Solution Approach

1. Read the data

2. Identified the Categorical Data column & Numerical Data column.

3. Visualized the categorical data & numerical data columns

4. In Categorical data columns identified the ordinal data column and nominal data column

5. Converted the ordinal data column into the dummy variable.

6. Delete the respective columns upon which we have created the dummy variable.

7. Split the data frame into two parts Train set 70% & Test set 30%.

8. Scaled all the columns other then dummy variable column by either MinMax Scaler or by standardization.

9. Divided the train data into dependent(y) & independent variable(X) variable.

10. To build the model need to fit a regression line through the training data using `statsmodels`. Added a constant in that `statsmodels`, to explicitly fit a constant using `sm.add_constant(X)` because if we don't perform this step, `statsmodels` fits a regression line passing through the origin, by default.

11. Now need to verify the variable using p-value & VIF value of the model for **Feature Elemination**.

12. 1st need to drop the variable which have higher p-value and higher VIF then need to drop the variable which have higher p-value & lower VIF then lower p-value & higher VIF. We need to repeate the step until we get the p-value <0.05 and VIF less then 5.

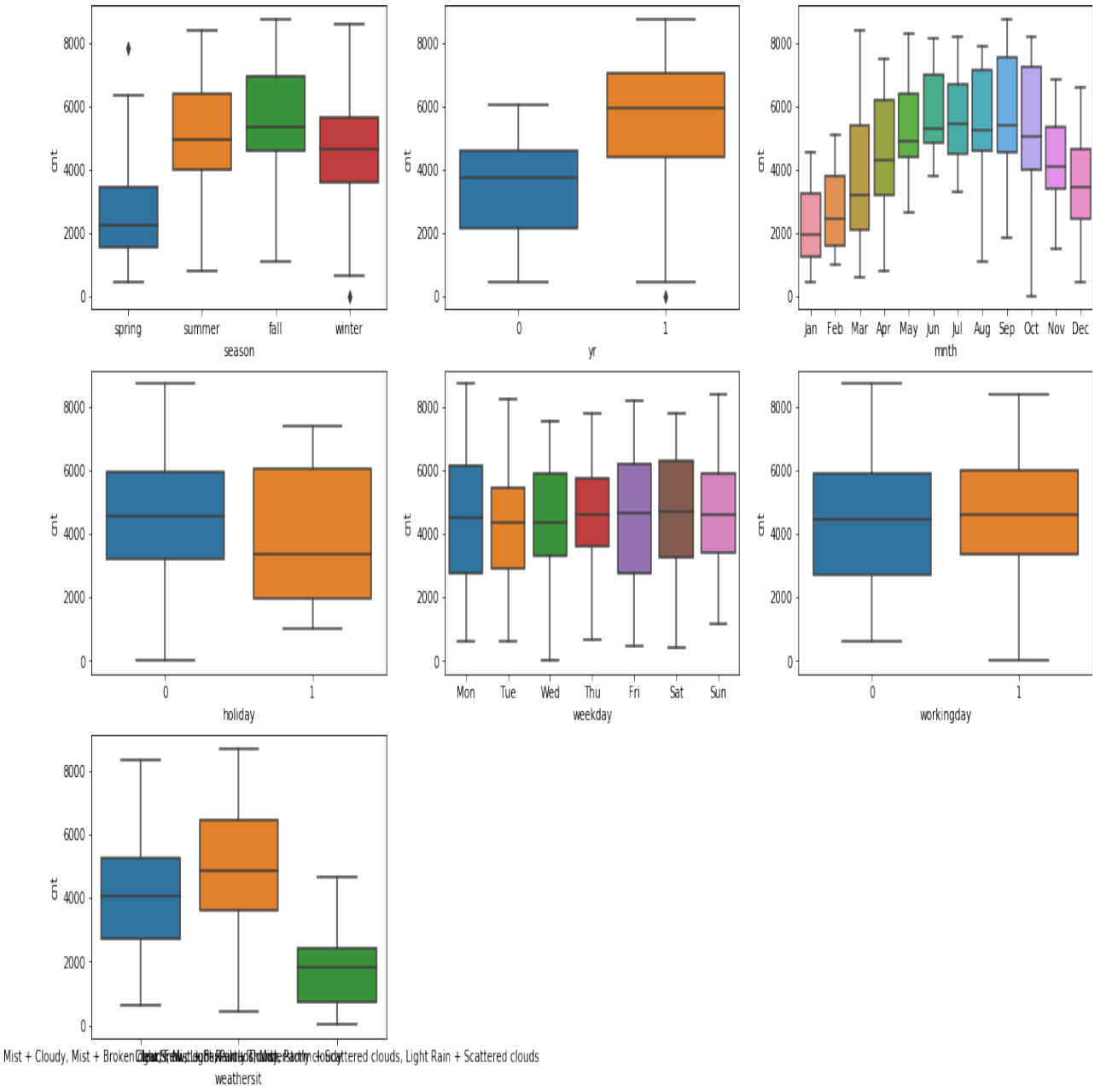
13. Plotted the residual analysis to validate the model.

14. Checked the linearity between actual data & predicted data.

15. After apply the same step as the train data link scaling, data splitting & dropping the non significant columns.

16. Evaluate the model using r2score.

We can infer from the below boxplot of categorical variable are,



1. season- From the above box-plot that the ‘fall’ has higher median means higher count of total rental bikes compared to other seasons.
2. Yr- We can infer that the median of 2019 is higher so there is significant rise of count of total rental bikes
3. mnth- ‘Jul’ mas maximum median so ‘Jul’ has height count of total rental bikes.
4. ‘holiday’- The count of total rental bikes is higher in non-holidays
5. ‘weekday’- All median values are almost same but data spread on Monday & Friday are more.
6. ‘workingday’- The median values are almost s the data spread of non-working is more.
7. ‘weathersit’- Count of total rental bikes are more in the Clear, Few clouds, Partly cloudy, Partly cloudy weather condition.