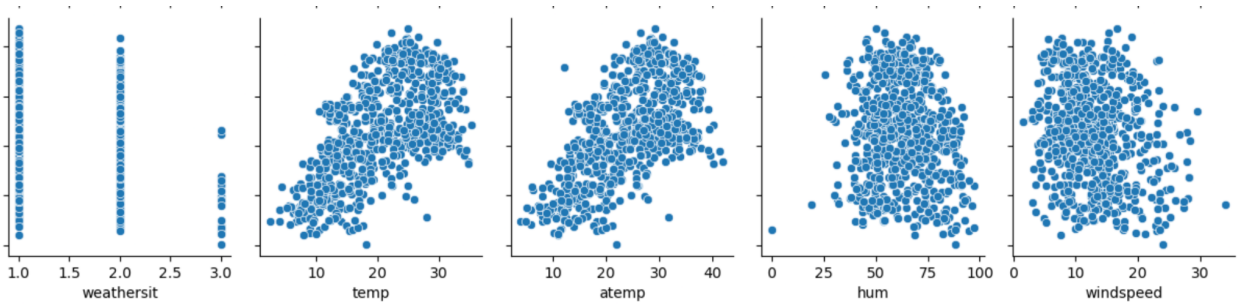
1

I noticed one thing, which is pretty clear maybe, that the dependent variable has a good correlation with temp and atemp.
It also showed that the dependent variable is also dependent on wind speed, as most of the sales was done only when the wind speed is less.
It also showed the great difference between clear and snow.



2

We need to use drop_first while creating the dummy variables because if all other variables are 0, it means that the first variable is by default 1. And not dropping the first variable can add the extra load on model and VIF values.

3

Temp, atemp, casual and registered has the highest correlation with the target variable.

4

To validate the training model, we can see that the r-squared value is 0.819 which is a pretty good value.

5

Top 3 features contributing significantly towards explaining the demand of the shared bikes are holiday, weathersit and weekdays.

General Questions

1. **What is a linear regression model?**
   Linear regression is the model where we compute the relation between independent variable and dependent variable. Basically Independent variables are the past data for which we have the target variable which is dependent on the values of independent variables.
   Independent variables are denoted as X and dependent variables are denoted as y.
   To plot the proper trend line on the graph, we use the line formula which is:
   $y = mx + c$
   Where m is the slope of line and c is the y-intercept.

2. **Explain the Anscombe's quartet in detail.**
   Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x,y) points.

3. **What is Pearson's R?**

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**
   The process of scaling means to reduce the scale of numbers without changing its original value. Scaling is performed to keep the scale of multiple variables equal, which makes our model easier to plot the points on a graph (basically it makes it easier to calculate the values).
   Normalization scaling means to scale our values between 0 and 1.
   Standardize values between -1 and 1.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**
   It happens because the variables share the perfect correlation between them.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**
   Q-Q plots are also known as Quantile-Quantile plots. As the name suggests, they plot the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential.