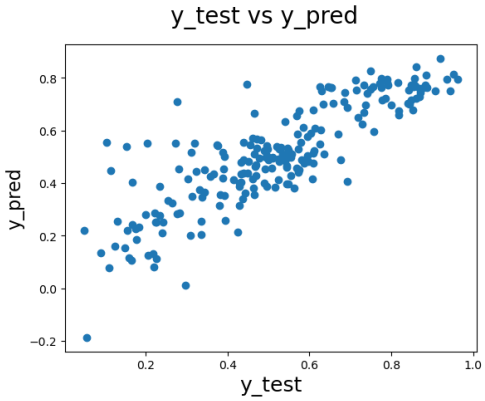
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)
    a. The season seems to have good impact on bike rental with lower demand in spring.
    b. When it's not a holiday the rental is higher. OR we can say, on a working day the rental is slightly higher.
    c. Looking at weathersit, the clearer the weather the number of rental increases.
    d. The months of Dec, Feb, Jan and Mar have negative impact on rental.
    e. The month of Sep has a positive impact on rental.
    f. Sat has a positive impact on rental.

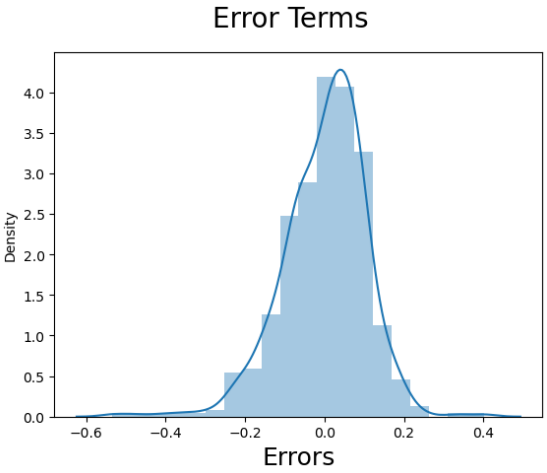
2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
   When we have a categorical variable with only two possible e.g. Yes, No, then we don't need to create dummy variable for it the two values No=0,Yes=1 are sufficient to express them in a single variable. In case when we have a variable with 3 values A,B,C then we can express them using two variables say A,B. For value A will be represented as 1,0. For B values will be 0,1 and C can be represented as 0,0. So we only need n-1 dummy variables to represent n values of a categorical variable. If we don't remove the extra dummy variable then it can introduce multicollinearity and it can also impact time of model evaluation. So we should use drop_first=True. However, in certain cases instead of dropping the first dummy variable we can drop a specific one based on business understanding.

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

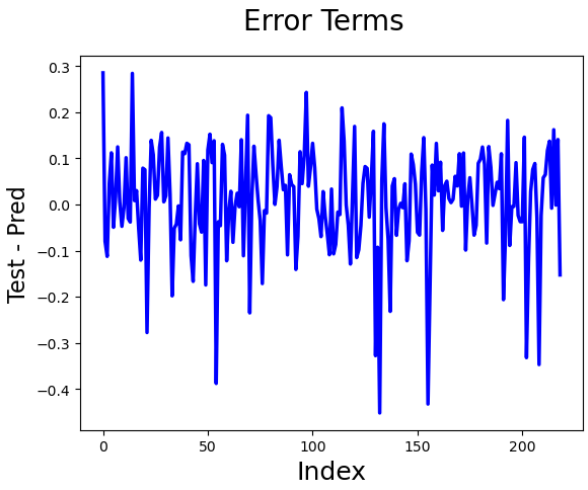
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
    a. Assumption of linearity:
       The y test and y pred shows a linear relationship so our model seems to capture a linear relationship.



b. Errors should be normally distributed and centered around zero. Created a Histogram of error terms to validate that they are normally distributed.



c. The residuals are random variables i.e. there is no pattern observed in residual.
Create a chart of (y_test - y_pred) to validate that the errors are randomly distributed i.e. there is no pattern within them.



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

General Subjective

Questions

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

Linear regression is used to find the value of one numeric target variable based on one or more dependent variables. The dependent variables should be independent from each other. The algorithm finds the best fit line which takes the following form

$Y = C + B_0X_0 + B_1X_1 + \ldots$

Where Y is the target variable, C is a constant, B_0 , B_1 are coefficients of dependent variables X_0 , X_1 etc.

The best fit line is one which has least value of residuals i.e y_test – y_predicted . Since absolute value of residuals is difficult to quantify as good as bad it is better to use a ratio. That's why R-Square is used. Its defined as

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}$$

Where:

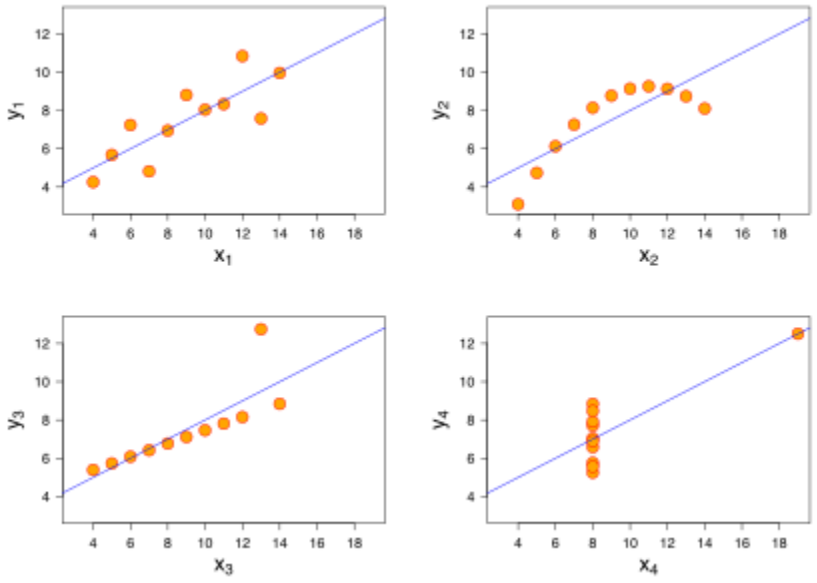
- RSS = sum of squared residuals

- TSS = total sum of squares

When R squared is 1 it means the model can explain 100% variance in the data.

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

The Anscombe's quartet consists of 4 datasets such that they have different distribution of data but the datasets have same statistical values like mean, standard deviation. Each dataset has 11 records of x and y values. The primary purpose of it was to demonstrate the significance of visualizing the data instead of only relying on statistical values.



- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two correlated variables, where y could be modelled as gaussian with mean linearly dependent on x.
- For the second graph (top right), while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third graph (bottom left), the modelled relationship is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier, which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

3. What is Pearson’s R? (3 marks)

The Pearson’s R is one of the correlations coefficients which is used in linear regression. A higher value of Pearson’s R does not guarantee that the relationship is linear. The other correlation coefficient is Spearman's R which is used to determine the correlation if the relationship between the variables is not linear. So even though, Pearson's R might give a correlation coefficient for non-linear relationships, it might not be reliable. For example, the correlation coefficients as given by both the techniques for the relationship $y = X^3$, for 100 equally separated values between 1 and 100 were found out to be: Pearson’s R(aprox 0.91) and Spearman's R(aprox 1). And as we keep on increasing the power, the Pearson's R value consistently drop whereas the Spearman's R remains robust at 1.

So, the takeaway here is that if you have some sense of the relationship being non-linear, you should look at Spearman's R instead of Pearson's R. It might happen that even for a non-linear relationship, the Pearson's R value might be high, but it is simply not reliable.

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

The dataset that we are given to analyze will have multiple numeric columns. Each numeric column will have different range of data. Example one column’s value can range from 1 to 100 while other columns level can range from 1 to 1 million. Scaling is the process of converting these values to bring them to same scale. Scaling is performed so that coefficients of features are not dependent on range of values. That will create an incorrect model with higher coefficients for variable with higher range. Another advantage of scaling is that it enables gradient descent to converge faster.

Normalized scaling uses min and max value while standardized scaling uses mean and standard deviation to scale the variable.

Normalized scaled value = $(X-X_{min})/(X_{max}-X_{min})$

Standardized scaled value = $(X- X_{mean})/X_{\text{Standard deviation}}$

Normalized Scaling	Standardized scaling
Rescales values to a range between 0 and 1	Centers data around the mean and scales to a standard deviation of 1
Useful when the distribution of the data is unknown or not Gaussian	Useful when the distribution of the data is Gaussian or unknown
Sensitive to outliers	Less sensitive to outliers
Retains the shape of the original distribution	Changes the shape of the original distribution

May not preserve the relationships between the data points	Preserves the relationships between the data points

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
VIF is defined as 1/(1-RSquare). Now if variance of a variable is fully explained by another variable or combination of multiple variables then RSquare will be 1 resulting in infinite value of VIF. E.g BMI is a ratio of two variables weight and height. So if we calculate VIF for BMI for a dataset having weight and height as two variables then its VIF will be infinite.

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

A QQ or quantile-quantile plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. For example, if we run a statistical analysis that assumes our residuals are normally distributed, we can use a normal QQ plot to check that assumption.