Bike Sharing Assignment

Submitted

by

Vicky Bachu

Assignment-based Subjective Questions

Q1- From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans- The categorical variables in the dataset were season, mnth, holiday, yr, weekday, workingday and weathersit. These were visualized using a boxplot . From these variables we can conclude the following effect on our dependent variable:-

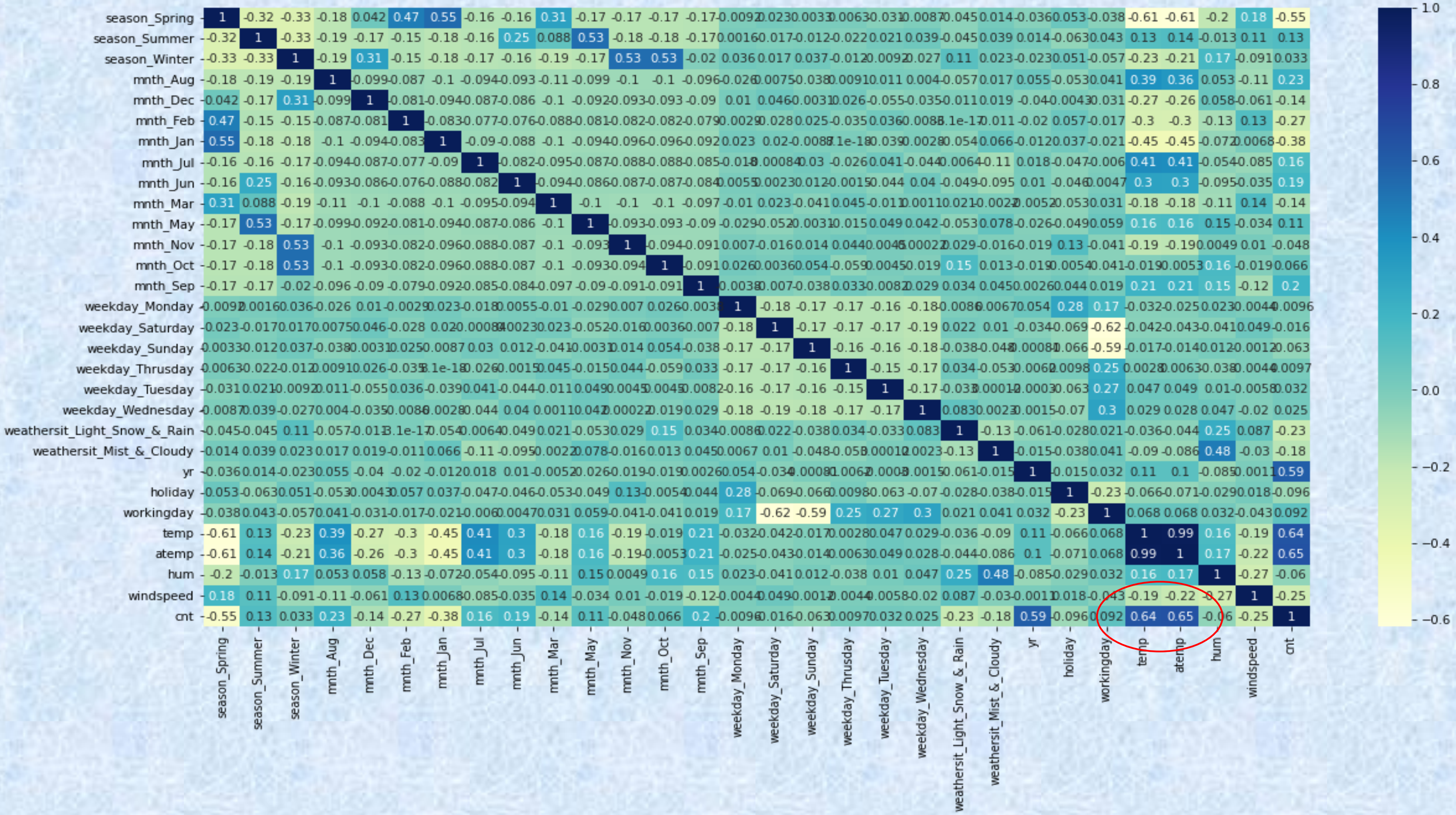
- The season box plots indicates the count of bike sharing is least for Spring
- The more bikes are rent during Fall season
- The number of bike shares increased in 2019 shown by yr value 1
- The cnt has less values for weather situation - 'Light_Snow_&_Rain' in Spring season and more in 'Clear' weather in Fall season
- The mnth box plots indicates the cnt values increases in Sep month
- The cnt values are less during holidays
- The weekday box plots indicates that more bikes are rent during Saturday

Q2- Why is it important to use drop_first=True during dummy variable creation?

Ans- New dummy column has at least one "1" within it. This is because every variable is accounted for with a True (1) indicator. The other way to identify the values is, if a row has all 0s. drop_first allows us to drop our first variable and identify it through all other columns being 0. If we don't drop the first column then our dummy variables will be correlated (redundant). This may affect some models adversely and the effect is stronger when the cardinality is smaller. For example iterative models may have trouble converging and lists of variable importance may be distorted. Another reason is, if we have all dummy variables it leads to Multicollinearity between the dummy variables. To keep this under control, we lose one column.

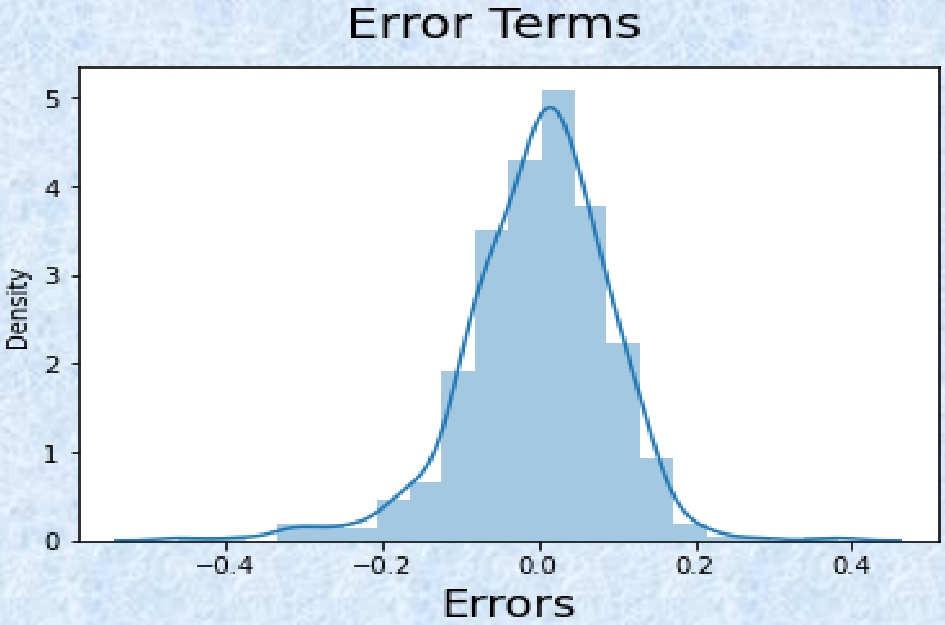
Q3- Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans- By looking at the pair-plot temp and atemp variable has the highest (0.64, 0.65) respectively correlation with target variable 'cnt'.



Q4- How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans-



Residuals distribution should follow normal distribution and centered around 0.(mean = 0). We validate this assumption about residuals by plotting a distplot of residuals and see if residuals are following normal distribution or not . The above diagram shows that the residuals are distributed about mean = 0.

Q5- Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans- The Top 3 features that are contributing significantly towards the demands of shared bikes are:

1. weathersit_Light_Snow_&_Rain - value as -0.260138
2. yr_2019 - value as 0.232999
3. temp - value as 0.302013

upGrad

General-based Subjective Questions

Q1- Explain the linear regression algorithm in detail.

Ans- Linear regression is a basic and commonly used type of predictive analysis mostly for regression analysis. It is a supervised Machine Learning algorithm. Regression is the most commonly used predictive analysis model.

Linear regression is based on the popular equation

$$x = c + my$$

where x = dependent variable, c = constant, m = regression coefficient, and y = predictor(s)/ independent variable.

In regression, we calculate the best fit line which describes the relationship between the dependent and independent variable with least error. Regression is performed when the dependent variable is of continuous data type and independent variables could be of any data type like continuous, nominal/categorical etc.

Types of Linear Regression

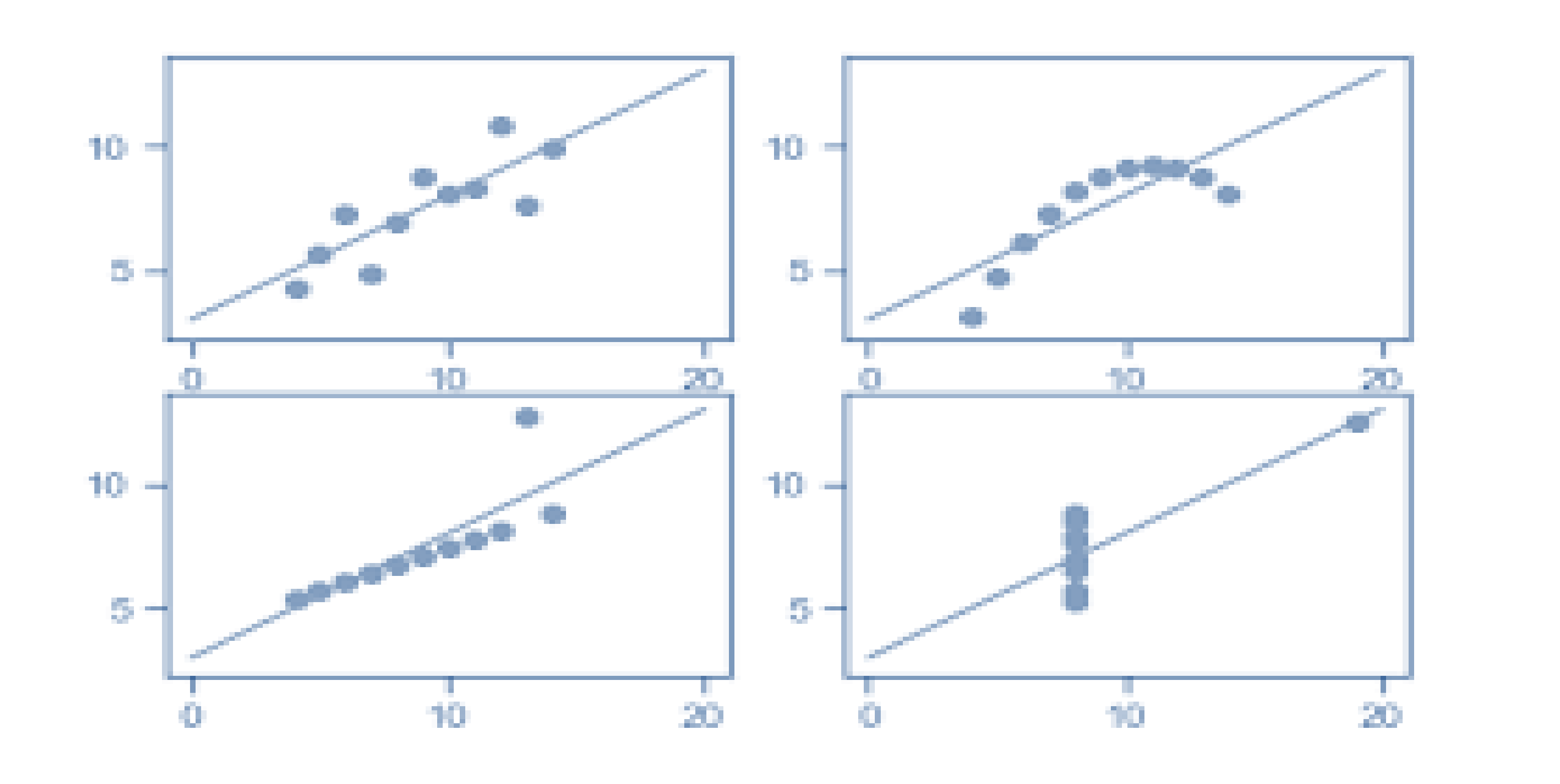
1. Simple Linear Regression : when the number of independent variables is 1.
2. Multiple Linear Regression : when the number of independent variables is more than 1.
3. The equation of the best fit regression line :

$$Y = \beta_0 + \beta_1 X$$

β1 = coefficient for X1 ….. And β0 is the intercept (constant term).

Q2- Explain the Anscombe's quartet in detail.

Ans- Anscombe's Quartet was developed by statistician Francis Anscombe. It includes four data sets that have almost identical statistical features, but they have a very different distribution and look totally different when plotted on a graph . It was developed to emphasize both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties



- The first scatter plot (top left) appears to be a simple linear relationship.
- The second graph (top right) is not distributed normally; while there is a relation between them , it’s not linear.
- In the third graph (bottom left), the distribution is linear, but should have a different regression line 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.

Q3- What is Pearson’s R?

Ans- Pearson's r is a numerical summary of the strength of the linear association between the variables. It value ranges between -1 to +1. It shows the linear relationship between two sets of data.
In simple terms, it tells us can we draw a line graph to represent the data
r = 1 means the data is perfectly linear with a positive slope
r = -1 means the data is perfectly linear with a negative slope
r = 0 means there is no linear association

Q4- What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans- Scaling is a method used to normalize or standardize the range of independent variables or features of data. It is performed during the data preprocessing stage to deal with varying values in the dataset.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, irrespective of the units of the values.

sklearn.preprocessing.MinMaxScaler

helps to implement normalization

Difference between normalized scaling and standardized scaling

Normalized scaling	Standardized scaling
Normalization is generally used when you know that the distribution of your data does not follow a Gaussian distribution.	Standardization, can be helpful in cases where the data follows a Gaussian distribution.
Scikit-Learn provides a transformer called MinMaxScaler for Normalization..	Scikit-Learn provides a transformer called StandardScaler for standardization.
Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.

Q5- You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans- The VIF gives how much the variance of the coefficient estimate is being inflated by collinearity

(VIF) =1/(1-R_1^2)

If there is perfect correlation, then VIF = infinity.
Where R-1 is the R-square value of that independent variable which we want to check how well this independent variable is explained well by other independent variables. If that independent variable can be explained perfectly by other independent variables, then it will have perfect correlation and it’s R-squared value will be equal to 1.

So, VIF = 1/(1-1) which gives VIF = 1/0 which results in “infinity”

Q6- What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans- Q-Q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. It is used to compare the shapes of distributions. A Q-Q 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.

The Q-Q plot helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

Importance:

a) It can be used with sample sizes

b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.