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?**
   - given data reveals a strong seasonal influence on demand, with peaks in Fall and summer.
   - A positive trend is seen across years, with 2019 boasting significantly higher demand compared to 2018.
   - Weather conditions also play a role in shaping demand. Clear weather is associated with high demand, while snowy days see a noticeable dip. This aligns with expectations, as certain weather conditions can influence consumer behavior and product use.

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

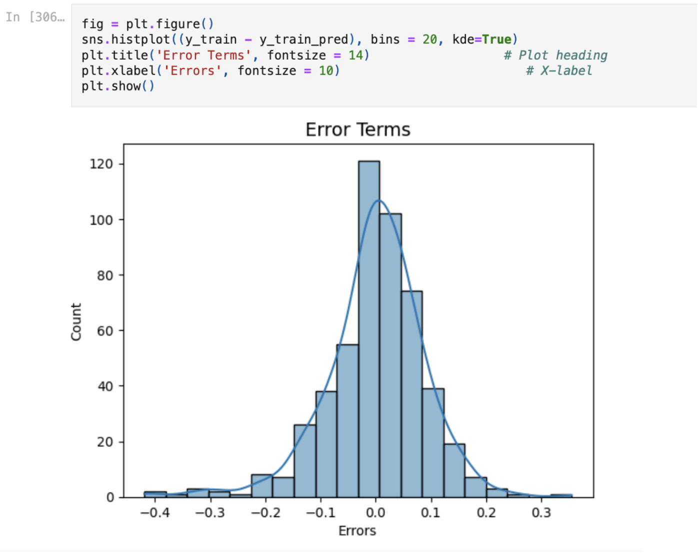
   → Using drop_first=True in dummy variable creation prevents multicollinearity, a situation where dummy variables are highly correlated, leading to inflated standard errors and unreliable coefficient estimates in multilinear regression.

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

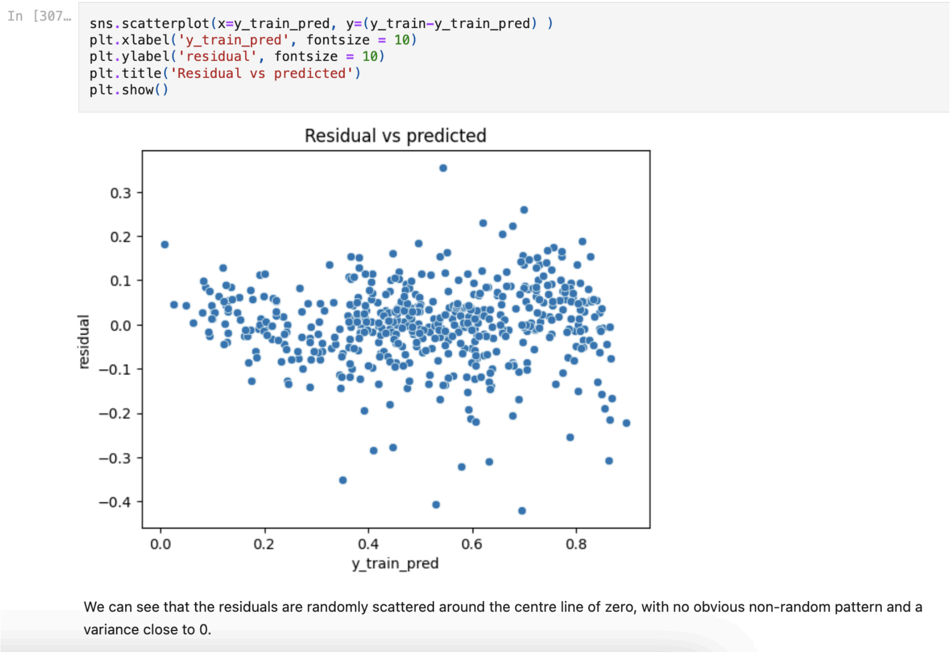
   → While "registered" users are a big factor in total rentals (cnt), we can't include both in the model. Since cnt is already counting registered users, temperature becomes the next most important influence to explore.

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

   → After building the model, we validate linear regression assumptions through residual analysis. Residuals, calculated by subtracting predicted values from actual values, are examined for normality. A normal distribution of residuals, visualized through a distribution plot, and a random scatter pattern support the validity of the linear model. Additionally, the sum and mean of the residuals should ideally be close to zero.



```
In [306…
fig = plt.figure()
sns.histplot((y_train - y_train_pred), bins = 20, kde=True)
plt.title('Error Terms', fontsize = 14)                  # Plot heading
plt.xlabel('Errors', fontsize = 10)                        # X-label
plt.show()
```

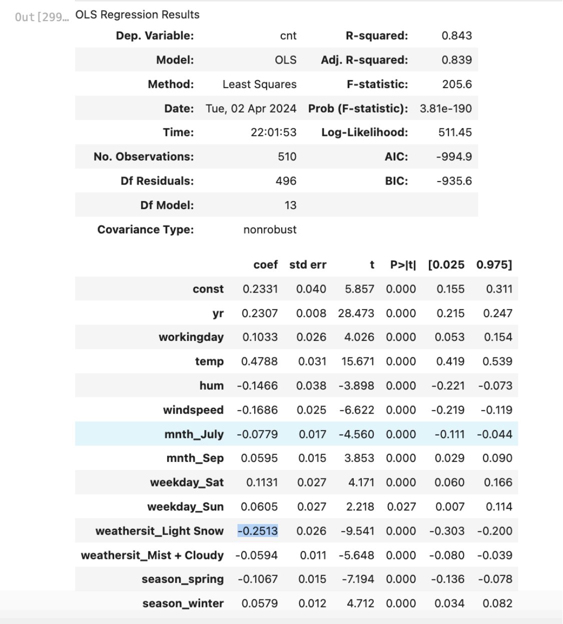


In [307…

```
sns.scatterplot(x=y_train_pred, y=(y_train-y_train_pred) )
plt.xlabel('y_train_pred', fontsize = 10)
plt.ylabel('residual', fontsize = 10)
plt.title('Residual vs predicted')
plt.show()
```

We can see that the residuals are randomly scattered around the centre line of zero, with no obvious non-random pattern and a variance close to 0.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
    a. Temperature :The coefficient for temperature is 0.48, which is the largest in the table. This indicates that temperature has a strong positive relationship with the final value.
    b. Year: The coefficient for yr is 0.23, which is the second largest positive coefficient in the table. This suggests that the year is also a positive contributor to the final value.
    c. Weathersit(Light Snow): The coefficient for windspeed is -0.246, which is the largest negative coefficient in the table. This indicates that Weathersit has a negative relationship with the final value.



Out[299…

OLS Regression Results

Dep. Variable:	cnt	**R-squared:**	0.843
Model:	OLS	**Adj. R-squared:**	0.839
Method:	Least Squares	**F-statistic:**	205.6
Date:	Tue, 02 Apr 2024	**Prob (F-statistic):**	3.81e-190
Time:	22:01:53	**Log-Likelihood:**	511.45
No. Observations:	510	**AIC:**	-994.9
Df Residuals:	496	**BIC:**	-935.6
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2331	0.040	5.857	0.000	0.155	0.311
yr	0.2307	0.008	28.473	0.000	0.215	0.247
workingday	0.1033	0.026	4.026	0.000	0.053	0.154
temp	0.4788	0.031	15.671	0.000	0.419	0.539
hum	-0.1466	0.038	-3.898	0.000	-0.221	-0.073
windspeed	-0.1686	0.025	-6.622	0.000	-0.219	-0.119
mnth_July	-0.0779	0.017	-4.560	0.000	-0.111	-0.044
mnth_Sep	0.0595	0.015	3.853	0.000	0.029	0.090
weekday_Sat	0.1131	0.027	4.171	0.000	0.060	0.166
weekday_Sun	0.0605	0.027	2.218	0.027	0.007	0.114
weathersit_Light Snow	-0.2513	0.026	-9.541	0.000	-0.303	-0.200
weathersit_Mist + Cloudy	-0.0594	0.011	-5.648	0.000	-0.080	-0.039
season_spring	-0.1067	0.015	-7.194	0.000	-0.136	-0.078
season_winter	0.0579	0.012	4.712	0.000	0.034	0.082

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**
    a. Linear regression is a fundamental algorithm in machine learning used for predicting continuous values based on input features. It works by finding the linear relationship between independent variables X and a dependent variable y
    b. We aim to find the best-fitting straight line through the data points. This line is represented by the equation: Y = C + m X
        Y → Dependent variable
        X → Independent variable
        C → Intercept
        m → Slope
    c. We use least squares method, This method minimizes the sum of squared errors between the predicted Y values and the actual Y values from the data.
    d. There are two types of Linear Regression Simple and multiple.
    e. Linear regression provides a equation which can be used to predict Y values for X values
    f. The accuracy of the model is calculated using metrics like R-squared, which indicates how well the regression line fits the data.

2. **Explain the Anscombe's quartet in detail**.
    a. Anscombe's quartet is a example where four datasets share identical statistics but reveal wildly different patterns when visualized. This highlights the crucial role of data visualization in uncovering hidden trends and the limitations of relying solely on numerical summaries.
        i. The datasets have identical statistical summaries
        ii. Despite that, their scatter plots show very different patterns
        iii. This emphasizes the importance of visualization for understanding data beyond just numbers
    b. It signifies the value of exploratory data analysis (EDA), which involves visualizing data before starting with complex analysis
    c. Visualization helps reveal trends, outliers, and other not obvious details hidden in statistics

3. **What is Pearson's R?**
    a. Pearson's R, also called the Pearson correlation coefficient, is a statistical measure that reflects the strength and direction of a linear relationship between two continuous variables. denoted by the symbol "r" and falls between -1 and +1.
    b. +1: perfect positive linear relationship (as one variable increases, the other increases proportionally).
        0: No linear relationship (changes in one variable are unrelated to changes in the other).

-1: perfect negative linear relationship (as one variable increases, the other decreases proportionally).

c. Pearson's R only measures linear relationships. It doesn't capture nonlinear patterns.
d. It doesn't imply causation, just correlation. Just because two variables are related (according to R), it doesn't mean one causes the other.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**
    a. Scaling adjusts features in your data to a similar range. Imagine stretching or shrinking features to fit a common scale.
    b. This improves the model's performance and avoids features with large values dominating the learning process.
    c. Normalization puts features between 0 and 1, like fitting all features to a 0-1 ruler.
    d. Standardization centers the data around a mean of 0 and scales it to have a standard deviation of 1, like centering features on a bell curve.

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

    → An infinite VIF (Variance Inflation Factor) in linear regression occurs when a predictor variable can be perfectly expressed as a linear combination of other predictor variables in the model. This indicates severe multicollinearity, where variables are highly correlated. In essence, the information one variable provides is redundant because it can be completely predicted from the others. This can lead to unreliable coefficient estimates and inflated variances, making interpretations difficult.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**
    a. A quantile-quantile plot, is a graphical tool used to assess if the errors in your linear regression model follow a normal distribution. It compares the quantiles of the actual errors in your model to the quantiles of a theoretical normal distribution.
    b. Q-Q plots are important in linear regression:
        i. Linear regression assumes the residuals are normally distributed. A Q-Q plot visually checks this assumption.
        ii. If the points on the Q-Q plot fall roughly along a straight diagonal line, it indicates the errors are likely normally distributed, suggesting a reliable model.
        iii. Deviations from the straight line suggest non-normality. This may prompt data transformation or the use of alternative regression techniques.
        iv. Addressing non-normality can lead to more accurate predictions and reliable statistical tests for your linear regression model.