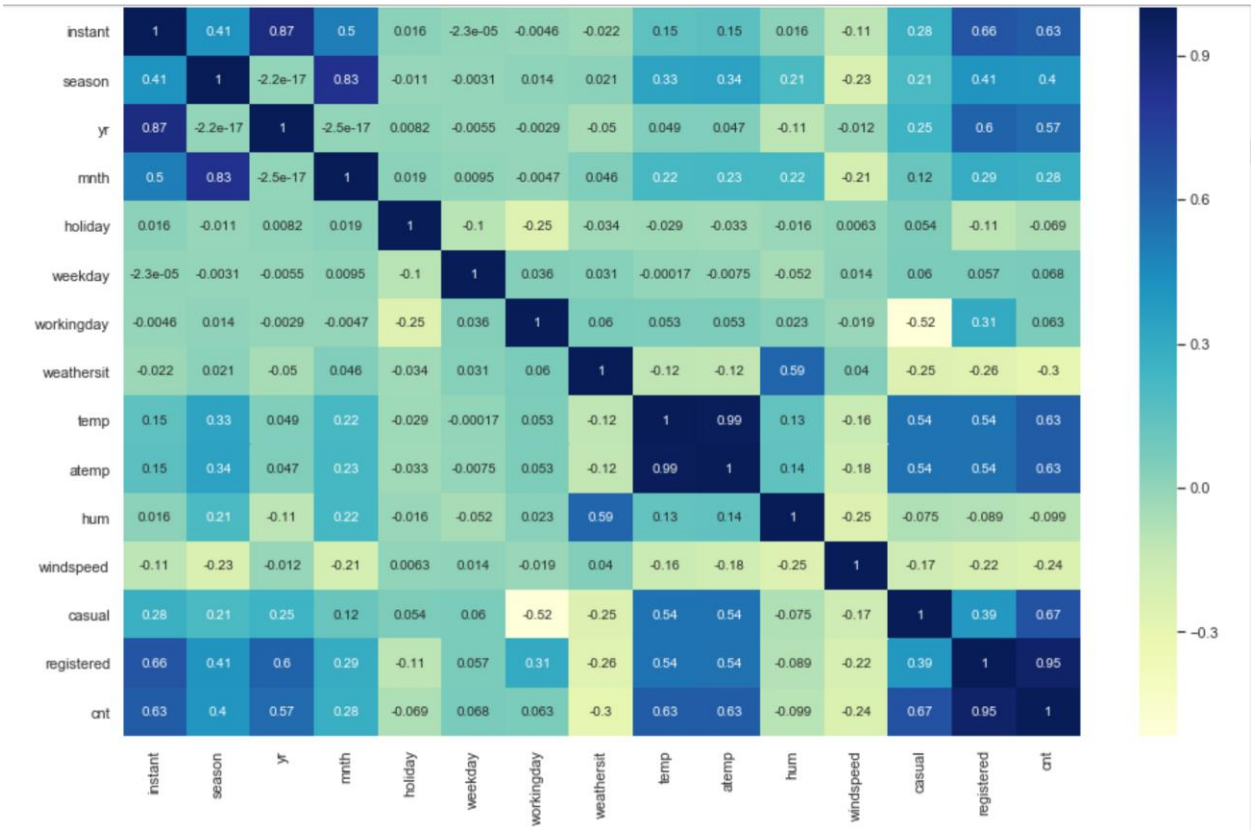
Assignment-based Subjective Questions

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: I plotted the heat map to check this.



And below are some of the observations initially came up by taking data.

Season and month have good Correlation which is 83%

2. atemp and temp have good Correlation which is 99%

3. temp and registered have little Correlation which is 54%

4. temp and casual have little Correlation which is 54%

3. registered and cnt have good Correlation which is 95%

4. Year and cnt have little Correlation which is 57%

5. weathersit and hum have little Correlation which is 59%

6. cnt and casual have little Correlation which is 67%

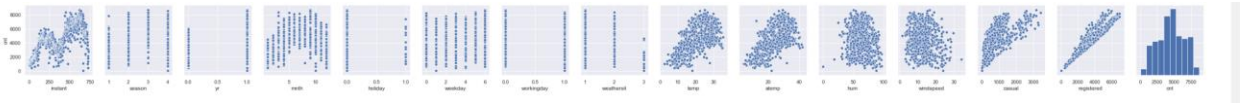
2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

Ans: It is better to use drop_first=True so that we can get **less some extra columns** created during dummy variable creation. And in that way it reduces the correlations created among dummy variables.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation

with the target variable? (1 mark)

Ans: cnt and temp have highest correlation it seems by seeing pair plot.

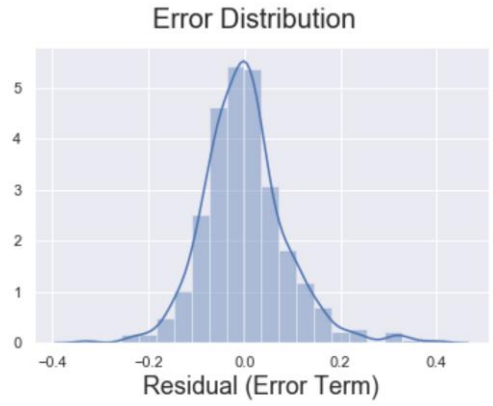


4. How did you validate the assumptions of Linear Regression after building the model on the

training set? (3 marks)

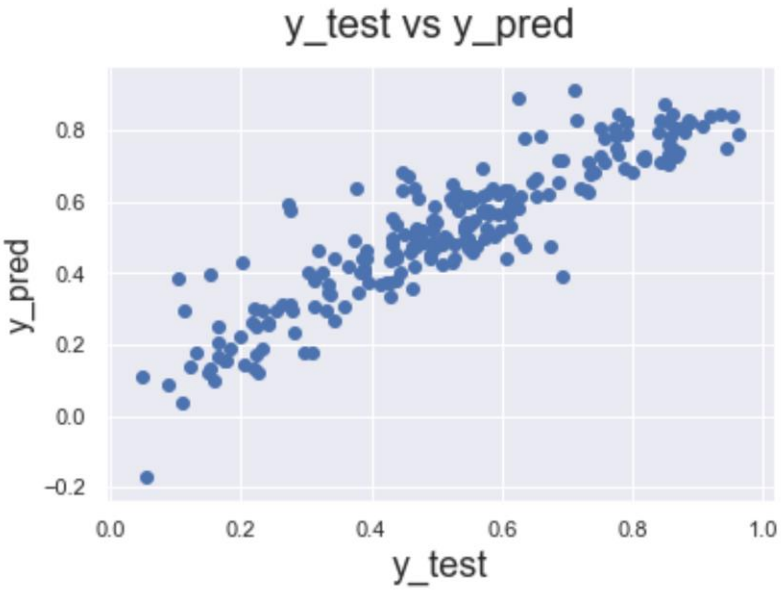
Ans: We can validate our assumptions in below some of the listed ways.

a. Checked for VIF and p-value if they are under control, for VIF minimum considered to be 5 and for p-value it's 0.05, if it is more than that droped all those.
b. Checked if the error terms are also normally distributed (which is in fact, one of the major assumptions of linear regression), and plotted the histogram of the error terms and see what it looks like. As After building model, we cannot finalise untill we proove the residual analysis wherein we check whether the distribution of Error is around 0 or not.So we found like below plot for it.



And we can see from the above graph that the Error Distribution Is Normallly Distributed Across 0, which shows that our model has handled the assumption of Error Normal Distribution properly.

c. Than we performed the model evaluation by plotting graph y_test vs y_pred to understand the spread. And From below Scatter Plot we observed Linear Relationship between Actual Test Data Points & Predicted Test Data Points



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 final model VIF , yr, holiday, windspeed are 3 features which contributed significantly towards explaining the demand of the shared bikes.Below is the snippet from my notebook for the same.

[87]:

	Features	VIF
0	const	71.18
1	season_spring	5.12
2	temp	3.95
3	season_winter	3.62
4	season_summer	2.65
5	hum	2.10
6	workingday	1.88
7	weekday_saturday	1.78
8	weathersit_mist	1.65
9	mnth_January	1.57
10	mnth_July	1.49
11	weathersit_light	1.36
12	mnth_September	1.30
13	windspeed	1.22
14	holiday	1.16
15	yr	1.04

General Subjective Questions

1. **Explain the linear regression algorithm in detail. (4 marks)**
   Ans: linear regression is a linear approach to modeling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression. Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.

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

Anscombe's quartet comprises of four data sets and that may 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. Anscombe's Quartet reminds us that

graphing data prior to analysis is good practice, outliers should be removed when analyzing data, and statistics about a data set do not fully depict the data set in its entirety.

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

Ans: Pearson's correlation coefficient is used to measure the relationship between two continuous variables. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0. I could not try it in my code yet due to lake of time, but with the help on this we can calculate correlation coefficients to check how predictors are impacting target variable.

4**. What is scaling? Why is scaling performed? What is the difference between normalized scaling**

and standardized scaling? (3 marks)

Ans:

Scaling means adjusting data that has different scales so as to avoid biases from big outliers. The most common techniques of feature scaling are Normalization and Standardization.

Normalized scaling technique is that in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

Here's the formula for normalization:

X' = X-Xmin/Xmax-Xmin

Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

Here is the formula for it:

X' = X-Mu/sigma

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

(3 marks)

Ans: An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

(3 marks)

Ans: The Q-Q plot, or quantile-quantile plot, is a graphical tool that helps us to assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential. For example, if we run a statistical analysis that assumes our dependent variable is Normally distributed, we can use a Normal Q-Q plot to check that assumption. It's just a visual check, not an air-tight proof, so it is somewhat subjective. But it allows us to see at-a-glance if our assumption are ok, and if not, how the assumption is violated and what data points contribute to the violation.

it helps to determine if two data sets come from populations with a common distribution. This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.