Assignment-based Subjective

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

A. Season: Demand is more during the Summer and Fall seasons
Month: Users prefer to opt for riding bike mainly around June, July and August months which coincides with the higher demand in Summer and Fall seasons
Weathersit: It plays an important factor in opting for riding bikes, most of the people prefer to ride bikes when its clear / cloudy weather or mist / cloudy weather.
Weekday: It does not have much impact as the average number of users is almost same for all 7 days of the week.

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

A. before we get on to understanding why we need to drop one of the dummy variables during the dummy variable creation, we need to understand two important concepts what is **Dummy variable** why we need to create one and what is **Dummy variable Trap in regression models**.

Dummy Variable and its significance: Dummy variables are created on categorical variables. Categorical variables represent data values with a fixed and unordered number of values. For example is gender (male/female) or season (summer/winter/spring/fall). As the **regression analysis is used with numerical variables** we need to convert categorical values to numerical values. Results only have a valid interpretation if it makes sense to assume that having a value of 2 on some variable is does indeed mean having twice as much of something as a 1 and having 50 means 50 times as much as 1 and hence we can not represent categorical variables as 1, 2, 3 and so on, rather we need to create dummy variables for categories with only two values zero and one.

Dummy variable Trap in regression models: The Dummy Variable trap is a scenario in which the independent variables are multicollinear - a scenario in which two or more variables are highly correlated; in simple terms, one variable can be predicted from the others.

To demonstrate the Dummy Variable Trap, take the case of gender (male/female) as an example. Including a dummy variable for each is redundant of male as 0, female as 1 doing so will result in the following linear model equation:

y = b + b1 * {0} male + b2 * {1} female.

The sum of all categorical dummy variables for each row is equal to the intercept value of that row and there is perfect multi-collinearity. In short, we have created a duplicate category: if we drop one of the categories multi-collinearity is taken care of.
The solution to the dummy variable trap is to drop one of the categorical variables if there are n number of categories, we have to use n-1 in the model and hence we are using **drop_frist=True** while creating dummy variables.

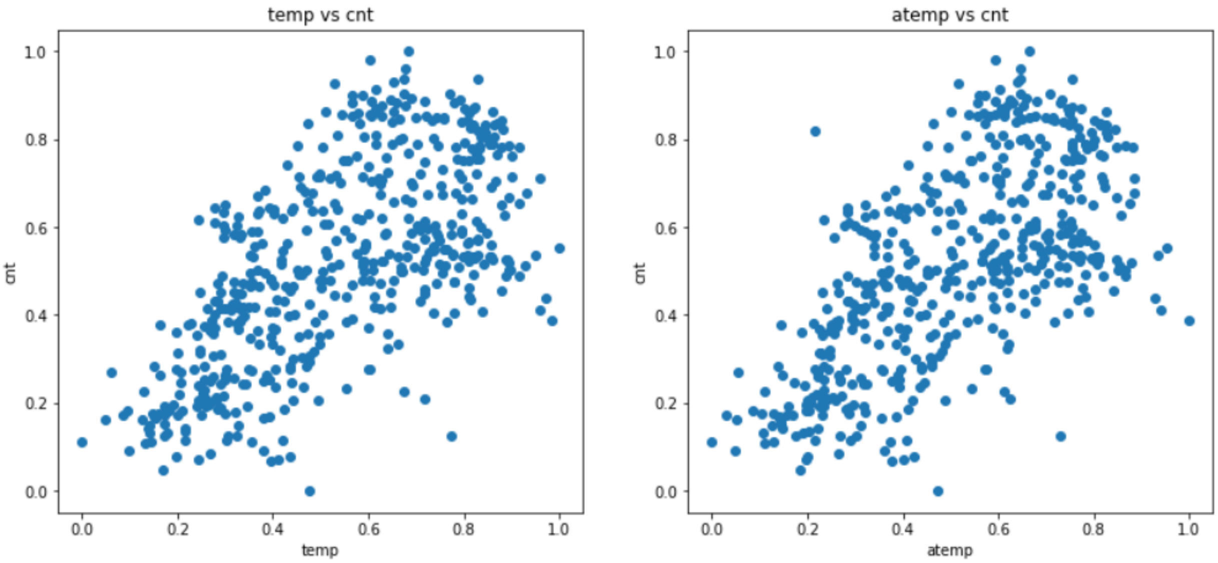
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

A. Temp/atemp

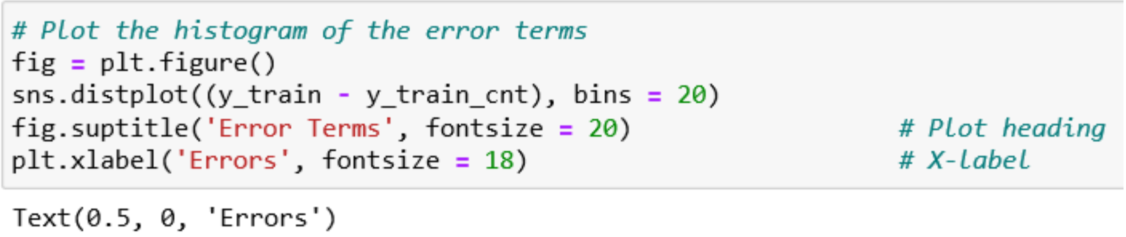
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

A. Below are the assumptions and related validations in detail:

1. One of the first assumptions in Multiple linear regression is we should have a linear relationship between the independent and dependent variables, the same has been verified part of scatter plot between cnt, temp and cnt, atemp.

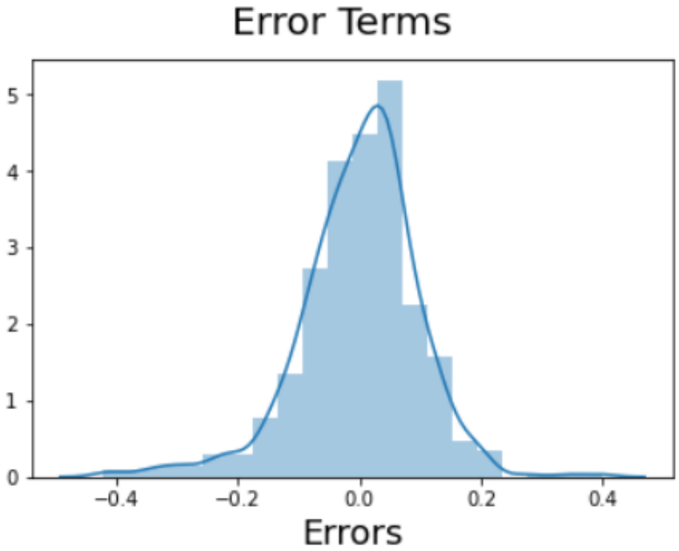


2. Multiple linear regression analysis requires that the errors between observed and predicted values should be normally distributed with mean zero. Same has been checked please find the evidence below:

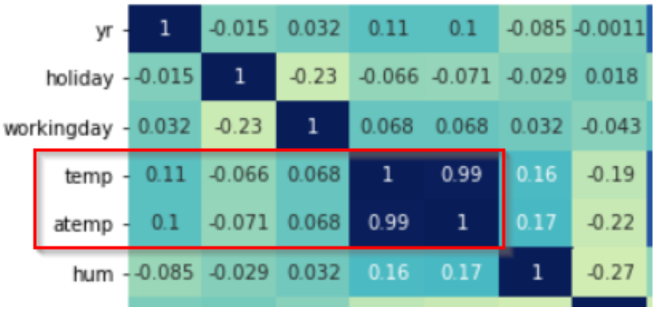


```
# Plot the histogram of the error terms
fig = plt.figure()
sns.distplot((y_train - y_train_cnt), bins = 20)
fig.suptitle('Error Terms', fontsize = 20)                  # Plot heading
plt.xlabel('Errors', fontsize = 18)                         # X-label
```

```
Text(0.5, 0, 'Errors')
```

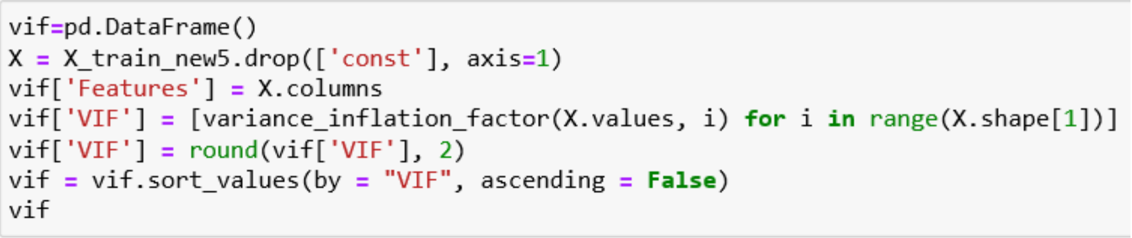


3. The third assumption is that there is no Multicollinearity in the data, this can be checked in two ways
    a. Using the correlation matrix we can evaluate for multicollinearity, the magnitude of the correlation coefficients should be less than 0.80. In this Bikes assignment, we can see multicollinearity between temp and atemp and hence we use only one of them as part of the final model.

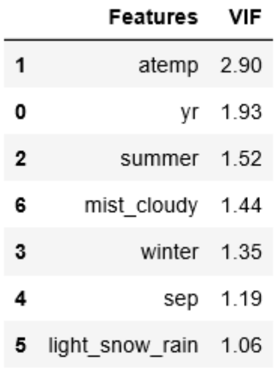


yr	1	-0.015	0.032	0.11	0.1	-0.085	-0.0011
holiday	-0.015	1	-0.23	-0.066	-0.071	-0.029	0.018
workingday	0.032	-0.23	1	0.068	0.068	0.032	-0.043
temp	0.11	-0.066	0.068	1	0.99	0.16	-0.19
atemp	0.1	-0.071	0.068	0.99	1	0.17	-0.22
hum	-0.085	-0.029	0.032	0.16	0.17	1	-0.27

    b. The second method of addressing multicollinearity is using the Variance Inflation Factor (VIF). The VIFs of the multilinear regression indicates that the degree of the variance in the regression estimates are increased due to multicollinearity. Hence, we need to maintain the VIF value no more than 5, the same has been verified, please find the final VIF data below from the model that I have built for your reference.



```
vif=pd.DataFrame()
X = X_train_new5.drop(['const'], axis=1)
vif['Features'] = X.columns
vif['VIF'] = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]
vif['VIF'] = round(vif['VIF'], 2)
vif = vif.sort_values(by = "VIF", ascending = False)
vif
```



	Features	VIF
1	atemp	2.90
0	yr	1.93
2	summer	1.52
6	mist_cloudy	1.44
3	winter	1.35
4	sep	1.19
5	light_snow_rain	1.06

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

A: We see 3 major variables that affect positively and negatively, details are as below:
1st one is temp the count of bikers will increase by 0.60 times every time atemp increases. Positive relationship.
2nd one is weathersit category 3 (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds) during these weather situation demands of bikes are reduced by 0.286 times.

3^{rd} one is a year, every year we can see the growth of demand in bikes by 0.246 times.

General Subjective

1. Explain the linear regression algorithm in detail.
   Ans: Linear regression is a machine learning algorithm which is generally built on supervised

   learning. Basically these algorithms models a prediction for target variables based on the independent variables, by finding the correlation between each independent variables and the target variable. There are different methods of building regression models – Simple regression, multiple regression, gradient descent, least squares and regularization etc.

2. Explain the Anscombe's quartet in detail.

   Ans: Anscombe's quartet is a set of 4 datasets which demonstrates the importance of visualizing the data and also highlights the effect of the outliers on a findings of a dataset. Each dataset has a set of X and Y points, which has different statistical properties. However, when the data points are graphed, all the dataset shows similar statistical relations. All the four datasets provides the correlation between X and Y points and the equation of linear regression line.

3. What is Pearson's R? Ans:
4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
   Ans: Scaling a method of normalizing the data used to standardize the range of features of dataset.

   Standardized Scaling: Is also knows as z score normalization, which transforms the data in such a way that the resulting distribution has mean of 0 and a standard deviation of 1. Formula→x=x−mean(x)/sd(x)
   Normalized Scaling: Is also know as min-max scaling. Which basically transforms the numerical data to scale between 0 and 1. Formula→x=x−min(x)/max(x)−min(x)

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
   Ans: Infinite VIF indicates that there is high correlation between two variables or in a regression model, it indicates that two X variables are perfectly correlated.

6.What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
Ans: The purpose of the quantile-quantile (QQ) plot is to show if two data sets come from the same distribution. Plotting the first data set's quantiles along the x-axis and plotting the second data set's quantiles along the y-axis is how the plot is constructed. In linear regression it is used to determine if the data is showing the normal distribution.