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?

Answer:

	coef
const	0.1580
yr	0.2341
temp	0.4828
windspeed	-0.1620
season_summer	0.0956
season_winter	0.1119
mnth_Aug	0.0568
mnth_Jan	-0.0461
mnth_Oct	0.0394
mnth_Sep	0.1162
weathersit_Cloudy	-0.0809
weathersit_Rainy	-0.2867

Based on the coefficients in the table, here's that can be inferred about the effects on the dependent variable:

a. <u>Seasonal Effects:</u>

- Summer (season_summer): Has a positive effect (+0.0956) on the dependent variable
- Winter (season_winter): Also has a positive effect (+0.1119), with a slightly larger impact than summer

b. <u>Monthly Effects:</u>

- August (mnth_Aug): Slight positive effect (+0.0568)
- January (mnth_Jan): Slight negative effect (−0.0461)
- October (mnth_Oct): Small positive effect (+0.0394)
- September (mnth_Sep): Positive effect (+0.1162), indicating higher impact compared to other months

c. <u>Weather Conditions:</u>

- Cloudy (weathersit_Cloudy): Negative effect (−0.0809)
- Rainy (weathersit_Rainy): Stronger negative effect (−0.2867)

d. **<u>Windspeed:</u>** Has a negative effect (−0.1620) on the dependent variable, indicating that as windspeed increases, the dependent variable decreases. The impact is moderate, meaning higher windspeed tends to reduce the value of the dependent variable.

In summary, winter, summer, and September positively influence the dependent variable, while January, cloudy, and especially rainy weather, along with increased windspeed, have a negative impact.

---

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

Answer:

Using drop_first=True when creating dummy variables is important to avoid a problem called multicollinearity.

Multicollinearity happens when one variable can be perfectly predicted from others, which can confuse the model and make it hard to understand the impact of each variable.

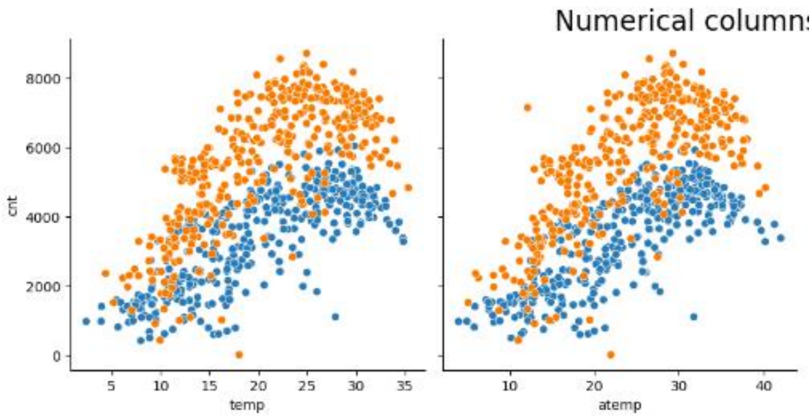
By setting drop_first=True, you drop one category from each set of dummy variables. This ensures that the model doesn't include redundant information, making it easier to interpret the results and keeping the analysis mathematically correct.

---

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

Answer:

Among all numerical variables temp and atemp has highest correlation with cnt target variable and they are also highly correlated with each other thus both can be assumed too similar.



---

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

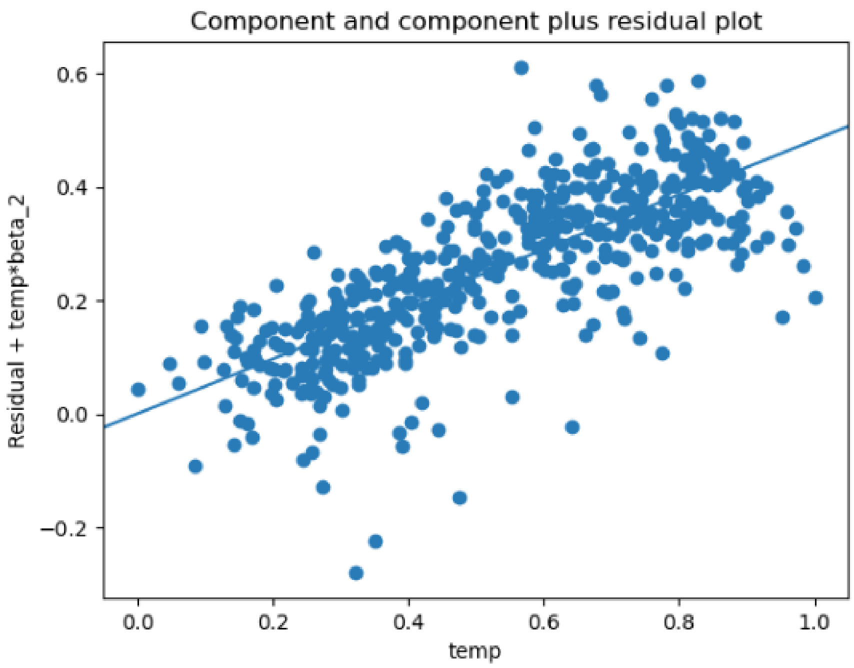
Answer:

The assumptions of linear regression are as follows:

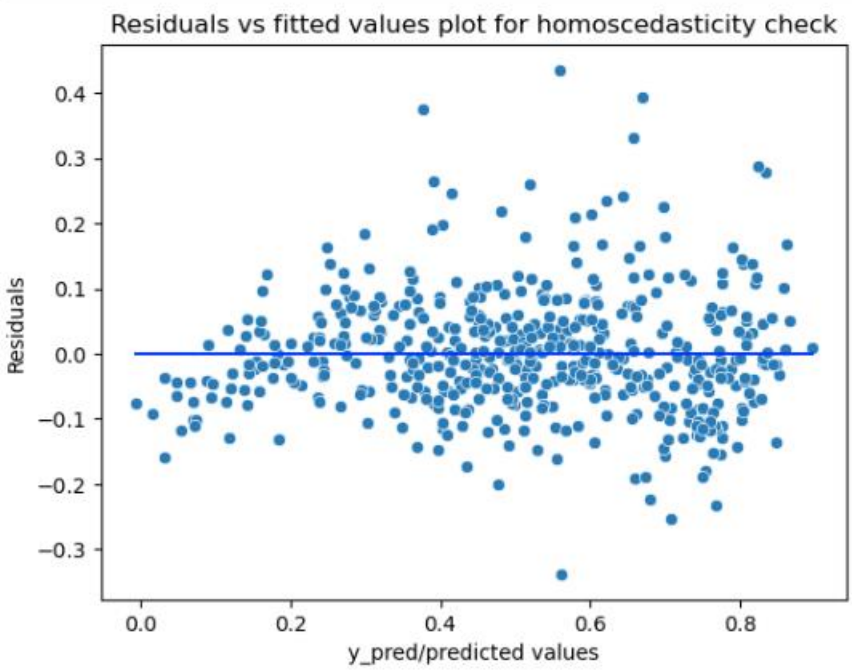
- Linear Relationship
- Homoscedasticity
- Absence of Multicollinearity
- Independence of residuals (absence of auto-correlation)
- Residuals are normally distributed

The evidence are as follows:

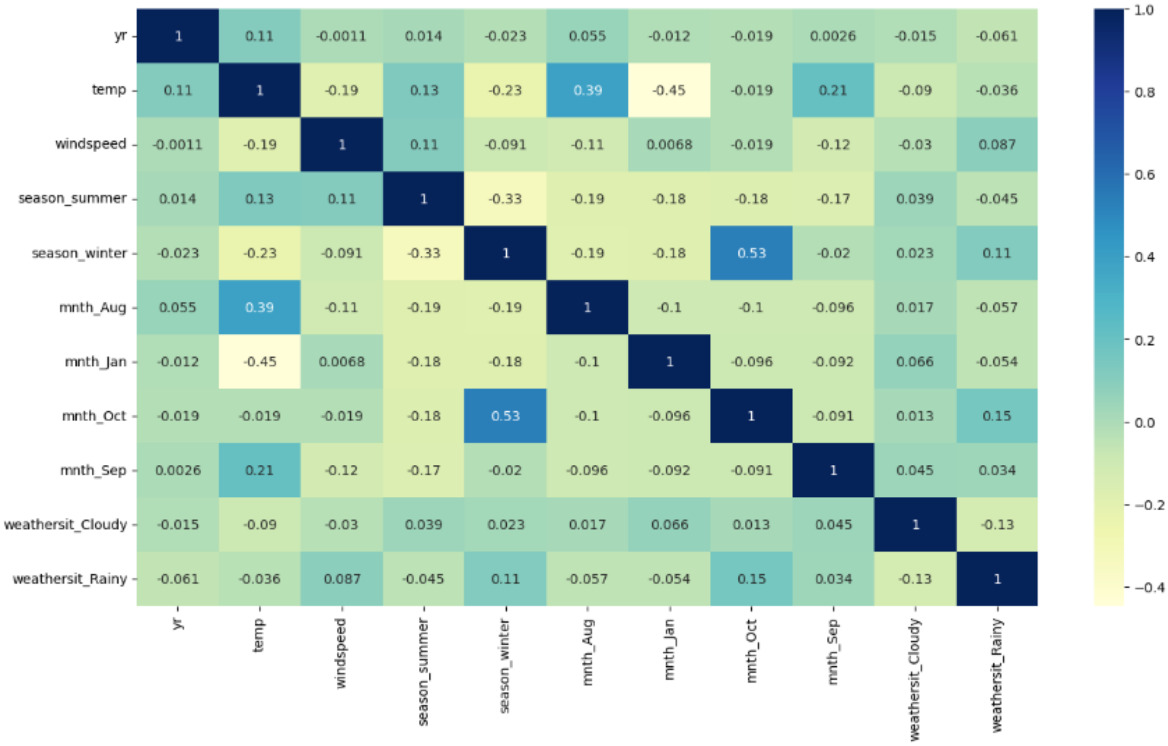
- The linear relationship was assessed using a partial residual plot (CCPR) from the statsmodels library. The CCPR plot allows us to evaluate the impact of a single regressor on the response variable while accounting for the influence of other independent variables. In this case, we plotted the target variable against 'temp' to demonstrate their linear relationship, considering all other variables.



- Homoscedasticity was tested by plotting residual vs predicted values and it shows no pattern in scatterplot thus verifying Homoscedasticity

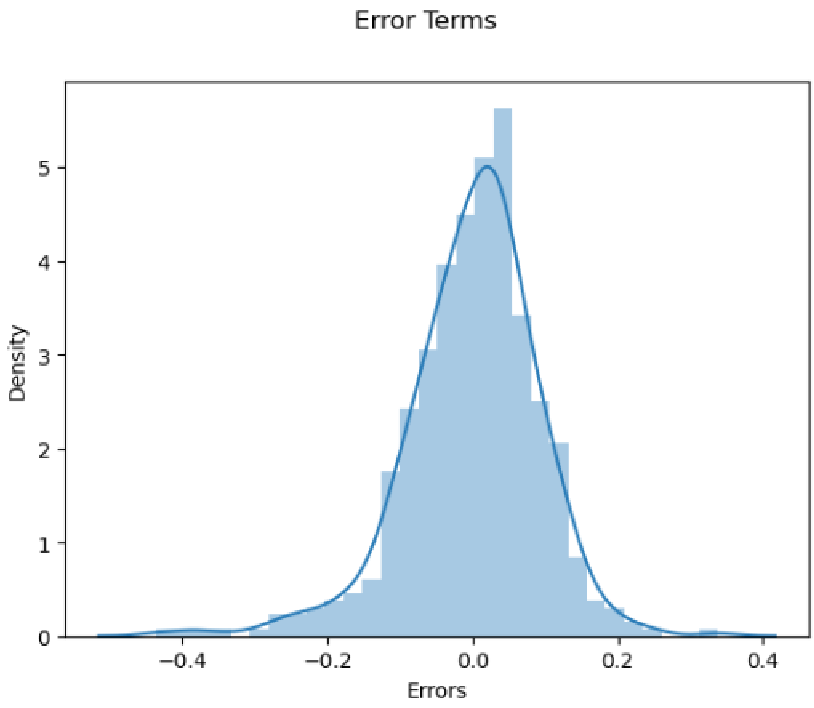


➢ Multicollinearity was checked via heatmap and VIF where no column had high correlation or VIF.



➢ Independence of residual was verified by Durbin-Watson statistic where value if final model is 1.9896 which is close to 2 which indicates non-autocorrelation.

➢ The distribution of residual was checked using histogram which is normally distributed.



---

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

Answer:

	coef
const	0.1580
yr	0.2341
temp	0.4828
windspeed	-0.1620
season_summer	0.0956
season_winter	0.1119
mnth_Aug	0.0568
mnth_Jan	-0.0461
mnth_Oct	0.0394
mnth_Sep	0.1162
weathersit_Cloudy	-0.0809
weathersit_Rainy	-0.2867

The top 3 features are:

- **Temperature (temp):** With a coefficient of +0.4828, temperature has the most substantial positive impact on bike demand.
- **Year (yr):** The year variable has a positive coefficient of +0.2341, indicating a notable increase in demand over time.
- **Weather Situation - Rainy (weathersit_Rainy):** This has a significant negative impact, with a coefficient of −0.2867, suggesting that rainy weather strongly reduces bike demand.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer:

Linear regression is a simple yet powerful algorithm used to model the relationship between a dependent variable and one or more independent variables. The goal is to predict the dependent variable based on the independent variables.

- **Equation**: The model is represented as $y=\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n+\epsilon$, where β_0 is the intercept, β_1, β_2, … are the coefficients, and ϵ is the error term.
- **Assumptions**: Linearity, independence of observations, homoscedasticity (constant variance of errors), normality of residuals, and no multicollinearity among predictors.
- **Coefficient Estimation**: The coefficients are estimated using Ordinary Least Squares (OLS), which minimizes the sum of squared errors between observed and predicted values.
- **Interpretation**: Coefficients represent the change in the dependent variable for a one-unit change in an independent variable.
- **Model Evaluation**: R-squared measures how well the model explains the variability in the data. P-values indicate the significance of each predictor.
- **Limitations**: Sensitive to outliers, assumes linear relationships, and can overfit with too many predictors.

In short, linear regression is widely used for predictive modelling, trend analysis, and understanding relationships between variables due to its simplicity and interpretability.

---

2. Explain the Anscombe's quartet in detail.

Answer:

Anscombe's quartet is a collection of four datasets that have nearly identical simple statistical properties—such as mean, variance, correlation, and linear regression line—but differ significantly when graphed. The quartet was created by statistician Francis Anscombe in 1973 to illustrate the importance of visualizing data before analysing it.

- **Identical Statistics**: All four datasets share similar summary statistics (e.g., mean, variance, correlation, and linear regression line).
- **Different Graphs**: When plotted, each dataset reveals very different patterns:
    1. The first dataset shows a typical linear relationship.
    2. The second dataset shows a non-linear relationship.
    3. The third dataset has a clear outlier that affects the regression line.
    4. The fourth dataset shows a vertical line, where most data points have the same x-value.
- **Lesson**: Anscombe's quartet demonstrates that relying solely on numerical summaries can be misleading, and emphasizes the importance of data visualization for a complete understanding of data.

In short, Anscombe's quartet highlights the critical role of graphical analysis in statistical data interpretation.

---

3. What is Pearson's R?

Answer:

Pearson's R, or Pearson's correlation coefficient, is a statistical measure that evaluates the strength and direction of a linear relationship between two variables. It ranges from -1 to 1:

- 1 indicates a perfect positive linear relationship,
- -1 indicates a perfect negative linear relationship,
- 0 indicates no linear relationship.

A positive R value means that as one variable increases, the other tends to increase as well. A negative R value means that as one variable increases, the other tends to decrease. Pearson's R is commonly used in various fields, including psychology, social sciences, and data analysis, to assess correlations between variables.

---

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

Answer:

Scaling in data analysis refers to the process of transforming data to fit within a specific range or to have particular statistical properties. It's often performed to prepare data for various algorithms or models that assume or benefit from data being on a common scale.

Scaling can help in:

1. Improving Model Performance: Many machine learning algorithms, like gradient descent-based methods or distance-based models (e.g., k-nearest neighbours), perform better when features are on a similar scale.
2. Ensuring Consistency: In datasets where features have different units or ranges, scaling ensures that no feature disproportionately affects the outcome due to its scale.
3. Enhancing Convergence Speed: Algorithms can converge faster if the data is scaled appropriately, especially when using methods that involve optimization.

Types of Scaling:

1. Normalized Scaling (Min-Max Scaling):
    - Purpose: Transforms data to fit within a specific range, usually [0, 1].
    - Method: For a feature x, the normalized value x' is computed as

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

    - Use Case: Useful when you need to bound the values within a specific range, often required for neural networks and algorithms that use distance metrics.
2. Standardized Scaling (Z-score Normalization):
    - Purpose: Centers the data around the mean with a standard deviation of 1.
    - Method: For a feature xxx, the standardized value x' is computed as

$$x' = \frac{x - \mu}{\sigma}$$

where μ is the mean of x and σ is the standard deviation of x.

    - Use Case: Useful when you need features with zero mean and unit variance, commonly used in algorithms assuming normally distributed data or when comparing features with different units.

In summary, normalization rescales data to a fixed range, while standardization shifts and scales data to have a mean of zero and a standard deviation of one. The choice between them depends on the specific requirements of your analysis or machine learning model

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

Answer:

The Variance Inflation Factor (VIF) measures how much the variance of a regression coefficient is inflated due to multicollinearity with other predictors. It is calculated for each predictor in a regression model, and a high VIF indicates that the predictor is highly correlated with other predictors.

A VIF value can become infinite in cases where there is perfect multicollinearity. This occurs when:

1. Perfect Multicollinearity: One predictor variable is a perfect linear combination of other predictor variables. In this scenario, the matrix used to estimate regression coefficients becomes singular, meaning it cannot be inverted. This results in the VIF calculation producing an infinite value.
2. Deterministic Relationships: If one predictor is entirely determined by a linear combination of other predictors, it creates a situation where the predictor's variance is infinitely inflated.

In practical terms, when you encounter an infinite VIF, it indicates that there is a redundancy in your predictors, and you may need to address multicollinearity by removing or combining predictors to resolve the issue.

---

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

Answer:

A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to assess if a dataset follows a specific theoretical distribution, typically the normal distribution. It plots the quantiles of the data against the quantiles of the theoretical distribution. Here's a brief explanation:

How It Works:

- Quantiles: The Q-Q plot compares the quantiles of the sample data against the quantiles of a specified theoretical distribution. For a normal Q-Q plot, it compares the quantiles of the data with the quantiles of the standard normal distribution.
- Plot: Points are plotted on a scatter plot where the x-axis represents the theoretical quantiles and the y-axis represents the sample quantiles. If the data follows the theoretical distribution, the points will approximately lie on a straight line (often a 45-degree line).

Use and Importance in Linear Regression:

- Assess Normality of Residuals: In linear regression, one of the key assumptions is that the residuals (the differences between observed and predicted values) are normally distributed. A Q-Q plot helps to visually assess this assumption by showing how well the residuals match the normal distribution.
- Model Validation: If the residuals deviate significantly from the straight line, it suggests that the normality assumption may be violated, which could affect the validity of hypothesis tests and confidence intervals derived from the model.
- Detect Outliers: Deviations from the straight line in the Q-Q plot can also indicate the presence of outliers or data points that do not conform to the normal distribution.

In summary, a Q-Q plot is a valuable diagnostic tool for validating the normality of residuals in linear regression, helping to ensure that model assumptions are met and that the results are reliable.