iiit-b

UpGrad

BIKE SHARING ASSIGNMENT

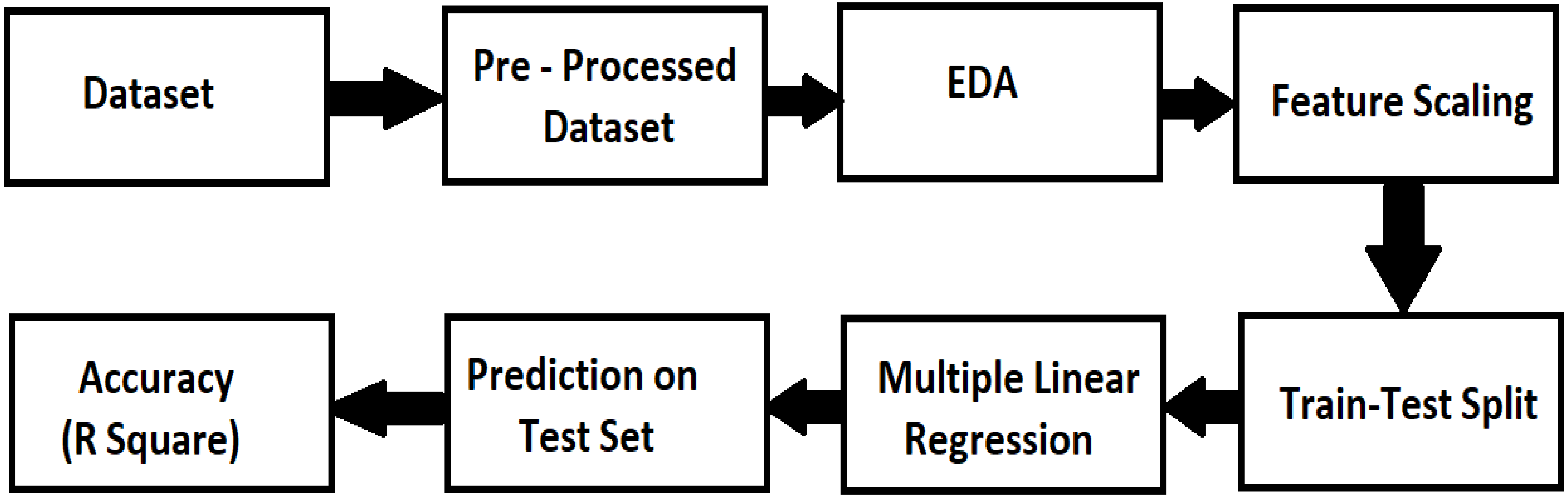
SUBMISSION

Name: Chandramouli Das

Overview

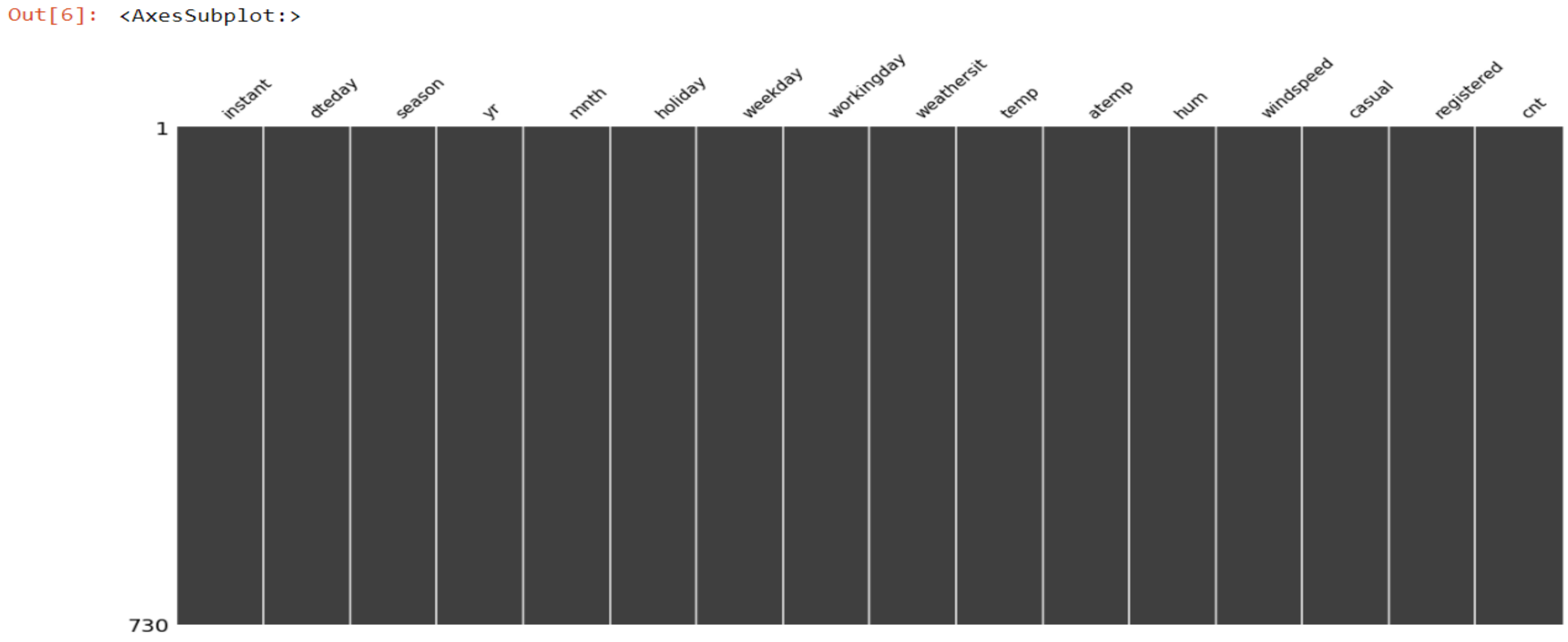
- **Problem Statement -** A bike-sharing system is a service in which bikes are made available for shared use to individuals on a short term basis for a price or free. Many bike share systems allow people to borrow a bike from a "dock" which is usually computer-controlled wherein the user enters the payment information, and the system unlocks it. This bike can then be returned to another dock belonging to the same system.
- **Business Goal –** I need to make model the demand for shared bikes with the available independent variables. It will be used by the management to understand how exactly the demands vary with different features.
- **Model Evaluation –** After successfully making the model I need to calculate the accuracy that indicates whether the prediction of my model is good or not.

Problem Solving Methodology



Data Cleaning

The given dataset was already clean. There was not a single missing value.



```
Out[6]: <AxesSubplot:>
```

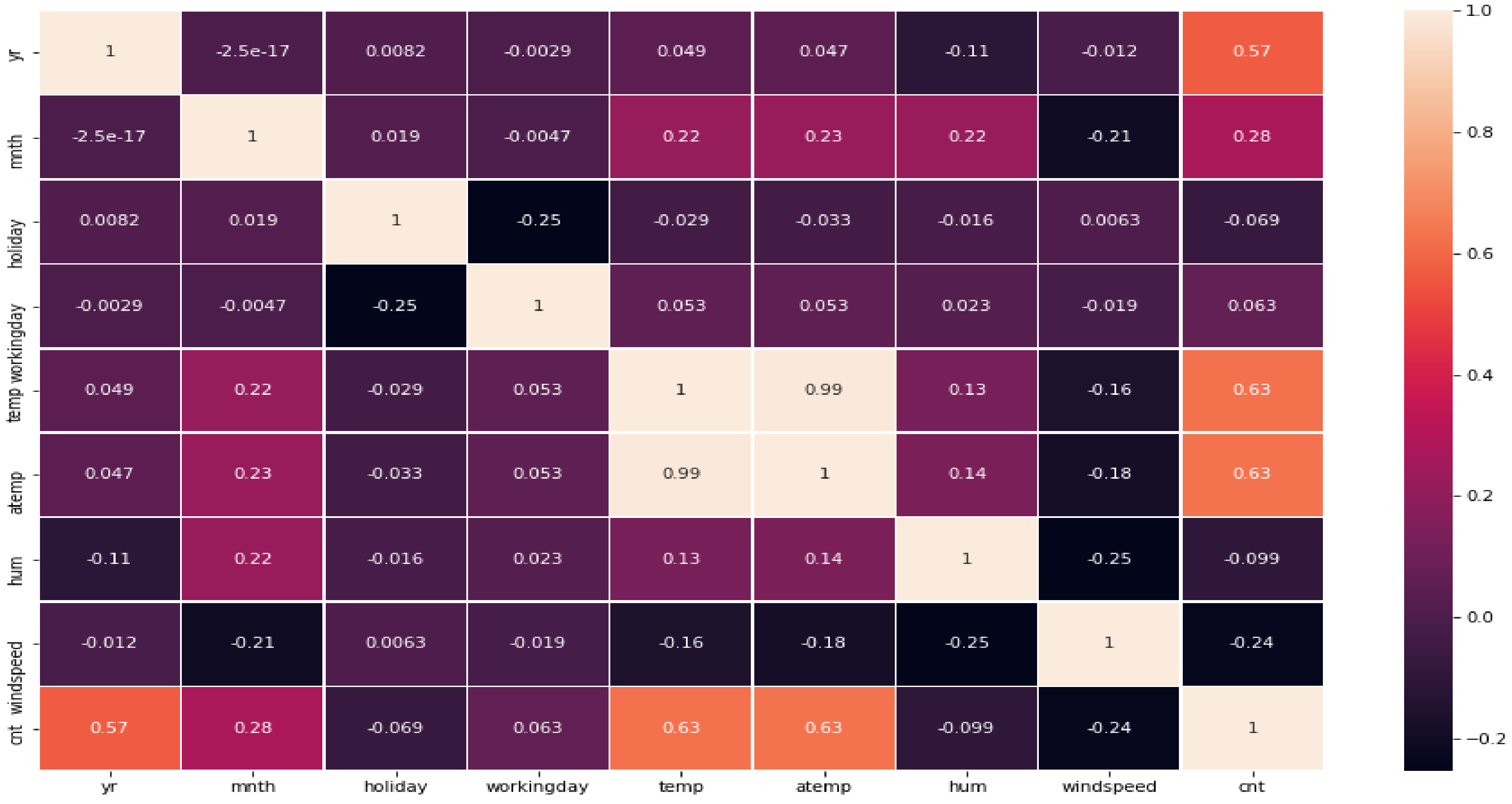
There is no Missing Value present in this dataset

Data Cleaning

- Dropped unnecessary columns "instant" and "dteday" columns
- Dropped 'casual' and 'regesterted’ column because it is related to our target variable.
- I have some categorical Variable column like ‘season’, ‘yr’, ‘mnth’, ‘holiday’, ‘weekday’, ‘workingday’,’weathersit’.
- Replaced those values of the columns with suitable and understandable values.

EDA

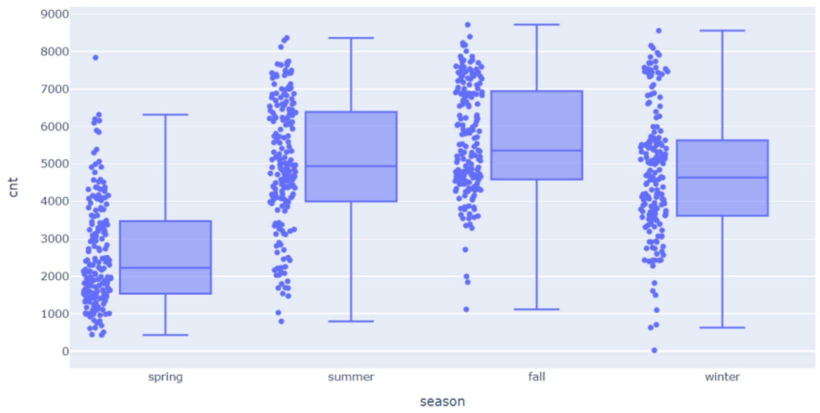
Plotted a heatmap with all the columns.

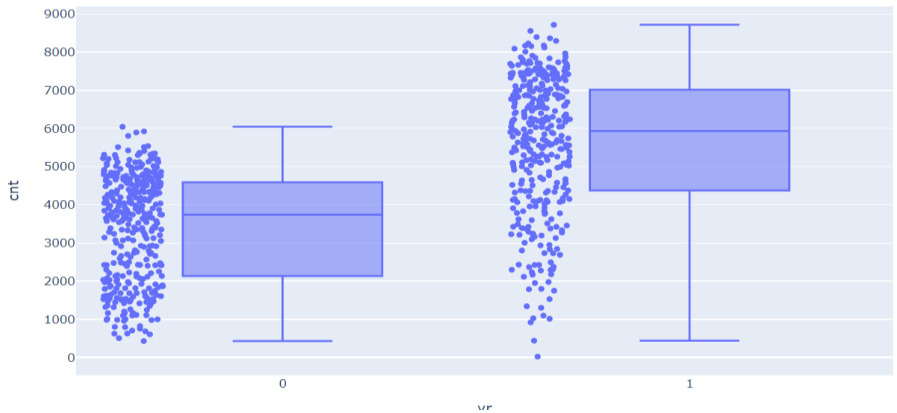


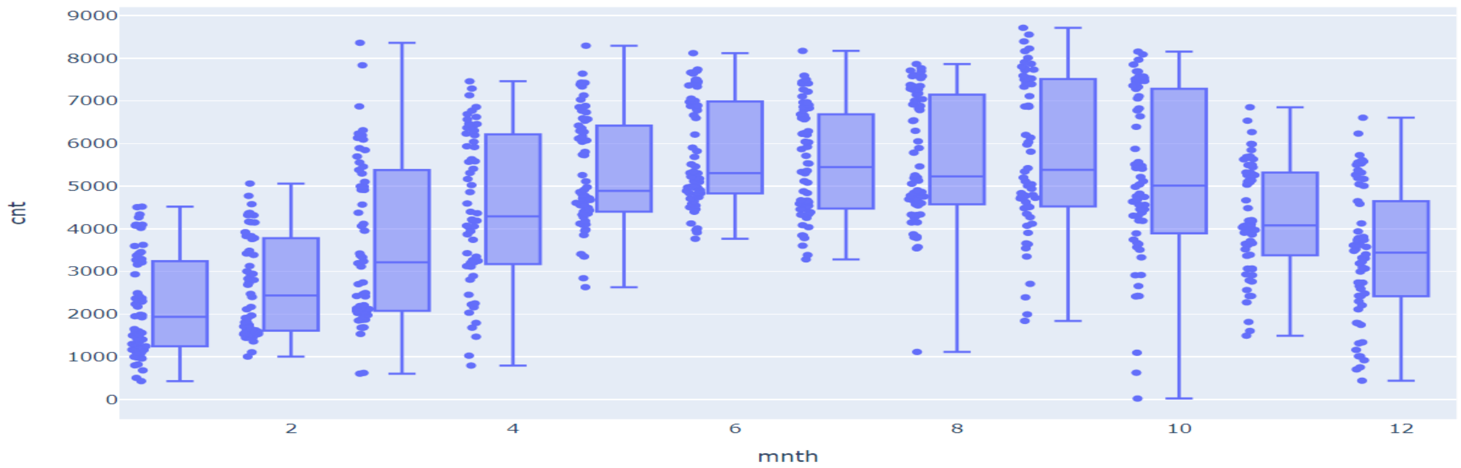
	yr	mnth	holiday	workingday	temp	atemp	hum	windspeed	cnt
yr	1	-2.5e-17	0.0082	-0.0029	0.049	0.047	-0.11	-0.012	0.57
mnth	-2.5e-17	1	0.019	-0.0047	0.22	0.23	0.22	-0.21	0.28
holiday	0.0082	0.019	1	-0.25	-0.029	-0.033	-0.016	0.0063	-0.069
workingday	-0.0029	-0.0047	-0.25	1	0.053	0.053	0.023	-0.019	0.063
temp	0.049	0.22	-0.029	0.053	1	0.99	0.13	-0.16	0.63
atemp	0.047	0.23	-0.033	0.053	0.99	1	0.14	-0.18	0.63
hum	-0.11	0.22	-0.016	0.023	0.13	0.14	1	-0.25	-0.099
windspeed	-0.012	-0.21	0.0063	-0.019	-0.16	-0.18	-0.25	1	-0.24
cnt	0.57	0.28	-0.069	0.063	0.63	0.63	-0.099	-0.24	1

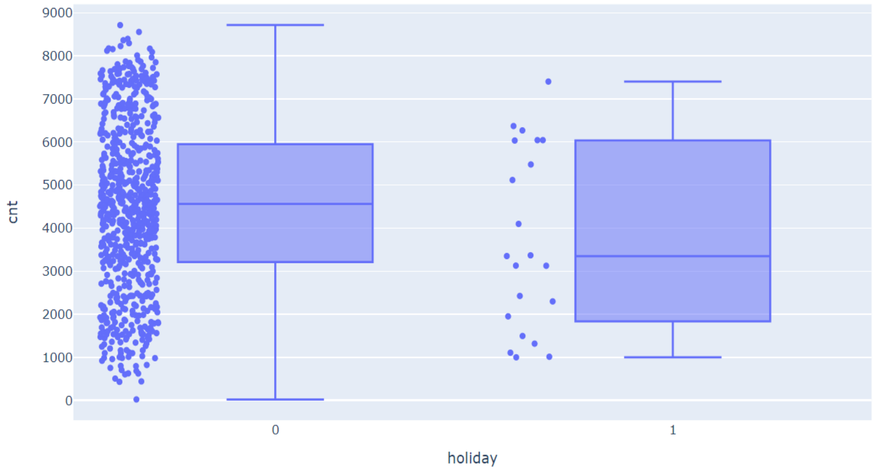
EDA

Boxplot analysis with all the categorical variables with out target variables.



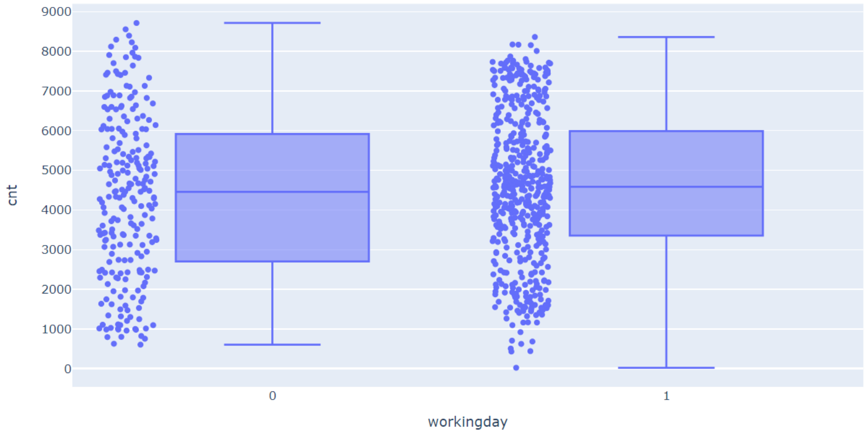


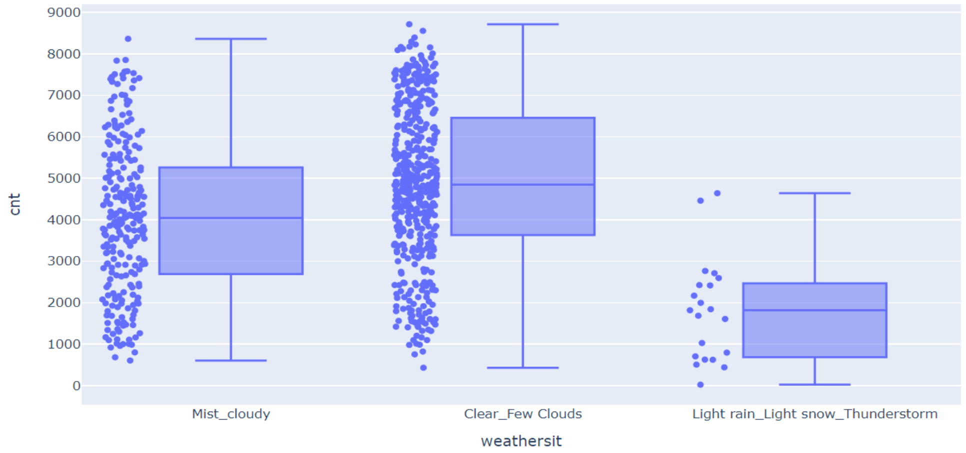


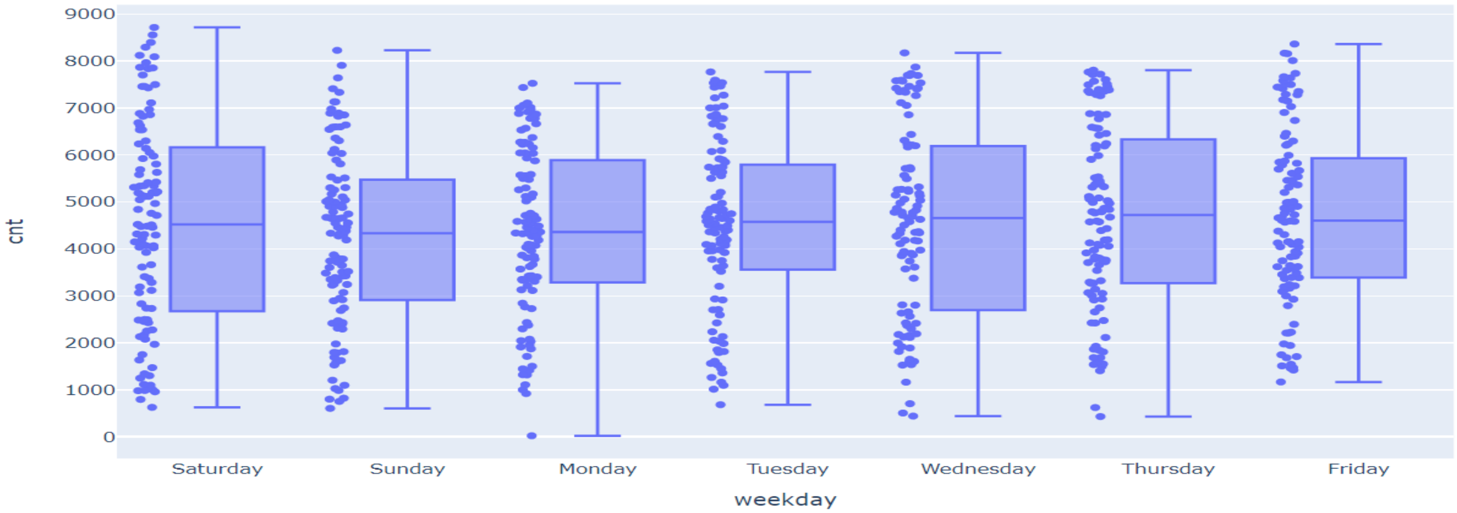


EDA

Boxplot analysis with all the categorical variables with out target variables.







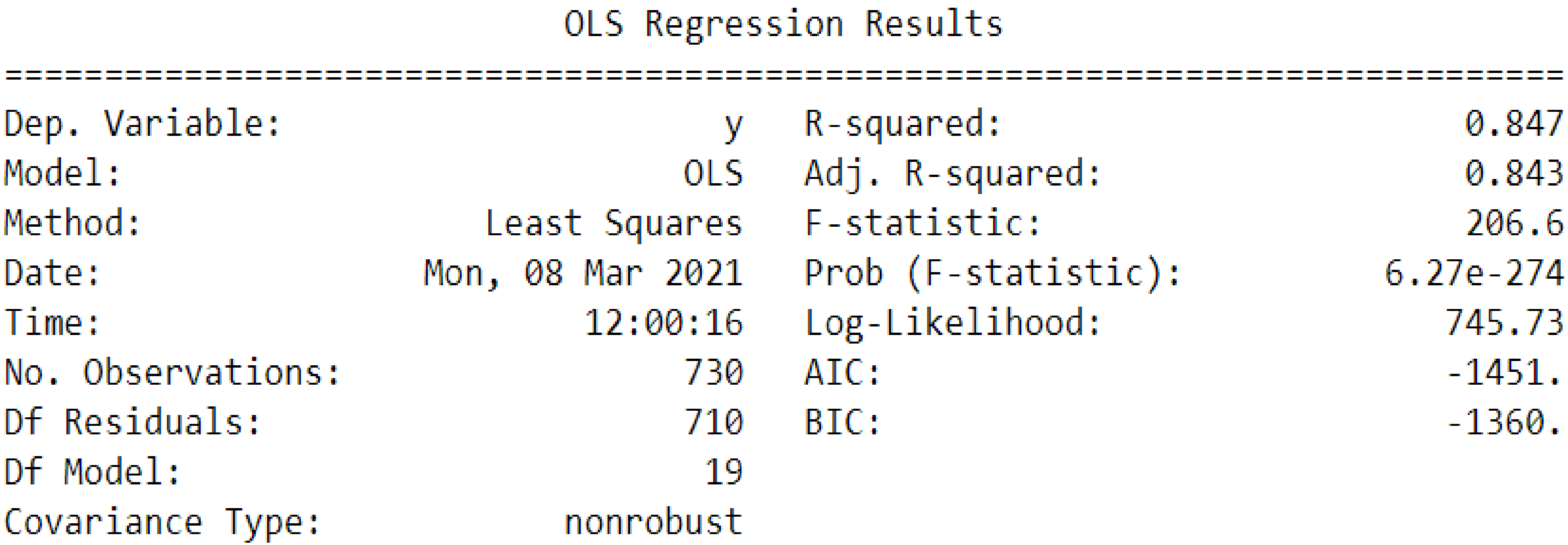
Data Analysis

- Created Dummy variable for every categorical variable
  - For the column “season” I have made four dummy columns
  - For the column “mnth” I have made twelve dummy columns
  - For the column “weekday” I have made seven dummy columns
  - For the column “weathersit” I have made three dummy columns
- Applied Feature Scaling in all the continuous variable
- I have applied MinMax Scaler for our dataset.
- Separated the target variable and stored that in a variable(y), and rest of the columns are stored in a different variable (X).

OLS Regression

- Now for making a model I have performed OLS regression by adding a single columns in the model.
- I have started with the ‘yr’ column and added the other columns one by one based on R – Square and Adj R- Square values.
- After adding all the column, I got a good accuracy. Got the R-Square value - 0.848 and Adj R-Square value - 0.842.
- Now I have discard some columns one by one based on the p-value.
- Then Got accuracy of R-Square value - 0.847 and Adj R-Square value - 0.843, which is a little improvement after dropping the columns.

OLS Regression Result



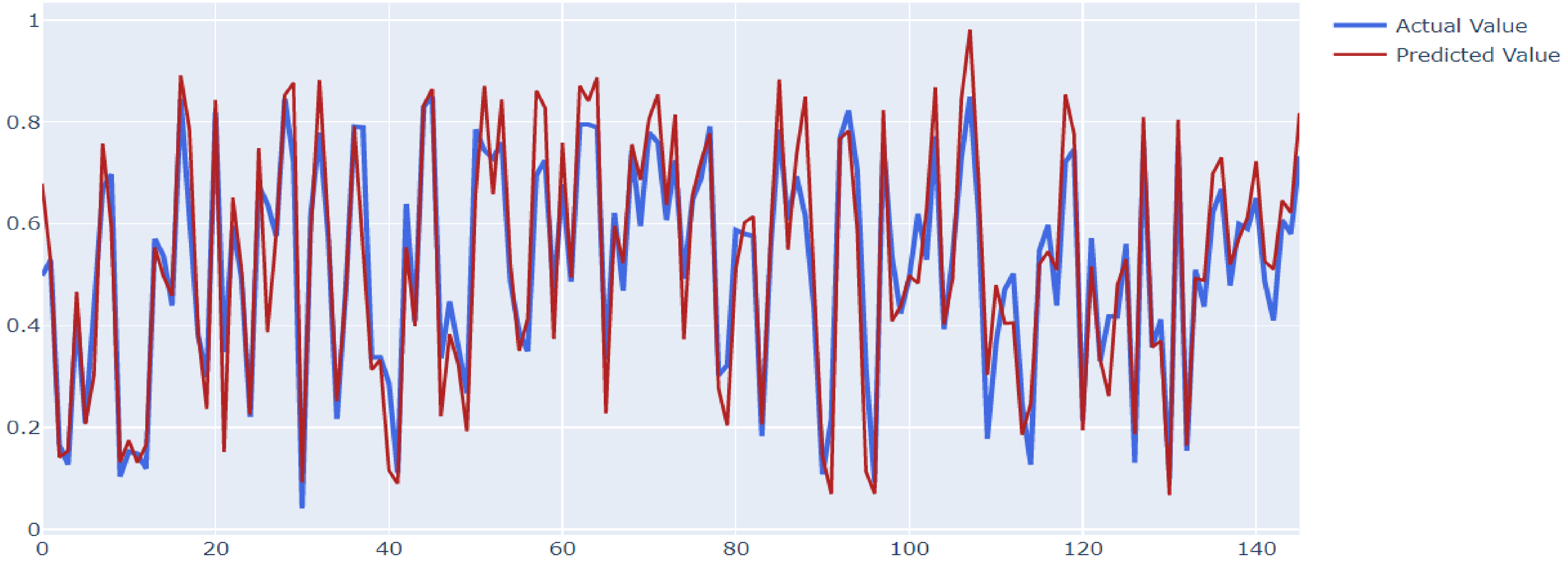
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.847
Model:                            OLS   Adj. R-squared:                  0.843
Method:                 Least Squares   F-statistic:                     206.6
Date:                Mon, 08 Mar 2021   Prob (F-statistic):          6.27e-274
Time:                        12:00:16   Log-Likelihood:                 745.73
No. Observations:                 730   AIC:                            -1451.
Df Residuals:                     710   BIC:                            -1360.
Df Model:                          19
Covariance Type:            nonrobust
```

Multiple Linear Regression

- Split the dataset into train and test set in the ratio of 80% and 20%.
- From sklearn library Linear Regression was imported.
- Fitted the training data into Linear Regression Model.
- After fitting, the model is now ready to predict new data.
- I have fitted the test data into the model by predict function.
- For every test data I got a prediction, now I need to ensure that the model is predicting correctly.
- After that I have calculated the R-square for the test dataset and accuracy of 88%.
- Then I have plotted a line graph of test data and predicted data for comparison and got a excellent line graph.

Result

The Blue line is the Actual data and the red line is the Predicted data from our model. This is clear that my model has predicted correctly for most of the cases.



Summery

- This model can predict the target variable ‘cnt’ with the accuracy of 88%.
- Some most useful columns are ‘yr’ ,’atemp’, ‘windspeed’ and ‘season’
- Some useless columns are ‘instant’ , ‘dteday’, ‘casual’, ’registered’, ‘holiday’ and some dummy variables.
- The accuracy difference between training dataset and test dataset is less than 5%, that signifies this model is perfect. (Neither Overfitted not Underfitted).
- This fact can be checked by the line graph of actual data and predicted data.