**Assignment-based Subjective Questions**

Q: 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans: Using Season, and we have created multiple dummy variables. Below variables from these dummy variable have -ve impact on target variable.

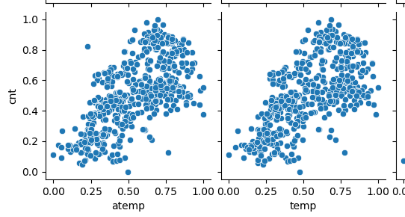
* Windspeed with coef of -0.108
* Sping with coef of -0.16
* Weathersit 3 with coef of -0.247

Q 2. Why is it important to use drop\_first=True during dummy variable creation?

Ans: By dropping we are removing redundancy of variables, Also we can delete any column which we think is ok to delete. For example if first columns have 70% of data and rest columns represent only 30 % of data then we can delete any columns except first.

Q 3: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

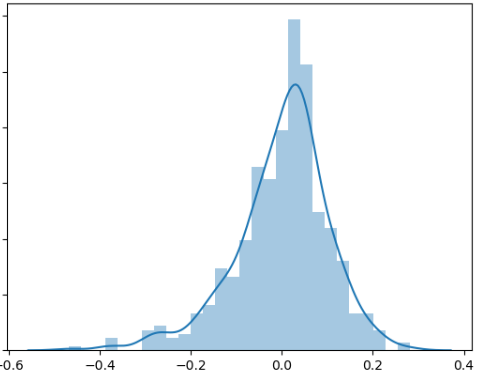
Ans: temp and atemp



Q 4: How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans: I followed below steps:

1. Calculated y\_train\_pred by using predict function of linear model on X\_train\_sm dataframe.
2. Then we subtract y\_train - y\_train\_pred to get residual score.
3. Then I used distribution plot to check bell shape curve and its mean which should be 0.



1. Now for all numerical variable I have used transform function to bring all test variables on same scale.
2. Now I stored cnt (target variable) into y\_test
3. Then I stored all selected features from x\_test in same order as X\_train
4. Added constant to X\_test
5. Now predicted y by using train model and predict method.
6. In the last calculated re\_score to compare it with train r2\_score

Q 5: Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans:

1. Yr with coefficient 0.232
2. Temp with coefficient 0.39
3. Weathersit 3 with coef of -0.247

**General Subjective Questions:**

Q 1: Explain the linear regression algorithm in detail.

Ans: Linear regression is a supervised machine learning algorithm in which we use training and test data set. On training data we create a model to get best fit line using OLS(ordinary least squares). Then this model we use on test data set which is representation of future data or we can say unseen data.

Q 2: Explain the Anscombe’s quartet in detail.

Ans: it’s a situation when we have multiple data sets and they all have similar mean coefficient and variance but when we plot them on graph they look totally different.

Q 3: What is Pearson’s R?

Ans: We measure linear correlation with person’s R. This is a number between -1 and 1. Where -ve values shows that it is a negative correlation and positive values shows positive correlation.

Q 4: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans: In Scaling we bring all numerical variables within one limited range.

For example in minmax scaling we transform numerical variables data between 0 and 1.

In Standardized scaling we bring numerical values within range by using standard deviation.

Q 5: You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans: did not get any infinite value, however it shows perfect correlation between variables.

Q 6: What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression

Ans: Q-Q plot we use to check if data is normally distributed or not.