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)

Answer : Statistical summary is shown below

OLS Regression Results

Dep. Variable:	cnt	R-squared:	0.707
Model:	OLS	Adj. R-squared:	0.698
Method:	Least Squares	F-statistic:	79.42
Date:	Tue, 11 Oct 2022	Prob (F-statistic):	5.17e-121
Time:	20:05:14	Log-Likelihood:	-4274.0
No. Observations:	510	AIC:	8580.
Df Residuals:	494	BIC:	8648.
Df Model:	15		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2864.7323	184.151	15.556	0.000	2502.916	3226.549
holiday	-543.9903	325.016	-1.674	0.095	-1182.575	94.594
workingday	427.3713	139.116	3.072	0.002	154.038	700.704
windspeed	-1936.1051	293.874	-6.588	0.000	-2513.503	-1358.707
winter	896.8104	146.171	6.135	0.000	609.617	1184.004
March	503.3699	178.661	2.817	0.005	152.340	854.400
April	1313.5939	194.947	6.738	0.000	930.567	1696.620
May	1974.3618	188.864	10.454	0.000	1603.286	2345.437
June	2176.2521	195.693	11.121	0.000	1791.759	2560.745
August	2148.3165	180.636	11.893	0.000	1793.408	2503.225
September	2282.8469	186.584	12.235	0.000	1916.252	2649.442
October	982.4936	204.192	4.812	0.000	581.302	1383.685
2019	2173.7447	95.570	22.745	0.000	1985.971	2361.519
Monday	469.6778	174.950	2.685	0.008	125.940	813.415
Weather_Good	-913.5451	101.847	-8.970	0.000	-1113.652	-713.439
Weather_OK	-2528.7975	290.617	-8.701	0.000	-3099.796	-1957.800

Omnibus:	20.354	Durbin-Watson:	1.861
Prob(Omnibus):	0.000	Jarque-Bera (JB):	46.225
Skew:	0.143	Prob(JB):	9.17e-11
Kurtosis:	4.447	Cond. No.	10.9

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

From above statistical summary, we can see March,April,May,June,August,September and October months have positive corelation with dependent vatiable (cnt)
Year 2019 has a positive corelation with dependent variable.
Weekday Monday have positive correlation with dependent variable.
Season winter has less positive impact on depedent variable.
weathersit (Weather_Good, Weather_OK) has negative impact on dependent

variable.
Holiday has a negative impact on dependent variable.

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

Answer : Use of drop=first during dummy variable creation will reduce one dummy variable and thus helps to reduce corelation between created dummy variables.

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

Answer : registered:has the highest corelation with target variable.

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

Answer : The assumption is "The error terms will be normalized". To show that, I did following (draw histogram of error terms)

```
sns.distplot((y_train - y_train_count), bins = 20)
plt.title("Error terms", fontsize=15)
plt.xlabel("Error")
plt.show()
```

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

Answer : Months (March,April,May,June,August,September and October) , Weather_OK (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds), yr (2019).

==========

General Subjective Questions
1. Explain the linear regression algorithm in detail. (4 marks)

Answer : Linear regression has following steps

a. Reading , understanding and visualizing the data : Here we need to study the data very well and try pair plot, boxplot and heat map to visualize data.
b. Preparing the data for modeling (Encoding, test train split, rescaling) : Convert categorical variable to dummy variable and Convert yes & no to 1 & 0 respectively. Spliting test train set and rescaling train set.
c. Traning the model : Here we need perform feature selection using RFE (recursive feature elemination) and take care of multicolinearility VIF (variance inflation factor). It has to be both manual and automated process. Then train the model.
d. Residual analysis : Here we check the error term are normally distributed or not using hsitohgram
e. Prediction and evaluation of test set : Applied scalling on test set.

Test the model with test data. Evaluate model using r2_scoreof test set and compaing with r2_score gotten in train set.

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

Answer : Anscombe’s quartet is four data set having similar type of data distribution but when graphed, they looks very different. We use pair plot to see the relation between dependent and independent variable.
It help us to see, if we can apply linear regression on that data set or not and also help us to detect outliears.

3. What is Pearson’s R? (3 marks)

Answer : Pearson’s R help us to find the strength of linear relationship of two variable. It value lies in range -1 to 1 . -1 means total negative corelation between two varriable, 0 mean no corelation between two variables
and 1 means total positive corelation between two variables.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling
and standardized scaling? (3 marks)

Answer : Feature scaling is the process of normalizing the range of independent variable.

In multiple regression, predictor variable can have very different scale. If we get all of them in same scale, then it is easy for us to compair the coefficient of one feature with coefficient of other feature. It also helps
gradiant descent to converge much faster.

In case of normalized scaling values are between (-1, 1) or (0,1) but in case of standardized scaling ,value are not bounded in certain range. normalized scaling is used when we do not know about the data distribution
where as standardized scaling is used when feature distribution is normalized.

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

Answer : In case of perfect corelation between two independent variable, VIF wil be 'inf'

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

Answer : Q-Q (Quantile-Quantile) is plot of two quantiles. It heps to check if two data set come from a common distribution or not. If we get a line with 45 degree anle in Q-Q plot, that signifies both data set from common distribution.

In leanr regression, we can confirm test abd train set are from same distribution or not using Q-Q plot.