Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 1 goes below this line> (Do not edit)

The following are the conclusions based on dataset for categorical variables.

a) **'season'** : **summer** and **fall** are the two seasons where the demand of bike rental increases, so we need to plan the business accordingly in **summer** and **fall season** where as in spring season has the lowest demand.
b) **'mnth'** : months June, July , August, September has peak demand of bike rentals as per the box plot
c) **'weakday'** : there is no significant pattern of users renting bike on a particular day. All the weekdays business is almost same.
d) **'weathersit'** : weathersit or weather situation column shows that if the weather conditions are good or clear there will be significant increase in Bike rental.
e) **'Year' or 'yr':** This shows that there is an increase in Bike rental from 2018 to 2019

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: <Your answer for Question 2 goes below this line> (Do not edit)

Using one-hot encoding the dummy variables are created to cover the range of values of categorical variable. Each dummy variable have 1 and 0 values. 1 is used to depict the presence and 0 for absence of the respective category. This means if the category variable has 3 categories, there will be 3 dummy variables.

The drop_first = True is used while creating dummy variables to drop the base/reference category. The reason for this is to avoid the multi-collinearity

getting added into the model if all dummy variables are included. The reference category can be easily deduced where 0 is present in a single row for all the other dummy variables of a particular category.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

Total Marks: 1 mark (Do not edit)

Answer: <Your answer for Question 3 goes below this line> (Do not edit)

Lokking at the pare-plot for numerical variable we can conclude that **temp** and **atemp** variables has the highest correlation but since **atemp** is derived variable we will not include this for our model.

Temp has correlation with target variable cnt with 0.64 and atemp has 0.65

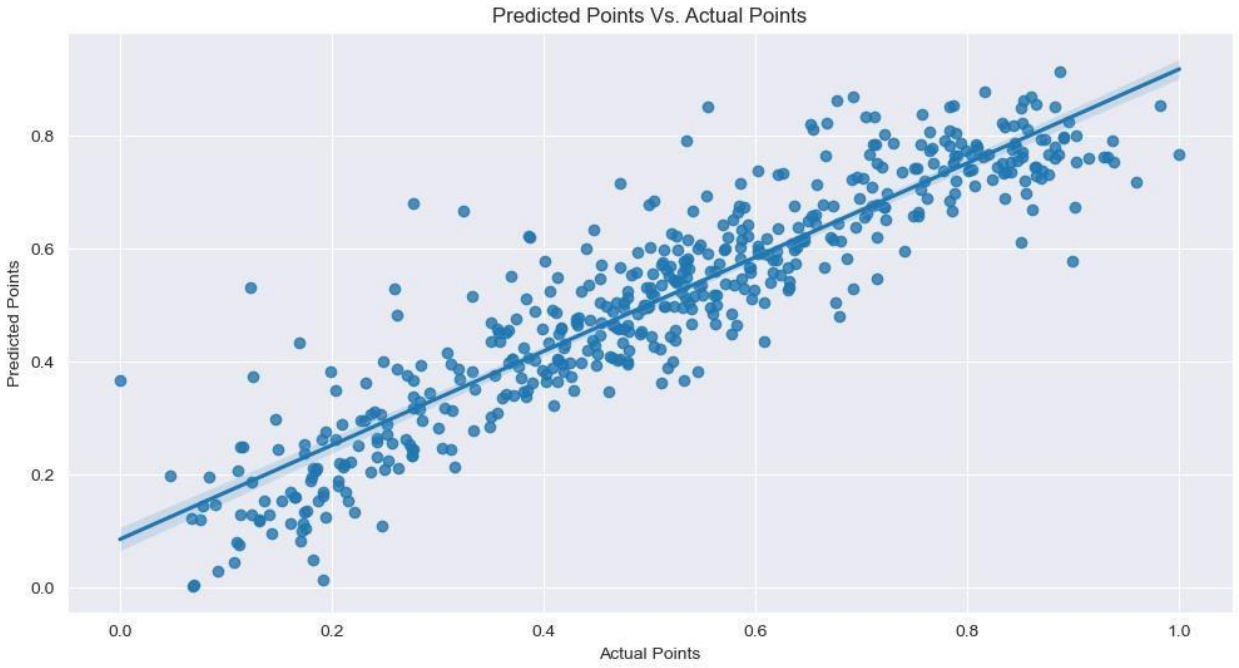
Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 4 goes below this line> (Do not edit)

Steps Validation of assumptions of linear regression

Step 1: The dependent variable and independent variable must have a linear relationship. To check this, we have used predicted vs actual points in regplot as below.



Step 2 : No Autocorrelation in residuals Use Durbin-Watson Test.
If DW = 2 would be the ideal case here and with summery we got DW as 2.03 which is 2 so no Autocorrelation.

```
print(lm5.summary())
```

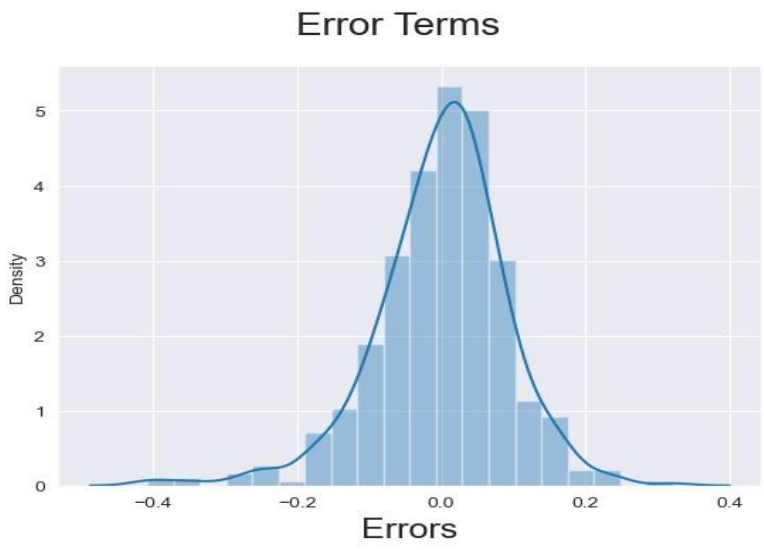
OLS Regression Results

Dep. Variable:	cnt	R-squared:	0.833
Model:	OLS	Adj. R-squared:	0.829
Method:	Least Squares	F-statistic:	248.2
Date:	Wed, 13 Sep 2023	Prob (F-statistic):	1.69e-186
Time:	20:40:51	Log-Likelihood:	494.31
No. Observations:	510	AIC:	-966.6
Df Residuals:	499	BIC:	-920.0
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-468.9122	16.735	-28.020	0.000	-501.791	-436.033
year	0.2324	0.008	28.033	0.000	0.216	0.249
temp	0.5721	0.022	25.693	0.000	0.528	0.616
windspeed	-0.1559	0.025	-6.176	0.000	-0.206	-0.106
summer	0.0810	0.011	7.366	0.000	0.059	0.103
winter	0.1271	0.011	11.993	0.000	0.106	0.148
july	-0.0366	0.018	-1.983	0.048	-0.073	-0.000
sep	0.0873	0.016	5.309	0.000	0.055	0.120
sun	-0.0457	0.012	-3.879	0.000	-0.069	-0.023
Light_snowrain	-0.2832	0.025	-11.361	0.000	-0.332	-0.234
Misty	-0.0808	0.009	-9.138	0.000	-0.098	-0.063

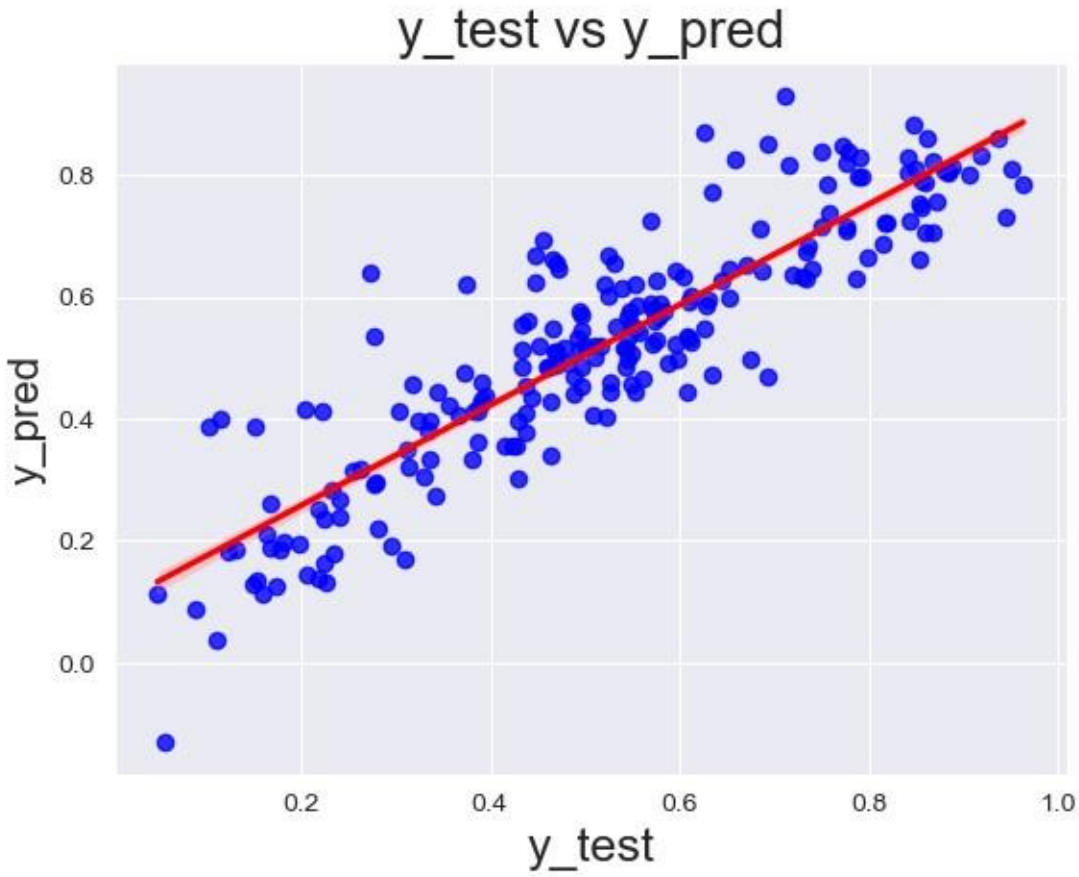
Omnibus:	67.621	Durbin-Watson:	2.023
Prob(Omnibus):	0.000	Jarque-Bera (JB):	159.667
Skew:	-0.699	Prob(JB):	2.13e-35
Kurtosis:	5.358	Cond. No.	8.22e+06

Step 3 : Residuals must be normally distributed. We plotted residual data with distplot and found Error Terms are normally Distributed with mean Zero



Finally Comparison

Let’s compare the two models and see if there is any improvement.



Comparision between Training and Testing dataset:
Train dataset R^2 : 0.833
Test dataset R^2 : 0.7922
Train dataset Adjusted R^2 : 0.829 Test
dataset Adjusted R^2 : 0.760

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: <Your answer for Question 5 goes below this line> (Do not edit)

The top 3 variables contributing the significantly towards the demand in shared bikes are **temp, year, season**. Every year there is an increase in demand of shared bikes. Also if the temperature is good demand for shared bike increases and lastly if the season is summer or fall there is a huge demand of shared bikes.

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between a dependent variable and one or more independent features. When the number of the independent feature, is 1 then it is known as Univariate Linear regression, and in the case of more than one feature, it is known as multivariate linear regression. The goal of the algorithm is to find the best linear equation that can predict the value of the dependent variable based on the independent variables. The equation provides a straight line that represents the relationship between the dependent and independent variables. The slope of the line indicates how much the dependent variable changes for a unit change in the independent variable(s).

Assumption for Linear Regression Model

Linear regression is a powerful tool for understanding and predicting the behavior of a variable, however, it needs to meet a few conditions in order to be accurate and dependable solutions.

1. **Linearity**: The independent and dependent variables have a linear relationship with one another. This implies that changes in the dependent variable follow those in the independent variable(s) in a linear fashion.
2. **Independence**: The observations in the dataset are independent of each other. This means that the value of the dependent variable for one observation does not depend on the value of the dependent variable for another observation.
3. **Homoscedasticity**: Across all levels of the independent variable(s), the variance of the errors is constant. This indicates that the amount of the independent variable(s) has no impact on the variance of the errors.
4. **Normality**: The errors in the model are normally distributed.
5. **No multicollinearity**: There is no high correlation between the independent variables. This indicates that there is little or no correlation between the independent variables.

Formula for Linear regression

- Simple Linear Regression – Single independent variable is used $Y = \beta 0 + \beta 1X$ is the line equation used for SLR.
- Multiple Linear Regression – Multiple independent variables are used

  $Y = \beta 0 + \beta 1X1 + \cdots + \beta pXp + \in$ is the line equation for MLR.

  Here

  $\beta 0$ is the $value\ of\ the\ Y\ when\ X=0$ ($Y\ intercept$)

$\beta 1, \beta 2, \ldots, \beta p$ are $Slope\ or\ the\ gradient$.

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Anscombe's quartet is a set of four data sets created by statistician Francis Anscombe in 1973. The key idea behind the quartet is to show that descriptive statistics like the mean, variance, and correlation can be the same for different data sets, but the data sets can look very different when graphed.

All four data sets in the quartet have:

- The same mean values for both X and Y
- The same variance for X and Y
- The same correlation (0.82) between X and Y
- The same regression line ($y = 3 + 0.5x$)

All four datasets have the same mean and variance for both X and Y, and the same correlation (0.82) between X and Y.

However, when plotted, the data sets reveal different patterns:

- One has a linear relationship.
- One shows a curved relationship.
- One contains an outlier that distorts the regression line.
- One appears linear but is affected by an outlier.

The purpose of Anscombe's quartet is to highlight that statistical summaries (like mean and correlation) can be misleading if you don't visually inspect the data. It reminds us that graphical methods are important in understanding the true nature of data.

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

The Pearson R also known as the correlation coefficient is a descriptive statistic, meaning that it summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables.

The **Pearson correlation coefficient (*r*)** is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.
If the r value is between 0 to 1 its positively correlated and if r values is in between -1 and 0 then the variables are negatively correlated and If r value is 0 then there is no correlation.

The Pearson correlation coefficient (r) is one of several correlation coefficients that you need to choose between when you want to measure a correlation. The Pearson correlation coefficient is a good choice when all of the following are true:

- Both variables are quantitative: You will need to use a different method if either of the variables is qualitative.
- The variables are normally distributed: You can create a histogram of each variable to verify whether the distributions are approximately normal. It's not a problem if the variables are a little non-normal.
- The data have no outliers: Outliers are observations that don't follow the same patterns as the rest of the data. A scatterplot is one way to check for outliers—look for points that are far away from the others.
- The relationship is linear: "Linear" means that the relationship between the two variables can be described reasonably well by a straight line. You can use a scatterplot to check whether the relationship between two variables is linear.

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: Please write your answer below this line. (Do not edit)

What - The scaling is the data preparation step for regression model. The scaling normalizes these varied datatypes to a particular data range.

Why – Most of the times the feature data is collected at public domains where the interpretation of variables and units of those variables are kept open collect as much as possible. This results in to the high variance in units and ranges of data. If scaling is not done on these data sets, then the chances of processing the data without the appropriate unit conversion are high. Also the higher the range then higher the possibility that the coefficients are impaired to compare the dependent variable variance. The scaling only affects the coefficients. The prediction and precision of prediction stays unaffected after scaling.

Normalization/Min-Max scaling – The Min max scaling normalizes the data within the range of 0 and 1. The Min max scaling helps to normalize the outliers as well.

$MinMaxScaling: x = x - \min(x) \max(x) - \min(x)$
Standardization converges all the data points into a standard normal distribution where mean is 0 and standard deviation is 1.
$Standardization: x = x - mean(x) sd(x)$

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

$VIF = 1/(1 - R2)$
The VIF formula clearly signifies when the VIF will be infinite. If the R2 is 1 then the VIF is infinite. The reason for R2 to be 1 is that there is a perfect correlation between 2 independent variables.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
(Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Q-Q plots are the quantile-quantile plots. It is a graphical tool to assess the 2 data sets are from common distribution. The theoretical distributions could be of type normal, exponential or uniform. The Q-Q plots are useful in the linear regression to identify the train data set and test data set are from the populations with same distributions. This is another method to check the normal distribution of the data sets in a straight line with patterns explained below

Interpretations

Similar distribution: If all the data points of quantile are lying around the straight line at an angle of 45 degree from x-axis.

- Y values < X values: If y-values quantiles are lower than x-values quantiles.
- X values < Y values: If x-values quantiles are lower than y-values quantiles.
- Different distributions – If all the data points are lying away from the straight line.

Advantages

- Distribution aspects like loc, scale shifts, symmetry changes and the outliers all can be daintified from the single plot.
- The plot has a provision to mention the sample size as well.