Q. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
Ans: Based on the analysis on the categorical variables, below are the effects on the dependent variabke

1. The demand for rental bikes is high in summer and fall seasons and less in spring season

2. The demand for rental bikes is high in the year 2019 increased significantly from the year 2018.

3. The demand for rental bikes is high in the months of June, July and August.

4. Bike demand is less in holidays compared to working days

5. The demand for rental bikes is high when the weather condition is clear, however demand is less in case of lightsnow and light rainfall

6. The demand for rental bikes remains more or less similar throughout the weekdays

Q. Why is it important to use drop_first=True during dummy variable creation?
Ans:

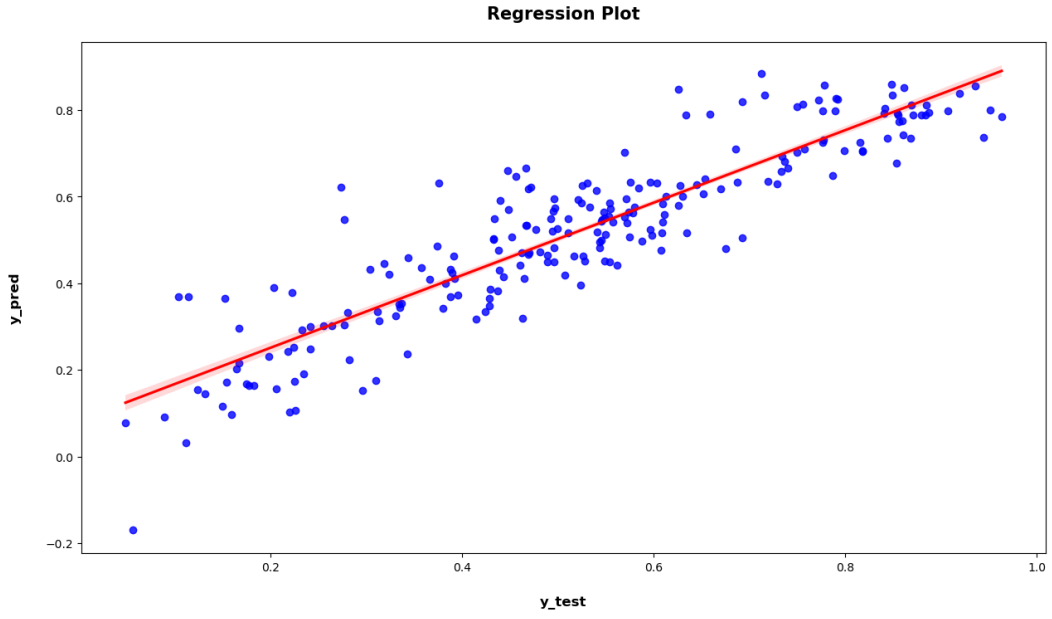
- It's essential to use **drop_first=True** which drops one of the dummy variables for each categorical level, as this avoids both multicollinearity and the dummy variable trap.
- In the case of dummy variables, if all of them are included, they will be perfectly correlated. This perfect correlation can lead to unstable coefficient estimates which is the case of multicollinearity.
- Since all the dummy variable values are known, it's easy to infer the value of the one and it can be dropped which can prevent the Dummy Variable Trap.
- For example, if there is a column gender that contains 4 variables- "Male", "Female", "Other", "Unknown". So, a person is either "Male", or "Female", or "Other". If they are not either of these 3, their gender is "Unknown". So, gender=unknown can be dropped

Q Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
Ans:

- "temp" variables are having the highest correlation with target variable "count" with correlation coefficient value 0.63
- "atemp" variable is also having same value however it's a derived variable from temp, humidity and windspeed. Hence, it's eliminated.

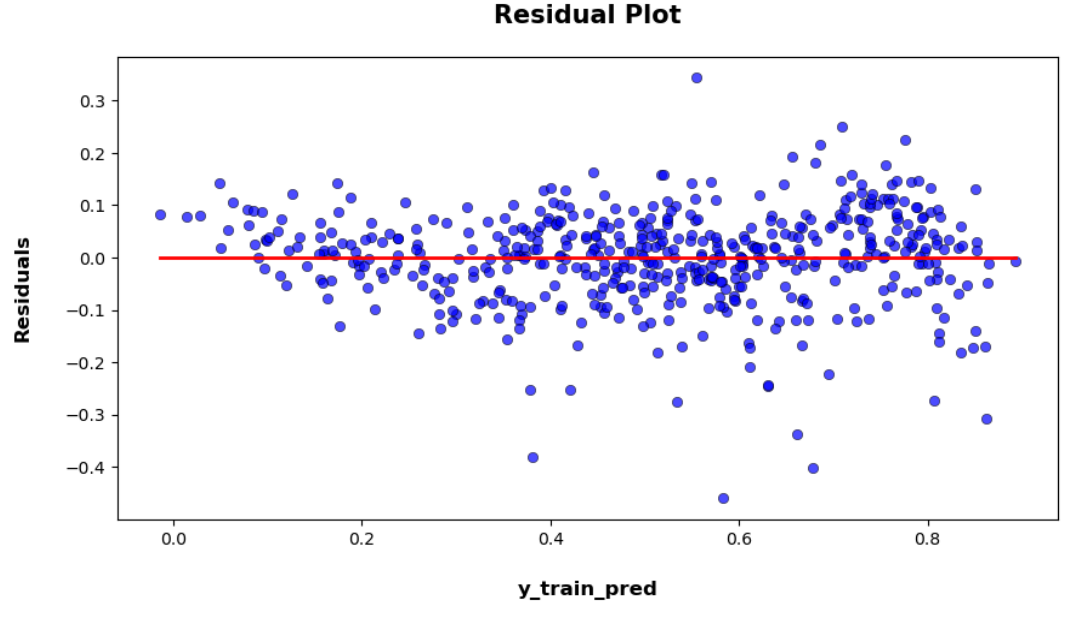
Q. How did you validate the assumptions of Linear Regression after building the model on the training set?
Ans: After constructing a linear regression model on the training dataset, it is imperative to validate the key assumptions of linear regression to ensure the reliability and accuracy of the model's predictions. Below, I outline the assumptions and the corresponding validation steps:

1. Linearity Assumption
- Validation: To assess the linearity assumption, I've examined regression plot illustrating the relationships between the dependent variable and each independent variable.



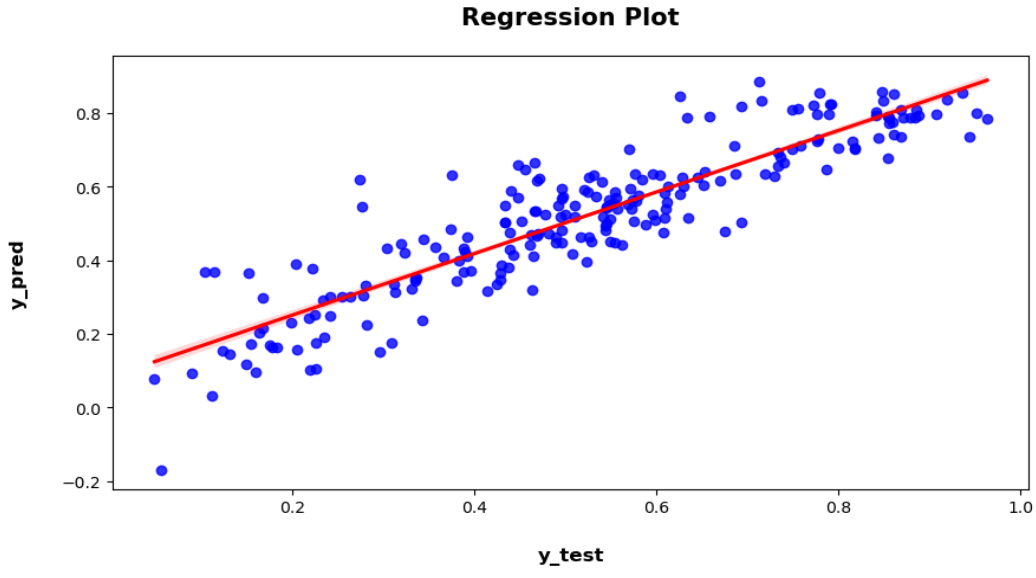
2. Independence of Errors:

- Validation: I've assessed the independence of errors by examining residual plots. The objective was to detect any autocorrelation in the residuals, with the expectation that no clear patterns or significant autocorrelation should be present.



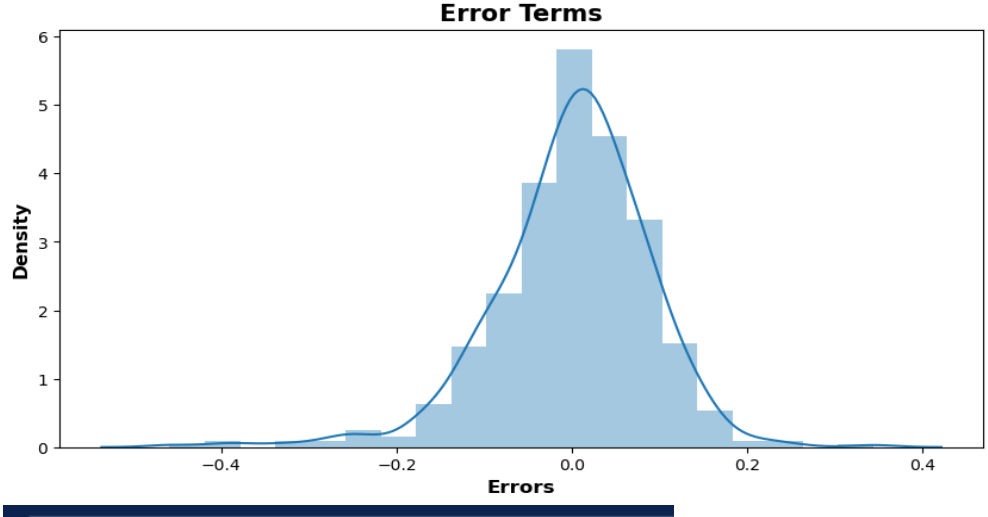
3. Homoscedasticity (Constant Variance):

- **Validation:** Homoscedasticity was evaluated through residual plots. I've aimed to confirm that the spread of residuals remained relatively constant across the range of predicted values.



4. Normality of Residuals:

- Validation: To assess the normality of residuals, I've generated histograms of the residuals. Additionally, I've calculated the mean of. The goal was to observe that the residuals approximated a normal distribution and mean value lying close to zero.




```
(y_train-y_train_pred).mean()
```

✓ 0.0s

```
-5.494515517948814e-16
```

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

Ans: Top 3 features contributing towards explaining the demand of shared bikes are

1. temp (positive correlation) - coefficient: 0.4197
2. year 2019 (positive correlation) - coefficient: 0.2352
3. light rain (negative correlation) - coefficient: -0.295

General Subjective Questions

Q. Explain the linear regression algorithm in detail.

Ans: The primary goal of linear regression is to find the best-fitting linear model that explains the relationship between the dependent variable (target) and one or more independent variables (features or predictors). This model allows us to make predictions and understand the strength and direction of relationships between variables.

Basic Assumptions:

Linear regression relies on several assumptions, including linearity, independence of errors, constant variance of errors, normality of residuals, and more. Validating these assumptions is crucial for the reliability of the model.

Linear Equation:

The linear regression model assumes a linear relationship between the independent variables (X) and the dependent variable (Y). The model can be represented as:

Simple Linear Regression – Single independent variable is used.

▪ $Y = \beta 0 + \beta 1X$ is the line equation used for SLR.

Multiple Linear Regression – Multiple independent variables are used. ▪

$Y = \beta 0 + \beta 1X1 + \cdots + \beta pXp +$ ∈ is the line equation for MLR.

$\beta 0 = value\ of\ the\ Y\ when\ X = 0\ (Y\ intercept)$ o $\beta 1, \beta 2, \ldots, \beta p = Slope\ or\ the\ gradient$

The cost functions helps to identify the best possible values for the $\beta 0, \beta 1, \beta 2, \ldots, \beta p$ which helps to predict the probability of the target variable. The minimization approach is used to reduce the cost functions to get the best fitting line to predict the dependent variable. There are 2 types of cost function minimization approaches – **Unconstrained and constrained**.

- Sum of squared function is used as a cost function to identify the best fit line. The cost functions are usually represented as
  - ▪ The straight-line equation is $Y = \beta 0 + \beta 1X$
  - ▪ The prediction line equation would be $Ypred = \beta 0 + \beta 1xi$ and the actual Y is as Yi.
  - ▪ $Now\ the\ cost\ function\ will\ be\ J(\beta 1, \beta 0) = \sum(yi - \beta 1xi - \beta 0)\ 2$
- The unconstrained minimization are solved using 2 methods
  - Closed form
  - Gradient descent

Ordinary Least Squares (OLS):

Linear regression aims to minimize the sum of squared differences between the observed values (Y) and the predicted values (Ŷ) generated by the linear equation. This technique is known as Ordinary Least Squares. The coefficients ($\beta_0, \beta_1, \beta_2, \ldots$) are estimated to minimize the sum of these squared differences.

- $ei = yi - ypred$ is provides the error for each of the data point.
- OLS is used to minimize the total e2 which is called as Residual sum of squares.
- RSS = = $\sum (yi - ypred)\ 2$

Model Evaluation:

The model's performance is assessed using various metrics such as the **coefficient of determination (R^2), mean squared error (MSE)**, and others. These metrics help gauge how well the model fits the data and makes predictions.

Interpretation of Coefficients:

Interpretation of coefficients is crucial in linear regression. A positive coefficient (β_i) indicates that an increase in the corresponding independent variable (X_i) leads to an increase in the dependent variable

(Y), assuming other variables are constant. A negative coefficient suggests the opposite. The magnitude of the coefficient represents the strength of the relationship.

Multivariate Linear Regression:

Linear regression can handle multiple independent variables simultaneously, resulting in multivariate linear regression. The model estimates coefficients for each independent variable, accounting for their combined influence on the dependent variable.

Regularization:

In some cases, linear regression can overfit the data. In this case adjusted R-square helps to prevent overfitting by adding penalty terms to the coefficient estimation process.

Predictions:

Once the model is trained, it can be used to make predictions on new, unseen data by plugging in values for the independent variables into the linear equation.

Applications:

Linear regression is widely used in various fields, including economics, finance, social sciences, and engineering, for tasks such as predicting stock prices, analysing the impact of variables on an outcome, and estimating the relationship between variables.

Limitations:

Linear regression assumes a linear relationship, which may not hold in all real-world scenarios. It is also sensitive to outliers and can be affected by multicollinearity.

In summary, linear regression is a powerful and interpretable tool for modelling and understanding relationships between variables. It forms the basis for more advanced regression techniques and provides valuable insights into data analysis and prediction tasks.

Q. Explain the Anscombe's quartet in detail

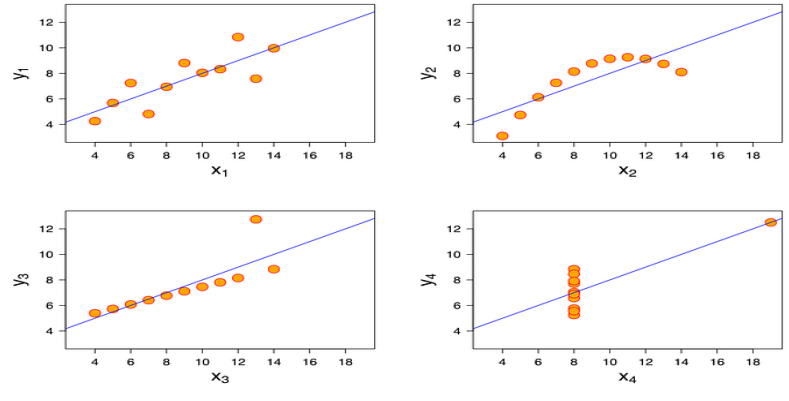
Ans: Anscombe's quartet is a set of four small datasets in statistics that were created by the British statistician Francis Anscombe in 1973. Anscombe's quartet consists of four distinct datasets, each containing 11 data points. These datasets are labelled I, II, III, and IV.

Identical Summary Statistics:

All four datasets share nearly identical summary statistics:

- **Mean of x (X-bar):** Approximately 9.0 for all datasets.
- **Variance of x (Var(x)):** Approximately 11.0 for all datasets.
- **Mean of y (Y-bar):** Approximately 7.5 for all datasets.
- **Variance of y (Var(y)):** Approximately 4.12 for all datasets.
- **Correlation between x and y (r)**: Approximately 0.816 for all datasets.

Different Data Patterns:



- Dataset I appear to have clean and well-fitting linear models.
- Dataset II shows a non-linear relationship with a single outlier.
- In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
- Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

Key Takeaways:

- Anscombe's quartet underscores the **importance of data visualization** in data analysis.
- It highlights that relying solely on summary statistics like means, variances, and correlations can be misleading, as different datasets can have similar summary statistics but different underlying structures.
- The quartet emphasizes the need to explore data graphically to understand relationships and patterns fully.
- It serves as a cautionary example when performing regression analysis, reminding analysts to be mindful of outliers and influential data points.

In summary, Anscombe's quartet serves as a compelling reminder that a deeper understanding of data often requires more than just numerical summaries and necessitates the use of graphical tools to uncover hidden patterns and relationships.

Q. What is Pearson's R?

Ans: Pearson's correlation coefficient(r) represents the relationship between two variables that are measured on the same interval or ratio scale. The Pearson coefficient is a measure of the strength of the association between two continuous variables. Pearson's correlation coefficient can take values between -1 and 1

- r = 1: Perfect positive linear relationship. As one variable increases, the other also increases, and they follow a straight-line relationship.
- r = -1: Perfect negative linear relationship. As one variable increases, the other decreases, and they follow a straight-line relationship but with a negative slope.
- r ≈ 0: Little to no linear relationship. The variables do not exhibit a clear linear pattern, and there is no strong association between them.
- r > 0 < 5 means there is a weak association
- r > 5 < 8 means there is a moderate association
- r > 8 means there is a strong association

The formula for Pearson's correlation coefficient (r) is:

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

Where:
Xi and Yi are the individual data points for the two variables (X and Y).
$\bar{X}$ and $\bar{Y}$ are the means (averages) of X and Y, respectively.

Key points about Pearson's correlation coefficient (r):

- It measures the linear relationship only. It may not capture non-linear associations between variables.
- It is sensitive to outliers, meaning that extreme values can influence the correlation substantially.

In summary, Pearson's correlation coefficient (r) provides insights into the direction and strength of the relationship but should be interpreted alongside domain knowledge and other statistical techniques for a comprehensive analysis.

Q. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
Ans:
What?
Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It aids in accelerating algorithmic calculations. Scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc are impacted.

Why?
Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

Types of Scaling:
Normalized scaling (MinMax):

X_normalized = (X - X_min) / (X_max - X_min)

Standardized scaling

X_standardized = (X - mean(X)) / std(X)

Normalized scaling vs Standardized scaling:

Feature	Normalized scaling	Standardized scaling
Range	Transforms data into a specific range (e.g., [0, 1])	Standardized scaling does not impose a specific range.
Preservation of Distribution	Preserves the shape of the original distribution but shifts and scales it to fit within the specified range.	Centers the data around zero and scales it, resulting in a distribution with a mean of 0 and a standard deviation of 1
Sensitivity to Outliers	Affected by outliers	Less affected by outliers
Usage	It is applied when we are not sure about the data distribution	It is used when the data is Gaussian or normally distributed

Q. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Ans: The Variance Inflation Factor (VIF) can become infinite in the context of multicollinearity when there is a perfect linear relationship among a set of independent variables in a regression model. This situation is known as "perfect multicollinearity." Perfect multicollinearity arises when one or more independent variables can be perfectly predicted from a linear combination of other independent variables in the model. Here's why VIF can become infinite in such cases:

$VIF = 1 / (1 - R^2)$

In cases of perfect multicollinearity, the R^2 in the VIF formula becomes equal to 1. This is because when one variable can be perfectly predicted from another, there is a perfect linear relationship, and R^2 is 1.

VIF = 1 / (1 - 1) = 1 / 0 which is infinite

In practical terms, infinite VIF indicates a severe issue of multicollinearity that needs to be addressed, usually by removing one of the perfectly correlated variables from the model.

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

Ans: Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. It helps you compare the quantiles of your data to the quantiles of a chosen reference distribution (typically the normal distribution) to determine if the data fits that distribution.

Advantages of using Q-Q plot:

1. **Assessing Distributional Assumptions:**
   - In linear regression and many other statistical techniques, it is often assumed that the follow a specific distribution, typically the normal distribution.
   - A Q-Q plot is used to visually check whether this assumption holds. It compares the quantiles of the observed residuals to the quantiles of a theoretical normal distribution.

2. **Detecting Departures from Normality:**
   - If the data points in the Q-Q plot deviate significantly from a straight line, it suggests that the data do not follow a normal distribution.
   - Departures from normality can indicate issues such as outliers, skewness, or other non-normal characteristics in the residuals.

3. **Identifying Outliers:**
   - In a Q-Q plot, extreme outliers may appear as points that deviate markedly from the expected straight line.

Importance of a Q-Q Plot in Linear Regression:

1. **Model Assumption Validation:**
   - A Q-Q plot is an important tool for assessing whether this key assumption of a regression holds true.

2. **Diagnostic Tool:**
   - Q-Q plots are part of the diagnostic toolkit in linear regression.
   - Departures from normality detected in a Q-Q plot may prompt further investigation into the source of non-normality and possible model refinements.

3. **Outlier Detection:**
   - Outliers can be seen as data points that deviate substantially from the expected straight line in the Q-Q plot which have significant impact on the regression results.

4. **Model Improvement:**
    - Q-Q plot provides valuable insights into areas for model improvement

In summary, a Q-Q plot is a valuable graphical tool in linear regression for assessing the distributional assumptions of the residuals, detecting deviations from normality, identifying outliers, and guiding model diagnostics and improvements. It helps ensure that the underlying assumptions of the regression analysis are met, which is crucial for the validity and reliability of the results.