LINEAR REGRESSION SUBJECTIVE QUESTIONS

SHASHWATH S
19/12/2023

Assignment-based Subjective Questions

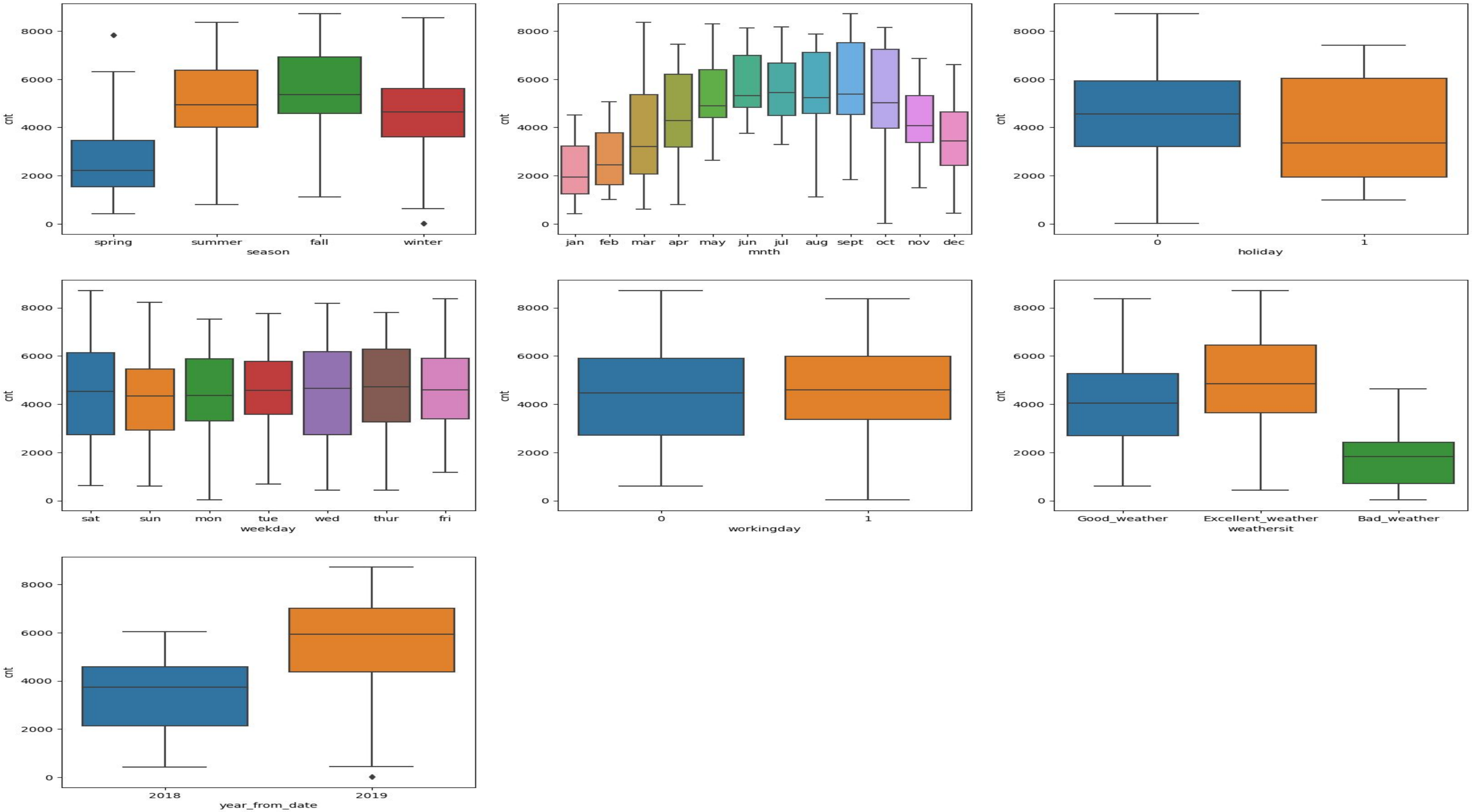
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer : According to Business Goal, the dependent variable is ‘cnt’ which indicates total number of bike rental count. The categorical Variable we have in our data are as follows: Season, Month, Holiday, Weekday, Workingday, Weather & Year.

After performing EDA on Data set we got following observation on behavior of cnt with other categorical variables.

- **cnt is more in the season of fall compared to spring,summer & winter.**
- **cnt is more in 2019 than in 2018 which means demand for rental is increasing gradually.**
- **cnt will be increased during mid of the years.**
- **cnt is more during working day than compared to holiday.**
- **cnt does not depend on which day of week.**
- **cnt depends on weather condition, the demand will be decreased as the weather getting worse.**

Please find the box plot which will indicates the behavior of cnt with categorical variables.

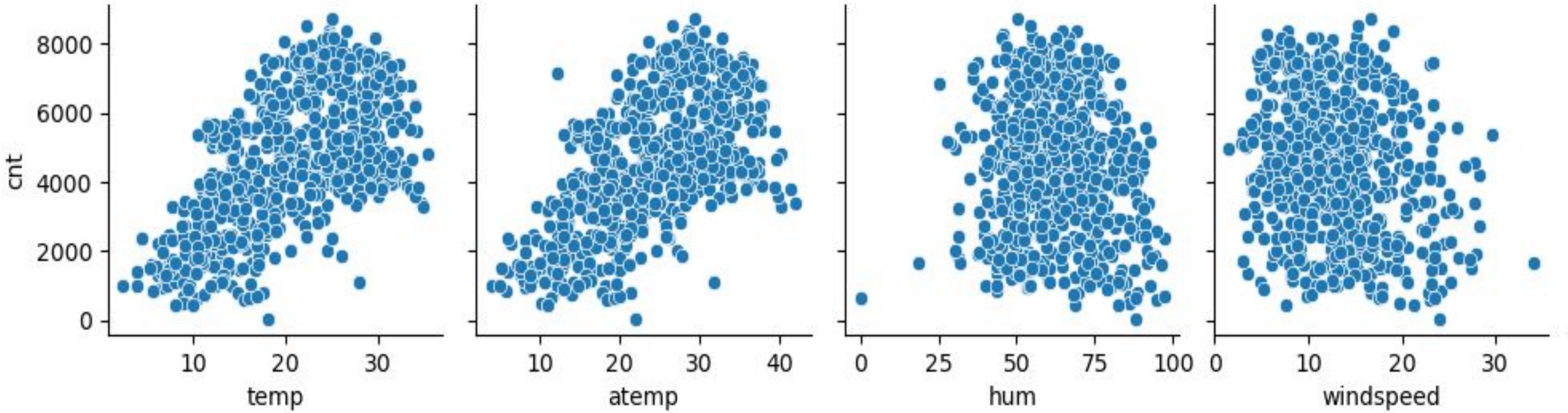

cnt
0
2000
4000
6000
8000
spring
summer
fall
winter
season
jan
feb
mar
apr
may
jun
jul
aug
sept
oct
nov
dec
mnth
0
1
holiday
sat
sun
mon
tue
wed
thur
fri
weekday
0
1
workingday
Good_weather
Excellent_weather
Bad_weather
weathersit
2018
2019
year_from_date

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

Answer: drop_first is used to avoid redundant of fields created by get_dummies function. For example in our data set we have a categorical variable called season which can take any one values from list summer,winter,spring & fall. If the season is not winter or spring or fall then the season is definitely be summer. We don't need a extra column to specify summer season. In general if there are ‘n’ categorical variable we just need ‘n-1’ columns to indicate a particular categorical variable. In order to avoid unnecessary column creation we use drop_first=True which will not create any columns for first value in a categorical variable.

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

Answer: Temp variable has highest correlation with cnt variable.



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

Answer: Once we the final model is done, we need to look into these 4 variables.

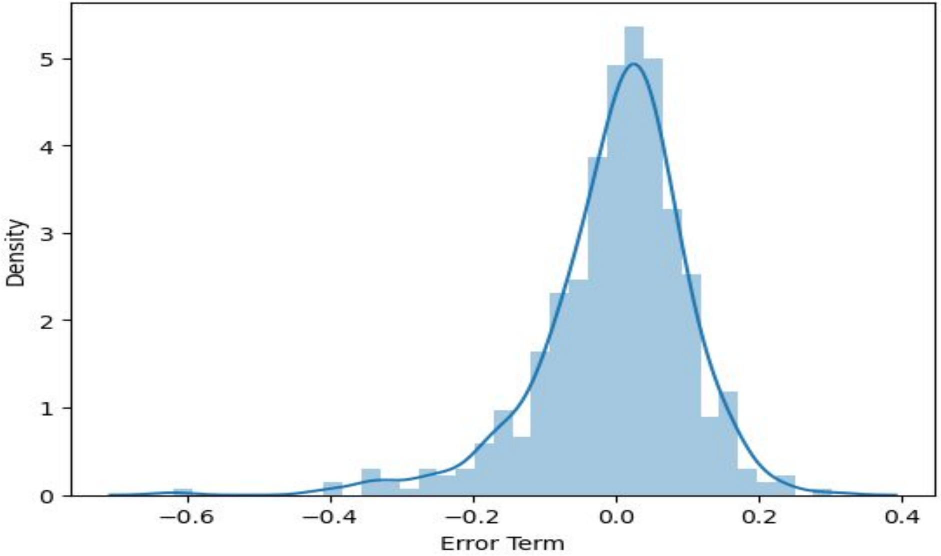
R-Squared : Which indicates the percentage of target variable explained by other variables. The value approaching 1 is considered as a good value. In our model we got a R-Squared value as 0.791 which is a good number.

F-Statistic : Which indicates the reliability to predict the future. Any number greater than 1 can be considered as goog. In Our model we got a F-Statistic as 210 which is a good number.

Porb (F-Statistic) : Which indicates the Overall Ratio of variance, Any number approaching 0 can be considered as good number, In our model we got this number as 1.38e-163 which is almost 0 which is a good number.

VIF: Indicates Variance of Independent variables with target variable. VIF of Independent variable should be < 5. From our final model all the independent variable used has a VIF < 5.

After analysing these 4 things we moved to residual analysis. We plotted a distplot for residuals calculated. We got a following graph.



The Error terms are pointed towards 0, which indicates the difference between actual value & predicted value is almost 0, so the model is capable of predicting value of target variable

After Residual analysis, we evaluated the model on test set & analysed the R2 Score. We got R2 Score on Test Set as 0.76 & The R2 Score on Training Set is 0.79. There no Significant difference exists between R2 Score of Training & Test Set. So we concluded that the model can able to predict value of target variable on test set as well.

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

- Temp (as the temp increases by 1 unit, the cnt seems to be increase by 0.55 units).
- Year (as the year increases, cnt seems to increase by 0.23 units).
- Winter (During winter season the count seems to increase by 0.11 units)

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer: Linear regression is a type of machine learning algorithm. Linear regression comes under supervised machine learning algorithm which means we have some past data available for analysis & building the model. In linear regression first we analyse the data and find out is there any linear relationship exists between our target variable and other independent variables. If the target variable has linear relation with only one variable then we perform simple linear regression, if it has linear relation with more than 1 variable then we perform multiple linear regression. The equation for finding relationship of target variable with other variable can be denoted as follows.

Y = B0 + (B1 * X1) + (B2 * X2) + (B3 * X3).......(Bn * Xn)

Where B0 -> Indicates the intercept

B1 to Bn -> Indicates the slope of different dependent variables

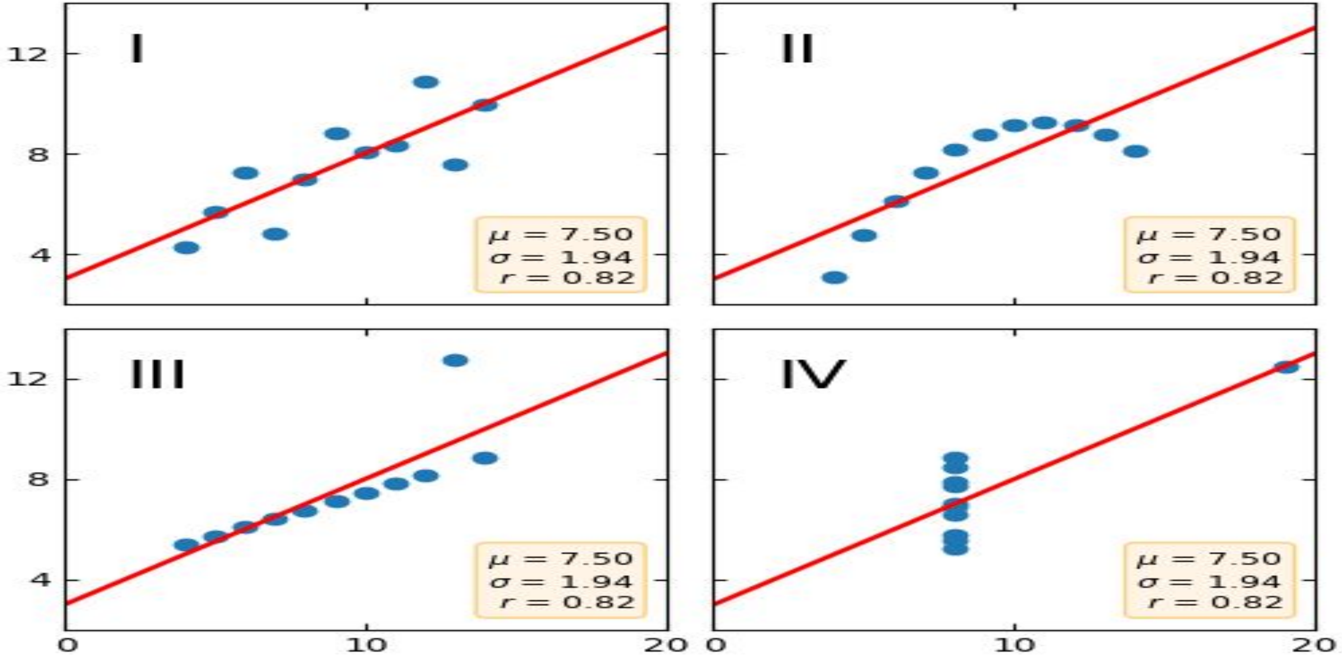
Y -> Target variable

X1 to Xn -> different dependent Variables

2. Explain the Anscombe’s quartet in detail.

Answer: Anscombe’s quartet means the data sets can have same mean, same standard deviation & regression line. But, they are qualitatively different. It is mainly used to visualize data in graphs rather than relying on statistics of data.

Following image indicates why it is important to visualize data rather than considering only statistics.



Even though the mean, standard deviation are same, the data points are scattered around regression line.

3. What is Pearson’s R?

Answer: Pearson’s R commonly called as Pearson Coefficient is a common way of finding linear relationship between variables. This coefficient will result a number between -1 to 1. Which can be indicated as follows.

From 0 to 1 : Indicates Positive correlation means when one variable changes other also changes in same direction.

0 : Indicates there is no correlation between variables.

From -1 to 0 : Indicates negative correlation means when one variable changes other changes in opposite direction.

Pearson’s R can be calculated using below formula:

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Where X & Y are the variables used to find correlation & n is the size of sample / size of records used.

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

Answer: Scaling means making all the values for a variable in a data set to same range. For example if a variable has random values ranging from 0 to 1000, scaling refers to change all these values, so that all the values range between 0 to 1.

Scaling is required because when we use any data for modeling, there will be no fixed range with the numeric variable, when this variable is used by modeling most of the values are missed by algorithm.

The two commonly used scaling methods are normalized scaling & standardised scaling.

In the normalized scaling all the values of variable in data set is scaled from minimum value of 0 to maximum value of 1. In this case all the values will range between 0 to 1.

Normalized Scaling can be calculated using this formula:

$$z = \frac{x - min(x)}{max(x) - min(x)}$$

In the case of standardized scaling, the values of variable is not restricted to 0 or 1 instead the mean of data set will be 0 and standard deviation will be 1. Standardized Scaling can be calculated using this formula:

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

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

Answer: If VIF is infinity which means there is a perfect correlation between two independent variables. When there is perfect correlation the value of R2 will be 1. From the concept of VIF, 1/1 - R2, since the R2 is 1 for perfect correlation the value of VIF becomes infinity. We may need to delete one of variable in order to avoid Multicollinearity.

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

Answer: Q-Q plot which stands for Quantile - Quantile Plot used to plot the quantile of distribution. This Plot can be used to find the distribution like normal, uniform or exponential.

In linear regression after modeling we can plot the Q-Q plot to find the quantile distribution to verify both test set & train set are having a same kind of distributions.