Bike Sharing Assignment

Agenda

- Business goal
- Steps for EDA
- Univariate and Bivariate Results
- Correlation Results
- Step took to get predict the data
- Residual virtual Analysis
- Final Prediction
- Final Result

Business Goal

We were required to 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. They can accordingly manipulate the business strategy to meet the demand levels and meet the customer's expectations. Further, the model will be a good way for management to understand the demand dynamics of a new market.

Steps for EDA

* First step was to load the data and check the data and get a overview.

* Once we get an overview check for the null values as there were non.

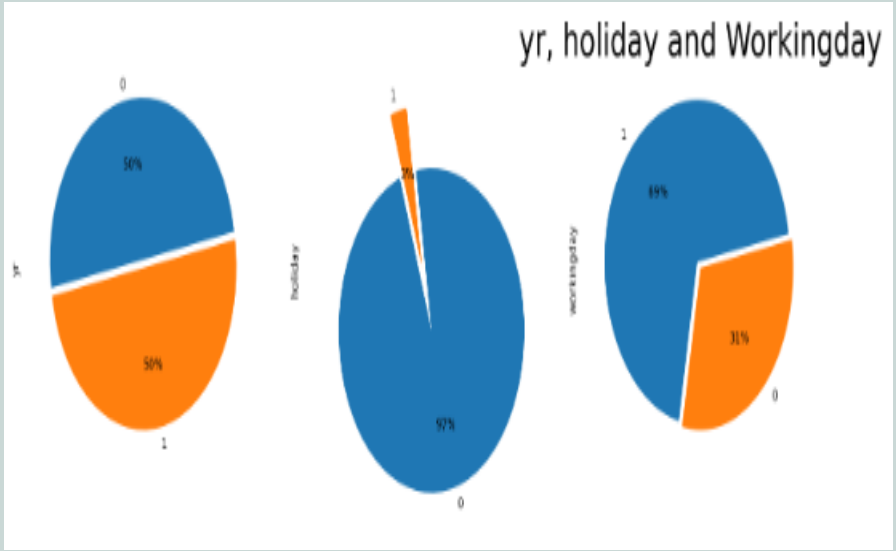
- Checked for duplicate values as well as the result was with no duplicates as well.

- We did drop some columns as we already have on reference and some were no needed for example 'instant' , 'dteday'.

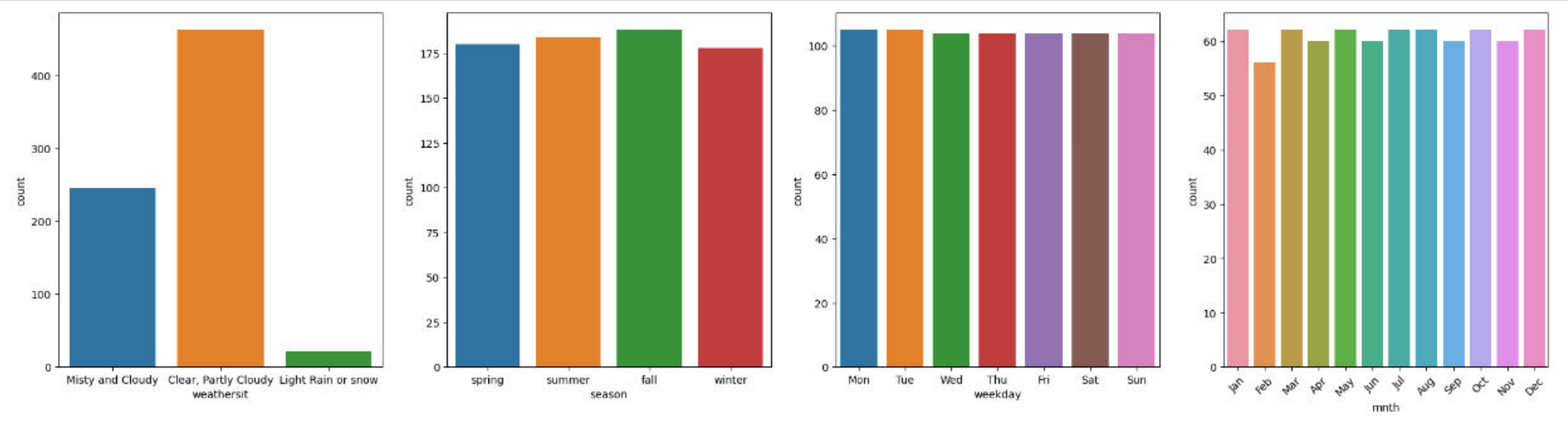
* Once this was done we have to change the numeric values to a string values for some of the columns

*Now our data was ready for analysis

Univariate Analysis Results

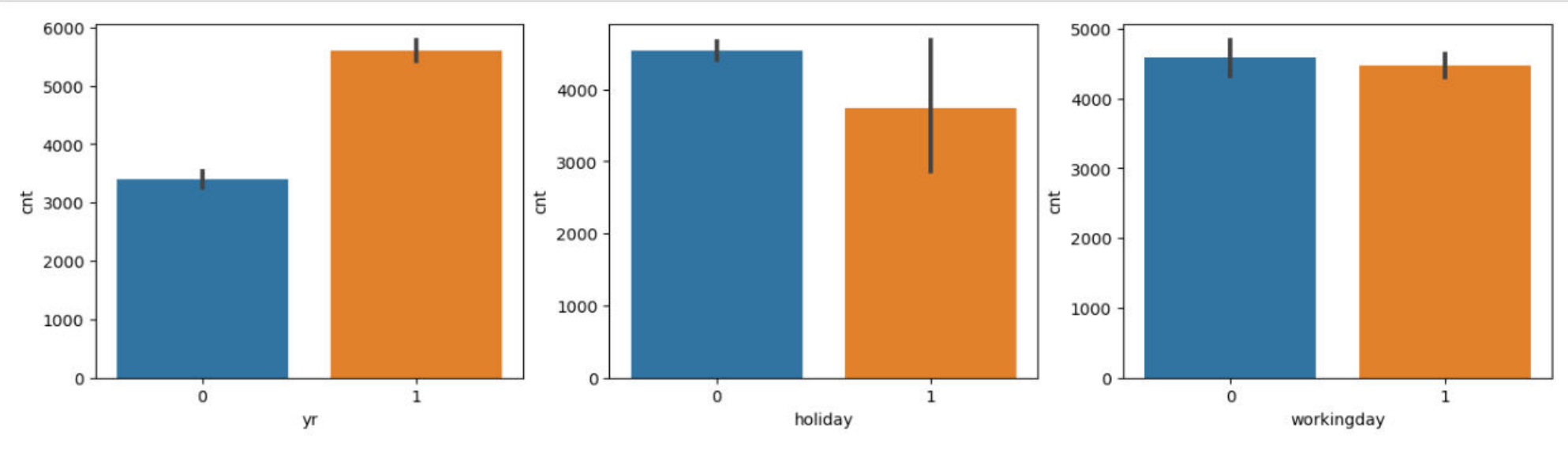


- Here in the first Pie chart we can see 'yr' is 50%-50% for both the years
- on the seconf Pie chart wecan see a huge difrent here we can say users of the bike is more on non-holidays(0)
- The same thing here as as 2nd pie chart 'workingday' is higher number as compare to non-working days.

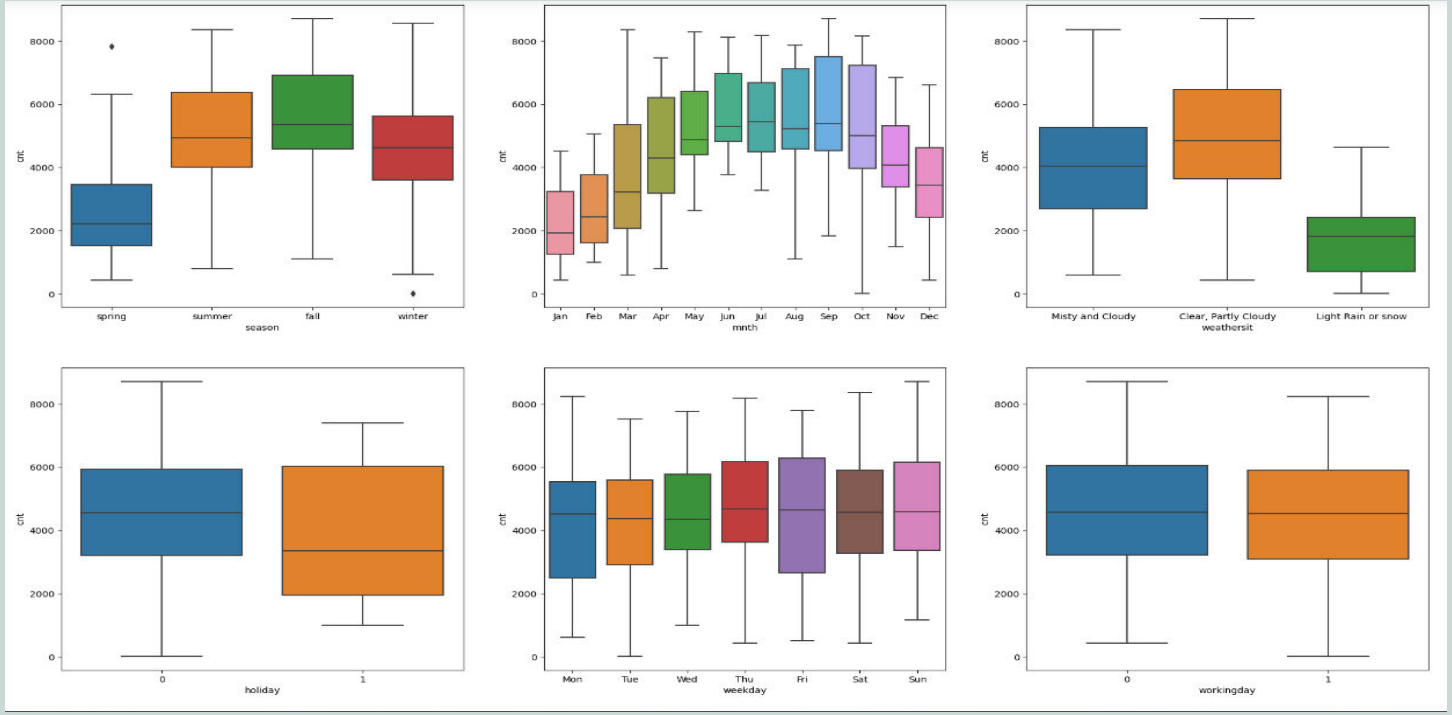


- We can say when where is clear and mist_sloudy the users is more than Light_snow_rain.
- Rest other values are almost same

Bivariate Analysis Results

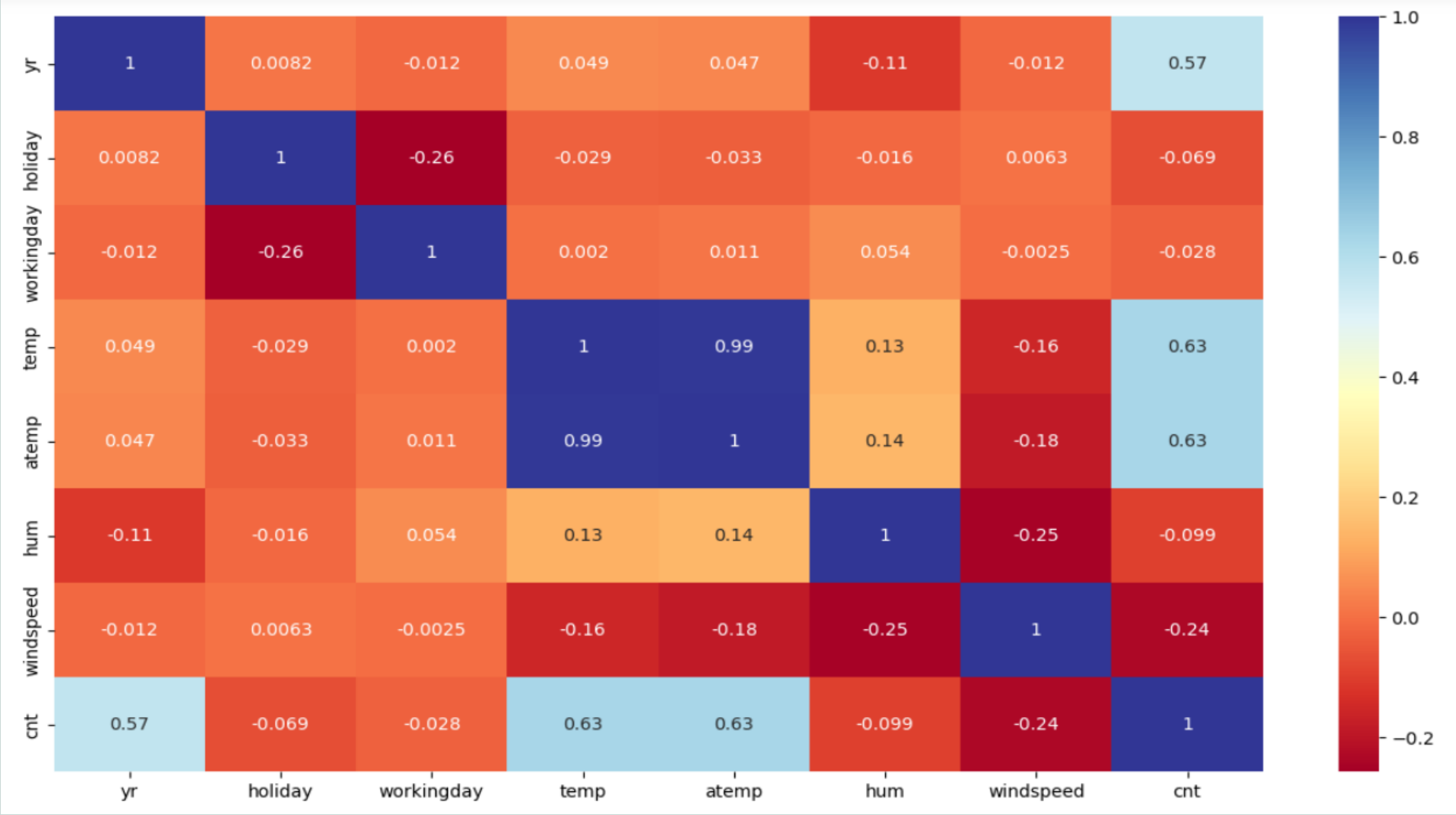


- 'Yr' we can say there is a increse of users on year
- More users on holidays(0) as compared to Non holidays(1).
- on Workingday we can say it's almote same



- here we can check the Season Spring have the low bookings, Fall have the higest bookings
- mnth we can check in few months the bookings are more
- clarely Light_snow_rain have the low bookings othe 2 have almost same bookings
- bike booking were happening when it is not a holiday which means this data is clearly biased. This indicates, holiday CANNOT be a good predictor for the dependent variable.
- weekday variable shows very close trend.
- most of the bike booking were happening in ‘workingday’.

Correlation Results



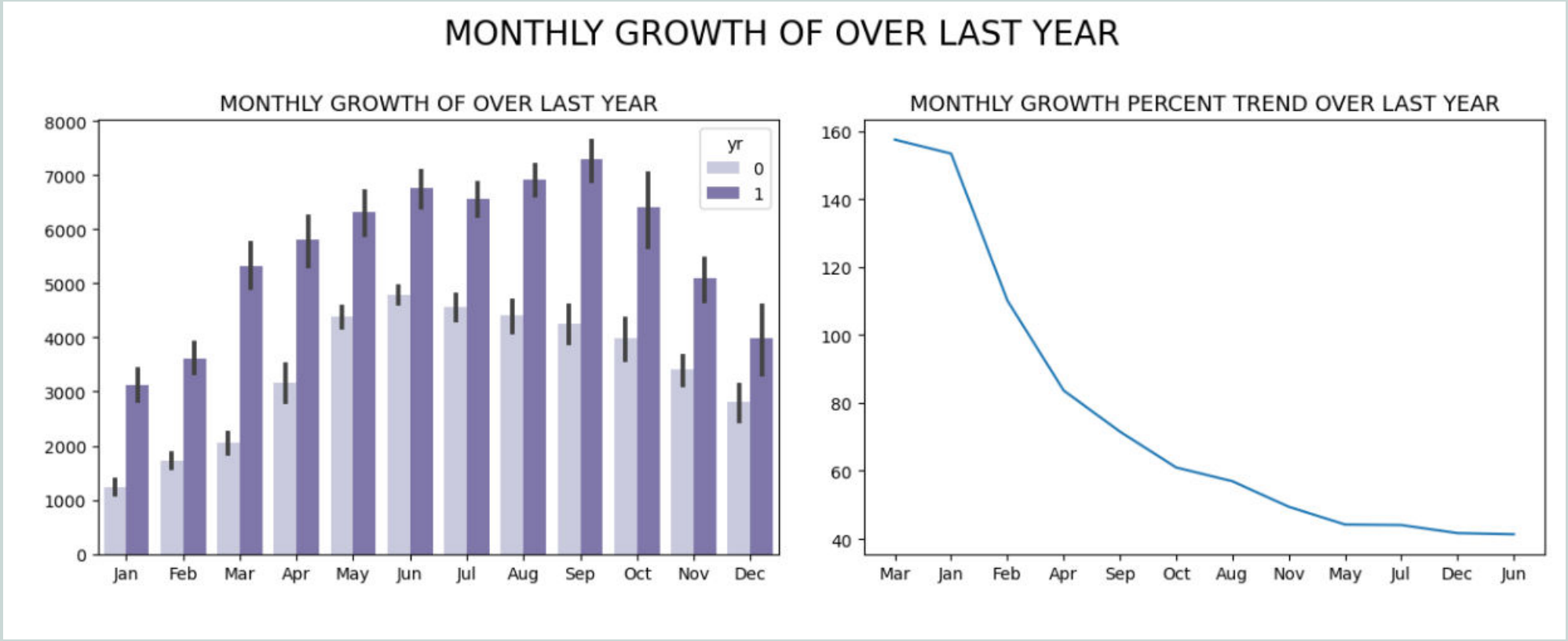
Correlation Results

- The heatmap clearly shows which all variable are multicollinear in nature, and which variable have high collinearity with the target variable.

Steps after EDA and checking Corelation.

- After EDA and Checking corelation between the Target value.
- Once it was done check the growth over the last year to get a better view so compare the with he final result
- Started with creating Dummy variables
- Once the Dummy variable was created now the data was splitted in to Train and Test data.
- Once it was done extract 15 most index intrain dataset to Build Function.
- Once the prefect model was found done Residual analysis then Predicted the data.

Growth over the last year



* We can check the growth shown in 2019 over 2018.
* Growth observer high in months of Jan, Feb and Mar.

Final ML Model

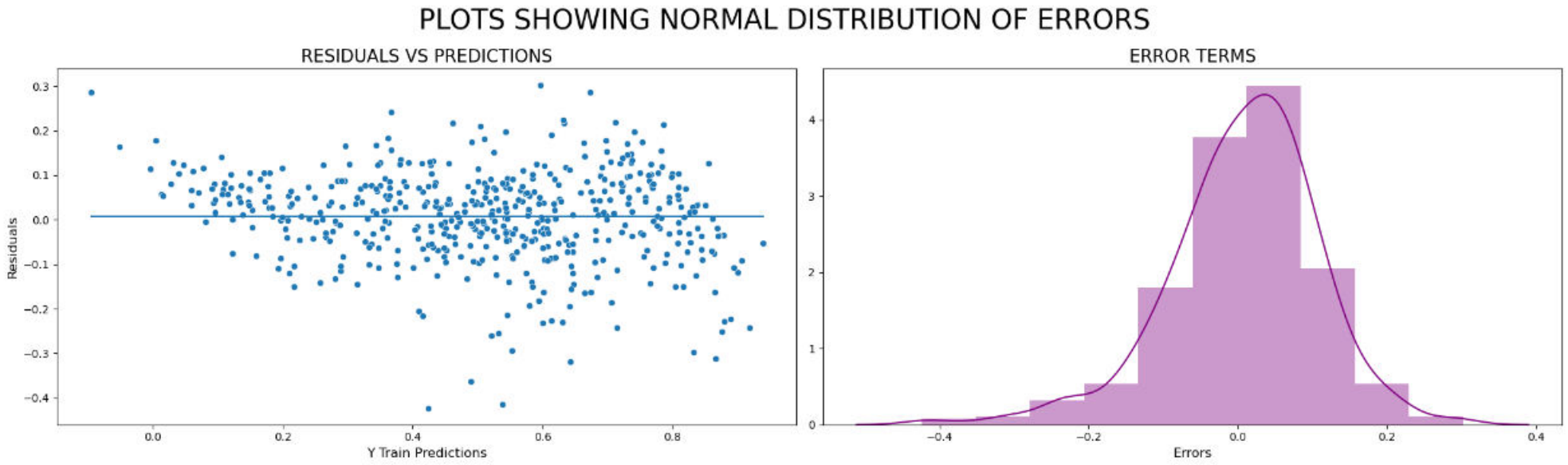
OLS Regression Results

Dep. Variable:	cnt	R-squared:	0.832
Model:	OLS	Adj. R-squared:	0.829
Method:	Least Squares	F-statistic:	248.0
Date:	Wed, 28 Jun 2023	Prob (F-statistic):	1.28e-186
Time:	23:19:05	Log-Likelihood:	494.32
No. Observations:	511	AIC:	-966.6
Df Residuals:	500	BIC:	-920.0
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2526	0.024	10.532	0.000	0.205	0.300
yr	0.2348	0.008	28.260	0.000	0.218	0.251
holiday	-0.0985	0.026	-3.741	0.000	-0.150	-0.047
temp	0.4505	0.031	14.690	0.000	0.390	0.511
windspeed	-0.1394	0.025	-5.528	0.000	-0.189	-0.090
spring	-0.1120	0.015	-7.333	0.000	-0.142	-0.082
winter	0.0462	0.012	3.714	0.000	0.022	0.071
Jul	-0.0732	0.018	-4.177	0.000	-0.108	-0.039
Sep	0.0570	0.016	3.584	0.000	0.026	0.088
Light Rain or snow	-0.2862	0.025	-11.459	0.000	-0.335	-0.237
Misty and Cloudy	-0.0801	0.009	-9.060	0.000	-0.097	-0.063

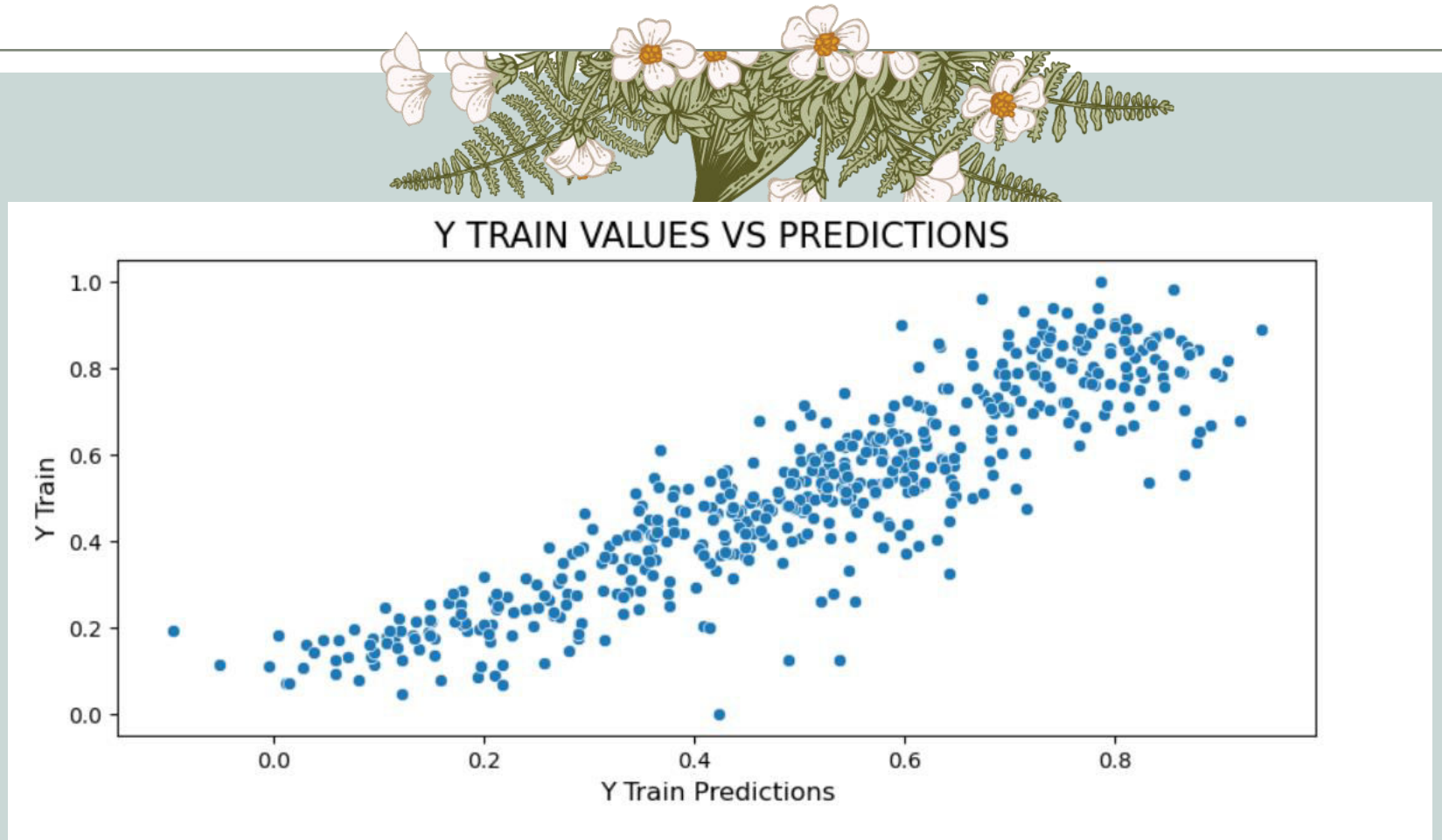
Omnibus:	57.757	Durbin-Watson:	2.001
Prob(Omnibus):	0.000	Jarque-Bera (JB):	136.435
Skew:	-0.600	Prob(JB):	2.36e-30
Kurtosis:	5.229	Cond. No.	14.0

Residual Analysis Result



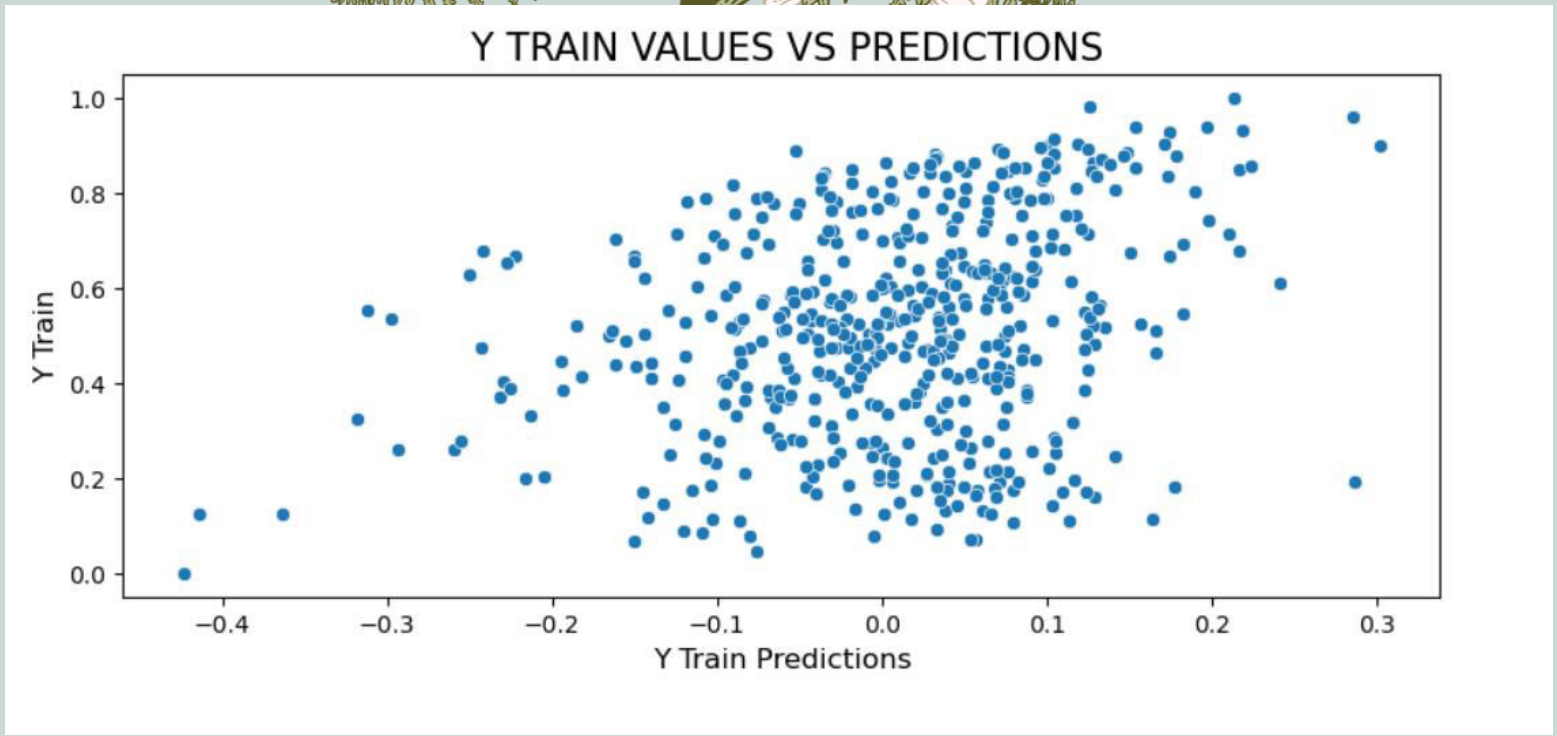
Mean of residuals are close to 0

Predicted values of the target variable on the test set.



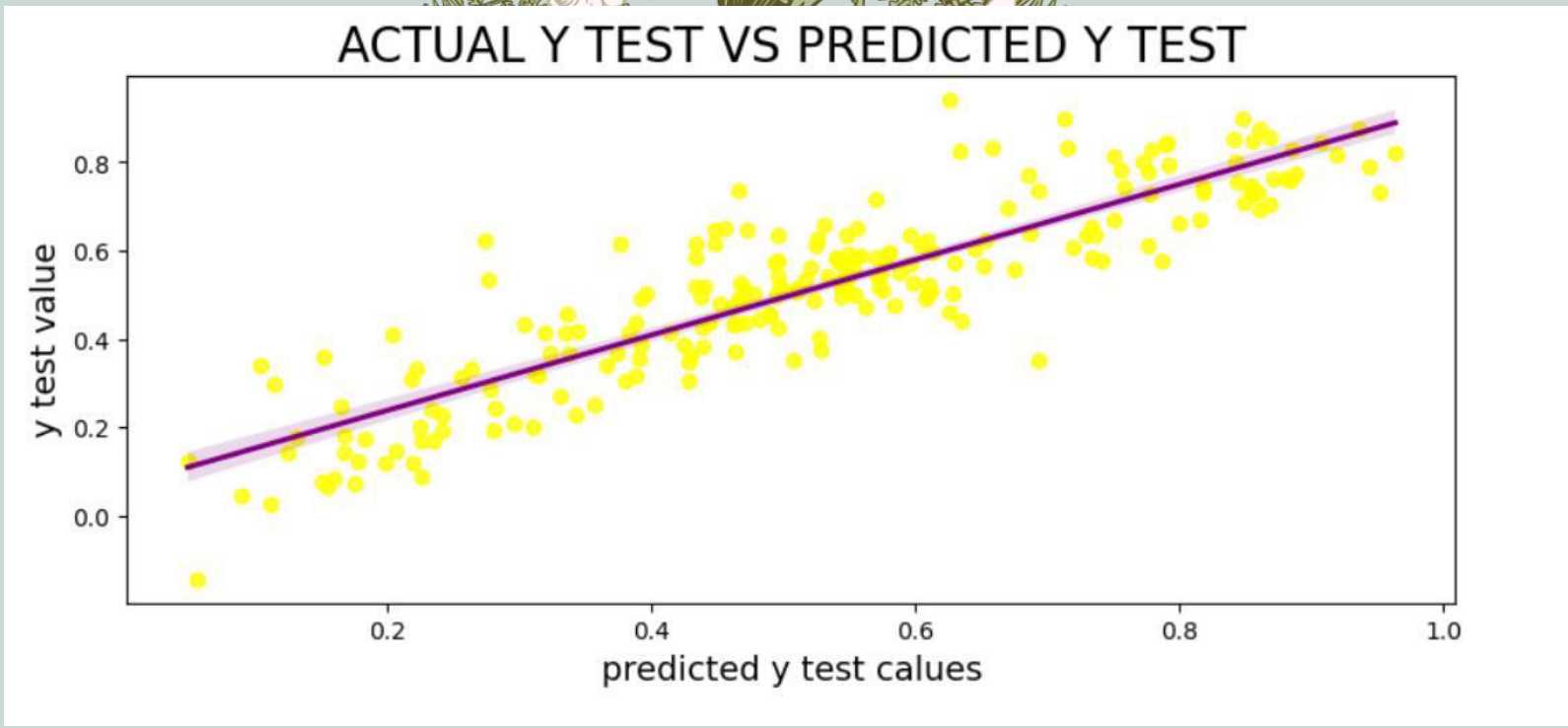
- The plot shows almost constance variance of prediction and thus the error validation the assumption

Testing for correlation between Error Terms



There is no correlation between Error Terms

Prediction



- Y Test values and Y Test predicted values have shown a strong visual semblance , Hence our predicted are evaluated as a healthy fit.

Final Report

- Most important factor affecting demand is temperature, With a coefficient of 0.73125, for every change in temperature of 1 degrees, demand increases by a factor of 0.73125 (temperature x 0.73125)
- Second most important factor is light rain or snow with a coefficient of -0.27750. Hence, if a particular day has light rains, it is expected to reduce the demand by 27.7%.
- Third most important factor is Year with a coefficient value of 0.24235. Rest all internal and external factors remain unchanged, the companies expected to see annual growth over last year at around 24%.
- Fourth most important factor is winter with a coefficient of 0.12793. This signifies the every winter , the demand is expected to increase by a factor of 0.12793 biased on other months.
- Fifth most important factor is month of July with a coefficient weightage of -0.10141 which is signifies that every July the demand is expected to drop by a factor of 0.10141 (around 10%).

Thank you

By:

Sachin Kumar Singh