Linear Regression Bike Sharing Assignment

Assignment Based Subjective Questions:

Q-1: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer: From our analysis of the categorical variable from the data set I am able to infer some of the following effects of categorical variables on the dependent variables:

- From the 'season' we can see that almost 5000 bookings are from the fall season compared to other seasons
- we can observe that the count of bike is increased in the year 2019.
- From the 'mnth' column we can see that the months are following a trends and could be a good predictor variable. The booking in mid months are above 4000.
- Most of the bike booking were happening when it is not a holiday. It means holiday can not be a good predictor for the dependent variables.
- There are seems no trend in the weekday dataset, so we can leave for prediction.
- we can see that bike rental was on the higher end on days which were marked as non-working days. Also, the median count of bikes on non-workings days equals the median count of bikes on working days.
- we can see that the bike rental was on the higher end on days which were marked as clear weather situation, and also the median count of the bikes on clear days are greater as compared to any other weather situation.

Q-2: Why is it important to use **drop_first=True** during dummy variable creation?

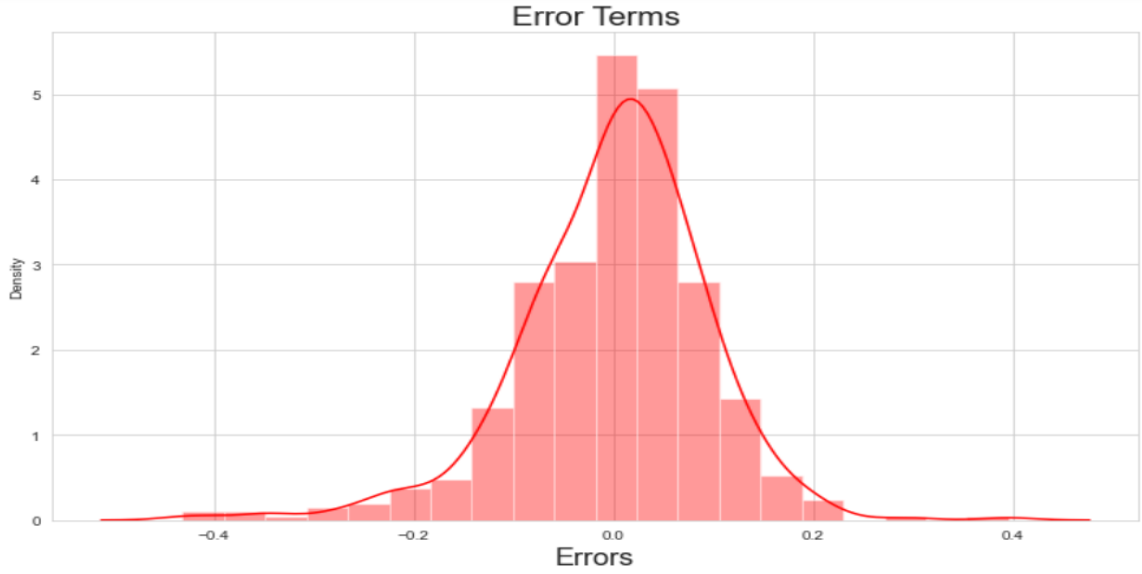
Answer: A dummy variable is a numerical variable used in regression analysis to represent subgroup of the sample.
It is important to use drop _first = True, as it helps to reduce the extra column created during dummy variable creation. Hence it reduce the correlation created among dummy variables.
If we do not drop one of the dummy variables created from a categorical variable then it becomes redundant with data set as we will have constant variable (intercept) which will create multicollinearity issue.

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

Answer: From the pair plot 'temp' has the highest correlation among the other numerical variables with 'cnt' as the target variable.

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

Answer: The assumption that we make after we building the linear regression model on the training data set is that error terms are distributed normally. In support of that we have doe the residual analysis. Residual represents the error of the different between actual y values and predicted y value by the model.



From the above diagram as we could see that the residuals are normally distributed and maximum of the error terms resolving around zero. Hence our assumption of linear regression is valid. Also ensured the overfitting by looking the R-square value and Adjusted R-square value.

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

Answer: From the final model the top 3 variables that needed for the prediction which are influence the counts are:

- **temp**: A coefficient value of 0.4711 indicates that a unit increase in temp variable, increases the cnt number by 0.4711 units.
- **2019**: A coefficient value of 0.2330 indicates that the unit increases in yr(2019) variable, increases the cnt number by 0.2330 units.
- **Sep**: A coefficient value of 0.0657 indicates that a unit increases in the mnth(Sep) variable, increases the cnt number by 0.0657 units.

General Subjective Questions

Q-1: Explain the linear regression algorithm in detail.

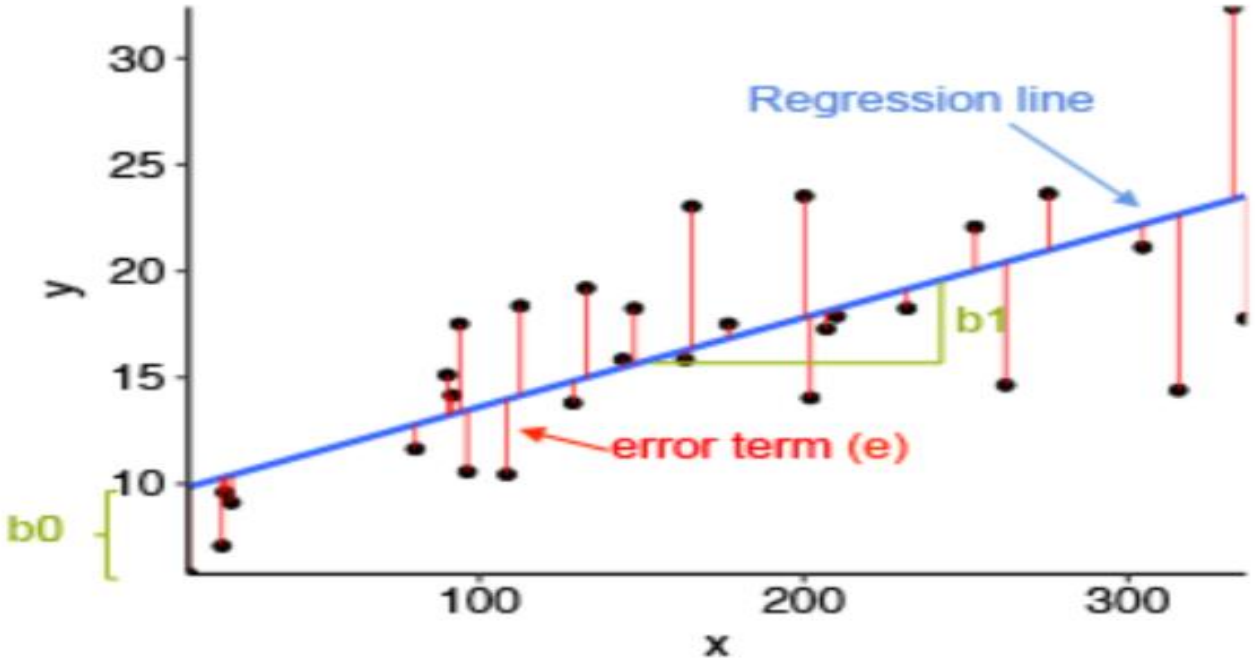
Answer: Linear regression is a supervised machine learning algorithm where the predicted output is continuous and has a constant slope. It's used to predict values within a continuous range, (e.g. sales, price) rather than trying to classify them into categories.

Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

Mathematically, we can write a simple linear regression equation as follow

$$Y = m*x + c$$

Where y is the predicted variable (dependent variable), m is slope of the line, x is independent variable, c is intercept(constant). It is cost function which helps to find the best possible value for m and c which in turn provide the best fit line for the data points.



Here x and y are the two variables on the regression line.

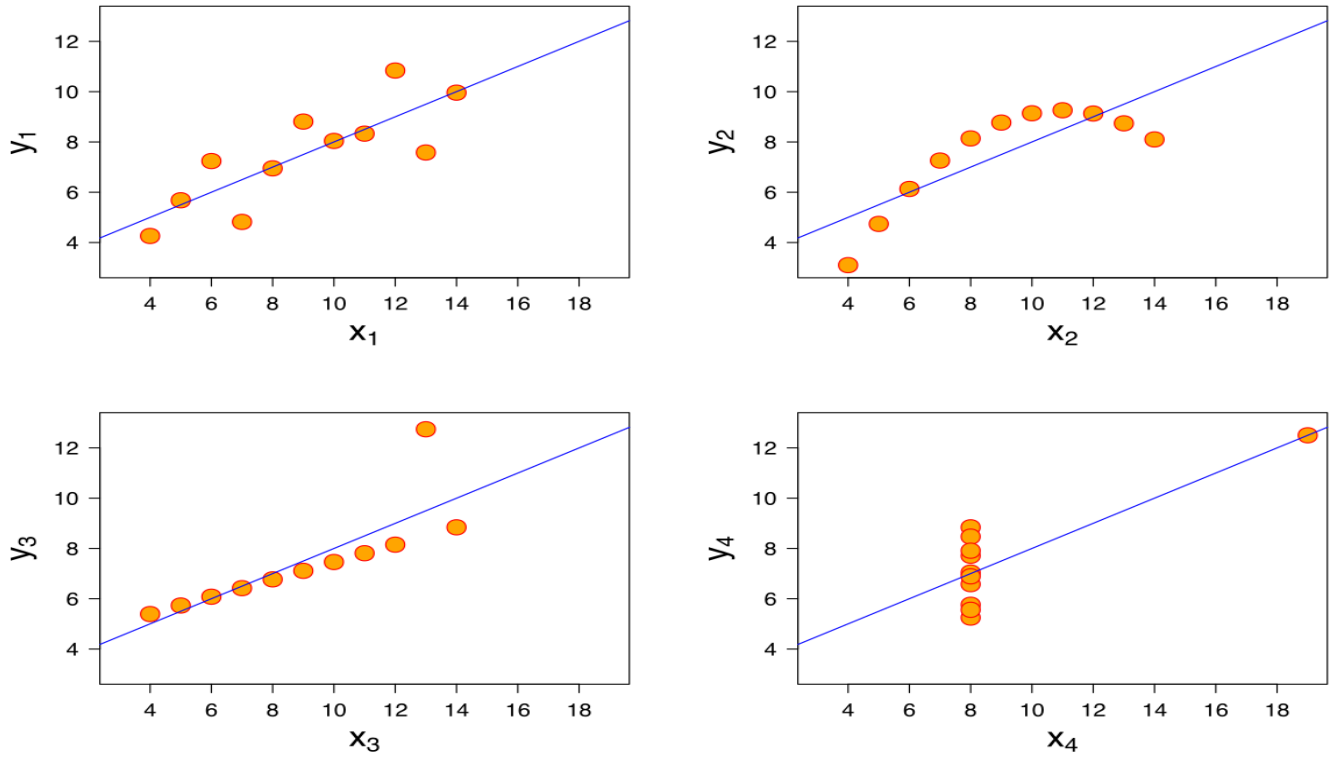
b1 – slope of the line, and **b0** – y-intercept of the line

x – independent variable, and **y** – dependent variable from the data set.

Q-2: Explain the Anscombe's quartet in detail.

Answer: Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed.

Each dataset has a set of X and Y points, which has different statistical properties. However, when the data points are graphed, all the dataset shows similar statistical relations. All the four datasets provides the correlation between X and Y points and the equation of linear regression line.



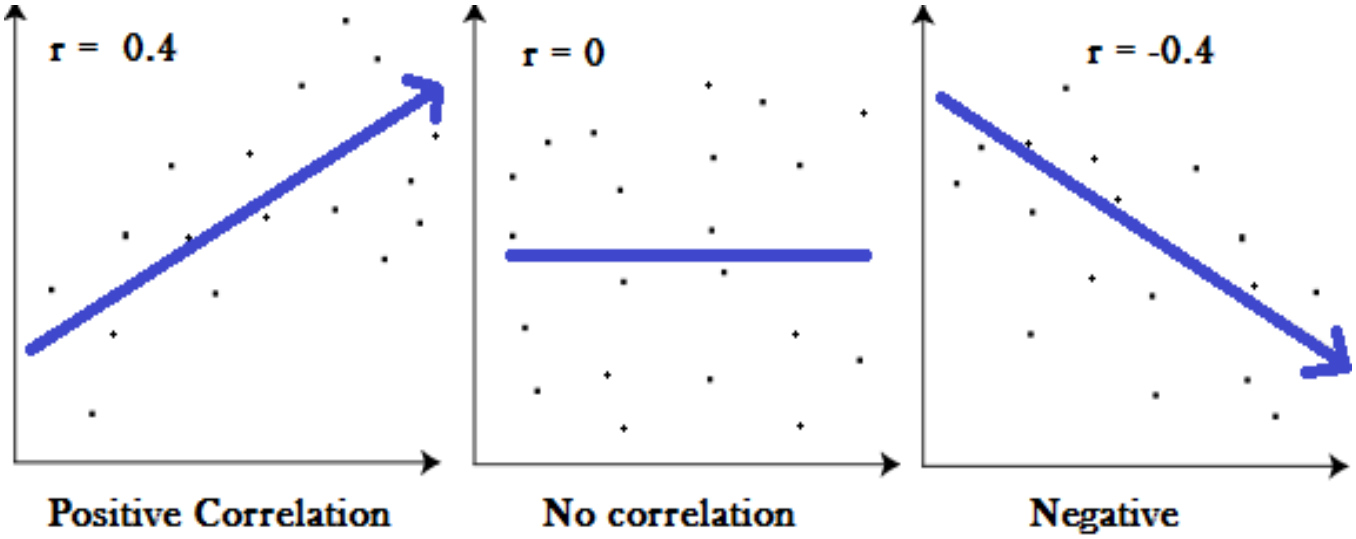
The four datasets can be described as:

- Dataset 1: this fits the linear regression model pretty well.
- Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.
- Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model
- Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model in the best.

Q-3: What is Pearson's R?

Answer: Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. Pearson's r measures the strength of the linear relationship between two variables.

Pearson's r always between -1 and 1.



- A correlation coefficient of 1 means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other. For example, shoe sizes go up in (almost) perfect correlation with foot length.
- A correlation coefficient of -1 means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other. For example, the amount of gas in a tank decreases in (almost) perfect correlation with speed.
- Zero means that for every increase, there isn't a positive or negative increase. The two just aren't related.

Q-4: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer: Scaling: Feature scaling in machine learning is one of the most critical steps during the pre-processing of data before creating a machine learning model. It is performed to bring all the independent variables on a same scale in regression. If Scaling is not done, then regression algorithm will consider greater values as higher and smaller values as lower values.

Why is scaling performed:

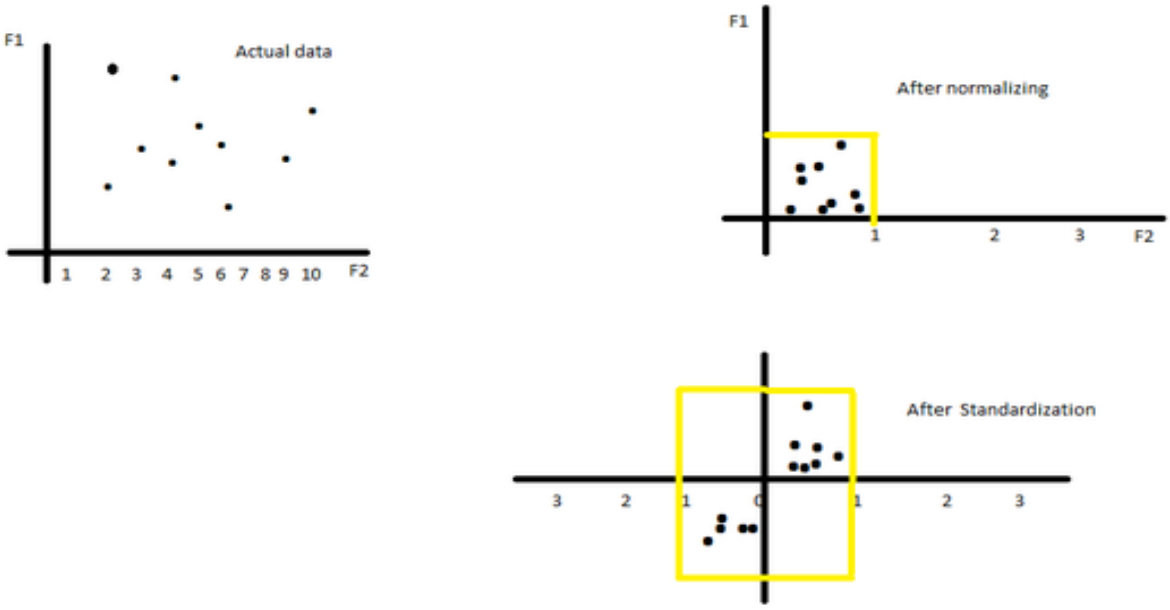
1. Machine learning algorithm just sees number, if there is a vast difference in the range say few ranging in thousands and few ranging in the tens, and it makes the underlying assumption that higher ranging numbers have superiority of some sort. So these more significant number starts playing a more decisive role while training the model.

2. To ensure that the gradient descent moves smoothly towards the minima and that the steps for gradient descent are updated at the same rate for all the features, we scale the data before feeding it to the model. Having features on a similar scale can help the gradient descent converge more quickly towards the minima

Difference between Normalization and Standardization:

Normalization is used when we want to bound our values between two numbers, typically, between [0,1] or [-1,1]. While Standardization transforms the data to have zero mean and a variance of 1, they make our data unit less. Refer to the below diagram.

Below diagram shows that hoe the data looks like after scaling:



Standardized Scaling: Is also knows as z score normalization, which transforms the data in such a way that the resulting distribution has mean of 0 and a standard deviation of 1.

Formula:

$$X_{new} = \frac{X_i - X_{mean}}{\text{Standard Deviation}}$$

Normalized Scaling: Is also know as min-max scaling. Which basically transforms the numerical data to scale between 0 and 1.

Formula:

$$X_{new} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

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

Answer: VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables.
If all the independent variables are orthogonal to each other, then VIF = 1.0. If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables.

The standard error of the coefficient determines the confidence interval of the model coefficients. If the standard error is large, then the confidence intervals may be large, and the model coefficient may come out to be non-significant due to the presence of multicollinearity. A general rule of thumb is that if VIF > 10 then there is multicollinearity.

An infinite VIF value indicates that the dependent variable may be expressed exactly by a linear combination of other variables. VIF = 1/ (1-R2), when R2=1 then VIF = Infinity
Example: In our Assignment, Registered Users + Casual Users = Total no. of Users If we fit the model including these 2 variables then VIF will be infinity because of this.

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

Answer: Q-Q plot is a probability plot, which is a graphical method for comparing two probability distribution by plotting their quantile against each other. First, the set of intervals for the quantiles is chosen. A point (*x*, *y*) on the plot corresponds to one of the quantiles of the second distribution (*y*-coordinate) plotted against the same quantile of the first distribution (*x*-coordinate). Thus the line is a parametric curve with the parameter which is the number of the interval for the quantile.

It is used for determining if two data sets come from populations with a common distribution. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. Whether the Distributions is Gaussian, Uniform, Exponential or even Pareto distribution, it can be found out.

The purpose of Q-Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q-Q plot; if the two data sets come from a common distribution, the points will fall on that reference line. ... It's being compared to a set of data on the y-axis.

Advantages:

1. The sample sizes do not need to be equal.
2. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

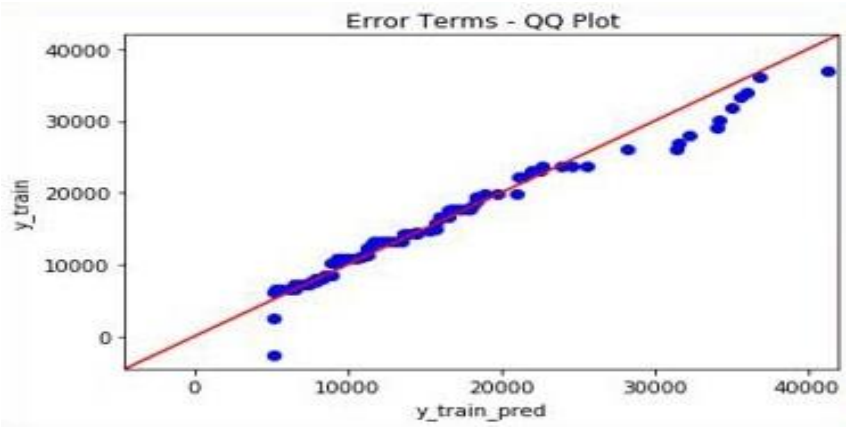
If two data sets —

- Come from populations with a common distribution.
- Have common location and scale.
- Have similar distributional shapes.
- Have similar tail behaviour.

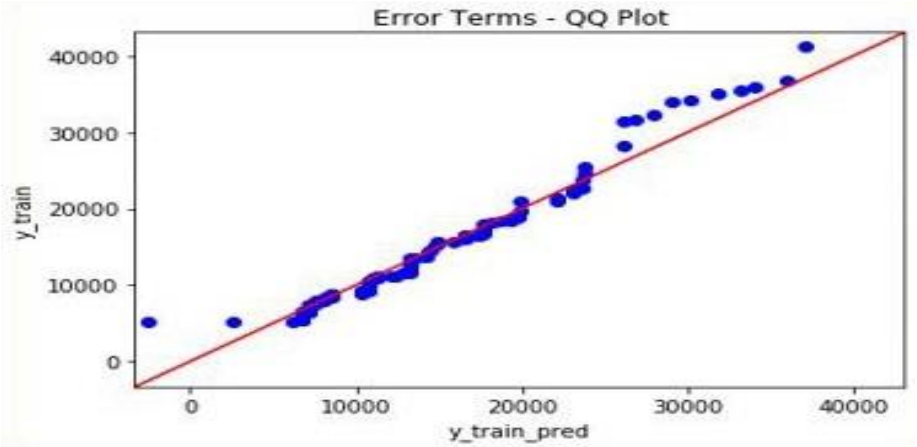
Below are the possible regressions for two data sets:

1. Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis

2. Y-values < X-values: If y-quantiles are lower than the x-quantiles.



3. X-values < Y-values: If x-quantiles are lower than the y-quantiles.



4. Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x –axis.

SUBMITTED BY:
ADARSHA SAHOO