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? (3 marks)

Answer:

- Seasonal Impact: Bike Rentals are highest in the fall season, followed by summer. September sees the most bike rentals, indicating peak demand during this month.
- Yearly Increase: There is an increase in bike rentals in 2019 compared to 2018.
- Weekday Variation: Bike rentals are fairly consistent across weekdays.
- Weather Influence: Clear weather conditions lead to more bike bookings, which is a predictable trend.
- Holiday Effect: Bike rentals are less on holidays compared to regular working days.

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

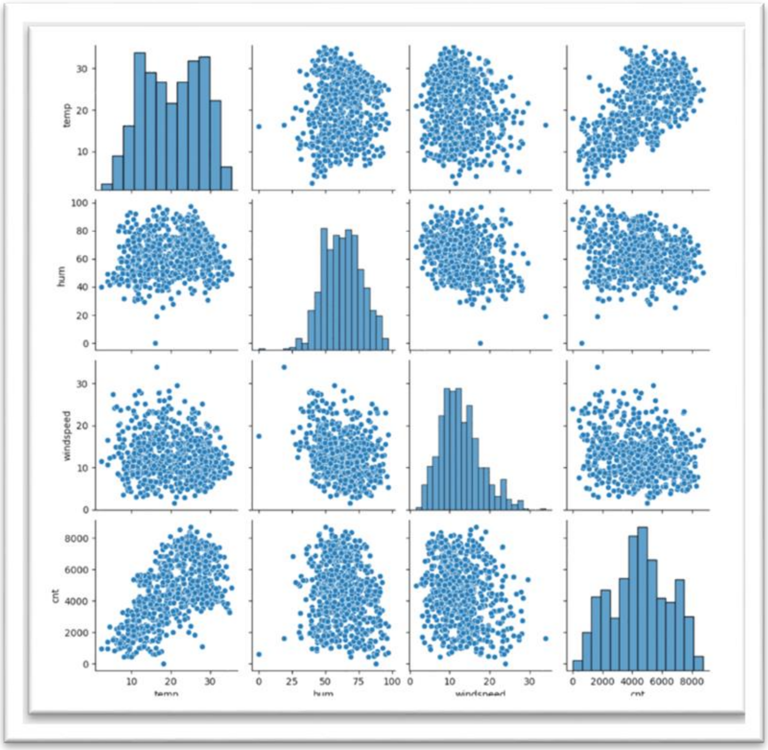
Answer: drop_first=True is used in dummy variable creation to avoid multicollinearity. When we convert categorical data into numbers using dummy variables, we often end up with one variable for each category. When creating dummy variables, we drop one of the categories of a categorical feature by setting drop_first=True. If we include dummy variables for all categories of a categorical feature without dropping one, it can cause a problem called multicollinearity.

Multicollinearity occurs when some variables in a model are highly correlated, meaning that one variable can be predicted from the others. This makes it hard for the model to distinguish the individual effect of each variable.

For example, if you have a categorical variable "Color" with three categories: Red, Blue, and Green, creating dummy variables with drop_first=False would result in three variables: "Color_Red," "Color_Blue," and "Color_Green." Using drop_first=True would drop one of these, such as "Color_Green," leaving "Color_Red" and "Color_Blue." The coefficients for "Color_Red" and "Color_Blue" would then be interpreted relative to "Color_Green."

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

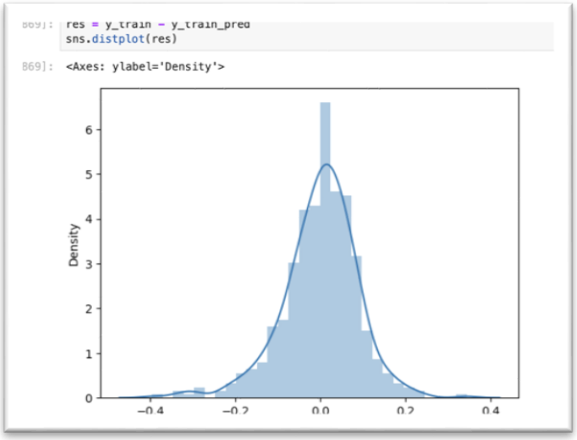
Answer: Looking at the below pair-plots among the numerical variables, it seems that "temp" variable has the highest correlation with the target variable "cnt".



4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

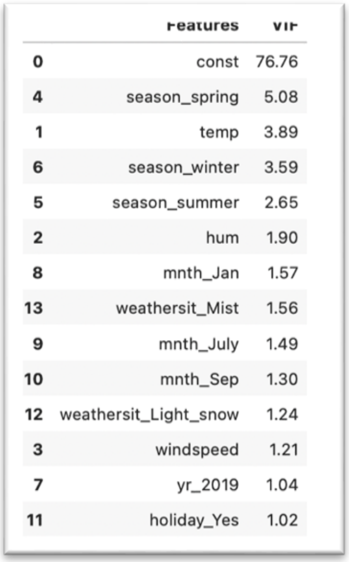
Answer: The key assumptions of Linear Regression and how they are validated:

1. **Normality of Residuals:** The residuals should be approximately normally distributed. A histogram of the residuals is used to visually inspect normality.



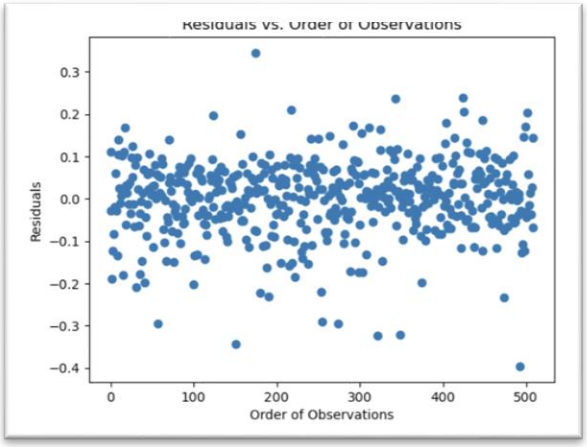
```
869]: res = y_train - y_train_pred
      sns.distplot(res)

869]: <Axes: ylabel='Density'>
```

2. **No Multicollinearity:** Independent variables should not be too highly correlated with each other. For this, we checked the Variance Inflation Factor (VIF) for each independent variable. A VIF above 10 suggests significant multicollinearity.

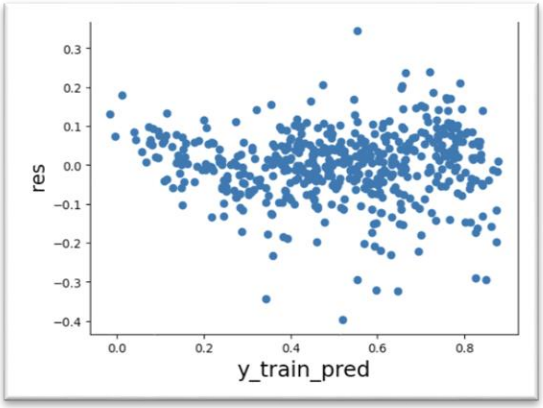


	Features	VIF
0	const	76.76
4	season_spring	5.08
1	temp	3.89
6	season_winter	3.59
5	season_summer	2.65
2	hum	1.90
8	mnth_Jan	1.57
13	weathersit_Mist	1.56
9	mnth_July	1.49
10	mnth_Sep	1.30
12	weathersit_Light_snow	1.24
3	windspeed	1.21
7	yr_2019	1.04
11	holiday_Yes	1.02

3. **Linearity**: The relationship between the independent variables and the dependent variable should be linear. Plot the residuals (the differences between the observed and predicted values) against the predicted values. The residuals should scatter randomly around zero without forming any specific pattern. A linear relationship can also be checked using scatter plots or by examining the coefficients and R-squared value for improvements.
4. **Independence**: The residuals (errors) should be independent of each other. Plotting residuals over time (if applicable) can also help check for patterns that suggest dependence.

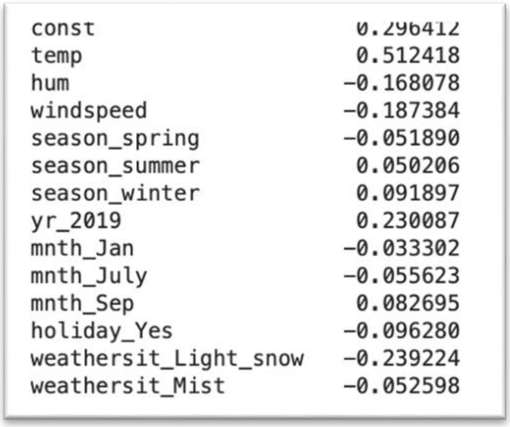


5. **Homoscedasticity**: The residuals should have constant variance at every level of the independent variables. Again, use a plot of residuals versus predicted values. If the spread of the residuals is roughly constant (no funnel shape or pattern), homoscedasticity is likely met.



5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Answer: To determine the top features contributing to the demand for shared bikes based on a linear regression model, we look at the absolute value of the coefficients of the features. Larger absolute values indicate a stronger impact on the dependent variable, which in this case is the demand for shared bikes. Below shared are the coefficients of variables:



```
const                    0.296412
temp                     0.512418
hum                     -0.168078
windspeed               -0.187384
season_spring           -0.051890
season_summer            0.050206
season_winter            0.091897
yr_2019                  0.230087
mnth_Jan                -0.033302
mnth_July               -0.055623
mnth_Sep                 0.082695
holiday_Yes             -0.096280
weathersit_Light_snow   -0.239224
weathersit_Mist         -0.052598
```

The top 3 features contributing significantly towards explaining the demand of the shared bikes seem to be -

1. Temp variable: It has coefficient value as 0.512. Indicates that as the temperature increases, the demand for shared bikes increases.
2. Year variable: It has coefficient value as 0.23. Suggests that there might be an upward trend in bike demand over the years.
3. Weathersit variable (Light snow, light rain, Thunderstorm): It has coefficient value as -0.23. The negative coefficient indicates that worse weather conditions are associated with a decrease in the demand for shared bikes.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Answer: Linear regression is a fundamental algorithm in statistics and machine learning used to model the relationship between a dependent variable (target) and one or more independent variables (features). The goal of linear regression is to find the best-fitting straight line through the data points that can be used to predict the target variable based on the feature(s).

Simple vs. Multiple Linear Regression:

- Simple Linear Regression: Involves one independent variable and models the relationship with a straight line.
- Multiple Linear Regression: Involves two or more independent variables and models the relationship with a plane or hyperplane in higher dimensions.

Equation of the Line:

- For simple linear regression, the equation of the line is: y=mx+c where:

y is the predicted value of the dependent variable.

m (or β1) is the slope of the line, representing the change in yyfor a one-unit change in x.

x is the independent variable.

c (or β0) is the y-intercept, representing the predicted value of y when x=0.

- For multiple linear regression, the equation generalizes to: y=β0+β1x1+β2x2+…+βnxn

Finding the Best-Fitting Line

The linear regression algorithm finds the best-fitting line by minimizing the "residual sum of squares" (RSS), which is the sum of the squared differences between the observed values and the values predicted by the model.

Assumptions of Linear Regression

- Linearity: The relationship between the independent and dependent variables is linear.
- Independence: Observations are independent of each other.
- Homoscedasticity: The residuals (errors) have constant variance across all levels of the independent variables.
- Normality: The residuals are normally distributed (for making statistical inferences).

2. Explain the Anscombe’s quartet in detail. (3 marks)

Answer: Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics (mean, standard deviation, correlation, etc.) but reveal distinct patterns when graphed. Created by statistician Francis Anscombe in 1973, the quartet illustrates the importance of visualizing data before making statistical inferences.

Key Characteristics of the Quartet

Each of the four datasets in Anscombe's quartet has:

- The same mean for both X and Y variables.
- The same standard deviation for both X and Y variables.
- The same correlation coefficient (Pearson's R) between X and Y.
- The same linear regression line equation for Y as a function of X.

Anscombe's quartet is a set of four groups of data that are used to show why looking at data in more than one way is important. Even though these groups have the same basic statistics, like average and correlation, they look very different when you plot them on a graph.

What Each Group Shows:

First Group:
- The data points form a straight line, showing a clear linear relationship.

Second Group:
- The points make a curve instead of a straight line. The same statistics don't show this curve.

Third Group:

  - Most points are close to a line, but there's one point far from the others, an "outlier," which can mislead you if you only look at the numbers.

Fourth Group:

  - Most points are the same, with one point very different. This can create a false impression of a pattern that isn't really there.

Why It Matters:

Anscombe's quartet teaches us that:

- Graphs Matter: Always look at data visually because patterns and outliers can change the interpretation.
- Numbers Can Mislead: Basic statistics don't always tell the whole story.
- Outliers: A single unusual point can greatly affect the results and interpretations.

In short, Anscombe's quartet reminds us to visualize data and think critically, not just rely on numbers.

3. What is Pearson's R? (3 marks)

Answer: Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It is denoted by rr and ranges from -1 to 1.

Interpretation:

r=1: Perfect positive linear correlation. As one variable increases, the other also increases proportionally.

r=−1: Perfect negative linear correlation. As one variable increases, the other decreases proportionally.

r=0: No linear correlation. There is no linear relationship between the two variables.

Use Case: Pearson's R is commonly used in statistics, data analysis, and research to identify and quantify the degree of linear association between variables, helping in understanding relationships and making predictions.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Answer: Scaling is a data preprocessing technique used in machine learning and data analysis to adjust the range of data features. The purpose of scaling is to ensure that different features contribute equally to the model's learning process, preventing features with larger ranges from disproportionately influencing the model's performance.

Reasons for Scaling:

- Uniform Contribution: Ensures all features contribute equally to the model's decision-making process.
- Improved Convergence: Helps gradient-based optimization algorithms converge faster by standardizing feature ranges.
- Performance Improvement: Can improve the performance of certain models, especially those sensitive to the scale of input features, like distance-based algorithms.

Normalized Scaling vs. Standardized Scaling:

Normalized Scaling (Min-Max Scaling):

  - Formula: $X'=(X-Xmin)/Xmax-Xmin$
  - Range: Transforms data to a fixed range, typically [0, 1] or [-1, 1].
  - Use Case: Useful when the data distribution is not Gaussian and the model assumptions are minimal about the data distribution.

Standardized Scaling (Z-score Scaling):

- Formula: $Z=(X-\mu)/\sigma$

- Mean (μ) and Standard Deviation (σ): Data is transformed to have a mean of 0 and a standard deviation of 1.
- Use Case: Commonly used when the data is normally distributed or when a Gaussian distribution is assumed by the model.

In summary, scaling adjusts feature ranges, normalized scaling maps data to a specific range, and standardized scaling centers data around the mean with unit variance.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer: An infinite Variance Inflation Factor (VIF) typically occurs when there is perfect multicollinearity among the independent variables in a regression model. Perfect multicollinearity means that one or more independent variables are exact linear combinations of the others, making it impossible to determine the unique contribution of each variable to the model. When there is perfect multicollinearity, Ri2 =1 for at least one of the variables because the variable is perfectly predictable from the others.

An infinite VIF indicates a serious problem in the model, as it suggests redundancy among the variables, making it difficult to estimate the regression coefficients accurately. This typically requires addressing the multicollinearity issue by removing or combining variables, using dimensionality reduction techniques, or other approaches to ensure the model's stability and interpretability.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Answer: A Q-Q (Quantile-Quantile) plot is a graphical tool used to assess whether a data set follows a particular theoretical distribution, such as the normal distribution. The plot compares the quantiles of the observed data with the quantiles of the theoretical distribution.

Key Components of a Q-Q Plot

- Quantiles: Points taken at regular intervals from the cumulative distribution function (CDF) of a distribution.
- Theoretical Quantiles: Quantiles from the theoretical distribution (e.g., normal distribution) that the data is being compared against.
- Sample Quantiles: Quantiles from the sample data.

How to Interpret a Q-Q Plot

- Linearity: If the points on a Q-Q plot fall approximately along a straight line, the data are approximately normally distributed.
- Deviations: Systematic deviations from the line indicate departures from the specified distribution. For instance, if the points form an S-shaped curve, it suggests a heavier or lighter tail than the theoretical distribution.

Importance in Linear Regression

- Normality of Residuals: One key assumption in linear regression is that the residuals (errors) are normally distributed. A Q-Q plot of the residuals can help verify this assumption. If the residuals are not normally distributed, it might indicate problems with the model, such as non-linearity, outliers, or heteroscedasticity (non-constant variance of errors).
- Model Diagnostics: By examining a Q-Q plot, one can detect anomalies or deviations from normality, which may suggest the need for data transformation, the inclusion of additional variables, or other remedial measures to improve the model.
- Assessment of Fit: A good fit in a Q-Q plot suggests that the chosen model and the assumptions underlying it are reasonable for the data. Conversely, a poor fit can indicate that the model needs to be revised.

In summary, a Q-Q plot is a crucial diagnostic tool in linear regression analysis, helping to validate the assumptions and assess the quality of the model fit.