Introduction

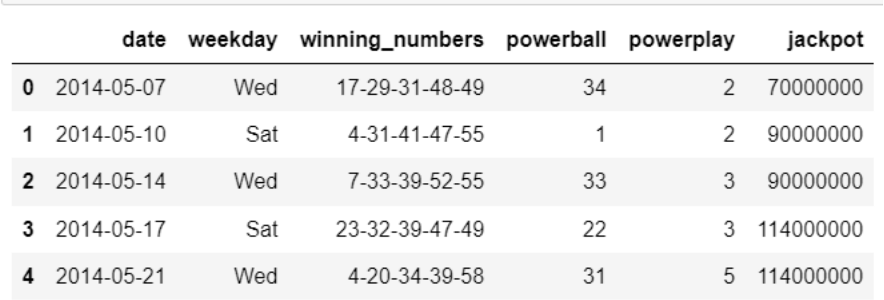
Lottery is a drawing of lots in which prizes are distributed to the winners among persons buying a chance.

Dataset

The dataset of lottery consists of 6 columns

- Date
- Weekday
- Winning-numbers
- Powerball
- Powerplay
- jackpot

The shape of dataset is (950,6). The head of dataset is shown in figure 1.

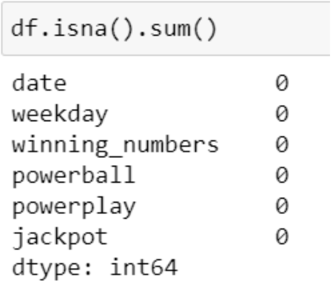


	date	weekday	winning_numbers	powerball	powerplay	jackpot
0	2014-05-07	Wed	17-29-31-48-49	34	2	70000000
1	2014-05-10	Sat	4-31-41-47-55	1	2	90000000
2	2014-05-14	Wed	7-33-39-52-55	33	3	90000000
3	2014-05-17	Sat	23-32-39-47-49	22	3	114000000
4	2014-05-21	Wed	4-20-34-39-58	31	5	114000000

Figure 1 Head of the Dataset

Null Value

The null value mean that some column has zero value in some rows. The solution to this problem is to drop that row. The figure 2 shows the null value count in each column



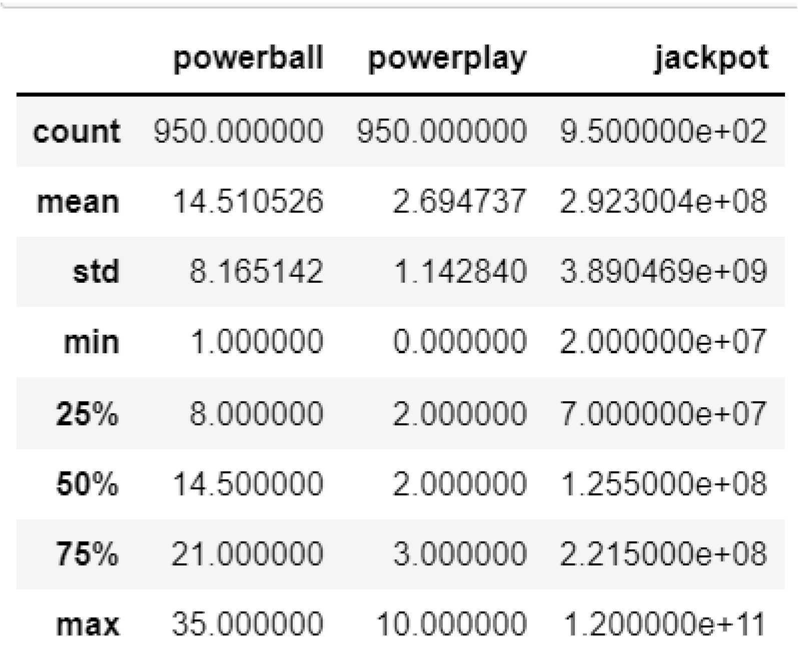
```
df.isna().sum()
```

```
date               0
weekday            0
winning_numbers    0
powerball          0
powerplay          0
jackpot            0
dtype: int64
```

Figure 2 null value in each column

Statistics of dataset

The statistics of dataset is showing statistics information of numerical column of the dataset shown in figure 3



	powerball	powerplay	jackpot
count	950.000000	950.000000	9.500000e+02
mean	14.510526	2.694737	2.923004e+08
std	8.165142	1.142840	3.890469e+09
min	1.000000	0.000000	2.000000e+07
25%	8.000000	2.000000	7.000000e+07
50%	14.500000	2.000000	1.255000e+08
75%	21.000000	3.000000	2.215000e+08
max	35.000000	10.000000	1.200000e+11

Figure 3 Statistical value of numerical columns

Outlier Detection

A dataset's abnormal observations/samples that do not fit its typical/normal statistical distribution are found using an outlier detection technique (ODT). Simple approaches for outlier detection examine each distinct attribute of the dataset using statistical tools like boxplot and Z-score. The boxplot method is used in this report as shown in figure 4.It can be seen in figure 4 that jackpot column has outlier value which is removed by setting threshold.

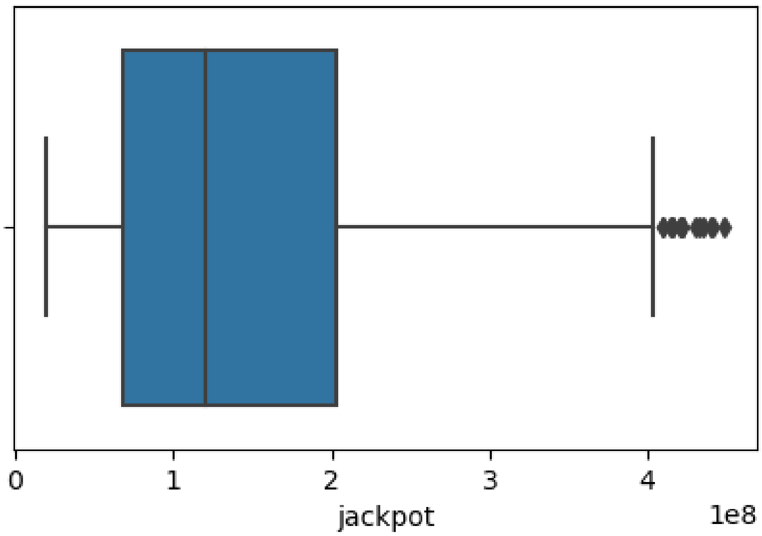


Figure 4 Outlier in jackpot

Analysis

Univariant

Unique values are the items that appear in a dataset only once. In the dataset, unique values of weather condition and wind direction is check and it was found that weather condition has 56 and wind direction has 10 unique values.

Weekday

This column has three type of values Wednesday, Saturday and Monday and it was found that Monday has the lowest value while other two values has same number of instances.

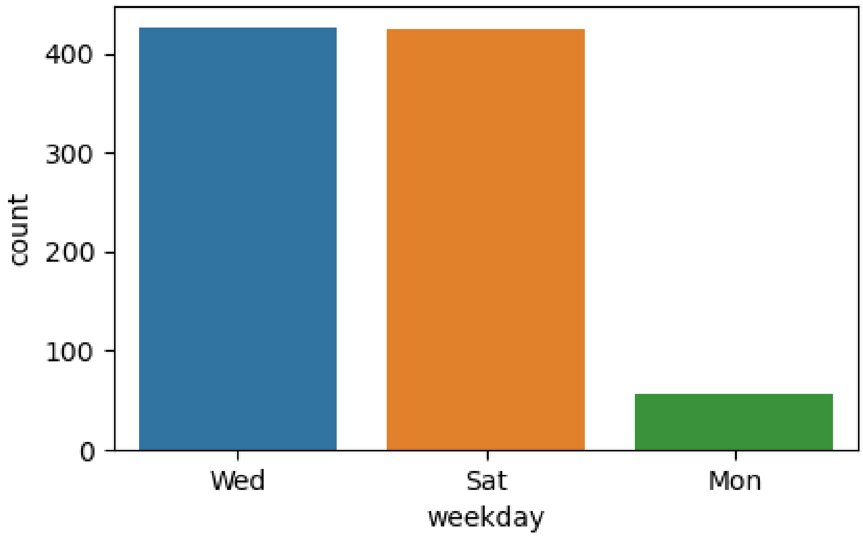


Figure 5 Weekday values

Powerball

Figure 6 shows the frequency of power ball value and it can been seen that highest frequency is for value of power ball 17 and value 30 has the lowest frequency.

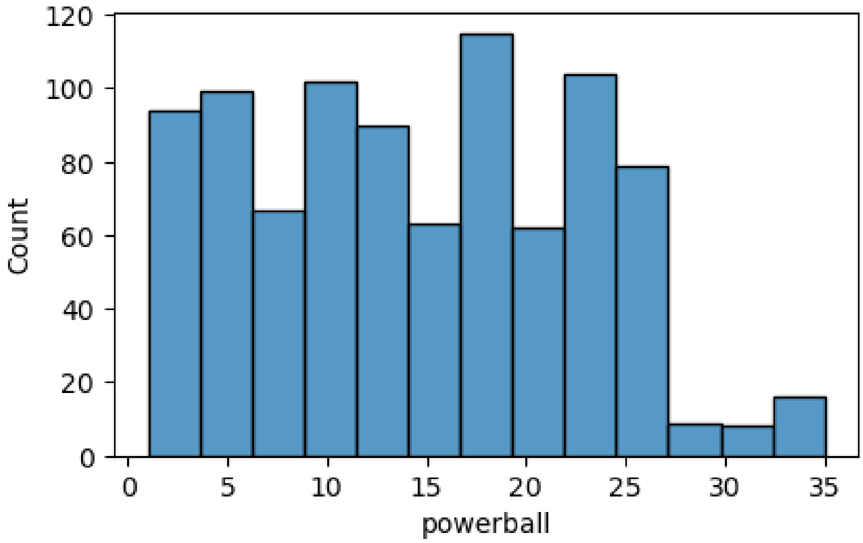


Figure 6 Frequency of Powerball value

Power Play

It can be seen in figure 7 that value 2 has most occurrence in dataset nearly 500 and value 10 has the lowest frequency of this.

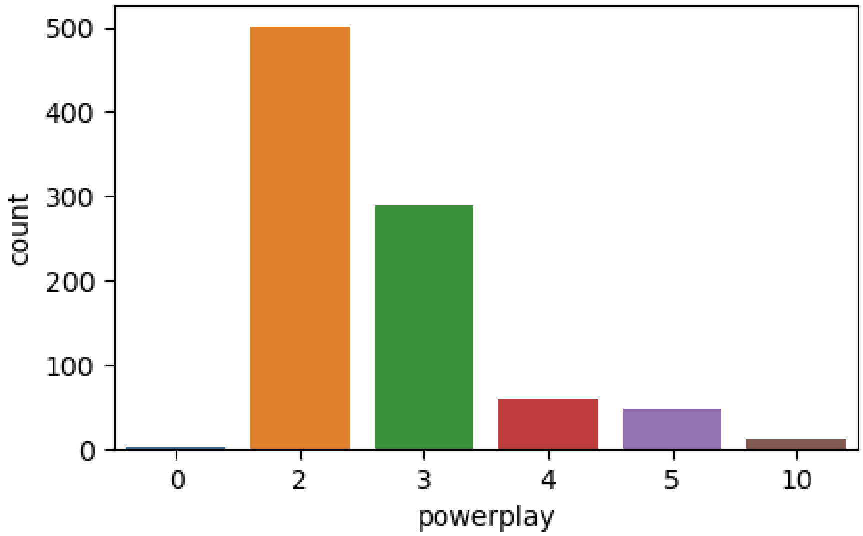
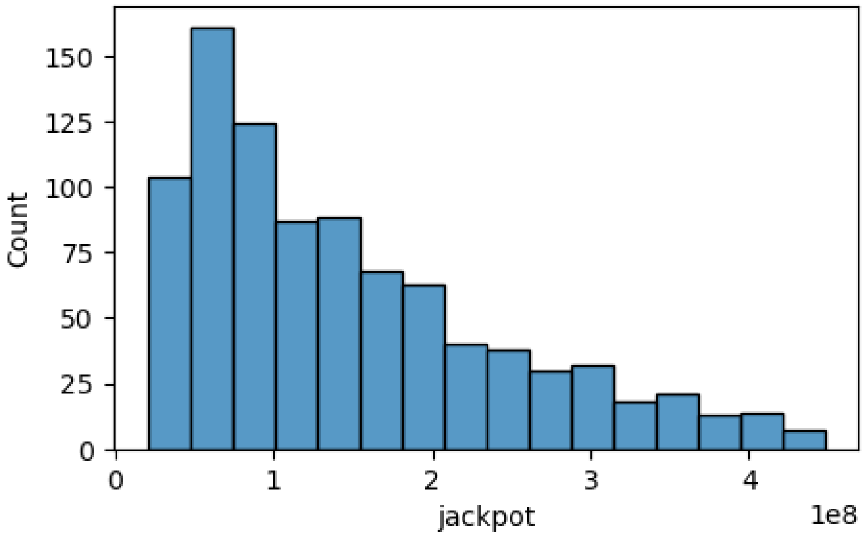


Figure 7 frequency of powerplay

Jackpot

Jackpot is the price money which was win by winner, In figure 8 the jackpot distribution of data is shown and it can be seen that there are more value for less jackpot and maximum value is for minimum jackpot as compare to maximum value. The least value oof occurrence is given for maximum jackpot value



ss

Figure 8 Frequency distribution of Jackpot

Bivariant Analysis

Bivariate analysis means the analysis of bivariate data. It is one of the simplest forms of statistical analysis, used to find out if there is a relationship between two sets of values. It usually involves the variables X and Y.

The figure 9 shows the year wise plot of jackpot count and it can be seen that most instances are from 2022 and least are from 2014 and it can be concluded that the instances are increasing continuously as the year passes. Moreover, month analysis shows that most of the date are from the month of June , July , April, and august. Figure 11 shows the frequency of occurrence of different days and it was noted that most frequency is found at the starting days of the months like first 10 days.

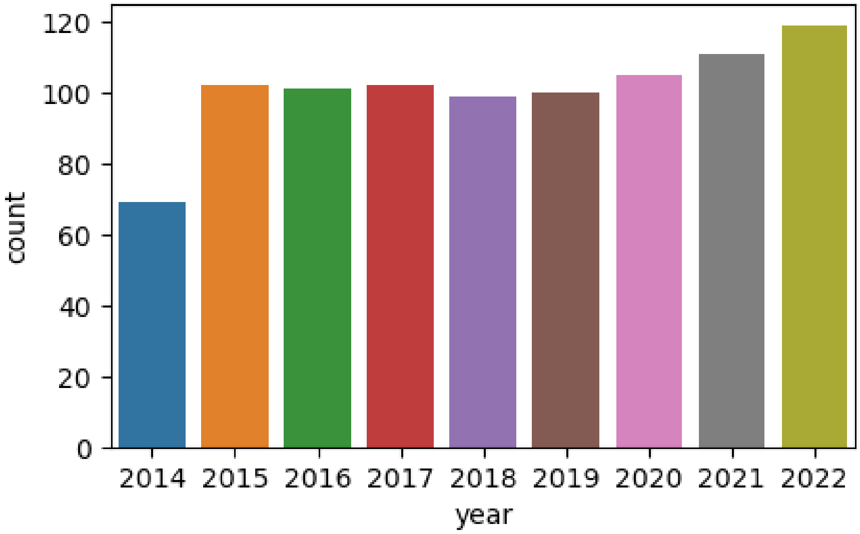


Figure 9 Year wise value of jackpot

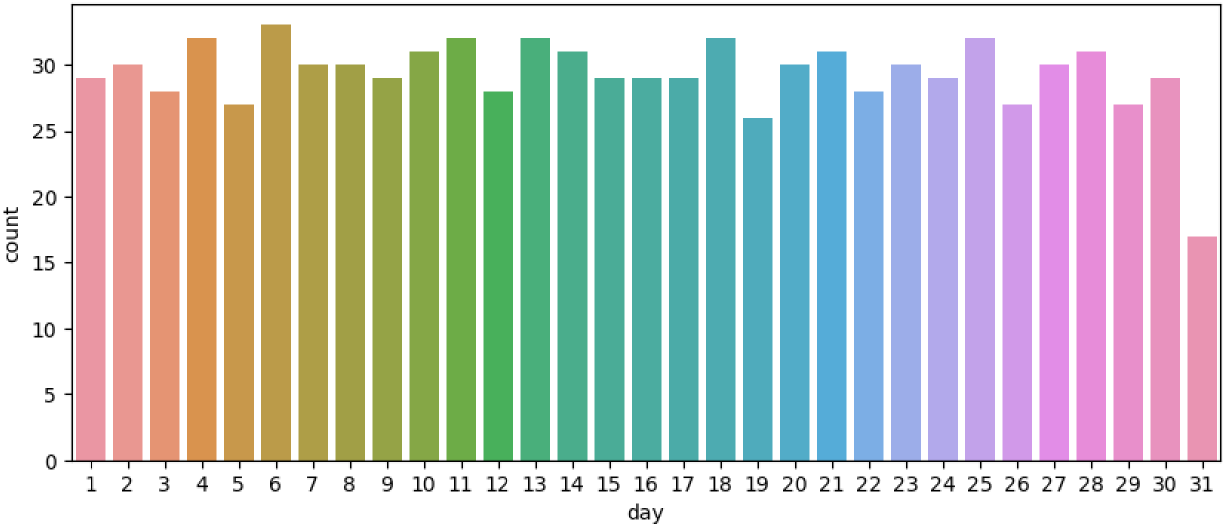


Figure 10 Frequency of day of month

Weekday and Jackpot

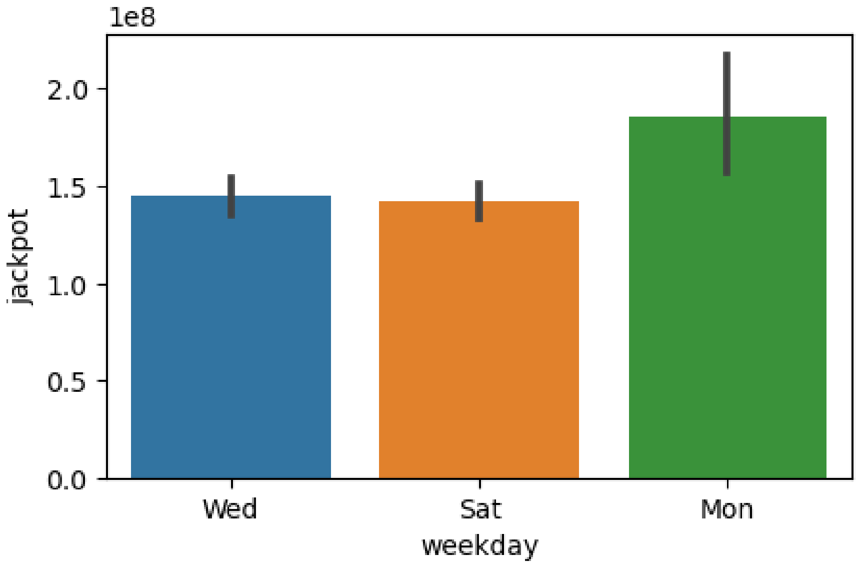


Figure 11 Weekday relation with Jackpot

Figure 11 shows that the mean value of Jackpot in each weekday. It can be seen that the highest mean value is on Monday. Though there is very less number of lotteries in that day, but it seems that on Monday there is a very bigger jackpot

Powerplay with Jackpot

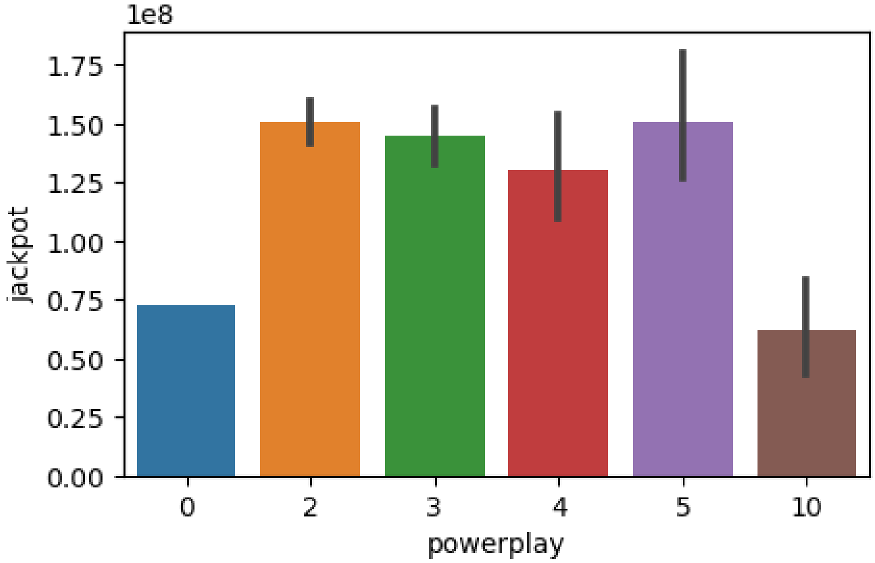


Figure 12 Powerplay with Jackpot

Figure 12 shows the mean jackpot against each powerplay and it can be seen that highest mean value of jackpot is from 2,3 and 5th powerplay while 0 and 10 has very low value of mean. If we talk about year wise jackpot value then it was found that 2022-year value has highest mean value and if we look at month then December month has the highest mean value.

Jackpot with Powerplay and Weekday

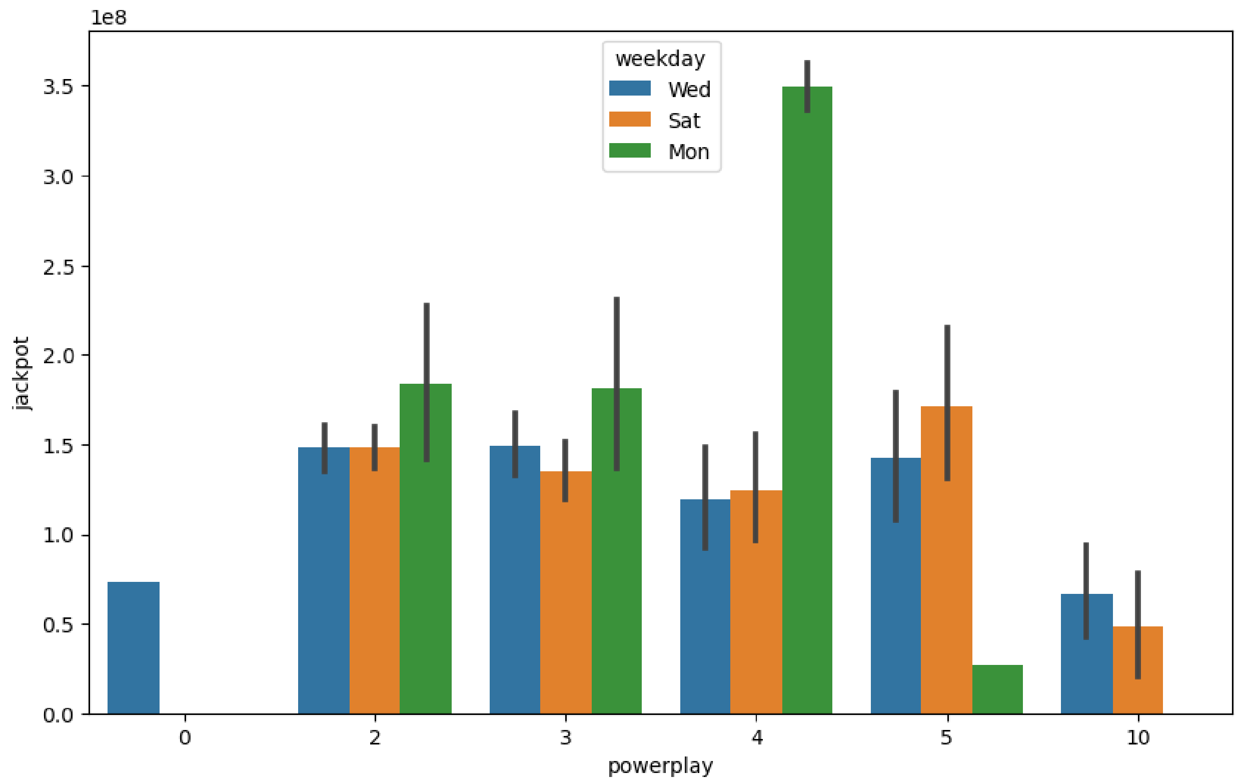


Figure 13 Jackpot and Powerplay with weekdays

This shows the mean Jackpot in each powerplay and each weekday in a particular powerplay. we can see that at 4th Powerplay and on Monday, there is a very high average of jackpot It can also be seen that 4th powerplay has value on Monday while powerplay value 5 is lowest among all values

Correlation

A correlation matrix is simply a table which displays the correlation coefficients for different variables. The matrix depicts the correlation between all the possible pairs of values in a table.

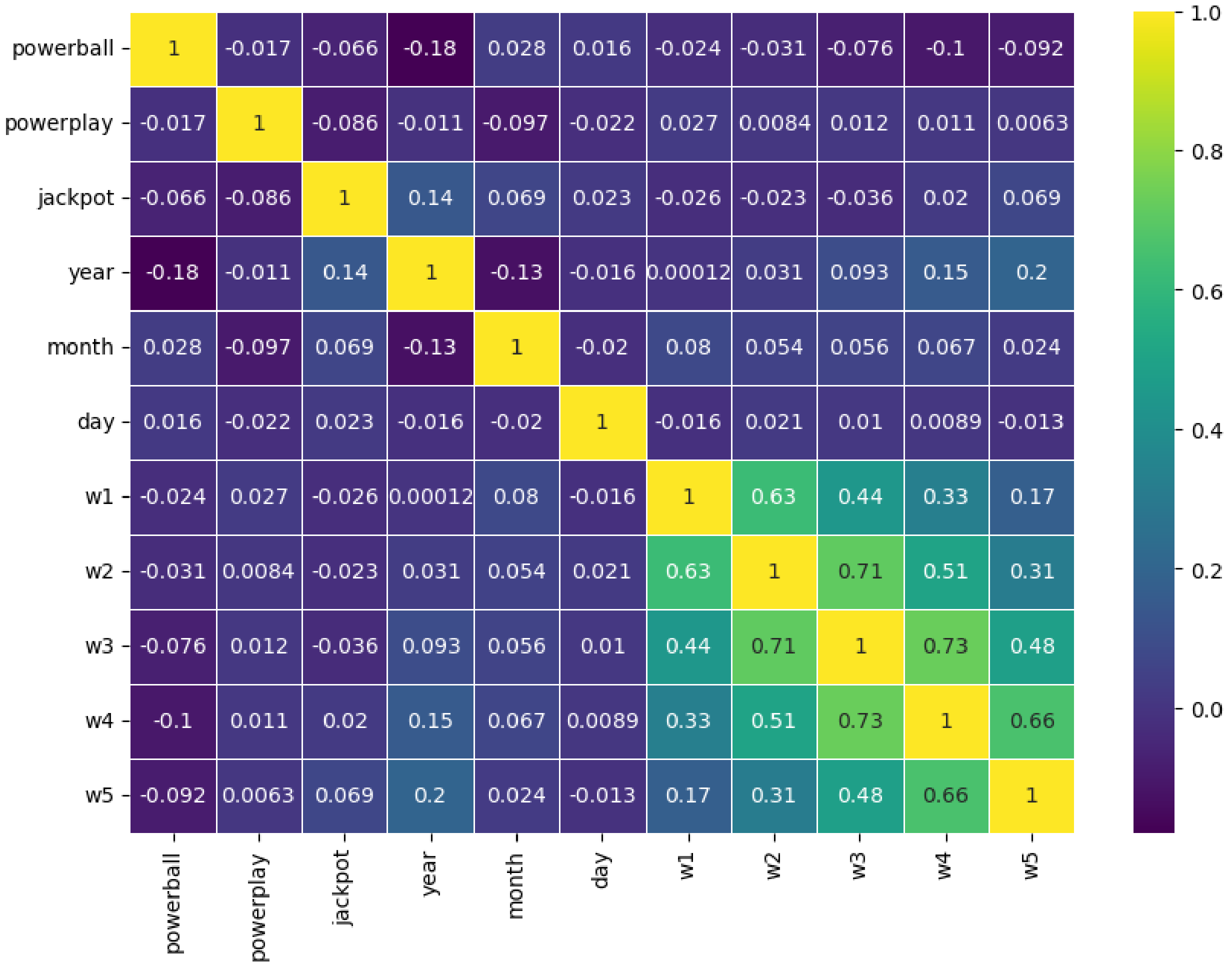


Figure 14 Corelation of dataset

Figure 14 shows that w1,w2,w3,w4 and w5 are the closely related to each other. It can see above in the correlation graph that the Winning combinations have high correlation with each other.

- W1 have high correlation with W2
- W2 have high correlation with W3
- W3 have high correlation with W4
- W5 have high correlation with W5

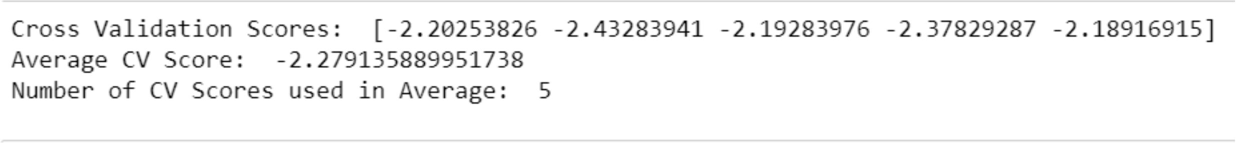
Algorithm

Following three algorithms were run on this repot

- Linear regression
- Ridge Rgression

- Lasso Regression

If we look at accuracy of regression algorithms. It can be seen that these algorithms have high MSE value which mean that their performance is not as much good. At last, Linear regression with cross validation is applied to better the model performance and it was noted that their performance increased by doing this,The result of cross validation linear regression model is given below



```
Cross Validation Scores:  [-2.20253826 -2.43283941 -2.19283976 -2.37829287 -2.18916915]
Average CV Score:  -2.279135889951738
Number of CV Scores used in Average:  5
```