Linear Regression Model :
cnt = 0.1922 + (0.2343 * yr) + (0.0254 * workingday) + (0.4638 * atemp) +
(-0.2879 * light_weather) + (-0.0733 * mist_weather) + (-0.1250 * spring) + (0.0455 * winter)

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?

- There are some positive and negative impact of the categorical variables on the dependent variable.
- Positive Impact (workingday , winter) - We can say that on workingday, demand for bike goes up by 0.0254 than on holiday. Most of the customers use bike for work. In winter season, demand goes up by 0.0455 keeping all other variables constant.
- Negative Impact (light_weather, mist_weather, spring) - When there is a bad weather (like mist, light snow, light rain, thunderstorm and spring season) demand for bike goes down.
- In spring season has negative impact where as winter season has positive impact on dependent variable.

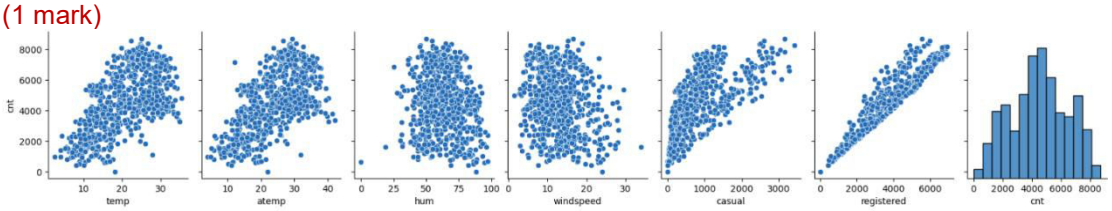
---

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

- If there are 'n' categories in a categorical variables, then we should introduce 'n-1' dummy variables only.
- With the help of (n-1) dummy variables, we can give different values for n categories.
- Ex. Here I have taken only 3 dummy variables for 4 categories in "Season".
- dummy variables - spring, summer, winter
  - 000 = fall (when all are 0, its fall season) base condition
  - 100 = spring
  - 010 = summer
  - 001 = winter

---

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

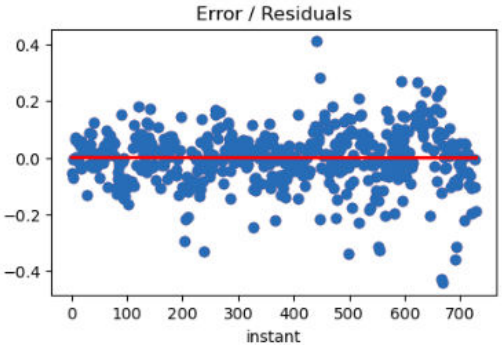


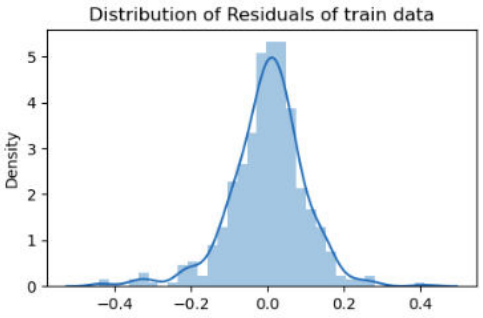
temp & atemp have high correlation with cnt

---

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

- There is a linear relationship between independent variables and "cnt" (dependent variable).





- We can say that residuals or errors are independent.
- And error are normally distributed as seen in graph.
- Errors have constant variance. There is no pattern in the error. There is a horizontal trend line passing through the errors. Which mean with change in x, there is no change or pattern for y=error.

---

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

1. yr : 0.2343
2. atemp : 0.4638
3. light_weather : -0.2879(though it has -ve coefficient, it has high significance towards explaining the demand)

General Subjective Questions

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

- It is used for predicting continuous variables from one or more input independent variables.
- General equation is y = b0 + b1*x1 + b2*x2....+ bn*xn
- b0 is an intercept. When all variables (x1,x2..xn) are null then output variable is equal to the intercept b0.
- We can say that y increases by b1 for 1 unit increase in x1 keeping all other variables constant. There is a linear relationship between x1 and y.
- Once the model is trained and the coefficients are determined, you can make predictions for new data points.
- And we can check the accuracy between actual value and predicted value.
- **Assumptions** : Linear regression assumes that the relationship between the features and the target variable is linear, the errors (residuals) are normally distributed, and the independence of errors (no autocorrelation).

---

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

- Anscombe's quartet is a set of four small datasets, each containing 11 data points.
- The remarkable feature of these datasets is that they have nearly identical simple descriptive statistics (such as means, variances, and correlation coefficients) but exhibit significantly different patterns when graphed.
- Anscombe's quartet serves as a strong reminder of the importance of data visualization.
- It demonstrates that relying solely on numerical summary statistics can lead to a misleading understanding of the data.
- Graphical exploration of the data can reveal important nuances, outliers, and relationships that are not apparent in summary statistics.

- This insight is particularly valuable in data analysis, where a visual examination of the data can lead to more accurate and meaningful interpretations.

---

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

- Pearson's correlation coefficient "r" is a statistic used to quantify the strength and direction of the linear relationship between two continuous variables.
- It measures the degree to which two variables are linearly related to each other.
- It provides insights into how changes in one variable correspond to changes in the other.
- Pearson's r is widely used in statistics and data analysis to assess the association between variables.
- Pearson's r can take values between -1 and 1,
  - r = 1 : indicates a perfect positive linear relationship. This means that as one variable increases, the other increases proportionally.
  - If r = -1: indicates a perfect negative linear relationship. This means that as one variable increases, the other decreases proportionally.
  - If r = 0 : suggests no linear relationship between the two variables. However, it's important to note that this does not rule out the possibility of other types of relationships (e.g., nonlinear relationships).

---

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

- **What is scaling?**

Scaling is a data preprocessing technique to transform the numerical features of a dataset into a specific range or distribution.

- **Why is scaling performed?**
  1. To ensure that all numerical features have a common scale or magnitude
  2. To make the model training process more efficient
  3. To reduce the sensitivity of algorithms
- **difference between normalized scaling and standardized scaling?**
- Normalized scaling scales the data to a specific range (usually between 0 and 1)
- while standardized scaling transforms the data to have a mean of 0 and a standard deviation of 1.

---

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

- The Variance Inflation Factor (VIF) is a measure used to assess multicollinearity in a multiple regression analysis.
- VIF can be infinite or undefined. This typically occurs when there is perfect multicollinearity in the dataset.
- **Scenarios:**
  1. Perfect multicollinearity happens when one or more independent variables can be expressed as exact linear combinations of other variables in the model.
  2. If we have two or more variables in regression model that are identical or can be expressed as a linear combination of one another, we will encounter perfect multicollinearity.
     - example - if we have both "length in meters" and "length in centimeters" as independent variables, VIF is infinite
  3. Adding a Constant Variable: When a constant or an intercept term is included in the regression model
     - Example - when a column of ones is added to represent the intercept
  4. Dummy Variable Trap: When dealing with categorical variables and creating dummy variables for them, if we include all possible dummy variables for a category (instead of using one less than the number of categories), it can result in perfect multicollinearity.

---

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

- **What is a Q-Q plot?**

  A Quantile-Quantile (Q-Q) plot is graphical representation of the quantiles (ordered values) of the data against the quantiles of the chosen theoretical distribution.

- **Use of a Q-Q Plot:**

1. primarily used to visually compare the distribution of dataset to the normal distribution.
2. Identifying Departures from Normality: Deviations from a straight line in the Q-Q plot can signal departures from normality, such as skewness or heavy-tailed distributions.
3. Detection of Outliers: Outliers in data may become apparent in a Q-Q plot

- **Importance of a Q-Q Plot in Linear Regression:**

1. Q-Q plot is a valuable diagnostic tool in linear regression analysis, helping us
2. assess the assumptions of normality
3. assess constant variance
4. identify outliers in the residuals.
5. It aids in ensuring the reliability and validity of your regression model