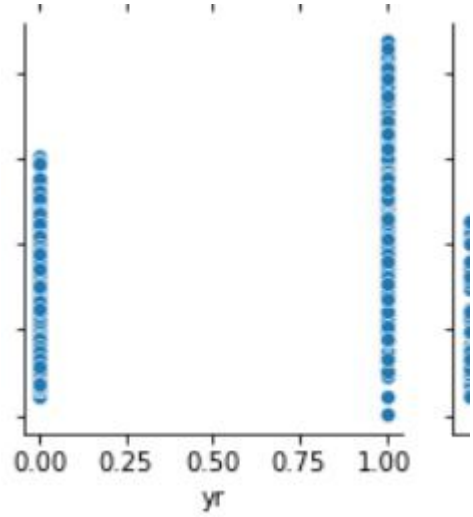
# Assignment-based Subjective Questions

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

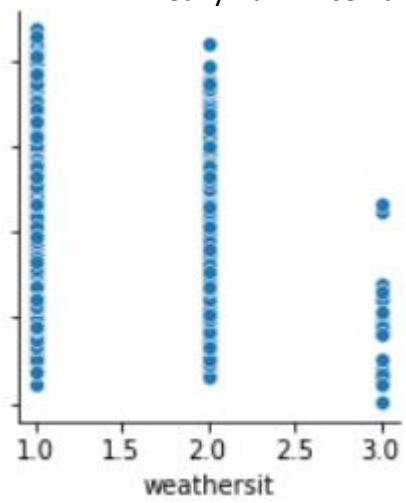
Ans:

Based on the subplots plotted on the categorical data we can say that in the yr 2018 there were less number of bikes that were rented




The second observation the “weathersit” shows that there are very less number of rentals on the weather of category 3 and no rentals on the weather category 4. where the category 3 and 4 represent the following

- 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
- 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog




Another worth while observation on the categorical variables is on the Holiday, that is on a holiday the rentals are way less then what we have on a non-holiday.

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

ANS:

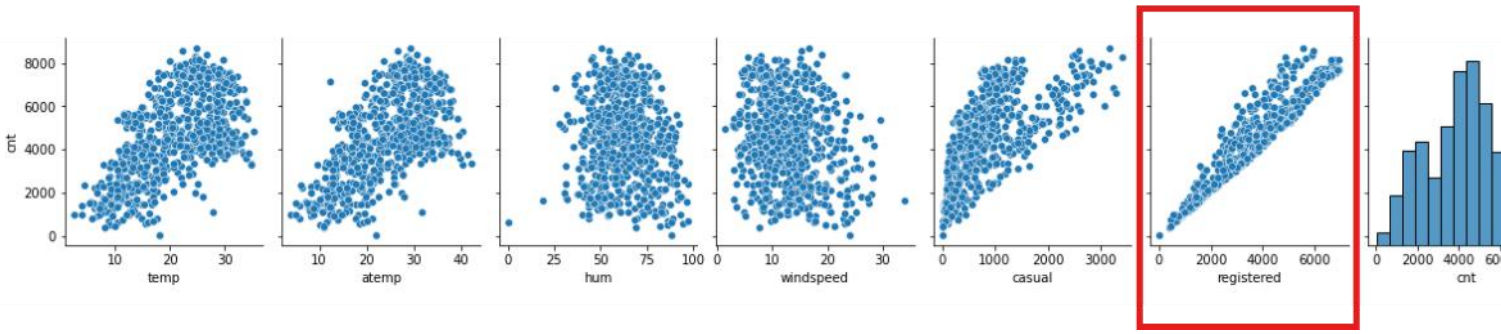
For representing any dummy variable we need n-1 variables for n number of features.

For example in our data we had categorical variable weather sit that had 4 outcomes 1,2,3,4 which we needed to show using dummy variables so in this case we would have got 3 variables when actually we only need 2 variables for 4 combinations (i.e. 00,01,10,11 )

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

ANS:

Registered has the highest correlation with the target value it indicates an exact linear correlation graph

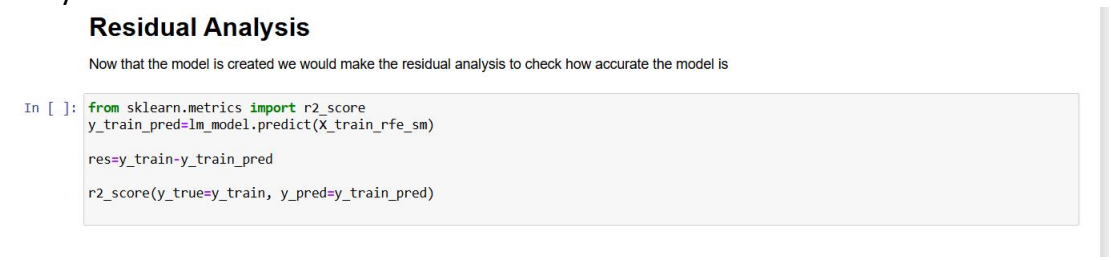


4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

ANS:

To validate the assumption we used the P-value of the predicted variables that were used in the model and based on the P-values if the P-value of the variable was higher than 0.5 then we have removed such variables from the model and then refitted the model this took many iterations, but at the end the model predicted perfect results.

We know this because we calculated the R squared value of the predicted variable using the model and the training variable of the model. Also known as the Residual Analysis



**Residual Analysis**

Now that the model is created we would make the residual analysis to check how accurate the model is

In [ ]:

```
from sklearn.metrics import r2_score
y_train_pred=lm_model.predict(X_train_rfe_sm)

res=y_train-y_train_pred

r2_score(y_true=y_train, y_pred=y_train_pred)
```

We also calculated the same on the test dataset to see how accurately our model can predict the results

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

ANS:

Based on the model the top 3 contributers are:

I. Registered: count of registered users

II. Casual: count of casual users

III. Season: season (1:spring, 2:summer, 3:fall, 4:winter)

# General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
ANS:
Linear regression is a supervised learning algorithm used to model the relationship between a dependent variable (target) and one or more independent variables (features).

The assumption is that the change in the dependent variable(X) is proportional to the independent variables(y).

$y=mx+c$

Where m is the slope or the ratio of change in x and c is the intercept.

Here we aim to find this line equation or in other words the value of “m” and “c” such that for a given value of x we can find the value of y.

The correct model should a predicted value with minimum error(difference between actual value and predicted value). We find the sum of squared differences fo this value.

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

ANS:

Anscombe's quartet consists of four datasets that have nearly identical simple statistical properties (mean, variance, correlation, and linear regression results), yet appear very different when graphed. It highlights the importance of visualizing data before analysis. Despite having similar summary statistics, the four datasets reveal varied distributions, outliers, and relationships when plotted. This emphasizes that relying solely on summary statistics can be misleading in machine learning and data science tasks.

3. What is Pearson’s R? (3 marks)

ANS:

Pearson’s R, or Pearson correlation coefficient, measures the strength and direction of a linear relationship between two variables. It ranges from -1 to 1, where 1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 means no linear correlation. The formula for calculating Pearson's R is:

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

Here, $x_i$ and $y_i$ are data points, and x bar and y bar are their means respectively.

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

In machine learning, scaling refers to adjusting the range of feature or independent variable values so that they fall within a specific range or distribution.

Without scaling, features with larger values can dominate the model, leading to biased results.

Normalized scaling rescales data to a specific range, typically between 0 and 1, maintaining relative proportions. Standardized scaling transforms data to have a mean of 0 and a standard deviation of 1, focusing on the distribution's shape.

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

The value of VIF (Variance Inflation Factor) becomes infinite when there is perfect multicollinearity between independent variables, meaning one variable is an exact linear combination of others. This makes the denominator in the VIF formula zero, leading to an undefined or infinite value, indicating that the regression model cannot distinguish the effect of the variables separately.

$$VIF_i = \frac{1}{1 - R_i^2}$$

Where,
$VIF_i$ is the VIF of the $i^{th}$ feature,
$R_i^2$ is the coefficient of determination from a regression of the $i^{th}$ feature on all other independent variables.

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

ANS:
A Q-Q (quantile-quantile) plot is a graphical tool used to compare the distribution of a dataset against a theoretical distribution, typically the normal distribution. In linear regression, it helps assess the normality of residuals. If the points closely follow a straight line, it indicates that the residuals are normally distributed, which is crucial for valid inference and hypothesis testing.