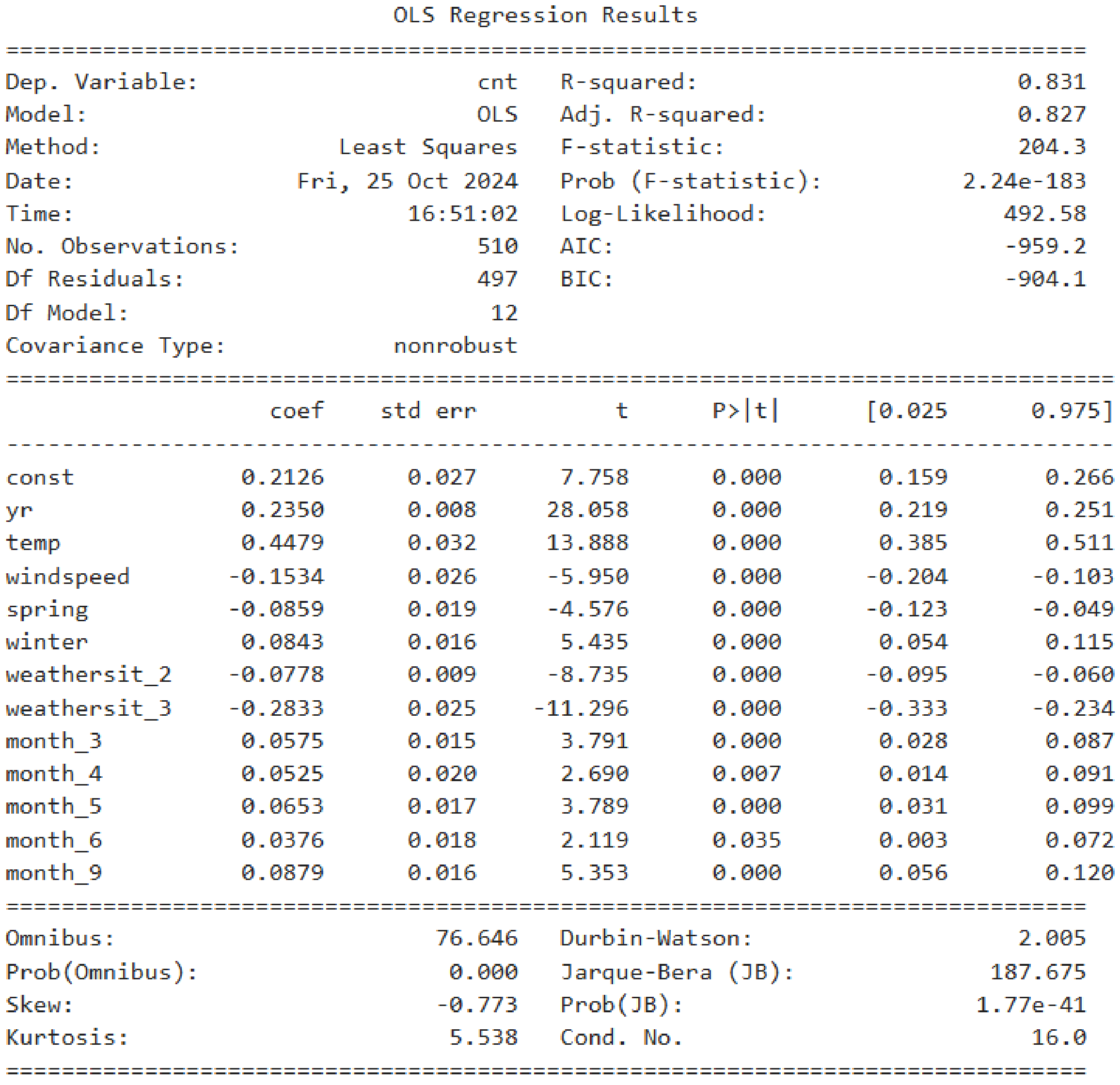
Assignment-based Subjective Questions

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 season and month are dummy coded, and below variables significant and explain the variance well.

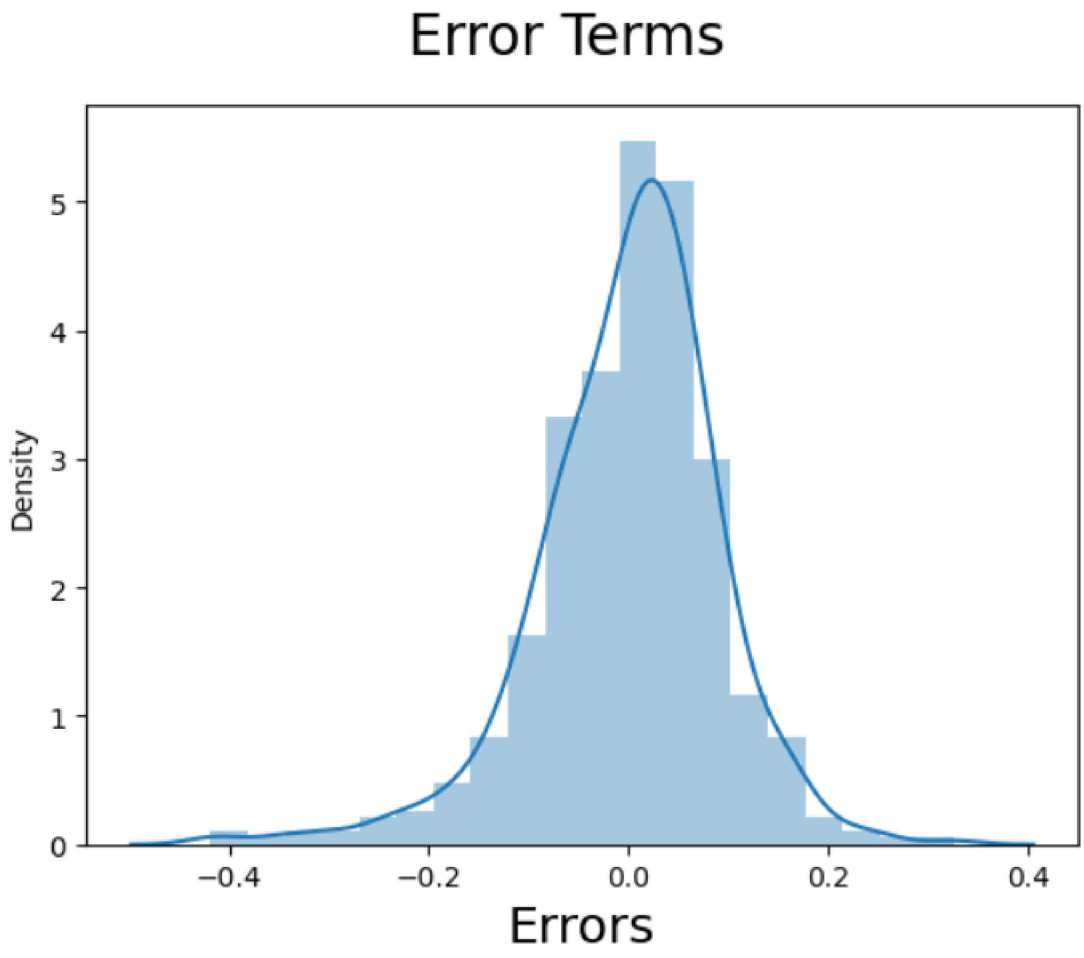


```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.831
Model:                            OLS   Adj. R-squared:                  0.827
Method:                 Least Squares   F-statistic:                     204.3
Date:                Fri, 25 Oct 2024   Prob (F-statistic):          2.24e-183
Time:                        16:51:02   Log-Likelihood:                 492.58
No. Observations:                 510   AIC:                            -959.2
Df Residuals:                     497   BIC:                            -904.1
Df Model:                          12
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const            0.2126      0.027      7.758      0.000       0.159       0.266
yr               0.2350      0.008     28.058      0.000       0.219       0.251
temp             0.4479      0.032     13.888      0.000       0.385       0.511
windspeed       -0.1534      0.026     -5.950      0.000      -0.204      -0.103
spring          -0.0859      0.019     -4.576      0.000      -0.123      -0.049
winter           0.0843      0.016      5.435      0.000       0.054       0.115
weathersit_2    -0.0778      0.009     -8.735      0.000      -0.095      -0.060
weathersit_3    -0.2833      0.025    -11.296      0.000      -0.333      -0.234
month_3          0.0575      0.015      3.791      0.000       0.028       0.087
month_4          0.0525      0.020      2.690      0.007       0.014       0.091
month_5          0.0653      0.017      3.789      0.000       0.031       0.099
month_6          0.0376      0.018      2.119      0.035       0.003       0.072
month_9          0.0879      0.016      5.353      0.000       0.056       0.120
==============================================================================
Omnibus:                       76.646   Durbin-Watson:                   2.005
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              187.675
Skew:                          -0.773   Prob(JB):                     1.77e-41
Kurtosis:                       5.538   Cond. No.                         16.0
==============================================================================
```

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)
Dummy variable trap might lead to multicollinearity issues and it will lead to violation of assumptions of linear regression. The model does not understand the text data so we need to convert them into numbers

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)
atemp has the highest correlation with target variable cnt

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)
We did the residual analysis to check if error terms are normally distributed. The histogram looks like below image where mean is 0 so residual follows a normal distribution.



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)

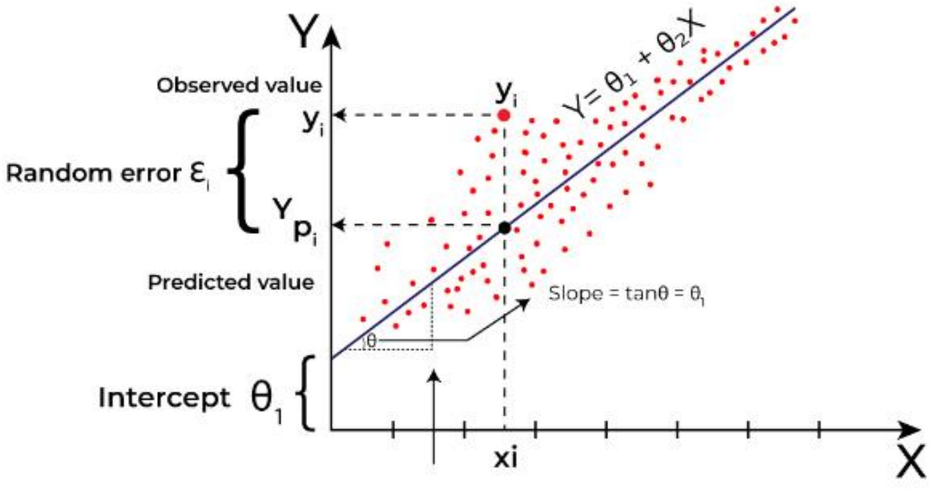
Top 3 features that contribute towards demand of the shared bikes are temp (coef -0.44), yr (coef – 0.23) and month Sep (coef – 0.08)

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 that computes the linear relationship between the dependent variable and one or more independent features by fitting a linear equation to observed data. When there is only one independent feature, it is known as Simple Linear Regression, and when there are more than one feature, it is known as Multiple Linear Regression. Similarly, when there is only one dependent variable, it is considered Univariate Linear Regression, while when there are more than one dependent variables, it is known as Multivariate Regression. Our primary objective while using linear regression is to locate the best-fit line, which implies that the error between the predicted and actual values should be kept to a minimum. There will be the least error in the best-fit line. The best Fit Line 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).


Y
Observed value
y_i
Random error ε_i
Y_{p_i}
Predicted value
$Y = \theta_1 + \theta_2 X$
y_i
Slope = tanθ = θ_1
θ
Intercept θ_1
xi
X

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)

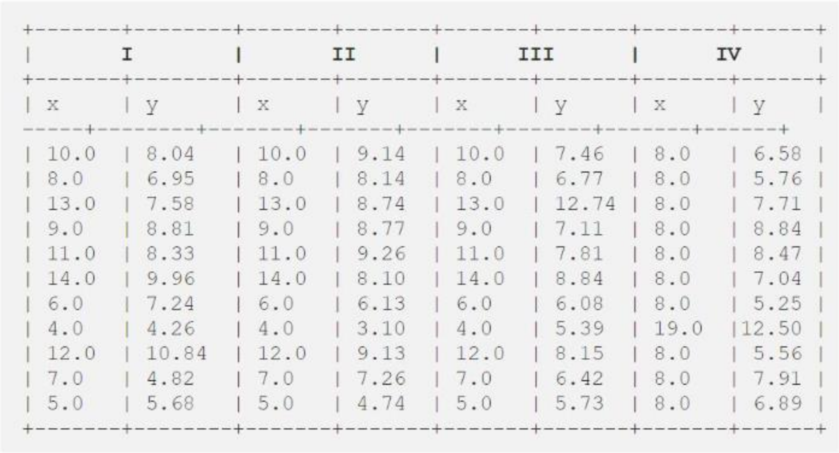
<Your answer for Question 7 goes here>

Anscombe's quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.

The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

The four datasets of **Anscombe's quartet.**



I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

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>

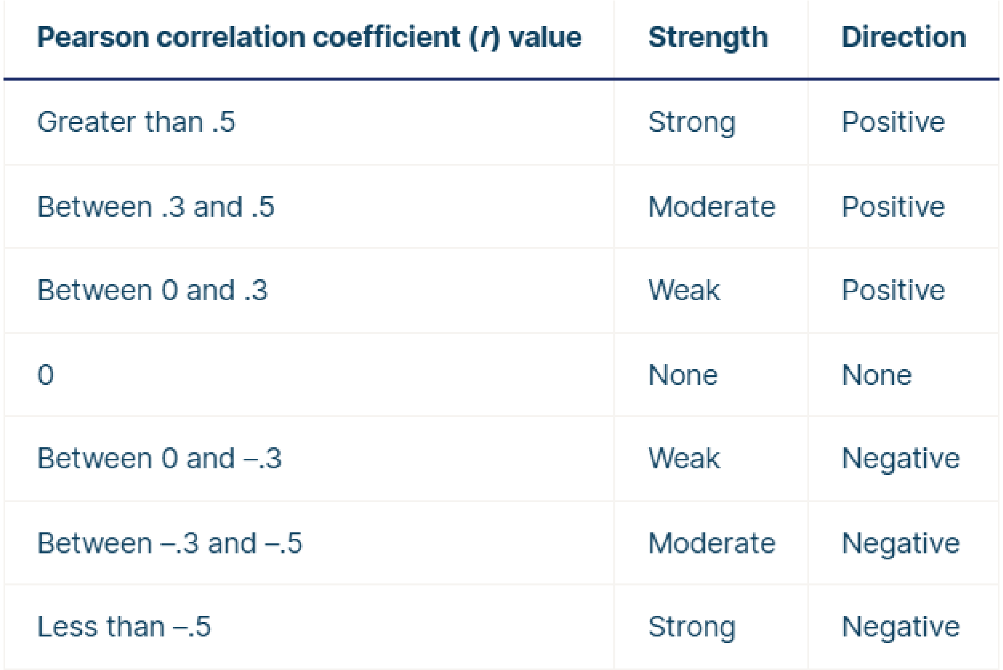
The **Pearson correlation coefficient (*r*)** is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

The Pearson correlation coefficient (*r*) is the most widely used correlation coefficient and is known by many names:

- Pearson's *r*
- Bivariate correlation
- Pearson product-moment correlation coefficient (PPMCC)
- The correlation coefficient

The Pearson 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.

Although interpretations of the relationship strength (also known as effect size) vary between disciplines, the table below gives general rules of thumb:



Pearson correlation coefficient (*r*) value	Strength	Direction
Greater than .5	Strong	Positive
Between .3 and .5	Moderate	Positive
Between 0 and .3	Weak	Positive
0	None	None
Between 0 and –.3	Weak	Negative
Between –.3 and –.5	Moderate	Negative
Less than –.5	Strong	Negative

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 a broader term that encompasses both normalization and standardization. Here are the key points:

1. Scaling: Adjusts the spread or variability of your data.
2. Normalization: Preserves the shape of the distribution but changes the scale (usually to a range of 0-1).
3. Standardized Scaling: Transforms data to have a mean of 0 and a standard deviation of 1 by subtracting the mean and dividing by the standard deviation

Normalization and Scaling are two fundamental preprocessing techniques when you perform data analysis and machine learning. They are useful when you want to rescale, standardize or normalize the features (values) through distribution and scaling of existing data that make your machine learning models have better performance and accuracy.

Normalization:

Min-max scaling, also known as rescaling, is a popular normalization technique that rescales the data to a common range, usually between 0 and 1. This is achieved by subtracting the minimum value and then dividing by the range of the data.
x’ = x-min(x)/ max(x)-min(x)

Standardization:

Standardization Scales features to have a mean of 0 and a standard deviation of 1.
x’ = x-xbar/sigma

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>

You can get inf values for VIF due to the perfect multicollinearity. This happens when two or more independent variables in a model are perfectly linearly dependent. That is, one independent variable in the model can be entirely predicted by another independent variable.

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>

The quantile-quantile(q-q plot) plot is a graphical method for determining if a dataset follows a certain probability distribution or whether two samples of data came from the same population or not. Q-Q plots are particularly useful for assessing whether a dataset is normally distributed or if it follows some other known distribution. They are commonly used in statistics, data analysis, and quality control to check assumptions and identify departures from expected distributions.