Contents

Subjective Questions

Assignment-based Subjective Questions

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

Dependent variable or Target variable: cnt

1. More for the Year 2019 compared to Year 2018. *{Based on box plot and bar plot}*
2. Most in clear/ less cloudy weather and least in light snow weather. *{Based on box plot and bar plot}*
3. With respect to seasons:
    - Much in the fall and Least in the Spring
    - Increases from spring to summer to fall and decreases from fall to winter. *{Based on box plot and bar plot}*
4. With respect to months:
    - The month September has highest, and Jan has lowest.
    - Increases from Jan to Jun, there is a dip in July then increase till September (peaked) decreases till the year end December. *{Based on box plot and bar plot}*
5. With respect to weekdays
    - Saturday, Wednesday, and Thursday bit more than other days however not much significant {Based on box plot – median}
    - There is an upward trend from Sun to Saturday (approximately) based on the bar plot visualization.

6. Average and Median of non-holidays rental is more than holidays *{Based on box plot and bar plot}*
7. Working day and non-working day shows almost same *{Based on box plot and bar plot}*
8. All variables exhibit upward trend from 2018 to 2019

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

A dummy variable is a variable created to assign numerical value to levels of categorical variables. Each dummy variable represents one category of the explanatory variable and is coded with 1 if the case falls in that category and with 0 if not.

Take our example Season, which has four categorical levels such as 1, 2, 3, 4 and as per the dictionary provided these are mapped {1:'spring', 2:'summer', 3:'fall', 4:'winter'}

If a column has the value spring, we can encode the value as 1 for spring 0 for summer 0 for fall and 0 for winter. This can be done in python use get_dummies method from pandas library.

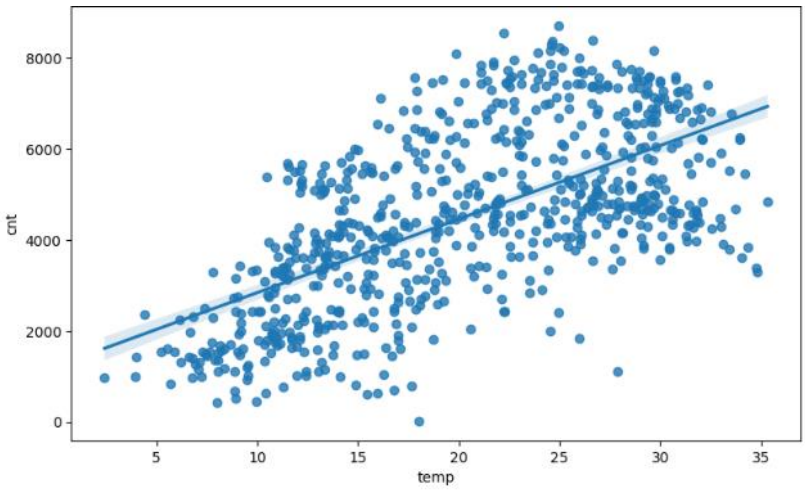
As we know if there are n variables and n-1 variables are zero then obviously the value belongs to nth variable. As per our example if the value is 0 for summer fall and winter, we can easily predict that it is spring. This leads to dependency between the columns. i.e Multicollinearity.

To avoid the above scenario drop_first=True can be passed as one of the arguments in get_dummies method.

By doing this we drop the first categorical level hence we are reducing the number of columns and also the redundant information (multi collinearity.

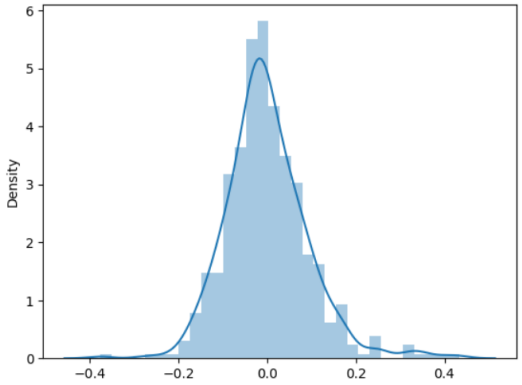
Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

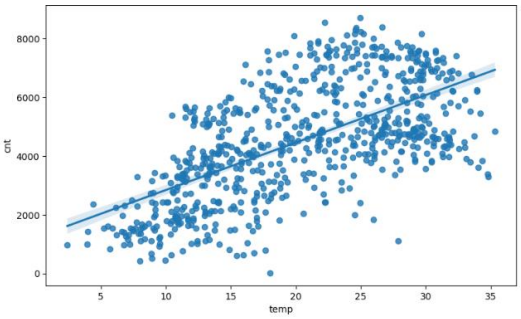
Temp indicator has the highest correlation. 0.63

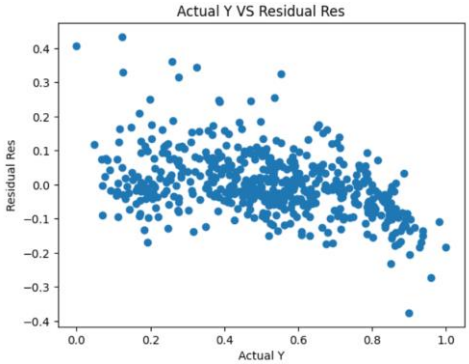


Note : Other indicators such atemp also has high correlation with cnt, but it has 99% i.e 0.99 correlation with temp, hence atemp is removed. Also registered, casual also have the correlation with cnt but those indicators were dropped in the model in the beginning itself (the 'cnt' variable indicates the total number of bike rentals, including both casual and registered)

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







- Target variable 'cnt' and prime impacting factor 'temp' are linear from the pair plot.
- Error terms are normally distributed (not X, Y), mean is zero (-4.4321844669349387e-16)
- Error terms are independent of each other i.e no pattern
- Error terms have constant variance (homoscedasticity)
- The indicators i.e independent variables are not correlated with each other, no multicollinearity among the independent variables. This is verified by checking the final VIF. All variables are less than 5
- Residual plots are done by keeping residues in Y-axis and target variable in X-Axis

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

Final Model

y = **(0.0902)** x *const* + **(0.0566)** x *workingday* + **(0.4914)** x *temp* + **(-0.0650)** x *season_spring* + **(0.0527)** x *season_summer* + **(0.0970)** x *season_winter* + **(0.0916)** x *mnth_Sep* + **(0.0645)** x *weekday_Sat* + **(-0.3041)** x *weathersit_LightSnow* + **(-0.0786)** x *weathersit_Mist* + **(0.2334)** x *yr_2019*

From the above model

Most positive impact i.e one unit increase in the indicator will increase the y value with their corresponding coefficient times.

- Temperature (0.4914)
- Year 2019 (0.2334)
- Winter Season (0.0970)

There are some negative impacting indicators also there such as

- LightSnow weather (-0.3041)

Negative impact i.e one unit increase in the indicator will decrease the y value their corresponding coefficient times.

General Subjective Questions

Explain the linear regression algorithm in detail