Bike Sharing Assignment

Mandeep Singh Sachdeva

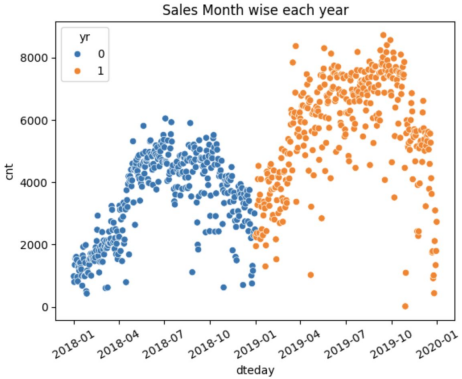
Objective

To implement a model which can predict the Bike Sharing number for next years.

- Helps the business to understand the patterns in the data.
- Prepare companies for the next year’s growth.

Problem Statement

- We want to see how Bike Rentals are affected due to different factors and see the trend for the next year.



Sales of 2018 & 2019

What will be the sales in 2020?

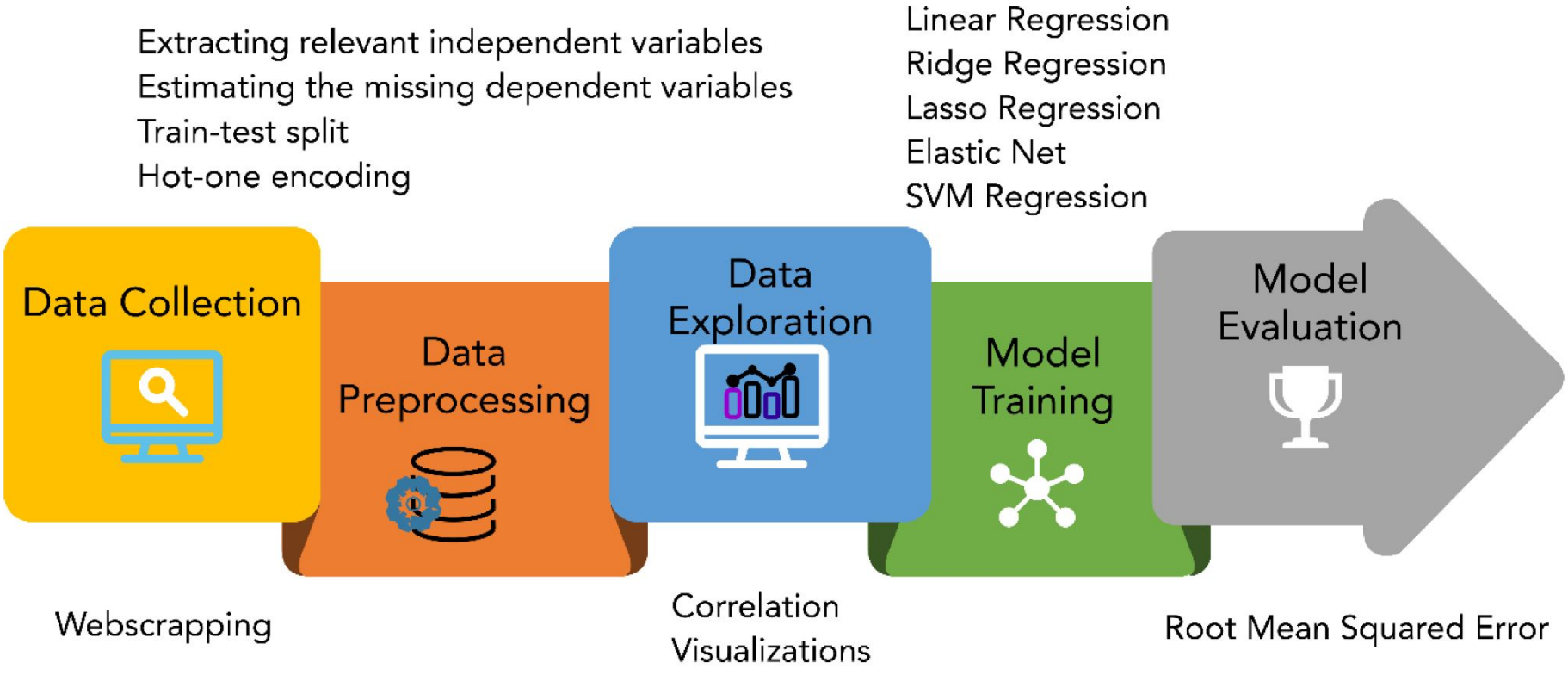
By analyzing the data of the past customers and number of variables associated to

it like:

- Temperature
- Windspeed
- Weather Situation
- Season
- Humidity

and many more..

Process of Modelling



Reference

Data Pre Processing

1. **Converted some columns data types**

   Eg: dte_days (object) -> Datetime

2. **Removing Unused / Derived Columns:**

   Eg: “instant”, “dte_days”

   “casual”: because it is derived from `cnt`

   “registered”: because it is also derived from `cnt`

3. **Scaling the features**

   It is done to get the features on same scale. The features that were off scale: “temp”, “hum”, “windspeed”

Interpretation Of Analysis / EDA

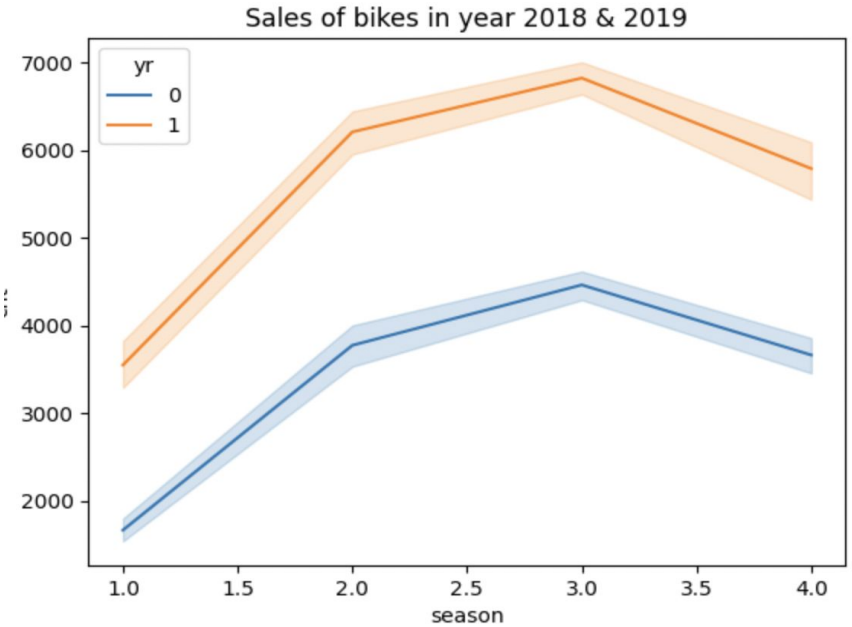
Sales vs Season

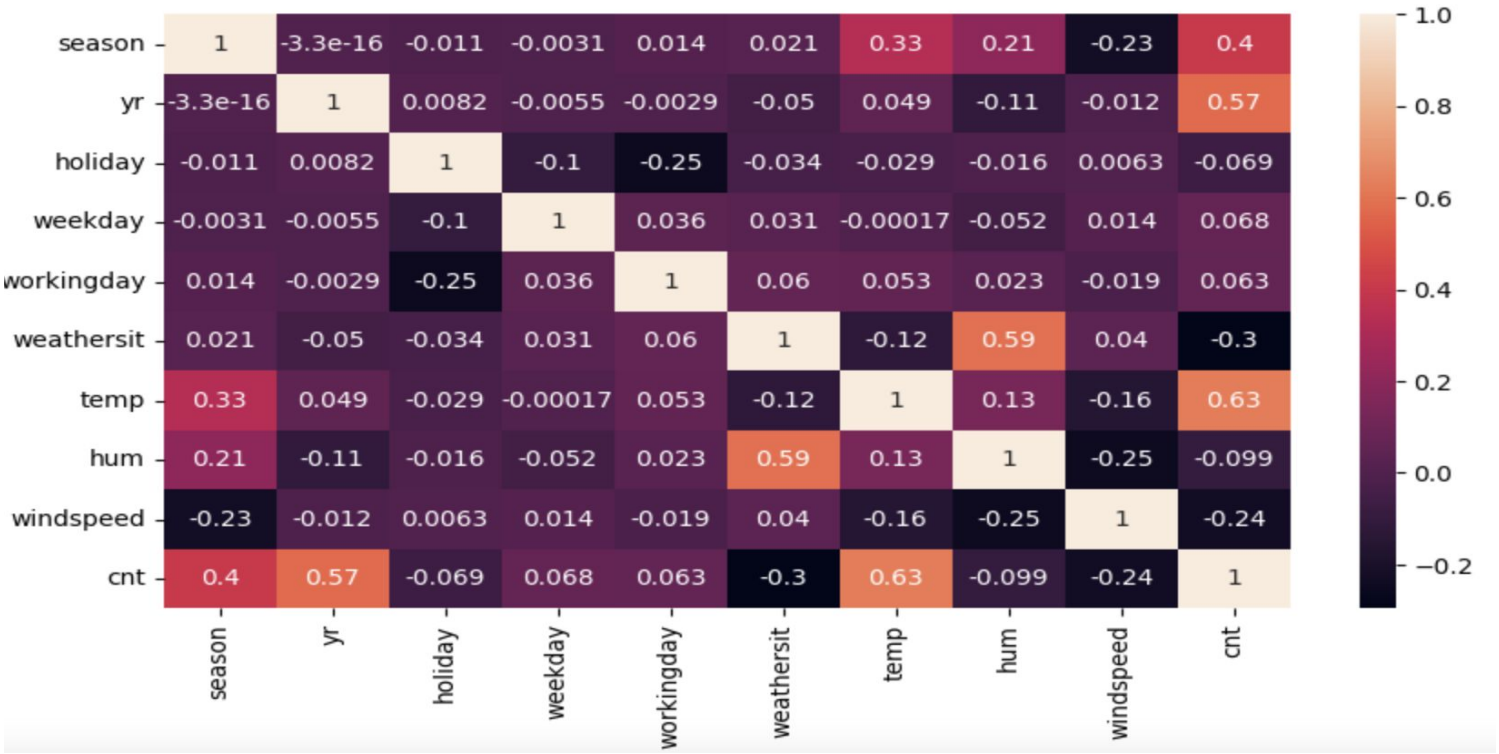
We can see the sales have increased

From 2018 to 2019. And for both the years

There's is this trend of the maximum

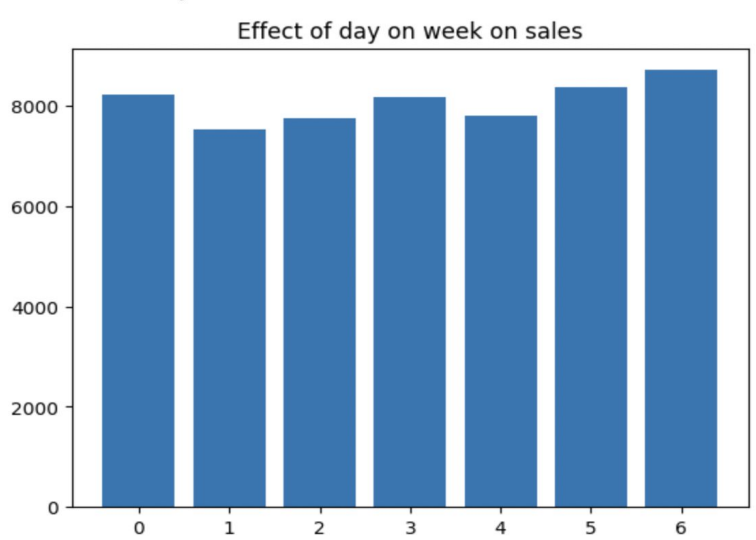
Rentals on the year of `Fall`.



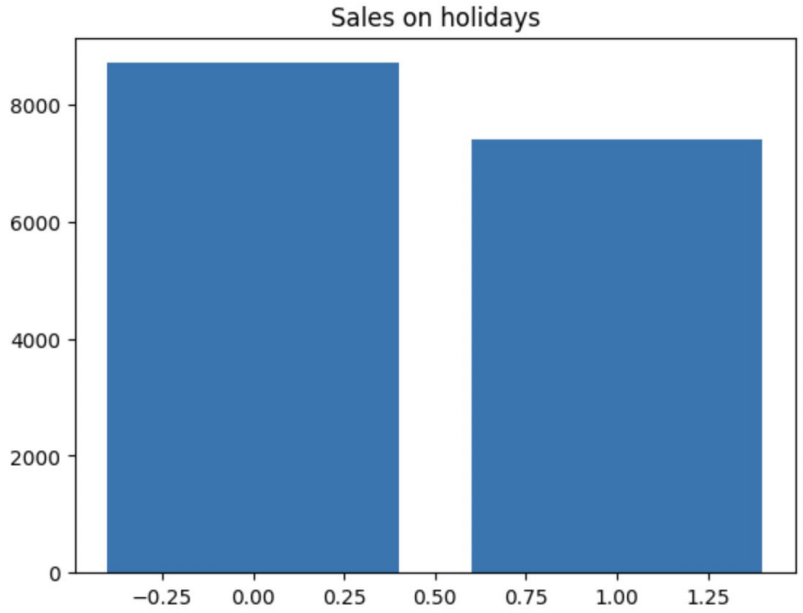


The correlation graph, shows us that

`Temperature` , `windspeed` and `year` are the most correlated features with the target variable.

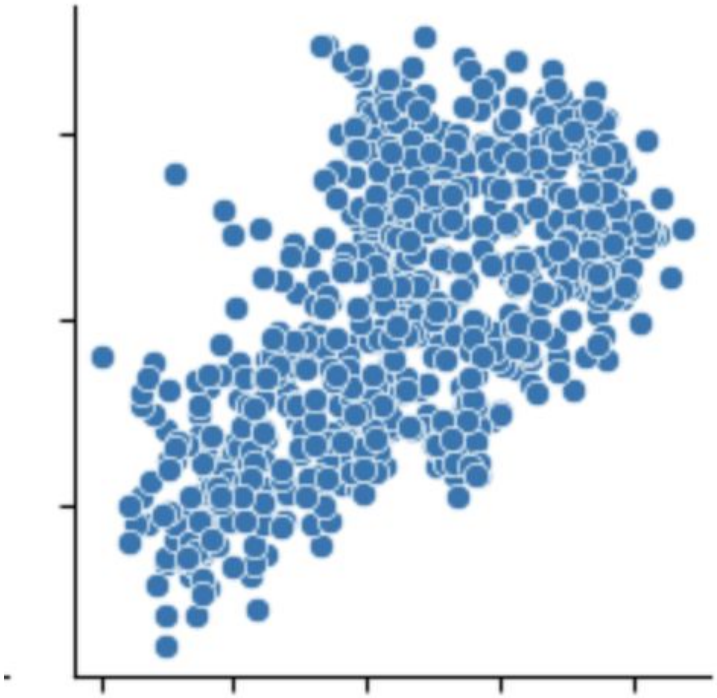


Interestingly, there's not much affect on which day it is of the week on the number of bike rentals



The interesting part comes here:

We as humans will have some intrusive Thoughts, that the on holidays the bike rental Would be more! **BUT** that's not the case actually. On holidays, the demand for bike rents is lesser compared to weekdays.



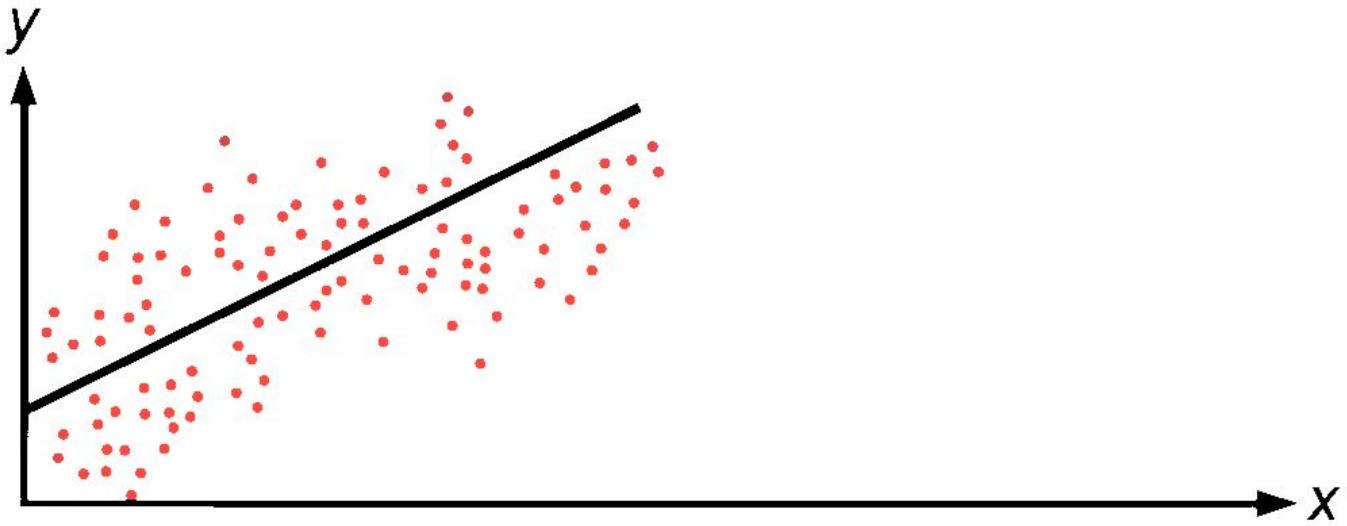
This one is from pairplot.

It is the plot of `Temperature` vs `cnt`

We can clearly see here, that a **linear relationship** is being formed with the target **(cnt)** and the independent variable(**temp**)

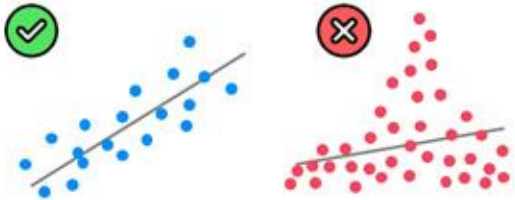
Building Model - Linear Regression

Imagine you have a magic line that can predict things! This line can tell you what your score might be based on how many hours you study. This magic line is what we call a "linear regression line."

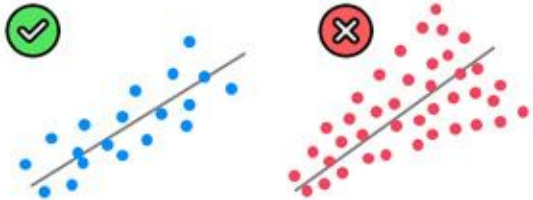


Assumptions of Linear Regression

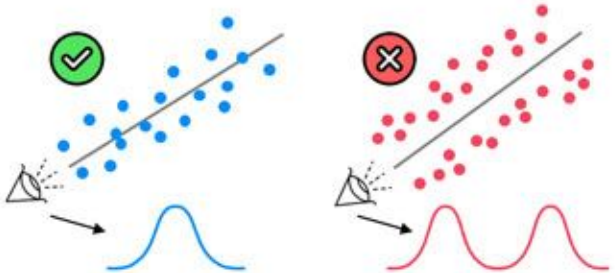
1. Linearity
(Linear relationship between Y and each X)



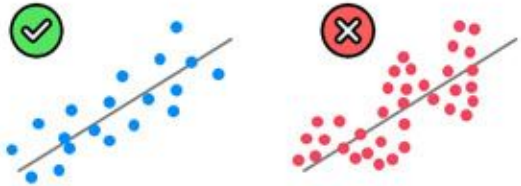
2. Homoscedasticity
(Equal variance)



3. Multivariate Normality
(Normality of error distribution)



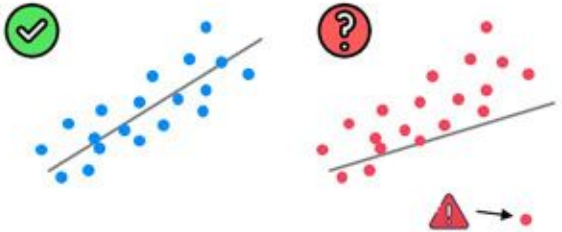
4. Independence
(of observations. Includes "no autocorrelation")



5. Lack of Multicollinearity
(Predictors are not correlated with each other)

$X_1 \not\sim X_2$ $X_1 \sim X_2$

6. The Outlier Check
(This is not an assumption, but an "extra")



Reference

Feature Selection - RFE

- One way is to manually look into p-values everytime and and drop the features which have **more** p-values.
  - More p-value means the feature is less important.
  - This we can do through **statsmodel** library.

	coef	std err	t	P>|t|	[0.025	0.975]
const	2013.3033	265.877	7.572	0.000	1490.930	2535.676
season	378.3204	38.043	9.945	0.000	303.577	453.064
yr	2042.2298	78.524	26.008	0.000	1887.951	2196.508
holiday	-624.6527	254.770	-2.452	0.015	-1125.205	-124.101
weekday	65.8284	19.477	3.380	0.001	27.562	104.095
workingday	168.7813	85.583	1.972	0.049	0.634	336.929
weathersit	-693.4608	90.931	-7.626	0.000	-872.115	-514.807
temp	4129.4580	189.502	21.791	0.000	3757.140	4501.776
hum	-815.7666	359.277	-2.271	0.024	-1521.646	-109.888
windspeed	-1512.3544	257.348	-5.877	0.000	-2017.971	-1006.738

- Another method is using **RFE.**
- It automatically sees what features to be included

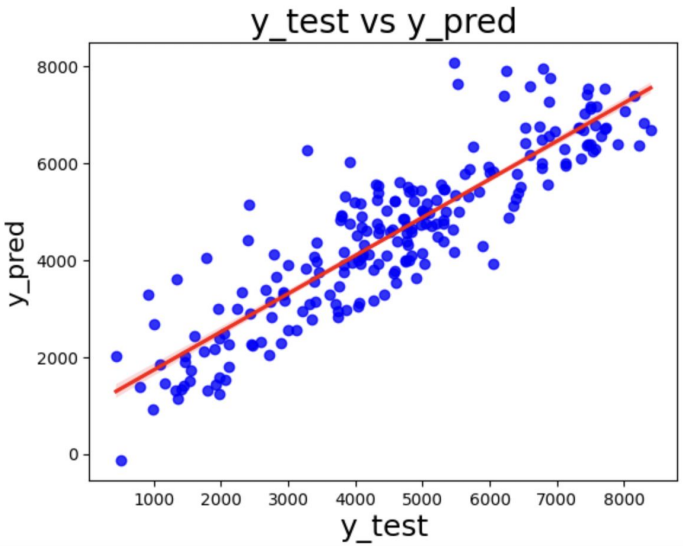
Predictive Analysis

We can use many methods for evaluation our model. But here we are using **r2_score** for evaluation purposes.

The **r2_score** with our final model is coming out to be: **0.7845**

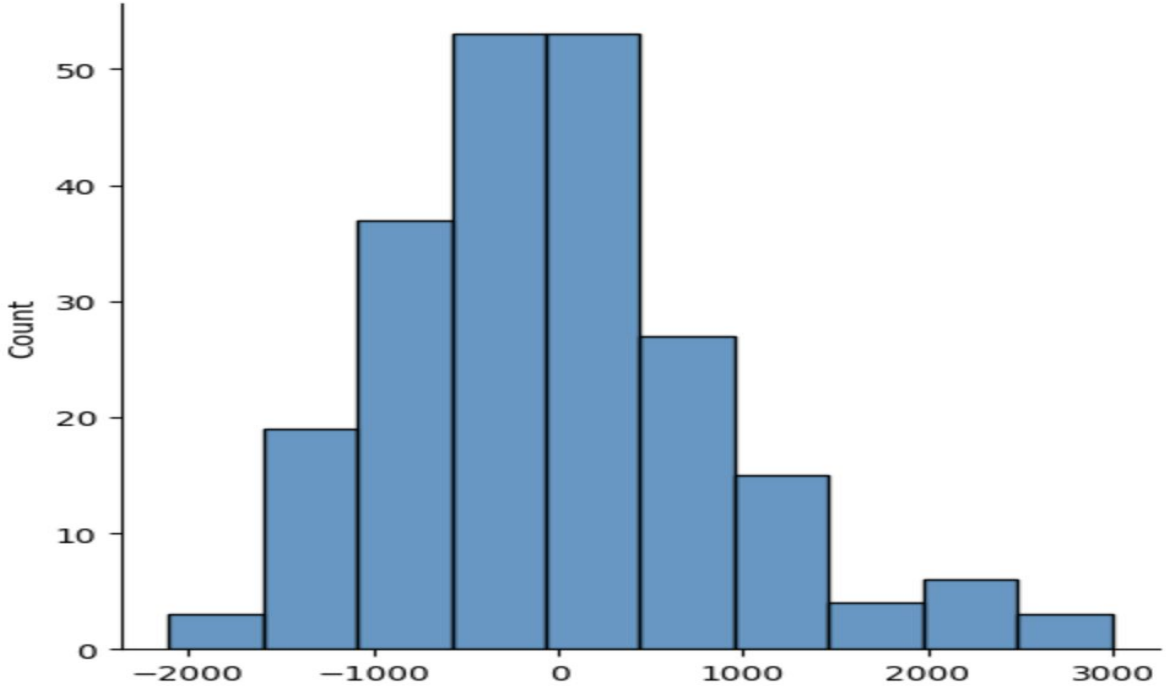
This plot shows the plot of y_predictions and y_test (actual).

This plot helps to understand how well the model's predictions align with the actual outcomes.



Residual Analysis

One of the main assumption for the linear regression is, the residuals (y_prediction - y_test) follows a normal distribution (also known as Multivariate Normality).



As we can see, the residuals here follows the normal distribution, hence the assumption that the linear regression model is a good fit seems to be True

Final Insights

- The most correlated feature with the target variable is `temp`.
- The assumptions for the Linear Regression holds true.
- Demand of the number of bikes to be rented depends on:
  - Temperature
  - Year
  - Humidity
  - Weather Situation