distribution. Q-Q plots are useful in linear regression to check if the residuals are normally distributed.