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)**

Categorical variables in our dataset played important role on dependent variables, as they are situational and affected the human behaviour to make a decision. As in case of working day which we converted in dummy variable affects the public gathering on certain place, and same phenomenon goes with other categorical variable like “holiday” and “weekday”.

For Example: It is visible in our final summery Seasons (spring, summer, winter): The coefficients for the seasonal variables are -0.0811, 0.0389, and 0.0774 for spring, summer, and winter, respectively. All of these variables have low p-values, indicating statistical significance. These coefficients represent the change in the count of rental bikes relative to the reference season (fall).

Spring has a negative effect, suggesting a decrease in bike rentals compared to fall.

Summer and winter have positive effects, suggesting an increase in bike rentals compared to fall.

OLS Regression Results

Dep. Variable:	cnt	**R-squared:**	0.829
Model:	OLS	**Adj. R-squared:**	0.825
Method:	Least Squares	**F-statistic:**	241.2
Date:	Wed, 10 Jan 2024	**Prob (F-statistic):**	6.16e-184
Time:	16:46:23	**Log-Likelihood:**	488.65
No. Observations:	510	**AIC:**	-955.3
Df Residuals:	499	**BIC:**	-908.7
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2291	0.029	7.776	0.000	0.171	0.287
yr	0.2340	0.008	27.876	0.000	0.217	0.250
holiday	-0.0956	0.027	-3.593	0.000	-0.148	-0.043
temp	0.4669	0.033	13.952	0.000	0.401	0.533
windspeed	-0.1544	0.026	-6.036	0.000	-0.205	-0.104
spring	-0.0811	0.021	-3.950	0.000	-0.121	-0.041
summer	0.0389	0.014	2.820	0.005	0.012	0.066
winter	0.0774	0.017	4.661	0.000	0.045	0.110
light_snow	-0.2842	0.025	-11.268	0.000	-0.334	-0.235
mist	-0.0773	0.009	-8.676	0.000	-0.095	-0.060
monday	-0.0510	0.012	-4.273	0.000	-0.074	-0.028

Omnibus:	65.805	**Durbin-Watson:**	2.039
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	171.485
Skew:	-0.648	**Prob(JB):**	5.79e-38
Kurtosis:	5.527	**Cond. No.**	17.0

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

As per formula for number of dummy variable (n-1), where n represent the number of distinct values present in categorical variable.

When one or more dummy variables can be predicted with high accuracy from the values of the other dummy variables in the dataset. This situation introduces multicollinearity in the regression analysis, which can lead to issues in estimating the regression coefficients and interpreting the results.

Setting drop_first=True when creating dummy variables helps to eliminate the multicollinearity issue by dropping one of the dummy variables for each categorical feature. The dropped variable becomes the reference category, and the presence or absence of the other dummy variables implicitly conveys information about that reference category.

```
status = pd.get_dummies(dayData['season'], dtype='int', drop_first=True)
status.head()

dayData = pd.concat([dayData, status], axis=1)
dayData = dayData.drop('season', axis=1)
dayData.head()
```

✓ 0.0s

	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt	spring	summer	winter
0	0	1	0	sunday	0	mist	14.110847	18.18125	80.5833	10.749882	331	654	985	1	0	0
1	0	1	0	monday	0	mist	14.902598	17.68695	69.6087	16.652113	131	670	801	1	0	0
2	0	1	0	tuesday	1	clear	8.050924	9.47025	43.7273	16.636703	120	1229	1349	1	0	0
3	0	1	0	wednesday	1	clear	8.200000	10.60610	59.0435	10.739832	108	1454	1562	1	0	0
4	0	1	0	thrusday	1	clear	9.305237	11.46350	43.6957	12.522300	82	1518	1600	1	0	0

000 will correspond to Fall

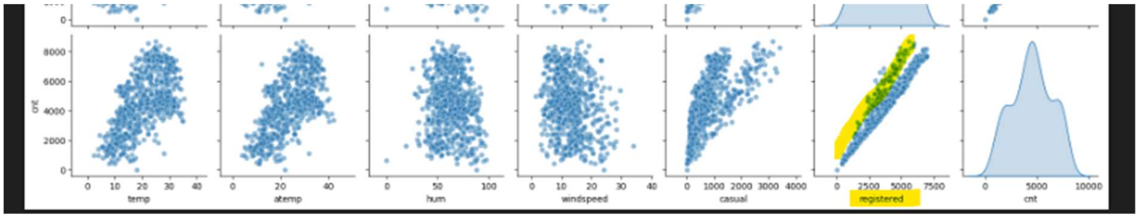
100 will correspond to Spring

001 will correspond to Winter

010 will correspond to Summer

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

As per our pair-plot “registered” numerical variable has the highest correlation with target variable (cnt).



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

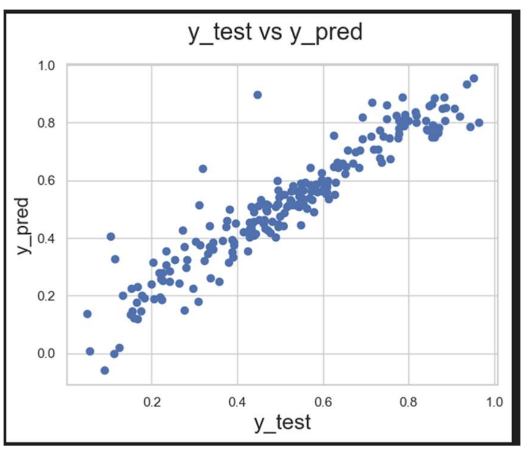
Validating the assumptions of linear regression is a crucial step after building the model on the training set. Here are several common approaches for validating the assumptions

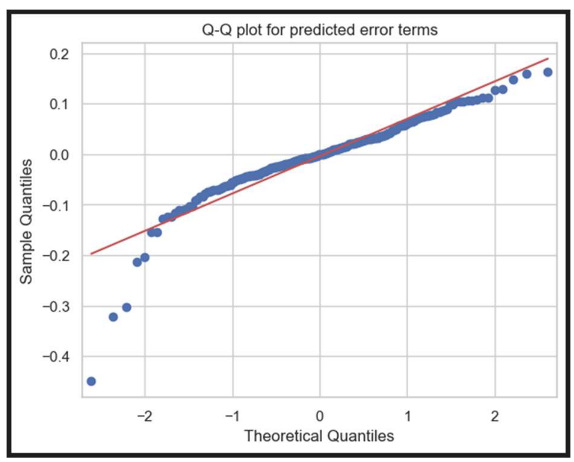
We have used the residual analysis technique to validate our assumptions.

We have examined the residuals (the differences between the observed and predicted values) for dayData prediction model created using RFE.

Below is the plotted scatterplot of residuals against predicted values to check for linearity, and against each predictor variable to identify potential heteroscedasticity.

Utilize residual plots, such as prediction plot and a Q-Q plot, to assess the normality assumption of residuals.





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

As per our final model the top 3 feature are:

1) Temp
2) Light_snow
3) Yr

The coefficient of these three features is very high in order to predict the shared bikes count

As it is visible in below model summery.

OLS Regression Results

Dep. Variable:	cnt	R-squared:	0.829
Model:	OLS	Adj. R-squared:	0.825
Method:	Least Squares	F-statistic:	241.2
Date:	Wed, 10 Jan 2024	Prob (F-statistic):	6.16e-184
Time:	16:46:23	Log-Likelihood:	488.65
No. Observations:	510	AIC:	-955.3
Df Residuals:	499	BIC:	-908.7
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2291	0.029	7.776	0.000	0.171	0.287
yr	0.2340	0.008	27.876	0.000	0.217	0.250
holiday	-0.0956	0.027	-3.593	0.000	-0.148	-0.043
temp	0.4669	0.033	13.952	0.000	0.401	0.533
windspeed	-0.1544	0.026	-6.036	0.000	-0.205	-0.104
spring	-0.0811	0.021	-3.950	0.000	-0.121	-0.041
summer	0.0389	0.014	2.820	0.005	0.012	0.066
winter	0.0774	0.017	4.661	0.000	0.045	0.110
light_snow	-0.2842	0.025	-11.268	0.000	-0.334	-0.235
mist	-0.0773	0.009	-8.676	0.000	-0.095	-0.060
monday	-0.0510	0.012	-4.273	0.000	-0.074	-0.028

Omnibus:	65.805	Durbin-Watson:	2.039
Prob(Omnibus):	0.000	Jarque-Bera (JB):	171.485
Skew:	-0.648	Prob(JB):	5.79e-38
Kurtosis:	5.527	Cond. No.	17.0

Also it is clearly visible that all the P-values of feature is below 0.05 which makes it good fit in model.

Also below is the VIF for the same model

```python
vif = pd.DataFrame()
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
3	windspeed	4.60
2	temp	3.62
0	yr	2.07
4	spring	1.98
5	summer	1.80
6	winter	1.64
8	mist	1.53
9	monday	1.18
7	light_snow	1.08
1	holiday	1.04

As none of the feature has VIF > 5, makes it good fit for predictions with 82% accuracy.

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.