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?**
   The analysis of categorical variables in the dataset uncovers valuable insights about their impact on the dependent variable, likely associated with bike rental rates:

   Seasonal Influence: Bike rental rates are notably higher during the summer and fall seasons, especially when the weather is clear and pleasant. This suggests that demand for bike rentals is strongly influenced by seasonal variations, with more people renting bikes when the weather is favorable for outdoor activities.

   Monthly Patterns: The peak in rental rates occurs in the months of June, July, August, September, and October, underscoring the seasonal nature of bike rentals, with the highest demand during the warmer months.

   Day of the Week: Saturdays and Wednesdays stand out as the days with the highest demand for bike rentals. This indicates specific days of the week when people are more inclined to rent bikes, possibly linked to leisure or recreational activities.

   Yearly Growth: The dataset also reflects a rise in the number of bike rentals from 2018 to 2019. This growth could be attributed to factors like increased awareness of bike-sharing services or enhancements in infrastructure and promotional efforts.

   Holiday Impact: Bike rental rates are elevated on holidays, particularly for casual users. This implies that holidays, when people have more free time, drive increased demand for bike rentals. It's possible that tourists or casual riders are more likely to use the service on holidays for leisure activities.

   In conclusion, the analysis of categorical variables in the dataset highlights the significant influence of seasonality, month, day of the week, year, and holidays on bike rental rates. This understanding can inform decisions about resource allocation, marketing strategies, and service enhancements for bike rental businesses.

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

   Using `drop_first=True` during dummy variable creation is important to avoid multicollinearity and improve the interpretability of regression models, particularly in situations where categorical variables are represented using one-hot encoding (dummy variables).

   Here's why it's important:

1. **Multicollinearity**: When you create dummy variables for a categorical variable without dropping one category, you introduce multicollinearity. Multicollinearity occurs when two or more independent variables in a regression model are highly correlated, making it difficult to distinguish their individual effects on the dependent variable. By dropping one category (using `drop_first=True`), you leave out one reference category, which serves as the baseline for comparison, and the multicollinearity issue is mitigated.

2. **Interpretability**: Including a dummy variable for every category of a categorical variable can make the model less interpretable. When you set `drop_first=True`, one category is implicitly treated as the reference category, making it easier to interpret the coefficients of the remaining dummy variables. The coefficients represent the difference in the outcome variable between the reference category and each of the other categories.

3. **Redundancy**: Including all dummy variables for a categorical variable can lead to redundancy in the model, as the information about one category can be inferred from the values of the other categories. By dropping one category, you maintain the necessary information without redundancy.

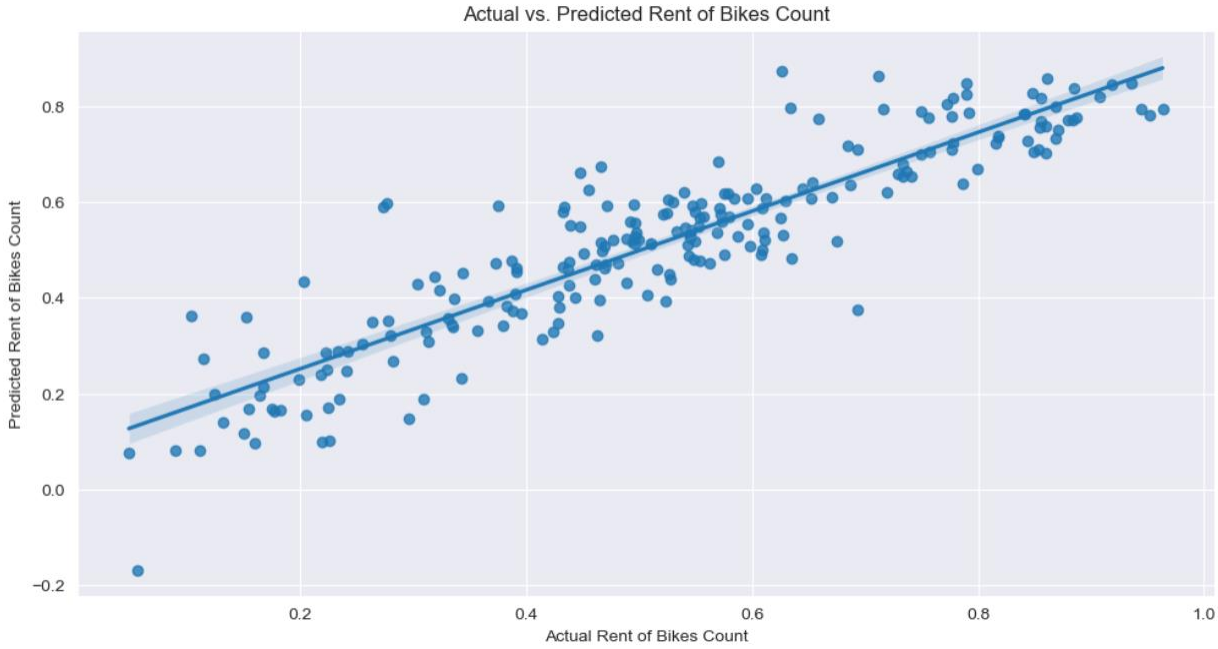
In summary, using `drop_first=True` in dummy variable creation helps improve the performance and interpretability of regression models, and it avoids issues like multicollinearity and redundancy that can arise when including all dummy variables for a categorical variable.

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

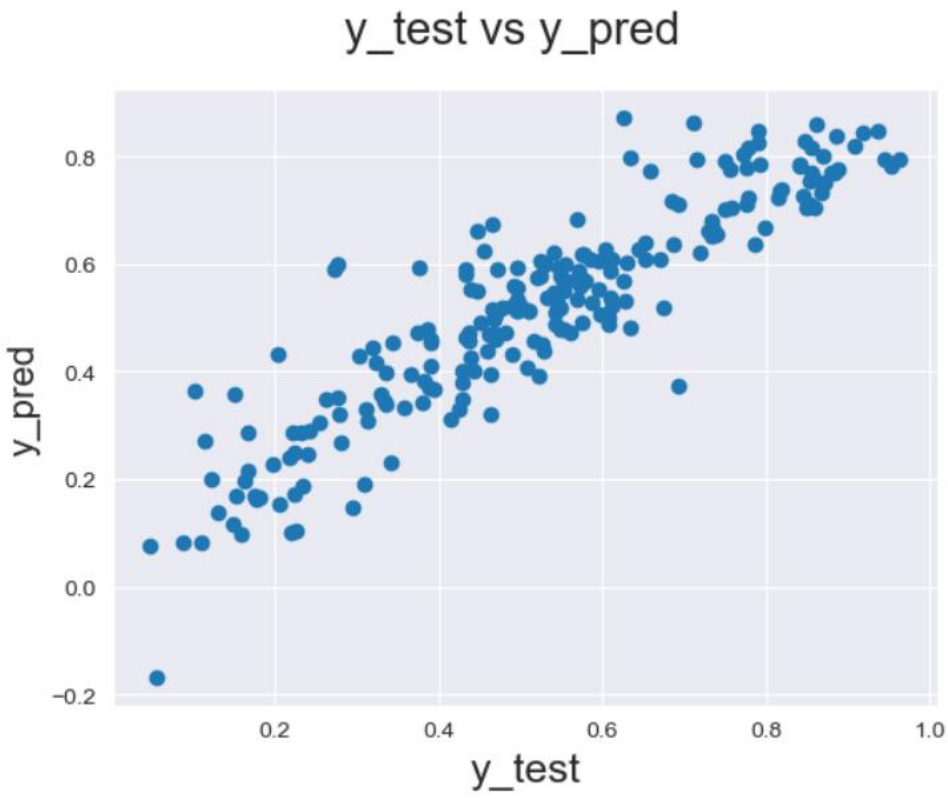
- The temp variable has the highest correlation with the target variable.
- Causal and Registered variables have a high correlation with target variable but since the summation of them make the target variable it is being ignored

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

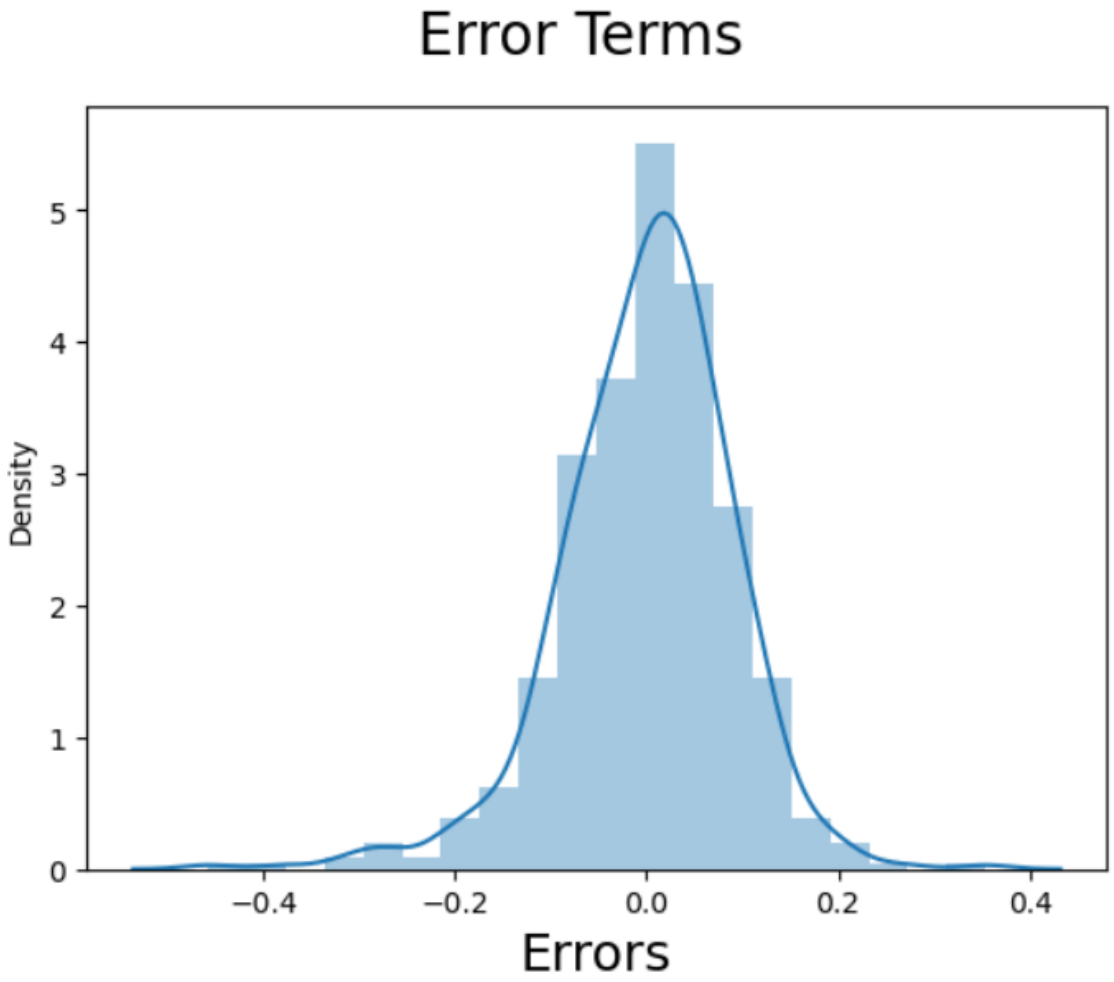
- Linear relationship between independent and dependent variables



- Error distribution of residuals



Distribution of Error Terms



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

   The top 3 variables are:

   • Temp: Higher temperatures and clear weather appear to encourage more biking activity.

   • yr: The business appears to experience organic year-on-year growth.

   • season: Summer and fall exhibit higher demand.

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Linear regression is a fundamental supervised machine learning algorithm used for modeling the relationship between a dependent variable (also known as the target or outcome) and one or more independent variables (predictors or features) by fitting a linear equation to the observed data. The goal of linear regression is to find the best-fitting linear relationship that can be used for prediction and understanding the dependencies between variables. Linear regression explained in detail:

1. **Simple Linear Regression**:
   Simple linear regression is used when there is only one independent variable that affects the dependent variable. The linear equation for simple linear regression is:
   $Y = \beta 0 + \beta 1 X + \epsilon$
2. **Multiple Linear Regression**:
   Multiple linear regression extends the concept to situations where there are multiple independent variables. The linear equation for multiple linear regression is:
   $Y = \beta 0 + \beta 1 X1 + \beta 2 X2 + \ldots + \beta p Xp + \epsilon$
3. **Fitting the Model**:
   The goal in linear regression is to estimate the values of $\beta 0, \beta 1, \beta 2, \ldots, \beta p$ that minimize the sum of squared errors between the predicted values (Y) and the actual values of the dependent variable in the training data. This is typically done using the method of least squares, which minimizes the sum of the squared residuals (the differences between the predicted and actual values).
4. **Assumptions**:
   - The relationship between the independent variables and the dependent variable is linear.
   - The error terms (ϵ) are independent of each other.
   - The variance of the error terms is constant across all levels of the independent variables.
   - The error terms follow a normal distribution.
   - Independent variables are not highly correlated with each other.
5. **Interpretation**:
   Linear regression allows for the interpretation of the coefficients $\beta 0, \beta 1, \beta 2, \ldots, \beta p$. For example, $\beta 1$ represents the change in the dependent variable Y for a one-unit change in $X1$, holding all other variables constant.
6. **Evaluation**:
   To assess the performance of a linear regression model, various metrics can be used, such as Mean Squared Error (MSE), R-squared ($R2$), and others, to measure how well the model fits the data and makes predictions.
7. **Predictions**:
   Once the model is trained, it can be used to make predictions on new or unseen data by plugging in the values of the independent variables into the linear equation.
8. **Extensions**:

Linear regression has several extensions and variations, including ridge regression, lasso regression, and polynomial regression, which address issues like multicollinearity and allow for more flexible modeling.

Linear regression is a powerful and interpretable tool commonly used for tasks such as predicting house prices, analyzing the impact of variables on an outcome, and understanding relationships in data. However, it has its limitations, and its effectiveness depends on the assumptions being met and the nature of the data being modeled.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is a famous example in statistics and data visualization created by Francis Anscombe in 1973 created by Francis Anscombe in 1973 to demonstrate that summary statistics alone may not provide a complete understanding of the data. It consists of four datasets, each containing 11 data points, with two numerical variables (X and Y). What makes Anscombe's quartet remarkable is that these four datasets have nearly identical simple descriptive statistics (such as means, variances, and correlation coefficients) for X and Y, yet they have vastly different graphical representations and relationships between the variables.

It emphasis on the importance of data visualization and graphical exploration thus revealing hidden patterns, outliers, and relationships that might not be apparent from summary statistics alone. This quartet is often used in statistics and data science education to emphasize the importance of visualizing data before drawing conclusions or making decisions based on statistical summaries.

3. What is Pearson's R?

The Pearson correlation coefficient (PCC)[a] is a correlation coefficient that measures linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1.

An absolute value of exactly 1 implies that a linear equation describes the relationship between X and Y perfectly, with all data points lying on a line. The correlation sign is determined by the regression slope: a value of +1 implies that all data points lie on a line for which Y increases as X increases, and vice versa for −1. A value of 0 implies that there is no linear dependency between the variables.

the formula for ρ can also be written as

$$\rho_{X,Y} = \frac{\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]}{\sqrt{\mathbb{E}[X^2] - (\mathbb{E}[X])^2}\,\sqrt{\mathbb{E}[Y^2] - (\mathbb{E}[Y])^2}}.$$

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

   Scaling is a technique performed in pre-processing during building a machine learning model to standardize the independent feature variables in the dataset in a fixed range. The dataset could have several features which are highly ranging between high magnitudes and units. If there is no scaling performed on this data, it leads to incorrect modelling as there will be some mismatch in the units of all the features involved in the model. The difference between normalization and standardization is that while normalization brings all the data points in a range between 0 and 1, standardization replaces the values with their Z scores.

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

   Reasons why the VIF might become infinite:

   - Perfect multicollinearity: Perfect multicollinearity occurs when two or more independent variables in a regression model are perfectly correlated with each other. In this case, one variable can be exactly predicted by a linear combination of the others, leading to an R^2 value of 1 and an infinite VIF.

   - Nearly perfect multicollinearity: Even if multicollinearity is not perfect but very high, the R^2 value can still approach 1, resulting in a very large VIF.

   - Too small a sample size: When you have a small sample size relative to the number of independent variables in the model, it can lead to unstable estimates and high VIF values, including infinity.

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

   The quantile-quantile plot is a graphical method for determining whether two samples of data came from the same population or not. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

   - **Use in Linear Regression:**
     A. In linear regression analysis, assumptions about the error terms (residuals) are critical. One of these assumptions is that the residuals should follow a normal distribution.
     B. A Q-Q plot can be used to assess the normality of residuals. You create a Q-Q plot of the residuals, comparing them to the expected quantiles of a normal distribution.
     C. If the points in the Q-Q plot roughly align along the identity line, it suggests that the residuals are approximately normally distributed, which is an important assumption for linear regression models.
   - **Importance:**

A. A well-behaved Q-Q plot is an essential diagnostic tool for linear regression. It helps you check the assumption of normality of residuals, which, if violated, can affect the validity of statistical tests, confidence intervals, and predictions made by the regression model.
B. If the Q-Q plot shows deviations from the identity line, it may indicate that the residuals have a non-normal distribution. This information can prompt further investigation, such as looking for outliers, transforming variables, or considering alternative regression models that relax the normality assumption.
C. Addressing deviations from normality in the residuals can lead to more accurate and reliable regression results.