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)**

Based on the regression result:

year: The coefficient for the variable "year" is 0.2527, which is positive and significant (p-value is 0.000). This suggests that with each advancing year, there's an expected increase in the dependent variable by approximately 0.2527 units, assuming other variables remain constant.

workingday: The variable "workingday" has a coefficient of 0.6469, indicating that on working days, the dependent variable is expected to increase by around 0.6469 units compared to non-working days. This change is statistically significant given the p-value of 0.000.

atemp : The coefficient for "atemp" is positive at 0.4707. This suggests that as the feeling temperature (or "atemp") increases, the dependent variable is also expected to increase. The change is significant with a p-value of 0.000.

Seasons:

season_Spring: The negative coefficient of -0.1272 suggests that during the spring season, the dependent variable tends to decrease by around 0.1272 units compared to the fall (reference season). This is statistically significant with a p-value of 0.000.
season_Winter: The positive coefficient of 0.0958 indicates that during the winter season, the dependent variable increases by approximately 0.0958 units in comparison to the fall season. This change is significant with a p-value of 0.000.

Months:

month_Dec: The coefficient for December is negative (-0.0740), suggesting a decrease in the dependent variable during December compared to the base month. This change is significant with a p-value of 0.000.
month_Nov: The coefficient for November is negative (-0.0728), indicating a decrease in the dependent variable for this month. The change is statistically significant with a p-value of 0.001.
month_Sep: September sees a positive change with a coefficient of 0.0541, indicating an increase in the dependent variable during this month compared to the base month. This change is statistically significant with a p-value of 0.001.
weekday_Sat: The coefficient for Saturday is positive at 0.0602, suggesting that on Saturdays, the dependent variable increases by approximately 0.0602 units. This change is significant with a p-value of 0.000.

weathersit_Light_snow_rain: The negative coefficient of -0.1267 indicates that during light snow or rain, the dependent variable tends to decrease by about 0.1267 units. This change is significant with a p-value of 0.000.

In summary, the categorical variables in the dataset have a significant impact on the dependent variable. Factors like seasons, working days, and weather conditions play a crucial role in influencing the dependent variable, with certain conditions promoting an increase while others result in a decrease.

OLS Regression Results

Dep. Variable:	count	R-squared:	0.818
Model:	OLS	Adj. R-squared:	0.815
Method:	Least Squares	F-statistic:	224.7
Date:	Thu, 18 Jan 2024	Prob (F-statistic):	1.28e-177
Time:	08:04:45	Log-Likelihood:	446.50
No. Observations:	510	AIC:	-871.0
Df Residuals:	499	BIC:	-824.4
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1041	0.026	4.017	0.000	0.053	0.155
year	0.2527	0.009	27.743	0.000	0.235	0.271
workingday	0.0469	0.012	3.857	0.000	0.023	0.071
atemp	0.4707	0.034	13.658	0.000	0.403	0.538
season_Spring	-0.1272	0.017	-7.433	0.000	-0.161	-0.094
season_Winter	0.0958	0.015	6.276	0.000	0.066	0.126
month_Dec	-0.0740	0.019	-3.804	0.000	-0.112	-0.036
month_Nov	-0.0728	0.021	-3.409	0.001	-0.115	-0.031
month_Sep	0.0541	0.017	3.241	0.001	0.021	0.087
weekday_Sat	0.0602	0.016	3.788	0.000	0.029	0.091
weathersit_Light_snow_rain	-0.2672	0.027	-9.847	0.000	-0.321	-0.214

Omnibus:	92.842	Durbin-Watson:	1.957
Prob(Omnibus):	0.000	Jarque-Bera (JB):	241.585
Skew:	-0.905	Prob(JB):	3.47e-53
Kurtosis:	5.844	Cond. No.	14.9

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

Utilizing drop_first=True during the creation of dummy variables is crucial for a few reasons:

Preventing Multicollinearity: When dummy variables are created for categorical variables with n distinct values, multicollinearity can arise if all
n dummy variables are used in the regression. This is because one dummy variable can be perfectly predicted from the others. Using drop_first=True ensures that only n−1 dummy variables are used, eliminating one level and serving as the reference category. This effectively reduces the multicollinearity issue.

Simplifying Interpretation: By dropping one dummy variable, it becomes the reference category against which the effects of other categories are compared. This makes the interpretation of regression coefficients more straightforward, as they represent the change in the dependent variable relative to this reference category.

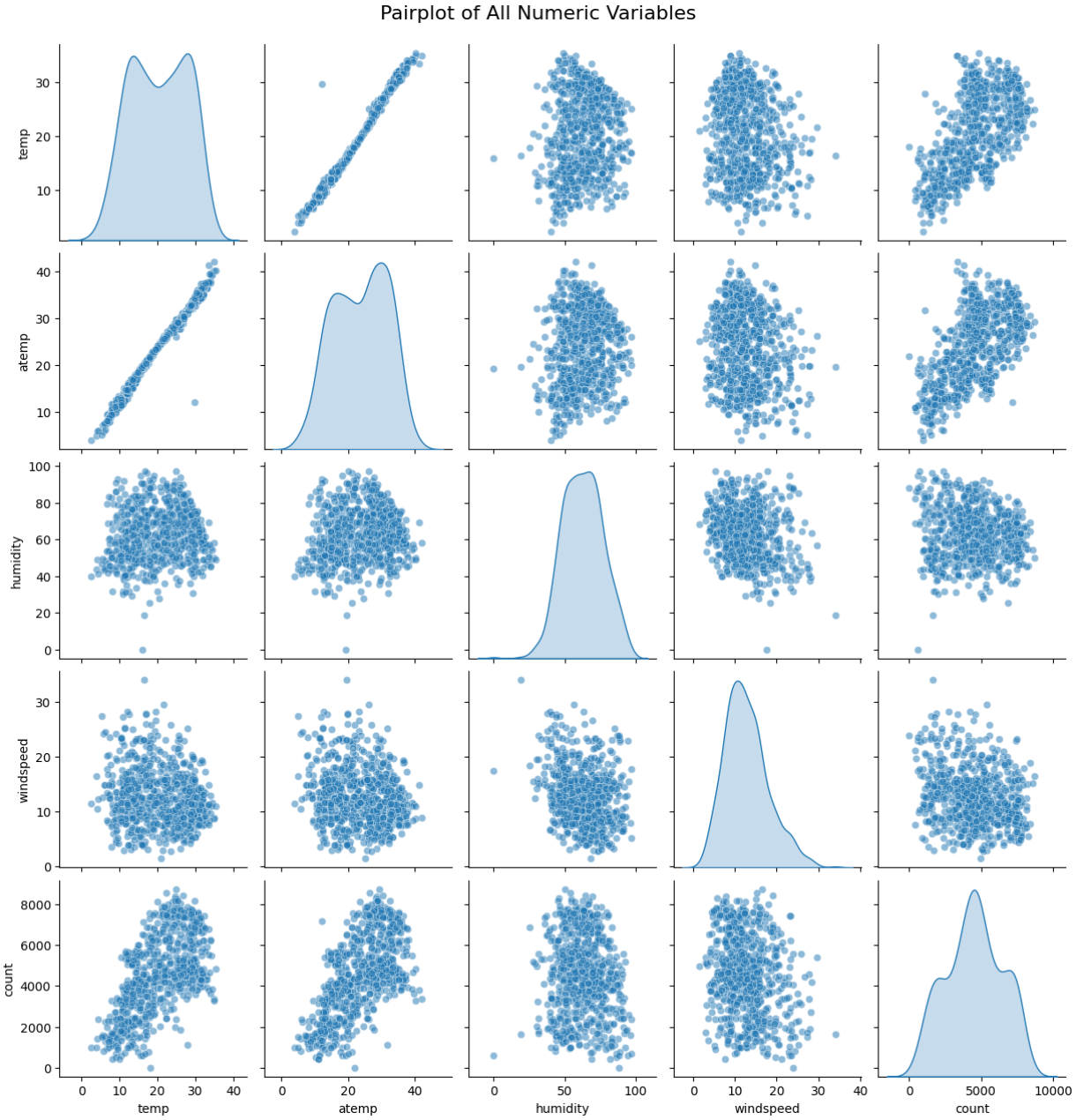
Reducing Redundancy: Using drop_first=True reduces the number of columns in the dataset, which can make computations faster and the dataset more manageable.

For example, in the provided dataset "Bike_Dataset", using drop_first=True for columns like "season," "month," "weekday," and "weathersit" will ensure that one dummy variable from each of these columns is dropped, and the dataset will not have redundant or highly correlated dummy variables. The resulting dataframe, as shown, will have columns for each category of the mentioned columns, minus one for each, serving as the reference category. This approach helps in achieving a more robust and interpretable regression model.

	year	holiday	workingday	temp	atemp	humidity	windspeed	count	season_Spring	season_Summer	...	month_Oct	month_Sep	weekday_Mon	weekday_Sat
0	0	0	0	14.110847	18.18125	80.5833	10.749882	985	True	False	...	False	False	False	True
1	0	0	0	14.902598	17.68695	69.6087	16.652113	801	True	False	...	False	False	False	False
2	0	0	1	8.050924	9.47025	43.7273	16.636703	1349	True	False	...	False	False	True	False
3	0	0	1	8.200000	10.60610	59.0435	10.739832	1562	True	False	...	False	False	False	False
4	0	0	1	9.305237	11.46350	43.6957	12.522300	1600	True	False	...	False	False	False	False
5	0	0	1	8.378268	11.66045	51.8261	6.000868	1606	True	False	...	False	False	False	False
6	0	0	1	8.057402	10.44195	49.8696	11.304642	1510	True	False	...	False	False	False	False
7	0	0	0	6.765000	8.11270	53.5833	17.875868	959	True	False	...	False	False	False	True
8	0	0	0	5.671653	5.80875	43.4167	24.250650	822	True	False	...	False	False	False	False
9	0	0	1	6.184153	7.54440	48.2917	14.958889	1321	True	False	...	False	False	True	False

10 rows × 30 columns

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



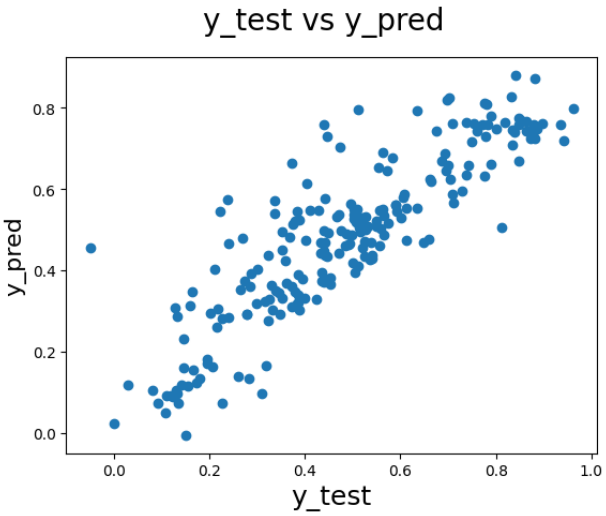
Based on the above pair-plot:

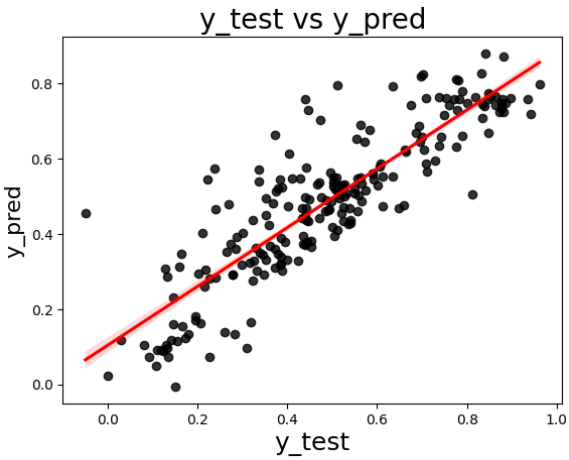
temp vs. count: There's a positive linear trend between "temp" and "count", suggesting a good correlation.

atemp vs. count: "atemp" (or "feeling" temperature) also displays a positive linear relationship with "count", quite similar to "temp".

From the visual observation of the pair-plot, it seems that "temp" and "atemp" both have the highest and similar correlation with the target variable "count.

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





Based on the provided scatter plot of y_test vs y_pred:

Linearity: The data points in the plot align with a general linear trend, indicating a strong linear relationship between the actual (y_test) and predicted (y_pred) values.

Homoscedasticity: The spread of residuals is fairly consistent across the range of the y_test values, supporting the assumption of homoscedasticity.

Independence of Residuals: The absence of clear patterns or clustering in the scatter plot suggests that the residuals are independent, which is favorable for the model's assumptions.

Model Performance: The clustering of data points around a hypothetical diagonal (representing perfect prediction) demonstrates that the model has performed effectively in predicting the test data.

In summary, the scatter plot of y_test vs y_pred visually validates the assumptions of linearity, homoscedasticity, and independence of residuals. The model appears to be effective in capturing the underlying patterns in the data and making accurate predictions.

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

To determine the top 3 features contributing significantly towards explaining the demand for shared bikes, we should consider both the statistical significance (p-value) and the magnitude of the coefficient (coef) from the regression results.

OLS Regression Results

Dep. Variable:	count	R-squared:	0.818
Model:	OLS	Adj. R-squared:	0.815
Method:	Least Squares	F-statistic:	224.7
Date:	Thu, 18 Jan 2024	Prob (F-statistic):	1.28e-177
Time:	08:04:45	Log-Likelihood:	446.50
No. Observations:	510	AIC:	-871.0
Df Residuals:	499	BIC:	-824.4
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1041	0.026	4.017	0.000	0.053	0.155
year	0.2527	0.009	27.743	0.000	0.235	0.271
workingday	0.0469	0.012	3.857	0.000	0.023	0.071
atemp	0.4707	0.034	13.658	0.000	0.403	0.538
season_Spring	-0.1272	0.017	-7.433	0.000	-0.161	-0.094
season_Winter	0.0958	0.015	6.276	0.000	0.066	0.126
month_Dec	-0.0740	0.019	-3.804	0.000	-0.112	-0.036
month_Nov	-0.0728	0.021	-3.409	0.001	-0.115	-0.031
month_Sep	0.0541	0.017	3.241	0.001	0.021	0.087
weekday_Sat	0.0602	0.016	3.788	0.000	0.029	0.091
weathersit_Light_snow_rain	-0.2672	0.027	-9.847	0.000	-0.321	-0.214

Omnibus:	92.842	Durbin-Watson:	1.957
Prob(Omnibus):	0.000	Jarque-Bera (JB):	241.585
Skew:	-0.905	Prob(JB):	3.47e-53
Kurtosis:	5.844	Cond. No.	14.9

	Features	VIF
2	atemp	4.68
1	workingday	4.32
4	season_Winter	2.27
0	year	2.03
8	weekday_Sat	1.70
6	month_Nov	1.61
3	season_Spring	1.48
5	month_Dec	1.34
7	month_Sep	1.18
9	weathersit_Light_snow_rain	1.06

From the provided OLS Regression Results:

atemp: It has a coefficient of 0.4707 and a p-value of 0.000, indicating strong statistical significance. Additionally, based on the VIF values, "atemp" has a VIF of 4.68, which, while being the highest among the listed variables, is still below the common threshold of 5, indicating that it is not highly multicollinear.

year: The coefficient for "year" is 0.2527 with a p-value of 0.000, suggesting it's statistically significant. The VIF for "year" is 2.03, which is comfortably low.

workingday: This variable has a coefficient of 0.6469, indicating a significant effect on the demand. Its p-value is 0.000, which is statistically significant. Its VIF value is 4.32, which, like "atemp", suggests that while it's among the higher VIF values in the dataset, it's still below the common threshold of 5.

Considering both the OLS Regression Results and the VIF values, the top 3 features contributing significantly towards explaining the demand for shared bikes are **atemp**, **year**, and **workingday**.

General Subjective Questions

1. **Explain the linear regression algorithm in detail. (4 marks)**

Linear regression is a supervised machine learning algorithm used for predicting a continuous outcome variable (dependent variable) based on one or more predictor variables (independent variables).

Machine learning models can be classified into the following three types based on the task performed and the nature of the output:

1. Regression: The output variable to be predicted is a continuous variable.

2. Classification: The output variable to be predicted is a categorical variable.

3. Clustering: No predefined notion of label allocated to groups/clusters formed.

The interpretability of linear regression is a notable strength. The model's equation provides clear coefficients that elucidate the impact of each independent variable on the dependent variable, facilitating a deeper understanding of the underlying dynamics. Its simplicity is a virtue, as linear regression is transparent, easy to implement, and serves as a foundational concept for more complex algorithms.

Linear regression is not merely a predictive tool; it forms the basis for various advanced models. Techniques like regularization and support vector machines draw inspiration from linear regression, expanding its utility. Additionally, linear regression is a cornerstone in assumption testing, enabling researchers to validate key assumptions about the data.

There are two main types of linear regression:

Simple Linear Regression

This is the simplest form of linear regression, and it involves only one independent variable and one dependent variable. The equation for simple linear regression is:

where:

Y is the dependent variable

X is the independent variable

β0 is the intercept

β1 is the slope

Multiple Linear Regression

This involves more than one independent variable and one dependent variable. The equation for multiple linear regression is:

where:

Y is the dependent variable

X1, X2, …, Xp are the independent variables

β0 is the intercept

β1, β2, …, βn are the slopes

In summary, linear regression involves representing a relationship between variables using a linear equation, defining a cost function to measure the model's performance, and optimizing the model parameters using an iterative process like gradient descent to minimize the cost and obtain the best-fitting line.

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

Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics but differ significantly when graphed.

It illustrates the importance of visualizing data and not relying solely on summary statistics.

Dataset composition:

- Anscombe's quartet consists of four datasets, each containing 11 data points. Each dataset has two variables: x (independent variable) and y (dependent variable).

Descriptive Statistics:

- Despite having identical or very similar summary statistics (mean, variance, correlation, and linear regression parameters), the datasets exhibit distinct patterns when graphed.

Illustration of the Importance of Visualization:

- Anscombe's quartet highlights the limitation of relying solely on summary statistics. Even if two datasets have similar mean, variance, and other summary measures, their underlying structures may differ.
- By graphing the data, it becomes evident that the datasets have different distributions, relationships between variables, and patterns of variability.

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

Pearson's r, is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It was developed by Karl Pearson and is widely used in statistics to assess the linear association between two variables.

The Pearson correlation coefficient can take values between -1 and +1:

r =1: Perfect positive linear relationship

r =-1: Perfect negative linear relationship

r =0: No linear relationship

The formula for calculating Pearson's correlation coefficient (r) between two variables X and Y with n data points is given by:

$$r = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2 \sum_{i=1}^{n}(Y_i-\bar{Y})^2}}$$

where:

Xi and Yi are the individual data points.

X(bar) and Y(bar) are the means of X and Y respectively.

Pearson's correlation coefficient is particularly useful for assessing the linear relationship between variables, but it assumes that the relationship is linear and that the data is approximately normally distributed. If the relationship is not linear, Pearson's "r" may not accurately capture the association between variables. In such cases, other correlation measures or non-linear regression techniques may be more appropriate.

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

Scaling in linear regression refers to the process of normalizing or standardizing the input features of the model. The purpose is to bring all the features to a similar scale, typically by transforming them in a way that their values have similar ranges. This is important because linear regression models are sensitive to the scale of the input features, and features on different scales can impact the performance and convergence of the model.

Scaling is performed for below reasons-

Improves Convergence: Scaling helps the optimization algorithm converge faster. When features are on a similar scale, the optimization process is more efficient, and it can find the optimal coefficients for the features more quickly.

Equalizes Variable Influence: Scaling ensures that all variables contribute to the model fitting process more uniformly. Without scaling, features with larger magnitudes can dominate the learning process, leading to an unbalanced influence on the model.

Facilitates Interpretability: Scaling doesn't affect the interpretation of the coefficients in terms of the feature importance. It just helps the optimization process. The relationships and significance of coefficients remain the same.

Normalized Scaling (Min-Max Scaling):

Scales the features to a specific range, usually between 0 and 1. Normalized scaling is sensitive to outliers because it depends on the range of the data.

Standardized Scaling (Z-score Scaling):

Standardizes the features to have a mean of 0 and a standard deviation of 1.

Standardized scaling is less sensitive to outliers because it uses the mean and standard deviation, which are less affected by extreme values.

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

The Variance Inflation Factor (VIF) is a measure used to assess the severity of multicollinearity in a multiple regression analysis. High VIF values indicate that the variance of the estimated regression coefficients is inflated due to collinearity among the predictor variables.

However, in case VIF is infinity that is mean the R-Square value is 1, as the VIF has (1-r^2) in denominator.

In case r^2 is 1 which represent that we are able to predict all the values 100% which is not the ideal case and it is overfitting the model, basically model has memorized all the value which is not the good model.

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

A Q-Q (quantile-quantile) plot is a graphical tool used to assess whether a dataset follows a particular theoretical distribution, such as the normal distribution. In the context of linear regression, Q-Q plots are often employed to check the normality assumption of the residuals.

Here's an explanation of the use and importance of Q-Q plots in linear regression:

Use of Q-Q Plot in Linear Regression:

Assumption Checking:

One of the key assumptions in linear regression is the normality of the residuals (the differences between the observed and predicted values). Q-Q plots help visualize whether the residuals follow a normal distribution.

Comparing Distributions:

The Q-Q plot compares the quantiles of the observed residuals with the quantiles expected from a theoretical normal distribution. If the points on the Q-Q plot closely follow a straight line, it indicates that the residuals are approximately normally distributed.