Data analysis:

Lottery Powerball Winning Numbers: Beginning 2010

El younsi Yousra
Jamhour Zaid

Questions to answer

1. What is the distribution of winning numbers and multipliers?
2. Which is the most commonly drawn number? Which is the least commonly drawn number?
3. Is there any correlation between the w inning numbers and the multiplier?
4. What are the top five most frequently drawn winning numbers for each position (1-5) in the Powerball lottery since 2010?
5. Based on the historical data from the Powerball lottery, which series of numbers have the highest probability of appearing in the winning combination?
6. What are the odds of winning from the first try and how many tries with random numbers would it take to win ?

Factors behind the subject choice

01

Personal interest in games of chance and probability

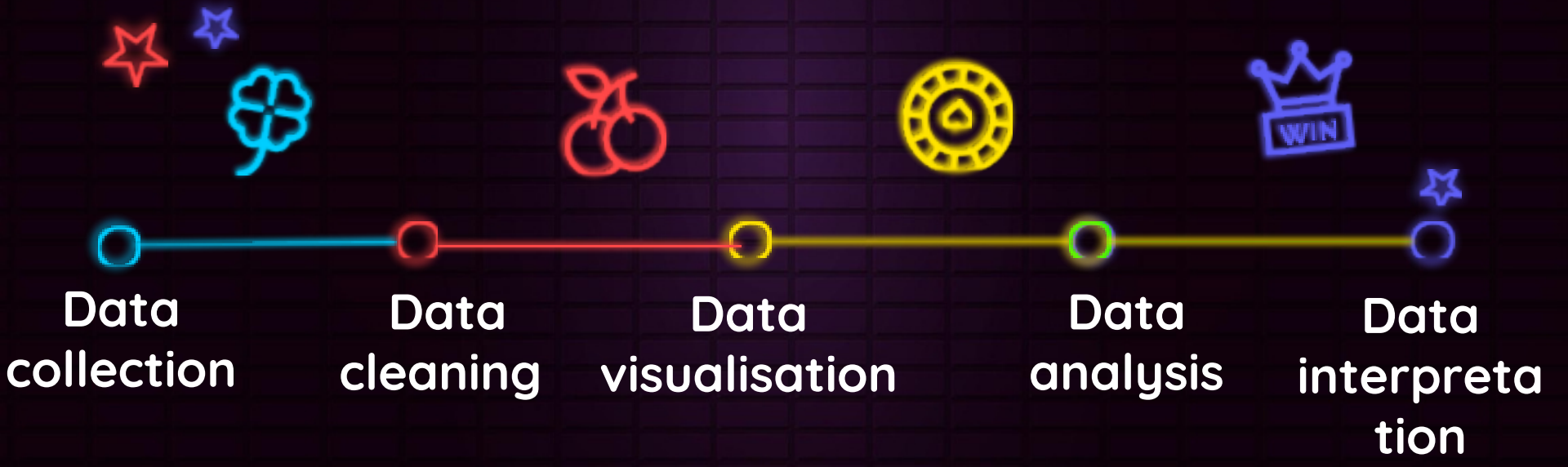
02

Relevance and timeliness of the subject

03

Risk assessment: gambling can be a serious addiction

Steps


WIN
Data collection
Data cleaning
Data visualisation
Data analysis
Data interpretation

1. Data collection

We imported that dataset from "https://data.ny.gov/api/views/d6yy-54nr/rows.xml?accessType=DOWNLOAD"

2. Data cleaning

Using python we did the following:

Check for missing values and duplicates

Split the winning numbers column into separate columns and name them appropriately

Check that the data types are correct

Display of the first 10 rows from the original Dataset

0	09/26/2020	11	21	27	36	62	24	3.0
1	09/30/2020	14	18	36	49	67	18	2.0
2	10/03/2020	18	31	36	43	47	20	2.0
3	10/07/2020	06	24	30	53	56	19	2.0
4	10/10/2020	05	18	23	40	50	18	3.0
5	10/14/2020	21	37	52	53	58	05	2.0
6	10/17/2020	06	10	31	37	44	23	2.0
7	10/21/2020	01	03	13	44	56	26	3.0
8	10/24/2020	18	20	27	45	65	06	2.0
9	10/28/2020	11	28	37	40	53	13	2.0

DATA CLEANING

Removing rows with missing values and checking for duplicates:

```
Number of rows: 1460
Number of rows after removing missing values (1460-210): 1250
Number of duplicates: 0
```

Missing values per column

Draw Date	0
Winning Numbers	0
Multiplier	210

Display of the new dataframe after splitting the winning numbers into different columns and naming them <num1>..<num6>. The numbers show their positions.

	Draw Date	Winning Numbers	Multiplier	num1	num2	num3	num4	num5	num6
0	09/26/2020	11 21 27 36 62 24	3.0	11	21	27	36	62	24
1	09/30/2020	14 18 36 49 67 18	2.0	14	18	36	49	67	18
2	10/03/2020	18 31 36 43 47 20	2.0	18	31	36	43	47	20
3	10/07/2020	06 24 30 53 56 19	2.0	06	24	30	53	56	19
4	10/10/2020	05 18 23 40 50 18	3.0	05	18	23	40	50	18
5	10/14/2020	21 37 52 53 58 05	2.0	21	37	52	53	58	05
6	10/17/2020	06 10 31 37 44 23	2.0	06	10	31	37	44	23
7	10/21/2020	01 03 13 44 56 26	3.0	01	03	13	44	56	26
8	10/24/2020	18 20 27 45 65 06	2.0	18	20	27	45	65	06
9	10/28/2020	11 28 37 40 53 13	2.0	11	28	37	40	53	13

Now we know there are no missing numbers amongst the winning numbers columns

```
Number of rows: 1250
Draw Date          0
Winning Numbers    0
Multiplier         0
num1               0
num2               0
num3               0
num4               0
num5               0
num6               0
```

```
Draw Date              object
Winning Numbers        object
Multiplier            float64
num1                   object
num2                   object
num3                   object
num4                   object
num5                   object
num6                   object
dtype: object
```

Because the data types aren't all correct, we convert the Dates to date type and the winning numbers to integers.

DATA Cleaned!

```
Draw Date           datetime64[ns]
Winning Numbers             object
Multiplier                   int32
num1                         int32
num2                         int32
num3                         int32
num4                         int32
num5                         int32
num6                         int32
dtype: object
```

The column types are now correct.

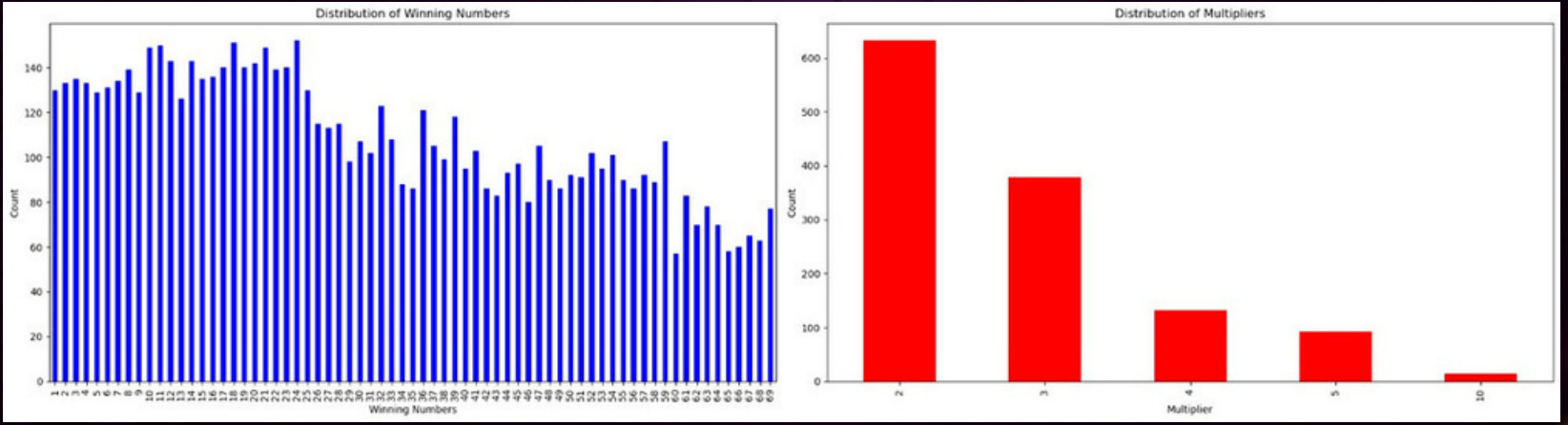
3.Data visualisation

4. Data analysis

5. Data interpretation

Question 1: What is the distribution of winning numbers and multipliers?

The distribution of the winning numbers and the multipliers:



Question 2:

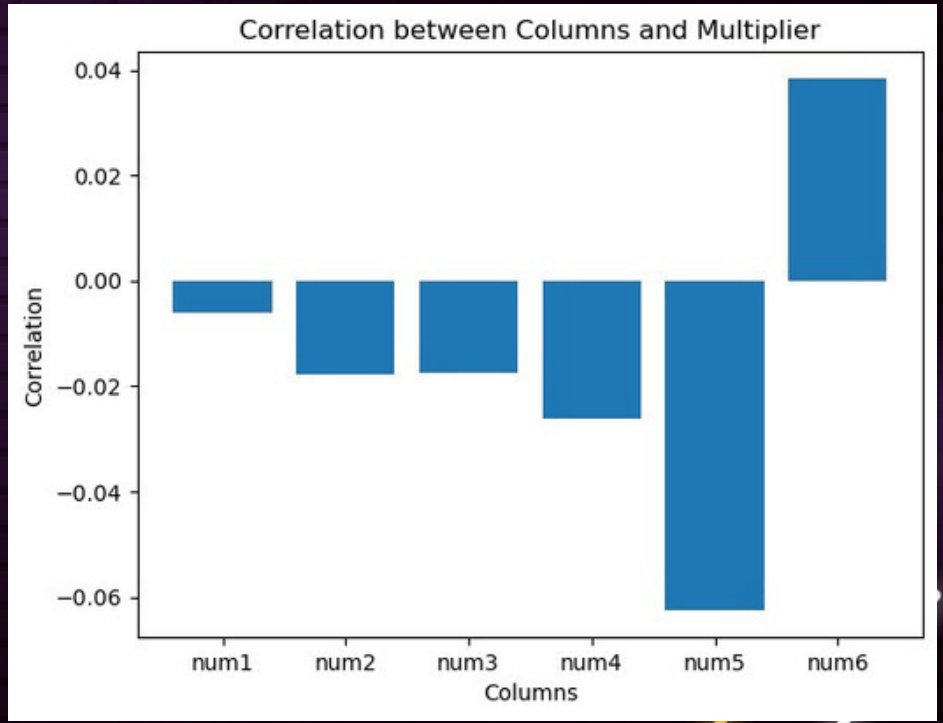
Which is the most commonly drawn number? Which is the least commonly drawn number?

Most commonly drawn number: 24

Least commonly drawn number: 60

Question 3:
Is there any correlation between the winning numbers and the multiplier?

Correlation between 'num1' and 'Multiplier': -0.006
Correlation between 'num2' and 'Multiplier': -0.018
Correlation between 'num3' and 'Multiplier': -0.017
Correlation between 'num4' and 'Multiplier': -0.026
Correlation between 'num5' and 'Multiplier': -0.062
Correlation between 'num6' and 'Multiplier': 0.038



Question 4:

What are the top five most frequently drawn winning numbers for each position (1-5) in the Powerball lottery since 2010?

Top five most frequently drawn numbers

for position 1:

1 90
2 85
3 77
5 76
4 69

for position 2:

21 48
12 48
15 46
28 46
20 45

for position 3:

39 44
37 42
36 40
32 40
34 39

for position 4:

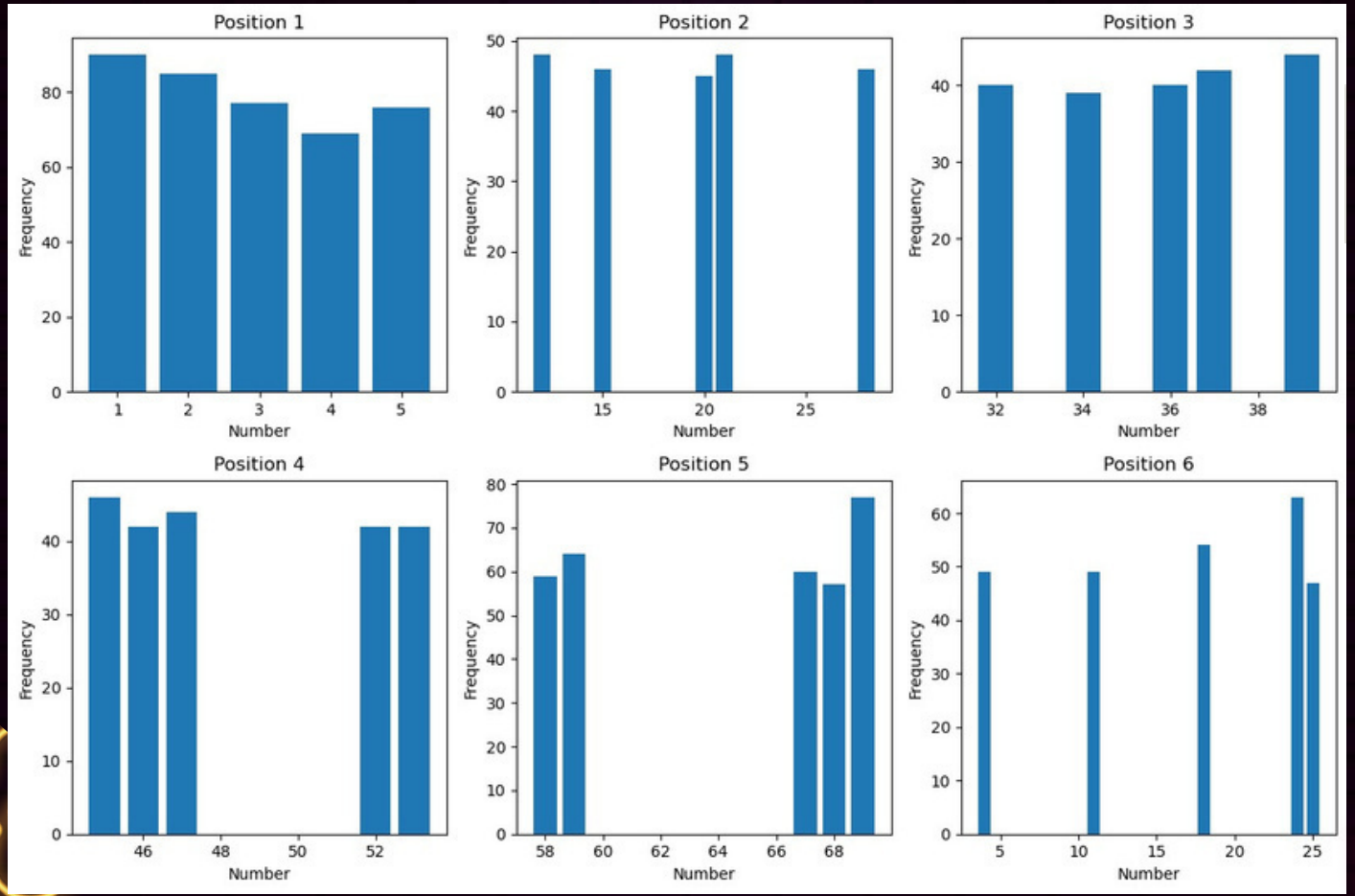
45 46
47 44
53 42
46 42
52 42

for position 5:

69 77
59 64
67 60
58 59
68 57

for position 6:

24 63
18 54
4 49
11 49
25 47


Position 1
Frequency
Number
Position 2
Frequency
Number
Position 3
Frequency
Number
Position 4
Frequency
Number
Position 5
Frequency
Number
Position 6
Frequency
Number

Question 5:

Based on the historical data from the Powerball lottery,
which series of numbers have the highest probability of appearing in the winning combination?

Series of numbers with the highest probability of appearing in the winning combination: ('11', '14', '18', '06', '05')

Question 6:

Based on the historical data from the Powerball lottery,
which series of numbers have the highest probability of appearing in the winning combination?

The estimated probability of winning from the first try is: 0.000100%
The estimated probability of winning after 1000000 (1 million)tries is: 63.212074%
The estimated probability of winning after 10000000 (TEN MILLION) tries is: 99.995460%

conclusions

Data Source:

“data.gov” website.
Especially: “https://catalog.data.gov/dataset/lottery-powerball-winnin g-numbers-beginning-2010”