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)

1.) Season:

Autumn season witnessed increase in shared bikes indicating this is the good season for many customers

2.) Weathersit:

Good weather witnessed increase in shared bikes indicating it as a appropriate time for customers

3.) Month:

September month witnessed increase in shared bikes indicating it as a good month for both the company and the customers

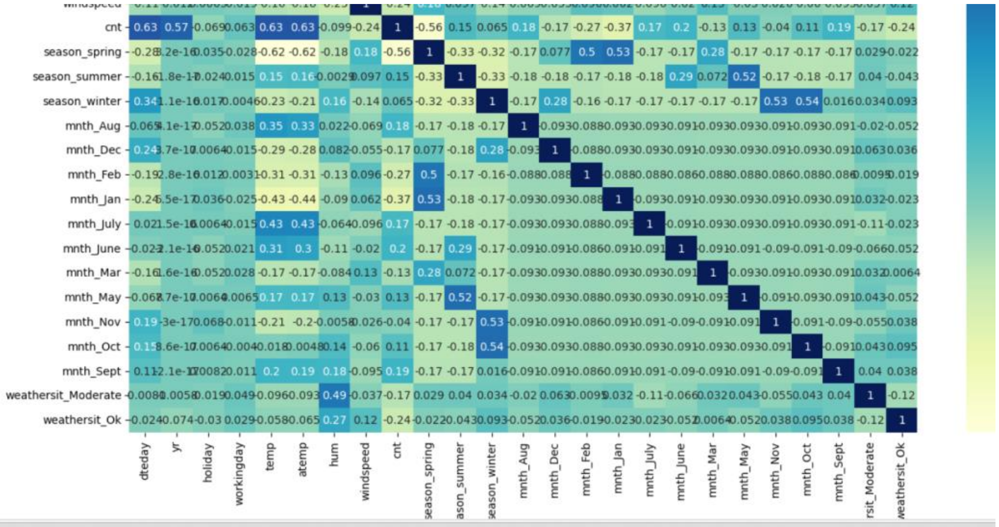
---

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)

It deletes the first column and reduces the correlation among dummy variables, as a result impact of one doesn't affect the another



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)

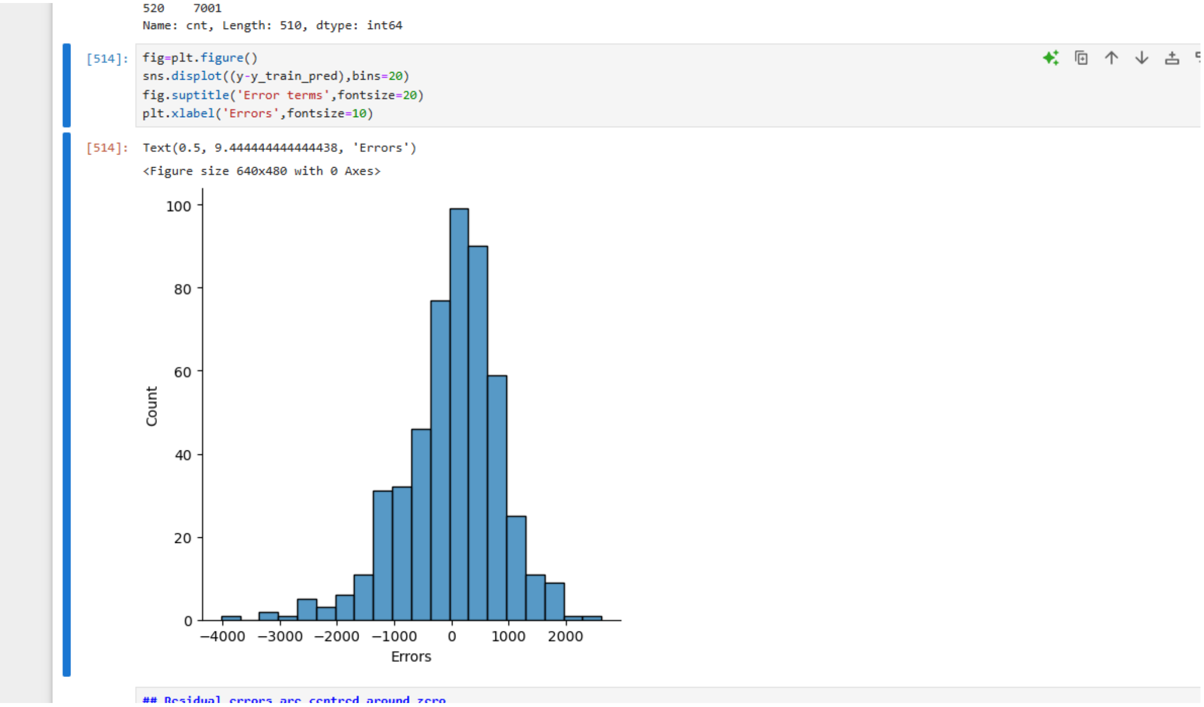
Answer: <Your answer for Question 3 goes below this line> (Do not edit)

Temp, atemp and windspeed as the increase in temperature is positively impacting the cnt variable

---

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)

1. Residual Analysis



```
520    7001
Name: cnt, Length: 510, dtype: int64
```

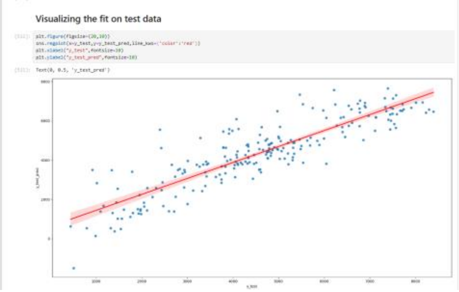
```
[514]: fig=plt.figure()
       sns.displot((y-y_train_pred),bins=20)
       fig.suptitle('Error terms',fontsize=20)
       plt.xlabel('Errors',fontsize=10)
```

```
[514]: Text(0.5, 9.444444444444438, 'Errors')
       <Figure size 640x480 with 0 Axes>
```

```
## Residual errors are centred around zero
```

Residuals are centered at zero, so Residuals are normally distributed, hence linear regression is valid

2. Linearity



Visualizing the fit on test data

```
plt.figure(figsize=(20,10))
sns.regplot(x=y_test,y=y_test_pred,line_kws={'color':'red'})
plt.xlabel("y_test",fontsize=18)
plt.ylabel("y_test_pred",fontsize=18)
```

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

Linear relationship between predicted and test variables

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)

1) Temp
2) Windspeed
3) Season_summer

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)

<Your answer for Question 6 goes here>

Linear Regression is a type of supervised machine learning algorithm, which contains independent and dependent variables(Output Variable) which is continuous

Dependent variables are predicted based on independent parameters by fitting a best line

Types:

1) Simple Linear Regression
2) Multiple Linear Regression

Simple Linear Regression:
Contains only one independent variable
Formula:
Y=b0+b1*x+e
Where b0 is the y-intercept when x=0
b1 is the slope of the line
x=independent variable
e= error term

Multiple Linear Regression:
Contains more than one independent variable
Formula:
Y=b0+b1x1+b2x2+…..+bnxn+e

Where (x1,x2,x3….xn) are independent variables
(b0,b1,b2…..bn) are coefficients
E is the error term

Objective Function

Goal is to find the value that minimizes the squared differences between actual and predicted values

Mean Squared Error(MSE)

$$L = \frac{1}{N}\sum_{i=1}^{N}(\hat{y_i} - y_i)^2$$

- *Yi bar* is the predicted value for the ith data point.

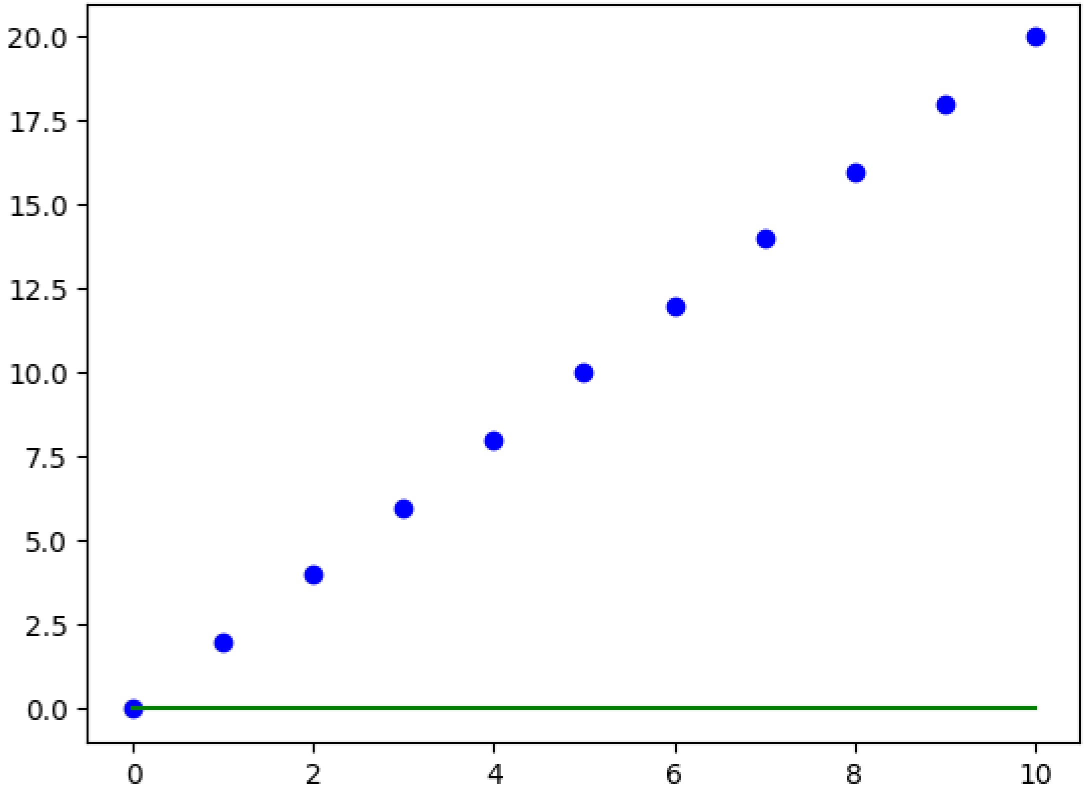
- *Yi is the actual value for the ith data point*
- *N= no.of data points*

Optimization

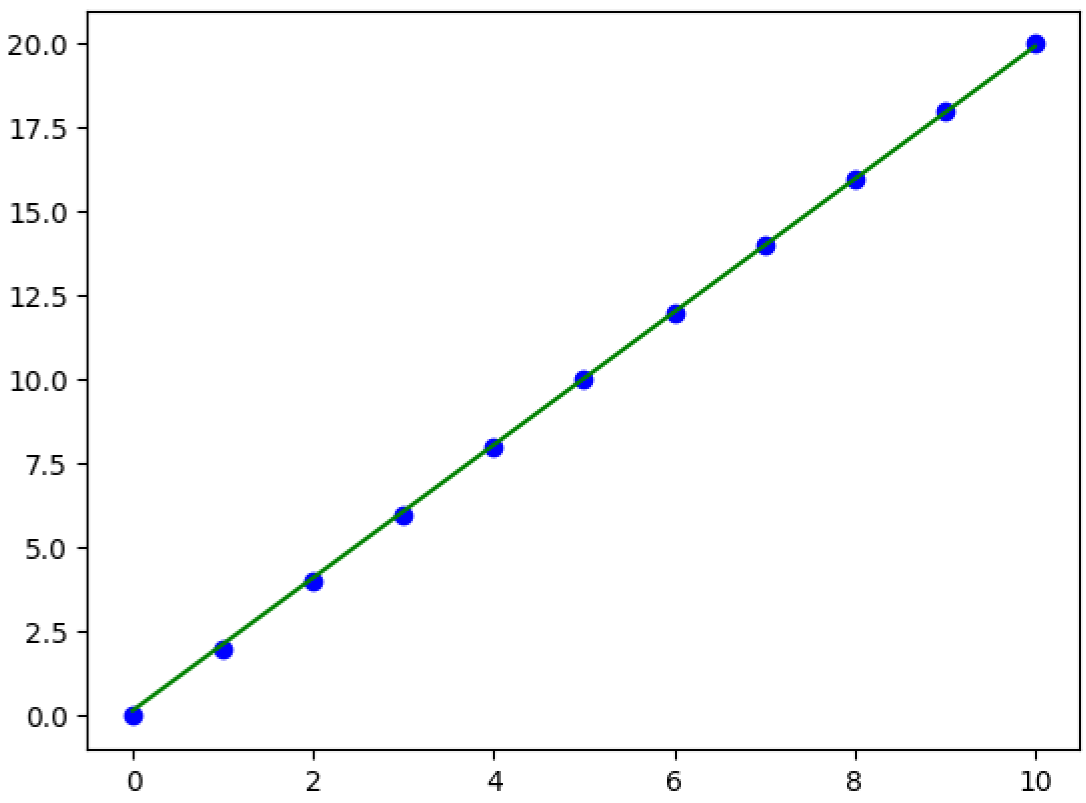
Gradient Descent is an iterative optimization algorithm that finds the optimal value(Minimum/ Maximum)

The main aim is to find the best parameter which gives high accuracy on train and test data

Regression line before gradient descent

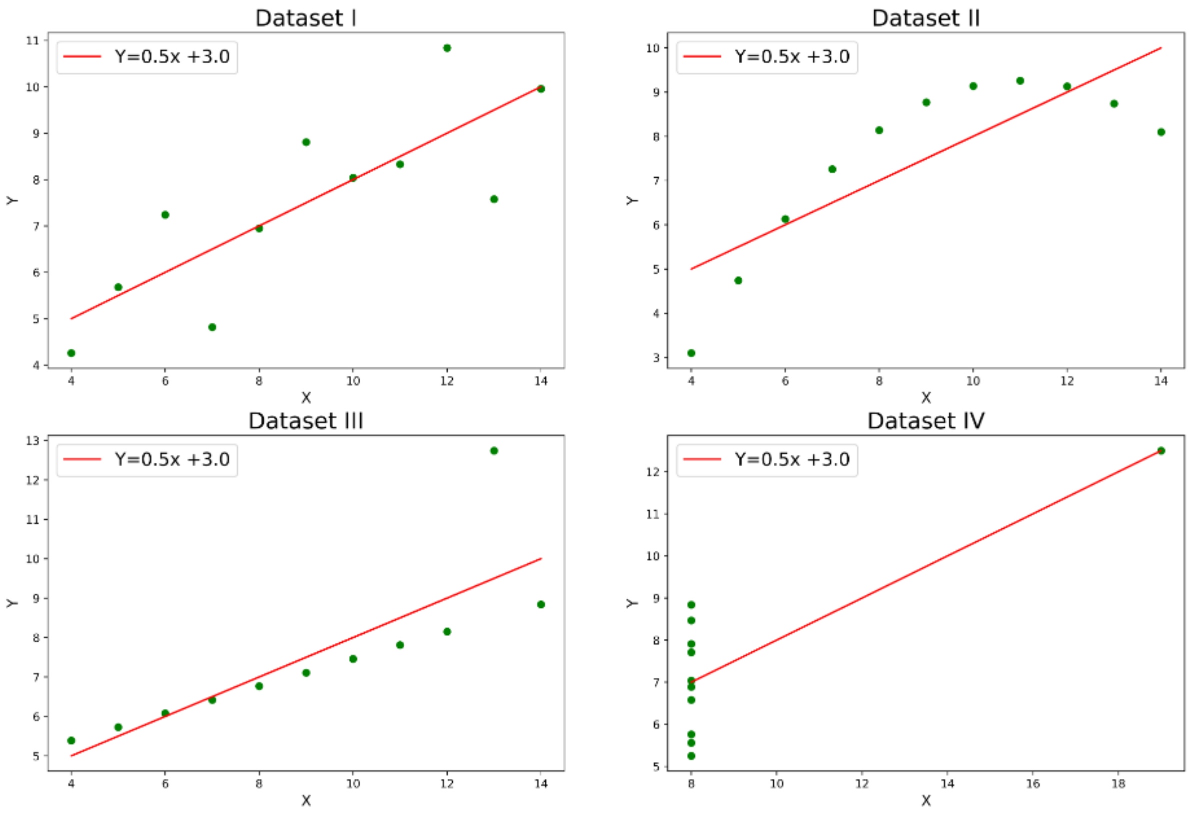


Regression Line after gradient descent



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)

<Your answer for Question 7 goes here>
Anscombe quartet consists of 4 datasets , having identical statistical properties in terms of mean, variance, R-squared and correlation, but differs in visual representation



Importance of Anscombe quartet:

- Plotting data to see distribution of samples
- Identifying anomalies in data, such as outliers
- Understanding the relationship between 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)

<Your answer for Question 8 goes here>

Pearson Correlation coefficient(r) is used for measuring linear correlation, it usually varies between -1 and 1

Positive correlation:
when one variable increases the other variable also increases ,value: 1

Negative correlation
when one variable increases the other variable also decreases, value: between 0 to -1

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

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)

<Your answer for Question 9 goes here>

Scaling is the process of transforming the independent features into fixed range, so that one independent feature will not dominate others in predicting the target variable

Types of scaling:

1) **Min Max**
2) **Standard**

Min Max(Normalized) scaling:
It fits the values between 0 and 1

$$X_{\text{scaled}} = \frac{X_i - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}$$

Advantages: Useful when variable distribution is unknown
Disadvantages: prone to outliers, doesn’t preserve the shape of distribution

Standard Scaling:
It rescales the features such that mean is 0 and standard deviation is 1

$$X_{\text{scaled}} = \frac{X_i - X_{\text{mean}}}{\sigma}$$

Advantages: less prone to outliers, preserve the shape of distribution
Disadvantages: Assumes variables follow gaussian distribution

---

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)

<Your answer for Question 10 goes here>

$$VIF_i = \frac{1}{1 - R_i^2}$$

If there is a perfect correlation r^2=1 then VIF value is infinite

Handling Infinite VIF:

- Remove redundant features and combine features
- Rebuild the model

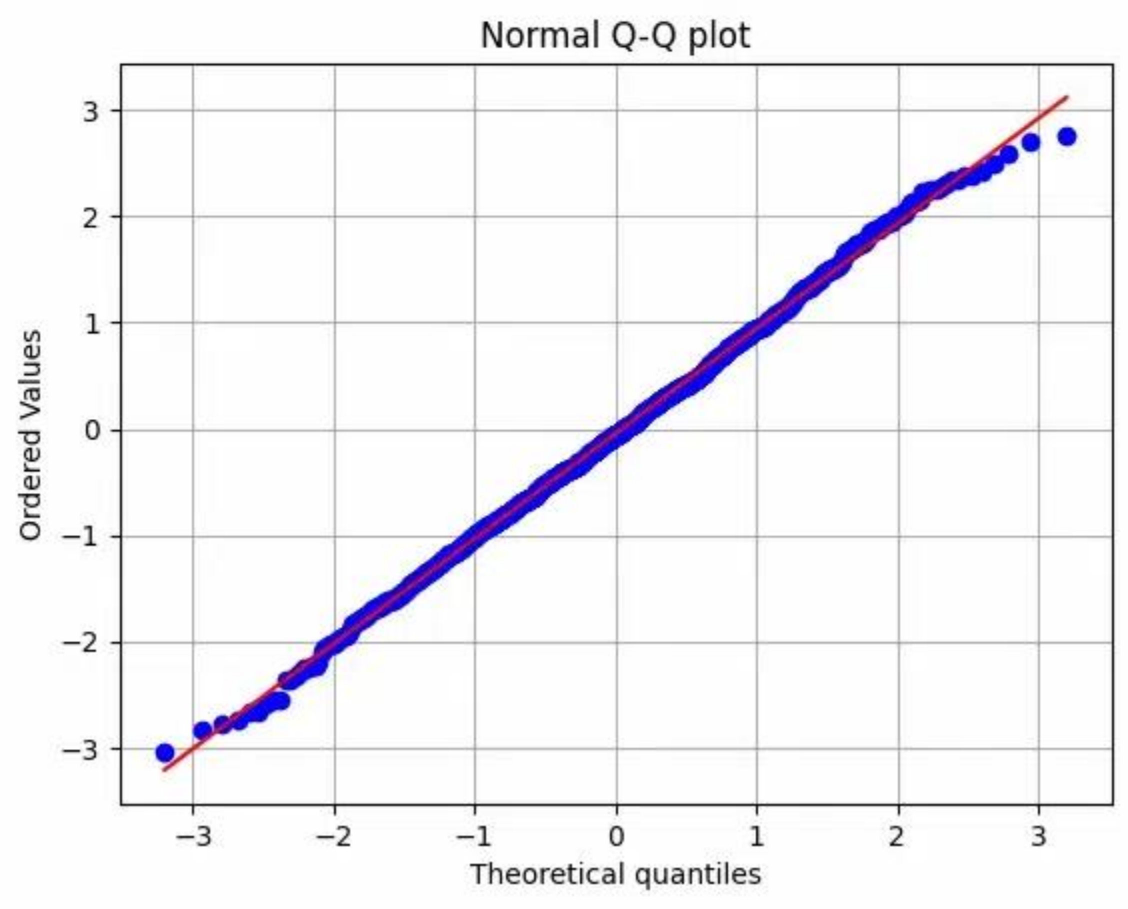
Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)

Total Marks: 3 marks (Do not edit)

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

<Your answer for Question 11 goes here>

Quantile-Quantile (Q-Q) plot is used for assessing whether the dataset Is normally distributed or not



Importance in Linear Regression:

1) Checks normality of residuals

2) Detecting deviations from normality

a)Fat tails

b) skewness