MULTIPLE LINEAR REGRESSION

Bike sharing Assignment

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?

Season: Majority of bikes were rented in the fall season, since the weather is not too hot or too cold so this could be one of the reasons why it's highly likely for people to rent bikes in the fall season.

Year: For the year of 2019, the growth in people renting bikes is substantial.

Holiday: People prefer to rent when there’s a holiday since they have time to perform leisure activities.

Weekday: For the days of “Friday”,”Saturday”,”Thursday” the total amount of rentals are higher as compared to other days.

wokingday: when it is not a working day, the count of bikes for rent is usually more compared to a working day

weathersit: the largest amount of bike rentals are observed when the weather is Clear, few clouds or Partly cloudy

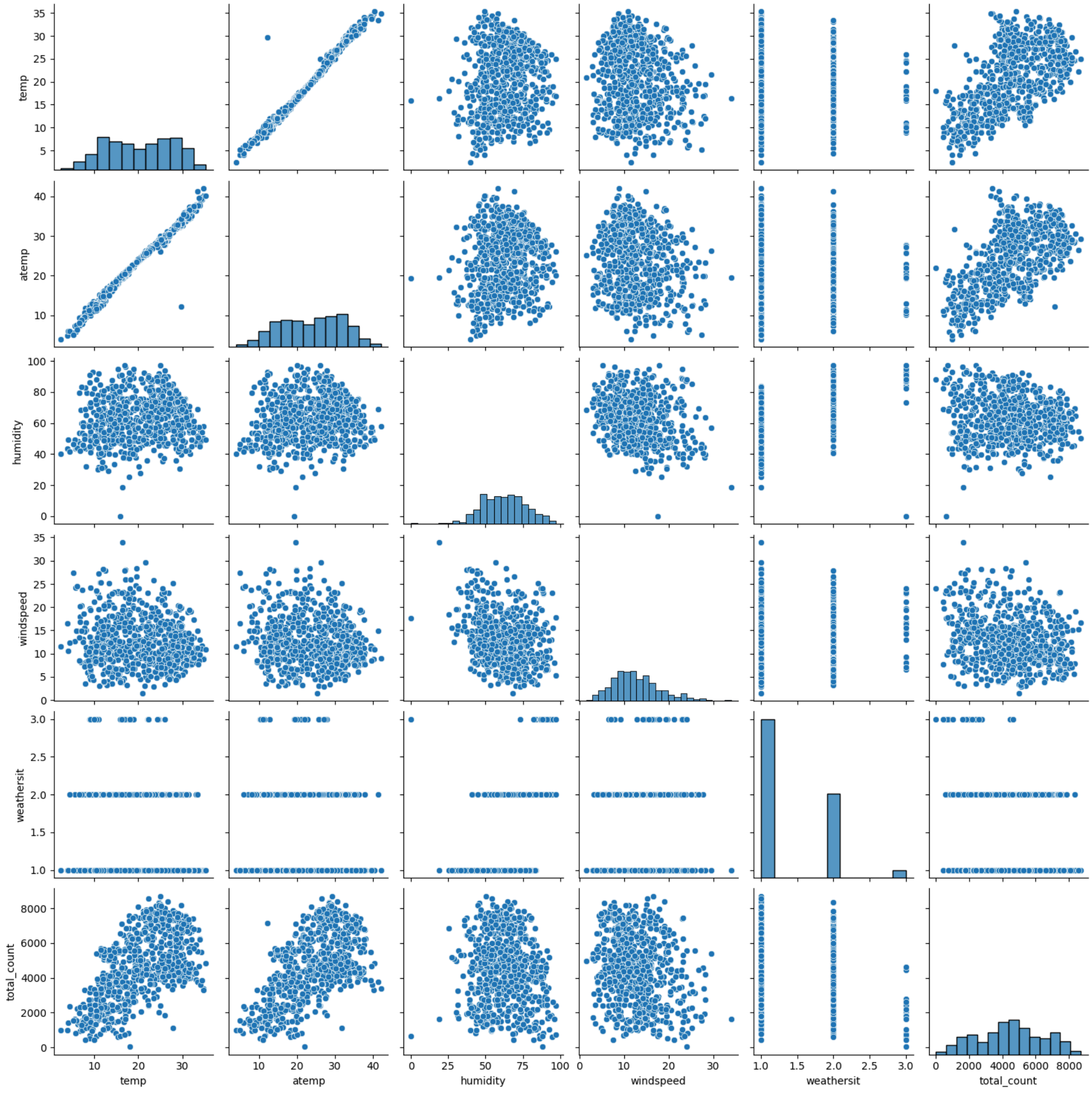
month: Majority of the bikes are being rented on the month of August-September compared to other months

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

Dummy variables are used to represent categorical variables in separate columns filled with binary values. It converts the categorical variables with n-levels into n-1 binary columns, which leads to prevention of multicollinearity.

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

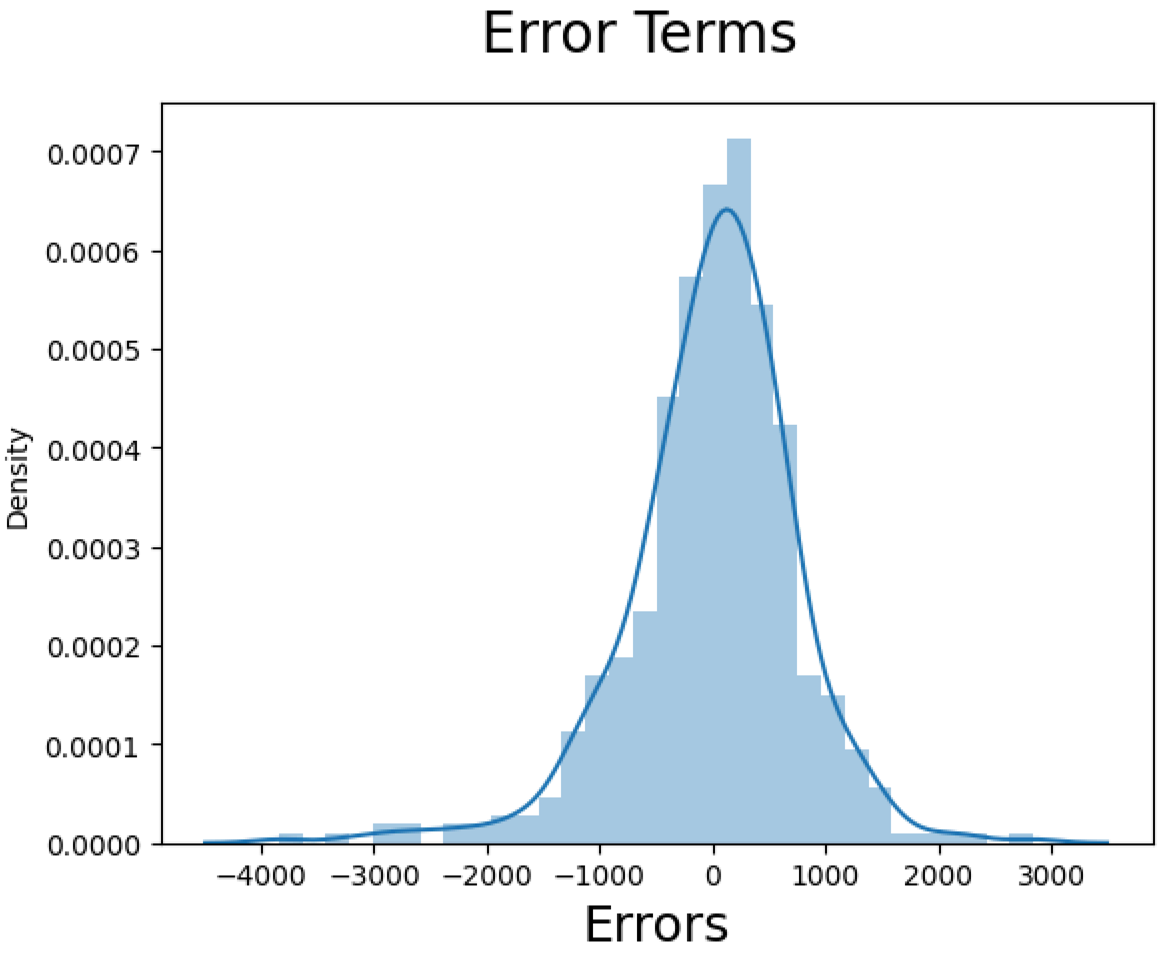
The pair-plot among the numerical variables showed a strong correlation between 'temp' and 'atemp' and the target variable, 'total_count'. The line plots showing the relationship between the two variables for the datasets used are very similar, and both of them indicate a linear relationship between 'total_count'.



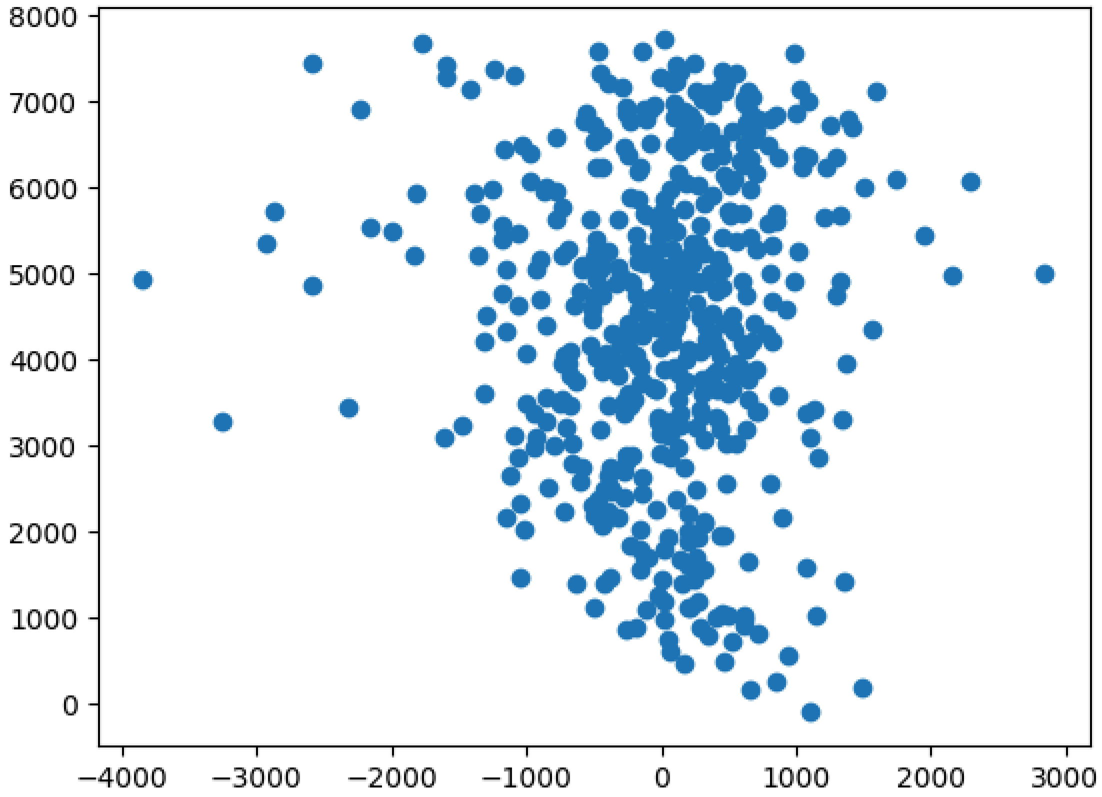
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

For validation of assumptions made after building the model in the training set, the following points are considered with regard to Linear Regression:

1. One of the factors that are to be observed for Linear Regression is that the Error terms should be normally distributed with a mean of 0. Thus, to validate this, we perform a residual analysis on the train set. Residuals are the difference between the actual y_train and the predicted y_train. A normally distributed pattern with a mean of approximately 0 is obtained when plotting a distribution plot of the residuals.



2. The following assumption is Multicollinearity, which means a very small correlation between independent features or no multicollinearity between them. This is verified looking at the VIF of the variables in the final model. A VIF below 5 is considered a good VIF.
3. The third assumption, Homoscedasticity, states that there should not be any pattern observed when graphed between residual and fitted values. The below plot validates the Homoscedasticity assumption.



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

Top three features which are highly significant towards explaining the demand of the shared bikes are:

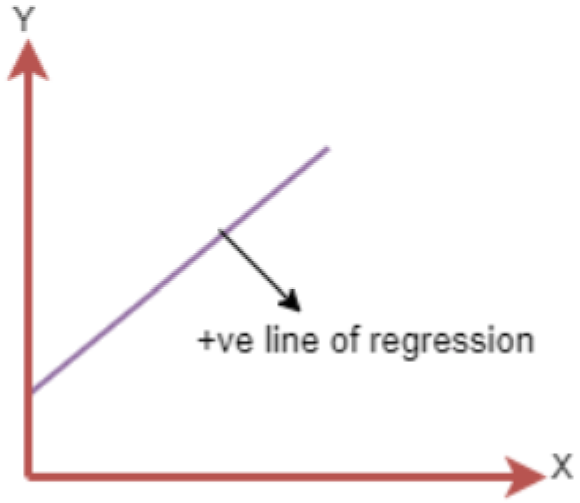
- atemp
- Humidity
- Wind speed

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a statistical method used to define the relationship between two or more variables. It is largely used to predict the outcome based on input variables. The algorithm assumes a linear relationship between the independent variables and the dependent one. For example, in the prediction of house prices, the independent variables could be square footage, number of bedrooms, and location; the dependent variable would be the price. The algorithm calculates the best-fit line, which aims to minimize the difference between the predicted values and actual values within the dataset. The line is represented by the equation y = mx + b, where y is the dependent variable, x is the independent variable, m is the slope of the line, and b is the intercept. Estimates of values of m and b are predicted by the training data so that the line will have minimal deviation from the training data.

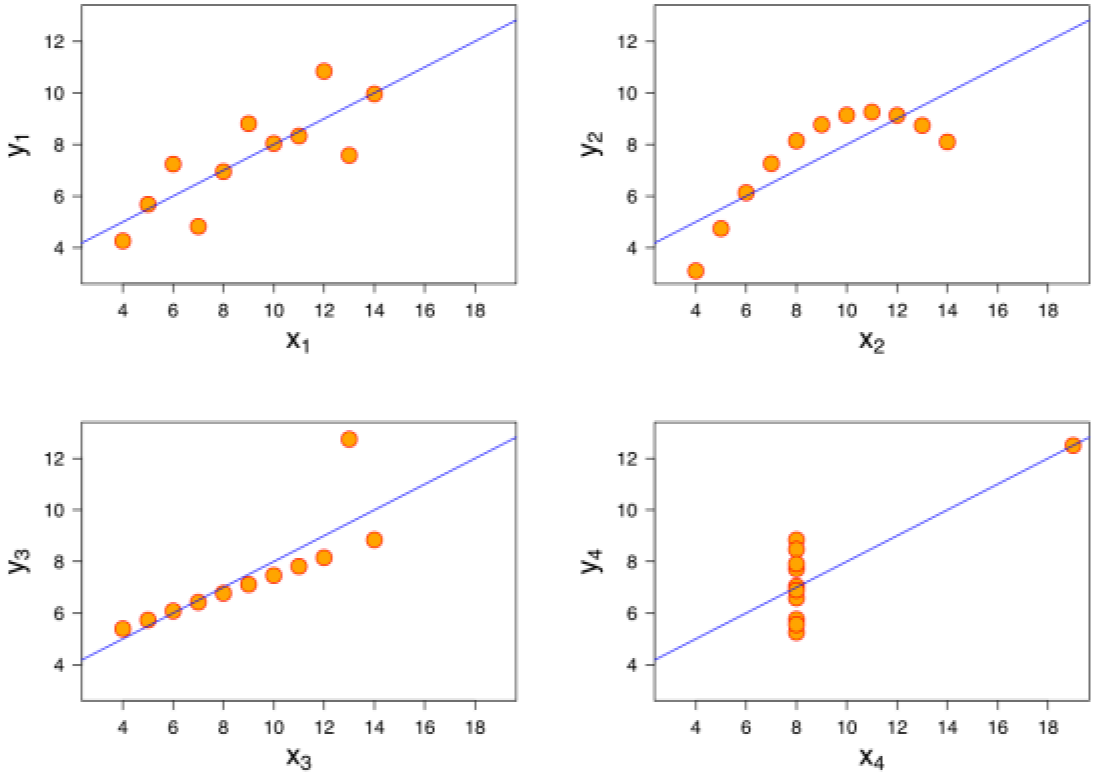
A line that demonstrates the relationship between the dependent and independent variable is called the line of regression. There are two varieties of RL: the positive line of regression and the negative line of regression.



The line equation will be: $Y= a_0+a_1x$

2. Explain the Anscombe’s quartet in detail.

Anscombe's quartet informs us that before putting together models using various algorithms, we should visualize our data. This means the data features must be plotted so as to see the distribution of samples, which can help identify various anomalies within the dataset (outliers, diversity of the data, linear separability of the data, etc.). Anscombe's quartet is a group of four data sets that are almost identical when it comes to simple descriptive statistics, but some very strange behaviors appear when plotted on scatter plots. Each dataset contains eleven (x,y) points.



Observations from above data sets:
Data Set 1: fits the linear regression model well.
Data Set 2: cannot fit the linear regression model because the data is nonlinear.
Data Set 3: shows the outliers involved in the dataset, which cannot be handled by the linear regression model.
Data Set 4: shows the outliers involved in the dataset, which also cannot be handled by the linear regression model.

3. What is Pearson's R?

The Pearson's Correlation Coefficient (r) is a statistical measure of the linear correlation between two variables. Like all correlations, it also bears a numerical value ranging from -1.0 to +1.0. Pearson's R cannot capture nonlinear relationships between two variables, nor can it differentiate between dependent and independent variables. In case there is a strong positive relationship, r equals 1. It means that if one increases, the other also increases; otherwise, r equals -1, which means that if one increases, the other decreases; otherwise, r equals 0, meaning that there is no linear relationship.

Formula for Pearson r:

$$r = \frac{\sum \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum \left(x_i - \bar{x}\right)^2 \sum \left(y_i - \bar{y}\right)^2}}$$

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

Scaling is what simply means that the numerical feature values should be put into the same range. Scaling is very important since few variables may have values on a different scale (smaller/higher) compared to other variables. Scaling also helps speed up the calculation of an algorithm.

Feature scaling is indispensable in the matter of machine learning, as it is important to standardize all the features (variables) so as to bring all of them up to the same scale. This is because when there are big differences between the scales of the features, the algorithms could pick up any of the big features and thus also lose some sense of proportion in the predictions. Scaling will improve the model's performance and convergence by making the features comparable, whether through the use of linear regression, k-nearest neighbors, or neural networks.

- Normalized Scaling, also known as min-max scaling, is a scaling that transforms data into a common range of between 0 and 1. It is calculated using the formula:

  $$x' = (x - x_{min})/(x_{max} - x_{min})$$

- Standardized Scaling, or z-score normalization scaling, is a way of scaling data such that it has a mean of 0 and a standard deviation of 1. It is calculated using the formula:

$$z = \frac{x - \mu}{\sigma}$$

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

When the Variance Inflation Factor (VIF) becomes infinity, it reflects perfect multicollinearity. This occurs when an independent variable can precisely be predicted by a linear combination of other predictors in the model. The VIF formula is as follows, using the coefficient of determination, or (R^2), representing how well a variable is predicted by other predictors in a model. If (R^2) approaches 1 (high predictability), then the VIF will increase to very large values. For VIF raning above 10 are deemed undesirable and often pose a problem. When sample size becomes large it mitigated the concerns related to multicollinearity.

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

The quantile-quantile (Q-Q) plot is a graphical way to determine whether two data sets come from populations with common distribution. Use of the Q-Q plot: The Q-Q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. A quantile refers to the fraction (or percentage) of points below the given value. So, the 0.3 (or 30%) quantile refers to the point at which 30% of the data fall below and 70% fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population having the same distribution, the points should approximately follow this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions. Importance of the Q-Q plot: If two sets of data exist, one usually desires to know if the assumption of common distributions is justified. In such a case, if the two samples differ, it would be helpful to gain some understanding of the differences. The Q-Q plot can yield more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.