BIKE SHARING ASSIGNMENT

By: Shreejith S

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?

The categorical variables in our model were season,yr,mnth,holiday,weekday,workingday and weathersit.

- □We could see that the workingday variable did not have much effect on the target variable ‘cnt’

- □When the weathersit was Clear the cnt of users was considerably higher compared to Light snow weather conditions.

- □The median of cnt of users on a holiday is just about 2700 while on other days the cnt of users was as high as around 4200

- □The median of cnt of users was similar for all the days of week , but the range of cnt of users was highest for Thursday

Contd from second slide

- □Year 2019 saw a higher number of users compared to 2018 and particularly in the month of September while the lowest number was in January

- □Fall saw a pretty good number of users compared to other seasons.

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

- drop_first=True helps in creating only the necessary number of dummy variables.

- If we see when there are n levels for a categorical variable, n-1 variables are sufficient to represent all the levels, as the False value of all other variables automatically implies the True value of 1 variable

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

- The highest correlation is between the cnt variable and registered variable but this doesn’t much value because the registered variable is a part of cnt variable.

- Keeping this in mind we can conclude that temp and cnt variable have highest correlation

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

- We performed the residual analysis to check if the error terms are normally distributed.

- We first calculated the predicted value of the model using predict() function and then found the error by taking difference of actual value and predicted value.

- Using the distplot of seaborn library we observed that the distribution of errors was a normal distribution.

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

- temp (Temperature) : As the temperature increases the count of users would increase by 0.5701 times.

- weathersit_Light Snow (Snowy weather) : Increase in the Snowy weather would reduce the count of users by 0.24 times.

- yr_2019 (Year 2019) : The demand for bikes increased in the year 2019 by 0.23 times

General Subjective Questions

1. Explain the linear regression algorithm in detail.

- Linear regression is a form of predictive modeling technique which tells us how the independent variables (predictors) influence the dependent (target variable) when they are in a linear relationship.
- If there is a single input variable (x) or predictor variable, such linear regression is called simple linear regression. And if there is more than one predictor variable, it is called as multiple linear regression. The linear regression model gives a sloped straight line describing the relationship within the variables.
- A regression line can be a Positive Linear Relationship or a Negative Linear Relationship. The goal of the linear regression algorithm is to get the best values for the intercept(β0) and the coefficient(β1,β2 etc) of the predictor variables to find the best fit line so that the best fit line has the least error.
- In Linear Regression, RFE or Mean Squared Error (MSE) or cost function is used, which helps to figure out the best possible values for β0 and β1, which provides the best fit line for the data points.

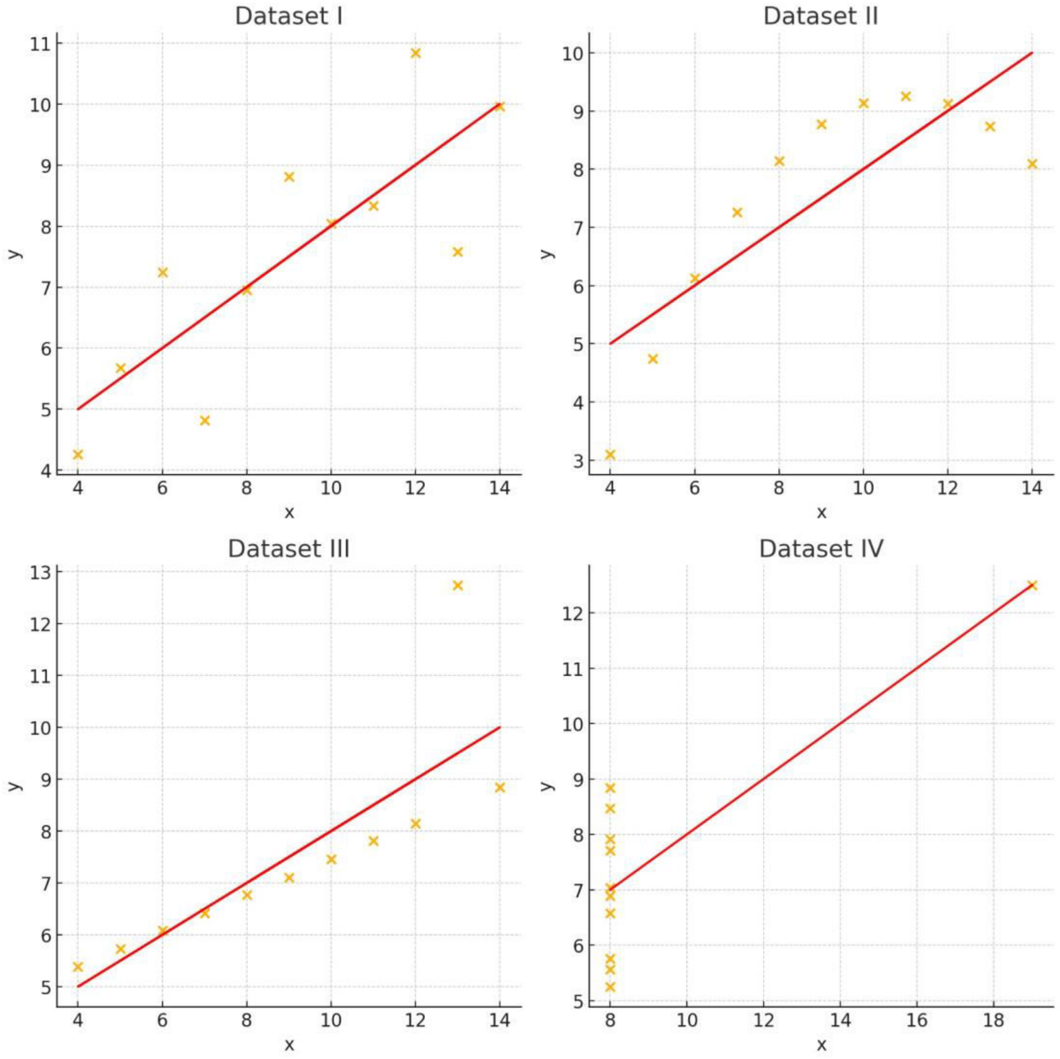
2. Explain the Anscombe’s quartet in detail.

Anscombe's quartet is a set of four datasets that have nearly identical simple statistical properties but appear very different when graph is plotted with this data. The quartet was created by the British statistician Francis Anscombe in 1973 to show the importance of graphing data before analyzing it and the effect of data anomalies,outliers on statistical properties.

Here are the four datasets, commonly referred to as I, II, III, and IV:

- Dataset I: Shows a simple linear relationship between x and y.
- Dataset II: Shows a non linear relationship , a curved trend.
- Dataset III: Contains a clear outlier that influences the regression line.
- Dataset IV: Consists of a vertical line (constant x value) with an outlier that affects the correlation and regression line.

Continued from slide 10



3. What is Pearson’s R?

Pearson’s R, also known as Pearson’s correlation coefficient, is a measure of the linear correlation between two variables. It quantifies the extent of linear relationship between them. The value of Pearson’s R ranges from -1 to 1, where:

1 indicates a perfect positive linear relationship where, as one variable increases other variable increases as well.

0 indicates no linear relationship.

-1 indicates a perfect negative linear relationship where, as one variable increases other variable decreases .

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

- Scaling is a data preprocessing technique used to adjust the range of data features to a standard range.
- This is done to improve the performance and training stability of machine learning algorithms. It helps ensure that each feature contributes equally to the model's predictions and prevents any single feature from dominating due to its scale.

Normalized Scaling:

- Adjusts the values of features to a common scale without distorting differences in the ranges of values. Min-max scaling is a normalized scaling method where data is transformed to fit within a range of [0, 1].
- More sensitive to outliers, as it directly scales based on the minimum and maximum values.

Standardized Scaling:

- Data is transformed to have a mean of zero and a standard deviation of one (unit variance). This process is often referred to as Z-score normalization.
- Less sensitive to outliers, as it uses mean and standard deviation, which are less influenced by extreme values.

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

- The Variance Inflation Factor (VIF) is a measure used in regression analysis to detect multicollinearity between independent variables.
- When one predictor variable is an exact linear combination of one or more other predictor variables, the VIF for that variable becomes infinite and this would be the case of perfect multicollinearity.
- This happens when you have redundant variables in your model.

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

- Q–Q plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other.

Use:

- If Q–Q plot is quite similar you can expect the QQ plot to be more linear.We can check linearity with the help of scatter plots.
- We can use Q-Q plot on two datasets to check if they have similar type of distribution shape,common scale etc

Importance:

- We can create Q–Q plot in linear regression by which we can confirm that both the data train and test data set are from the population with the same distribution or not