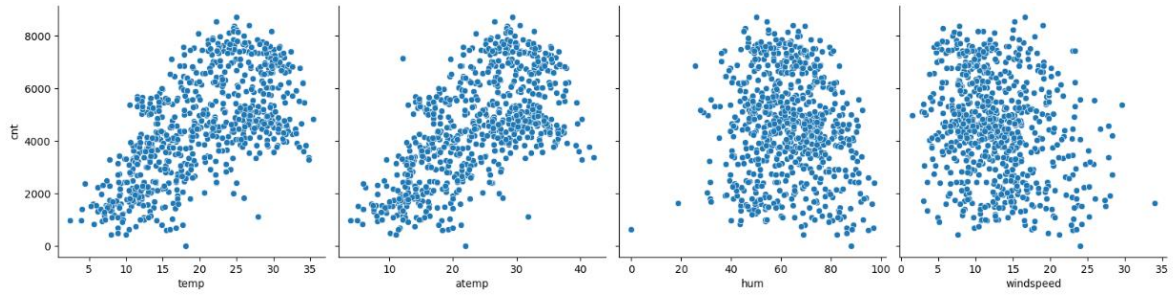
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

   A1. Based on the analysis, we find that categorical variables have a significant effect. Like we derived the weekend variable from variables working day defined as if day is neither weekend nor holiday is 1, otherwise is 0 and holiday : weather day is a holiday or not.

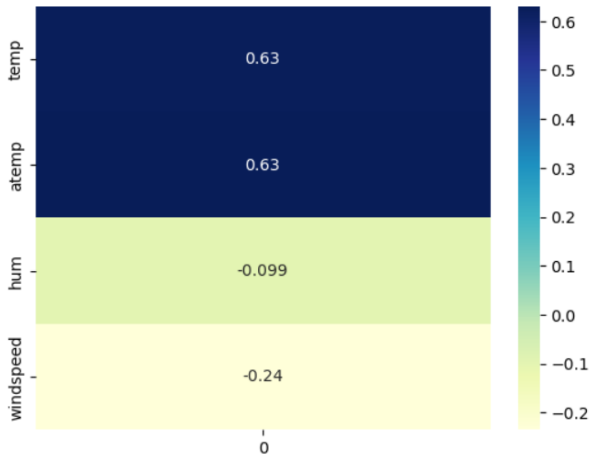
2. Why is it important to use **drop_first=True** during dummy variable creation?

   A2. In Python, the pandas library provides the get_dummies() function, which facilitates the creation of these dummy-coded representations. To avoid falling into the "Dummy Variable Trap," an undesirable situation where the predictors (dummy variables) become correlated, we employ the drop_first=True parameter. By setting this parameter, the first dummy variable is excluded, resulting in n-1 dummy variables out of n discrete categorical levels.
   In the absence of using drop_first=True, the function generates n dummy variables, leading to multicollinearity among the predictors. Multicollinearity occurs when the predictor variables are highly correlated with each other. This correlation can create issues in statistical models, making the interpretation of results more challenging. Therefore, it is recommended to use drop_first=True when applying one-hot encoding to prevent multicollinearity and mitigate the Dummy Variable Trap.

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



   A3. The variables temperature and atemp has a positive correlation while humidity and windspeed have negative correlation.



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

   A4. We can validate the assumptions of Linear Regression after building the model on the training set in the following ways:

1. Residual Analysis: Calculate the residuals (the differences between the predicted and actual values) using the test set. Plot the residuals against the predicted values and check for patterns or trends. Ideally, the residuals should be randomly scattered around zero without any discernible patterns.

2. Normality of Residuals: Check if the residuals follow a normal distribution using a histogram or a Q-Q plot. You can also perform statistical tests like the Shapiro-Wilk test or the Anderson-Darling test to formally assess the normality.

3. Homoscedasticity: Plot the residuals against the predicted values or independent variables and look for a consistent spread of the residuals. The residuals should have constant variance across the range of predicted values or independent variables.

4. Multicollinearity: Calculate the variance inflation factor (VIF) to assess multicollinearity among the independent variables. Higher VIF values indicate stronger multicollinearity.

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

   A5. Based on the analysis, year, temperature and winter contributing significantly towards explaining the demand of the shared bikes.

OLS Regression Results

Dep. Variable:	cnt	R-squared:	0.848
Model:	OLS	Adj. R-squared:	0.841
Method:	Least Squares	F-statistic:	118.1
Date:	Wed, 14 Jun 2023	Prob (F-statistic):	2.36e-182
Time:	23:22:27	Log-Likelihood:	-4106.5
No. Observations:	510	AIC:	8261.
Df Residuals:	486	BIC:	8363.
Df Model:	23		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1716.9349	272.319	6.305	0.000	1181.868	2252.002
yr	2005.5467	70.757	28.344	0.000	1866.520	2144.574
weathersit	-515.0457	95.340	-5.402	0.000	-702.375	-327.716
temp	3704.4326	1243.270	2.980	0.003	1261.584	6147.281
atemp	257.7593	1208.684	0.213	0.831	-2117.133	2632.651
hum	-1451.6548	336.966	-4.308	0.000	-2113.745	-789.564
windspeed	-1608.2392	231.388	-6.950	0.000	-2062.884	-1153.594
summer	756.6810	207.415	3.648	0.000	349.140	1164.222
fall	392.7221	263.732	1.489	0.137	-125.473	910.917
winter	1322.1000	227.860	5.802	0.000	874.387	1769.813
Mist	20.4900	94.758	0.216	0.829	-165.696	206.676
Light Snow	-1126.2353	119.903	-9.393	0.000	-1361.828	-890.642
Feb	231.7360	180.851	1.281	0.201	-123.610	587.082
March	552.0648	196.623	2.808	0.005	165.729	938.401
April	513.7269	294.919	1.742	0.082	-65.746	1093.200
May	728.7143	317.958	2.292	0.022	103.972	1353.456
June	509.6066	343.628	1.483	0.139	-165.574	1184.787
July	190.0712	387.020	0.491	0.624	-570.367	950.510
August	701.6909	369.114	1.901	0.058	-23.566	1426.947
Sep	1275.6075	329.548	3.871	0.000	628.092	1923.123
Oct	593.5682	302.719	1.961	0.050	-1.231	1188.367
Nov	175.7226	288.467	0.609	0.543	-391.073	742.519
Dec	169.8886	232.886	0.729	0.466	-287.700	627.477
weekend	668.2839	230.281	2.902	0.004	215.815	1120.753
holidays	805.7889	225.540	3.573	0.000	362.635	1248.942

Omnibus:	70.667	Durbin-Watson:	2.013
Prob(Omnibus):	0.000	Jarque-Bera (JB):	203.395

Subjective questions:

1. Explain the linear regression algorithm in detail

A1. Linear regression is a popular and widely used statistical algorithm for modeling the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the variables, aiming to find the best-fitting straight line that minimizes the differences between the observed data points and the predicted values.

Here is a detailed explanation of the linear regression algorithm:

1. Data Preparation: Gather the dataset consisting of both the dependent variable (target variable) and independent variables (features). Ensure the data is in a suitable format for analysis and handle any missing values or outliers.
2. Model Representation: In linear regression, the relationship between the dependent variable (y) and independent variables (x) is represented by the equation: y = b0 + b1*x1* + *b2*x2 + ... + bn*xn. Here, b0 is the intercept (the value of y when all x values

are zero), and b1, b2, ..., bn are the coefficients (slopes) representing the impact of each independent variable on the dependent variable.

3. Model Training: The goal is to find the optimal values for the coefficients that minimize the differences between the observed data points and the predicted values. This is achieved by estimating the coefficients using a training algorithm, most commonly the Ordinary Least Squares (OLS) method. OLS calculates the coefficients that minimize the sum of squared differences between the observed and predicted values.
4. Model Evaluation: Once the model is trained, it is important to assess its performance and evaluate its accuracy. Common evaluation metrics for linear regression include the coefficient of determination (R-squared), mean squared error (MSE), and root mean squared error (RMSE). These metrics provide insights into how well the model fits the data and the level of prediction accuracy.
5. Prediction: After the model is evaluated and deemed satisfactory, it can be used for making predictions on new, unseen data. Given the values of the independent variables, the model can generate predicted values for the dependent variable.
6. Assumptions of Linear Regression: Linear regression relies on certain assumptions to be valid. These assumptions include linearity (the relationship between variables is linear), independence of errors (residuals are uncorrelated), homoscedasticity (constant variance of residuals), and normality of errors (residuals follow a normal distribution).
7. Model Interpretation: Linear regression allows for the interpretation of the coefficients. The sign of each coefficient (+ or -) indicates the direction of the relationship with the dependent variable. The magnitude of the coefficient represents the impact or strength of the relationship. Statistical significance tests, such as t-tests or p-values, can be conducted to determine the significance of each coefficient.

Linear regression is a simple yet powerful algorithm that can provide insights into the relationship between variables and make predictions based on that relationship. It is widely used in various fields such as economics, finance, social sciences, and machine learning.

2. Explain the Anscombe’s quartet in detail.

A2. Anscombe's quartet refers to a set of four datasets that have nearly identical statistical properties but exhibit vastly different patterns when plotted. The quartet was introduced by the statistician Francis Anscombe in 1973 to emphasize the importance of visualizing data and not solely relying on summary statistics.

The four datasets in Anscombe's quartet share the following statistical properties:

The means and variances of the x and y variables are the same for all four datasets.

The correlation coefficient between x and y is approximately 0.816 for each dataset.

The linear regression line for each dataset has the same equation: y = 3 + 0.5x.

However, despite these similarities, the datasets reveal distinct patterns when plotted:

Dataset I:

- It is a simple linear relationship with no outliers.
- The linear regression line fits the data well.

Dataset II:

- It also follows a linear relationship, but with an outlier.
- The presence of the outlier affects the regression line, pulling it towards the outlier.

Dataset III:

- It consists of a non-linear relationship.
- The regression line is a poor fit for the data, as a linear model is inappropriate.

Dataset IV:

- It demonstrates a strong non-linear relationship, driven by an outlier.
- Removing the outlier would yield a perfect linear relationship.

The key purpose of Anscombe's quartet is to demonstrate that summary statistics alone, such as means, variances, and correlation coefficients, may not provide a complete understanding of the data. Visualizing the data through plots and graphs is crucial to gain insights into its underlying patterns, trends, and potential outliers. It serves as a reminder to not rely solely on numerical summaries but to explore the data visually for a more comprehensive analysis.

3. What is Pearson’s R?

A3. Pearson's R, also known as the Pearson correlation coefficient or Pearson's correlation, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It is named after Karl Pearson, who developed the coefficient in the late 19th century.

Pearson's correlation coefficient (denoted as r) takes values between -1 and +1, where:

- r = +1 indicates a perfect positive linear relationship,
- r = -1 indicates a perfect negative linear relationship,
- r = 0 indicates no linear relationship or complete absence of correlation.

The calculation of Pearson's correlation coefficient involves the following steps:

1. Standardization: Standardize both variables by subtracting the mean and dividing by the standard deviation. This step ensures that the variables are on a similar scale and helps in comparing their linear relationship.

2. Pairwise Calculation: Multiply the standardized values of the corresponding data points for both variables and sum them up. Then, divide the sum by the number of data points.
3. Interpretation: The resulting value is the Pearson correlation coefficient (r). It quantifies the strength and direction of the linear relationship between the variables. A positive value indicates a positive correlation (as one variable increases, the other tends to increase), while a negative value indicates a negative correlation (as one variable increases, the other tends to decrease).

Pearson's correlation coefficient is widely used in various fields, including statistics, social sciences, finance, and machine learning. It helps in understanding the relationship between variables, assessing the strength of association, and can be used to make predictions or guide decision-making processes. However, it is important to note that Pearson's correlation coefficient only measures the linear relationship between variables and may not capture non-linear relationships or other types of associations.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

   A4. Scaling: Scaling refers to the process of transforming data to a standardized range or distribution. It involves adjusting the values of variables to a specific scale, typically to make them comparable or suitable for certain analysis techniques or algorithms. Scaling does not change the underlying relationship between variables but rather changes their representation in terms of magnitude or spread.

   Purpose of Scaling: Scaling is performed for various reasons, including:

   Comparability: Scaling ensures that variables with different scales or units can be compared and analyzed together. It prevents variables with larger magnitudes from dominating the analysis.
   Algorithm Requirement: Some machine learning algorithms, such as k-nearest neighbors (KNN) or support vector machines (SVM), are sensitive to the scale of variables. Scaling helps to improve the performance and accuracy of these algorithms.
   Convergence: Scaling can help algorithms converge faster during the optimization process, particularly in gradient-based methods.
   Interpretability: Scaling makes the interpretation of coefficients or feature importance more meaningful in linear models or feature selection methods.
   Normalized Scaling vs. Standardized Scaling:

   Normalized Scaling (Min-Max Scaling): Normalization scales the values of variables to a fixed range, usually between 0 and 1. It is achieved by subtracting the minimum value and dividing by the range (maximum - minimum). The formula for normalization is: (x - min) / (max - min). Normalized scaling preserves the relative relationships and distribution of the data.

Standardized Scaling (Z-score Scaling): Standardization transforms variables to have zero mean and unit variance. It is achieved by subtracting the mean and dividing by the standard deviation. The formula for standardization is: (x - mean) / standard deviation. Standardized scaling centers the data around zero and adjusts the spread of the data.

The key differences between normalized scaling and standardized scaling are:

Range: Normalization scales the data to a specific range (e.g., 0 to 1), while standardization centers the data around zero with unit variance.
Preservation of Distribution: Normalization preserves the distribution and range of the data, while standardization standardizes the data irrespective of the original distribution.
Sensitivity to Outliers: Normalization is sensitive to outliers as it is based on the minimum and maximum values, while standardization is more robust to outliers due to the use of the mean and standard deviation.
The choice between normalized scaling and standardized scaling depends on the specific requirements of the analysis or the algorithm being used. Normalization is often preferred when the distribution of the data is non-Gaussian or when the range of the data needs to be fixed. Standardization is commonly used when the distribution is approximately Gaussian or when variables are required to have zero mean and unit variance.

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

A5. The occurrence of an infinite value of the Variance Inflation Factor (VIF) in the context of multicollinearity usually happens due to perfect multicollinearity. Perfect multicollinearity occurs when one or more independent variables can be expressed as a perfect linear combination of other independent variables in the model.
When perfect multicollinearity exists, the mathematical calculations involved in calculating VIF lead to an infinite value. Specifically, the VIF is calculated as the ratio of the variance of the estimated coefficient for a particular independent variable to the variance of the estimated coefficient when that variable is not included in the model but all other independent variables are present. Mathematically, the formula for VIF is $VIF = 1 / (1 - R^2)$, where R^2 represents the coefficient of determination for the regression model.
In the case of perfect multicollinearity, the determinant of the covariance matrix of the independent variables becomes zero. This causes the VIF calculation to involve division by zero, resulting in an infinite VIF value.
Perfect multicollinearity typically occurs in situations where two or more variables are perfectly correlated or when a variable is a linear combination of other variables. It can arise due to data errors, the inclusion of redundant variables, or the inappropriate inclusion of derived variables.
It is important to address perfect multicollinearity as it can cause issues in regression analysis. To handle this problem, one of the correlated variables should be removed

from the model to eliminate the perfect linear relationship. This can be done by carefully examining the variables and their relationships, considering domain knowledge, or using feature selection techniques to identify and remove redundant variables.

If perfect multicollinearity is not present but the VIF values are very high (typically above 10 or 5, depending on the context), it indicates high multicollinearity among the variables, which can still lead to unstable coefficient estimates and inflated standard errors. In such cases, it is recommended to assess the multicollinearity and consider techniques such as feature selection, dimensionality reduction, or regularization methods to mitigate the issue.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

   A6. A Q-Q plot (Quantile-Quantile plot), also known as a quantile plot, is a graphical tool used to assess the distributional similarity between a sample of data and a theoretical distribution. It helps to determine whether the data follows a specific distribution or if it deviates from the expected distribution. Q-Q plots are commonly used in linear regression to evaluate the assumption of normality of residuals.

   Here's an explanation of the use and importance of a Q-Q plot in linear regression:

1. Normality Assumption: Linear regression assumes that the residuals (the differences between observed and predicted values) follow a normal distribution. Normality of residuals is important for accurate parameter estimation, hypothesis testing, and producing reliable predictions.
2. Construction of Q-Q Plot: A Q-Q plot compares the quantiles of the sample data against the quantiles of a theoretical distribution, typically the normal distribution. It plots the observed quantiles on the y-axis and the expected quantiles (from the theoretical distribution) on the x-axis. The data points are then plotted to assess if they fall approximately on a straight line.
3. Interpretation of Q-Q Plot: If the data points on the Q-Q plot closely follow a straight line, it suggests that the sample data is approximately normally distributed. Deviations from the straight line indicate departures from normality.
4. Importance in Linear Regression: a. Assumption Checking: Q-Q plots provide a visual assessment of the normality assumption of the residuals in linear regression. If the Q-Q plot exhibits a reasonably straight line, it supports the assumption of normality. b. Outlier Detection: Q-Q plots can help identify outliers or extreme values that deviate significantly from the expected distribution. These outliers may indicate data issues or violations of assumptions. c. Model Evaluation: Deviations from the expected line in a Q-Q plot may suggest potential issues in the model, such as non-linear relationships, heteroscedasticity, or influential observations. It prompts further investigation and potential model refinement. d. Transformation Selection: If the Q-Q plot reveals a non-normal distribution, it may suggest the need for data transformation (e.g., logarithmic or power transformation) to improve the normality of residuals and meet the assumption of linear regression.

   By utilizing a Q-Q plot, analysts can visually assess the normality assumption, detect outliers, evaluate model performance, and guide decision-making in linear regression

analysis. It is a valuable diagnostic tool for checking the validity of assumptions and ensuring the reliability of regression results.