# Assignment-based Subjective Questions

# Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

# Total Marks: 3 marks (Do not edit)

# Answer: <Your answer for Question 1 goes below this line> (Do not edit)

# In data set we found multiple categorical variables like season, month, weekday, weather sit. Among those season and month have same graph, and weekday does not show too much variation in mean of sub-categories. So we can remove these 2 variable (month, weekday).

# Effect of other categorical variables on dependent variable as follows

# Categorical Variable – Season

# Fall – 0.0731 positive (Help to increase demand of shared bikes)

# Winter – 0.0948 positive (Help to increase demand of shared bikes)

# Categorical Variable – Weather sit

# Mist – 0.0298 Negative (Mist weather condition decrease the bike demand)

# 

**Question 2.** Why is it important to use **drop\_first=True** during dummy variable creation? (Do not edit)

**Total Marks:** 2 marks (Do not edit)

# Answer: <Your answer for Question 2 goes below this line> (Do not edit)

# While creating the dummy variables drop\_first=True help to reduce the number of column without compromise data

# For example,

# Season is categorical variable with 4 category Spring, Summer, Fall and Winter

# While creating dummy it will provide 4 columns with values of 0 and 1, since only one category could be 1 at the moment other should be 0 so by using Drop\_first=True we eliminate spring column and know that when all other rest categorical column is 0 then spring will be 1, by that method we reduce column without compromising data

**Question 3.** Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

**Total Marks:** 1 mark (Do not edit)

# Answer: <Your answer for Question 3 goes below this line> (Do not edit)

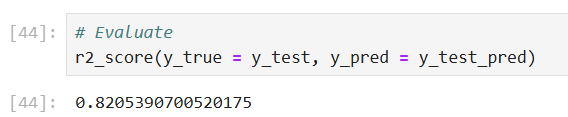
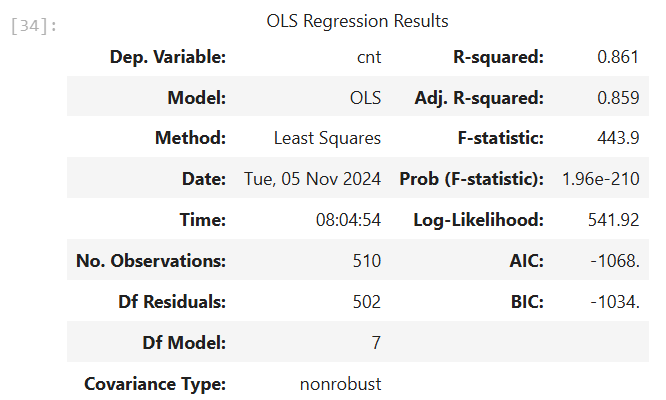
# By looking at pair plot we can say that registered variable has highest correlation with target variable.

**Question 4.** How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

**Total Marks:** 3 marks (Do not edit)

# Answer: <Your answer for Question 4 goes below this line> (Do not edit)

By analyzing it against test dataset and see the outcome with help of R square value. If difference between R square value of test data set and train data set is less then we can say our assumptions are valid



In our case R Square of test data set is 0.861 and train set is 0.82, so our model gives reasonable result.

**Question 5.** Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)

**Total Marks:** 2 marks (Do not edit)

# Answer: <Your answer for Question 5 goes below this line> (Do not edit)

# Top 3 variables are “Casual” “Working day” and “Year” shows significant positive impact on demand of the shared bikes.

# General Subjective Questions

**Question 6.** Explain the linear regression algorithm in detail. (Do not edit)

**Total Marks:** 4 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

# <Your answer for Question 6 goes here>

# First step is understanding of data and EDA, removing unnecessary variable columns

# Second step is preparation of data contains following steps

# Converting categorical variable in dummy variable

# Splitting into test data set and train data set

# Rescaling of numerical variables

# Third step is training of the model (Model building using train data set)

# Fourth step is Residual analysis

# Fifth step is prediction and evaluation on the test data set.

**Question 7.** Explain the Anscombe’s quartet in detail. (Do not edit)

**Total Marks:** 3 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

# <Your answer for Question 7 goes here>

# Anscombe’s state that 4 different data set might have identical descriptive statistic but still shows very different distribution and pattern when visualized, basically gives us importance of the visualization.

**Question 8.** What is Pearson’s R? (Do not edit)

**Total Marks:** 3 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

# <Your answer for Question 8 goes here>

Pearson’s R shows the linear relationship between two continuous variable, Its value ranges from 1 to -1

1 shows very perfect and positive linear relationship, other wards when x increase y increase in same factor

0 shows no relationship between variables

-1 shows vary perfect but negative liner relationship, other words when x increase y decrease in same factor

Values between 0 to 1 show positive relationship but not perfect

Values between 0 to -1 show negative relationship but not perfect

**Question 9.** What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)

**Total Marks:** 3 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

# <Your answer for Question 9 goes here>

Scaling is important because it helps to got all numerical variable on same range make it easy to compare and also improve outcome of algorithm

We have 2 types of scaling

First one is normalized scaling also know as Min Max scaling it convert all values between 0 to 1

Formula of Min Max scaling is Xnormal = (X-Xmin)/(Xmax-Xmin)

Second one is standardized scaling it convert all values in such way that their mean comes to 0 and standard deviation turn out to be 1.

**Question 10.** You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)

**Total Marks:** 3 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

# <Your answer for Question 10 goes here>

# When any variable is perfect combination of other variable then we found infinite VIF value for that variable its called Multicollinearity.

**Question 11.** What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

(Do not edit)

**Total Marks:** 3 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

# <Your answer for Question 11 goes here>