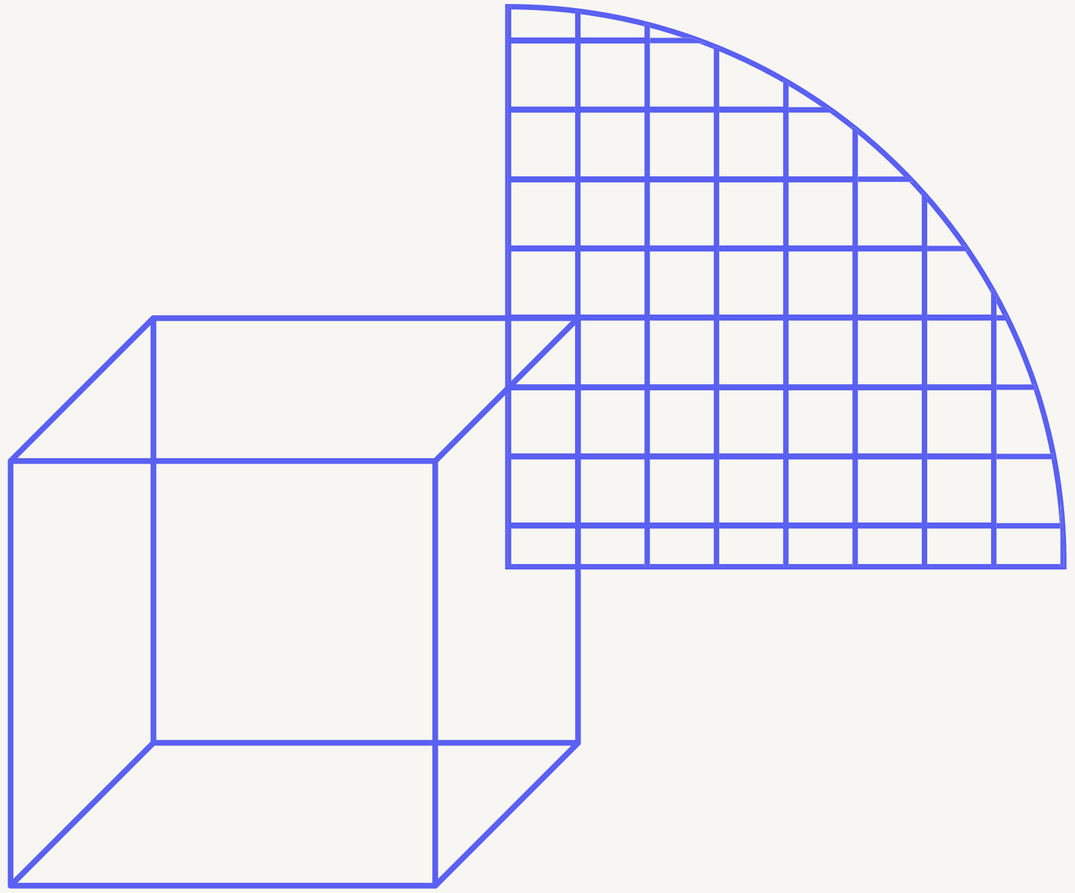
Classic

DATA VISUALIZATION



Learning Outcomes

At the end of this workshop, (almost all of) you will be able to answer:

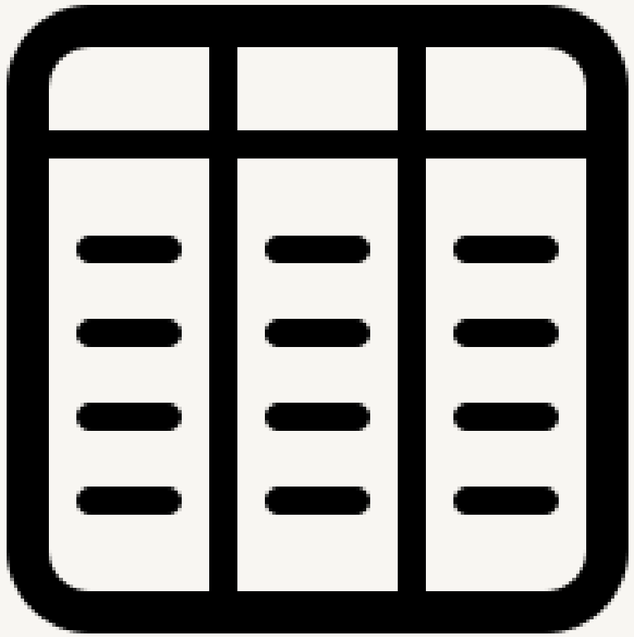
WHY do we visualize data?

WHAT data and visualizations?

HOW do we do it?

Part 0

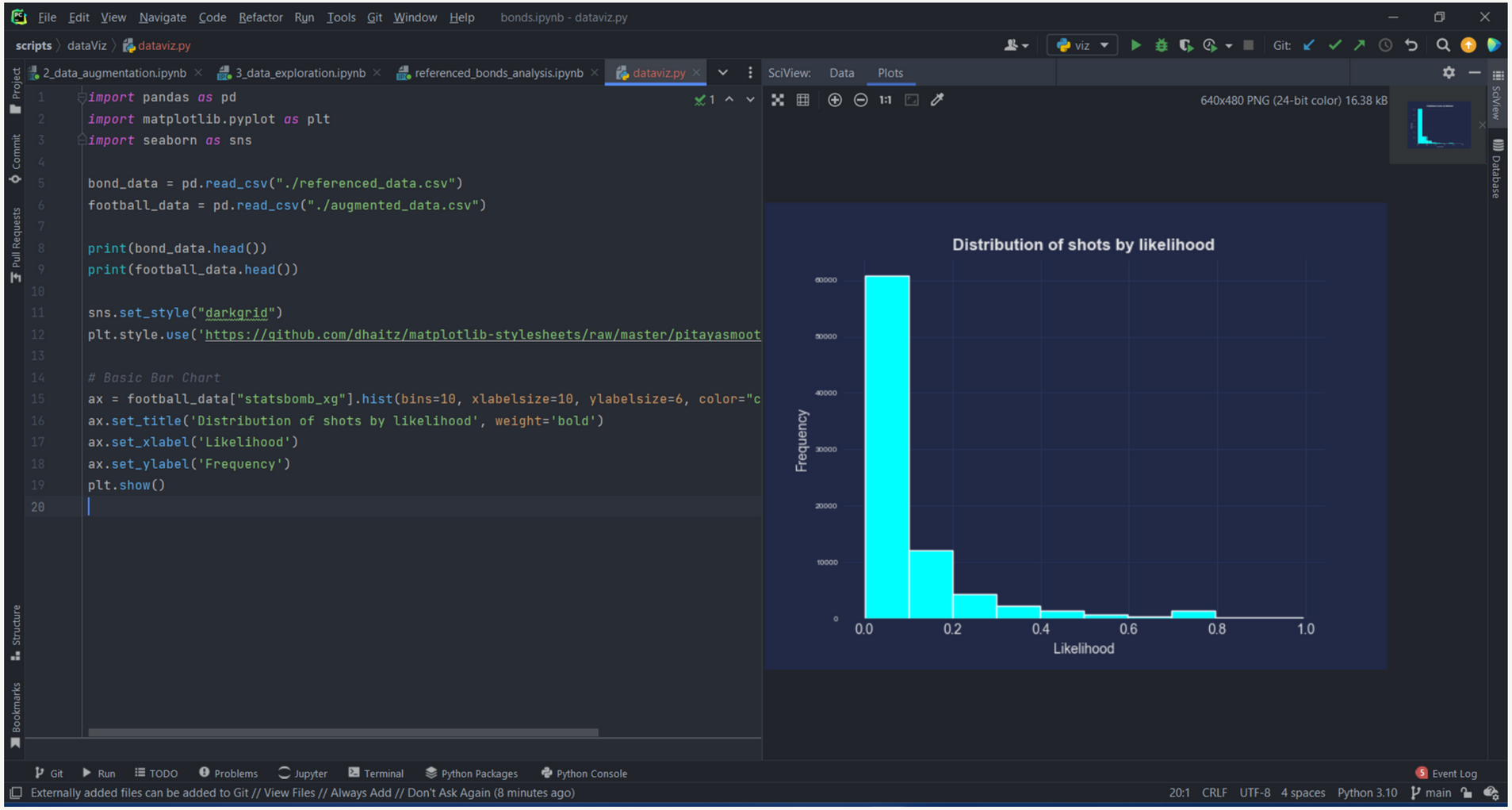
Imitation is the best form of learning



Replication

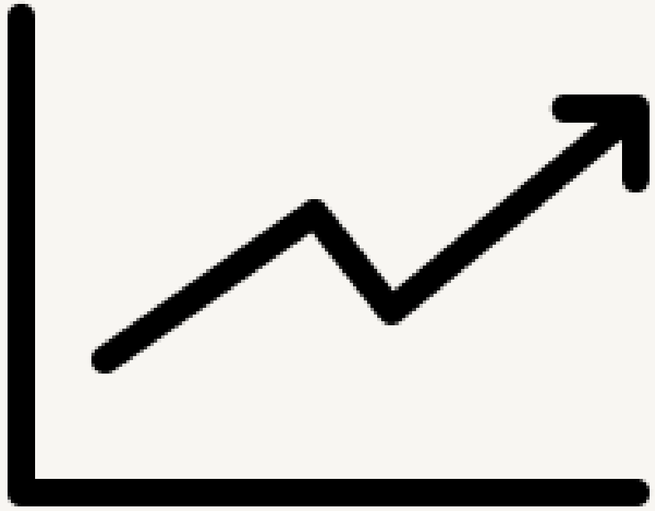
- github.com/kkkarnav/betterxG/data/augmented_data.csv
- github.com/kkkarnav/betterxG/3_data_exploration.ipynb
- github.com/kkkarnav/electoralBonds/referenced_data.csv
- github.com/kkkarnav/electoralBonds/referenced_bonds_analysis.csv
- github.com/kkkarnav/dataViz/a2.Rmd
- github.com/kkkarnav/dataViz/a2.html

github.com/kkkarnav/dataViz



Part I (why)

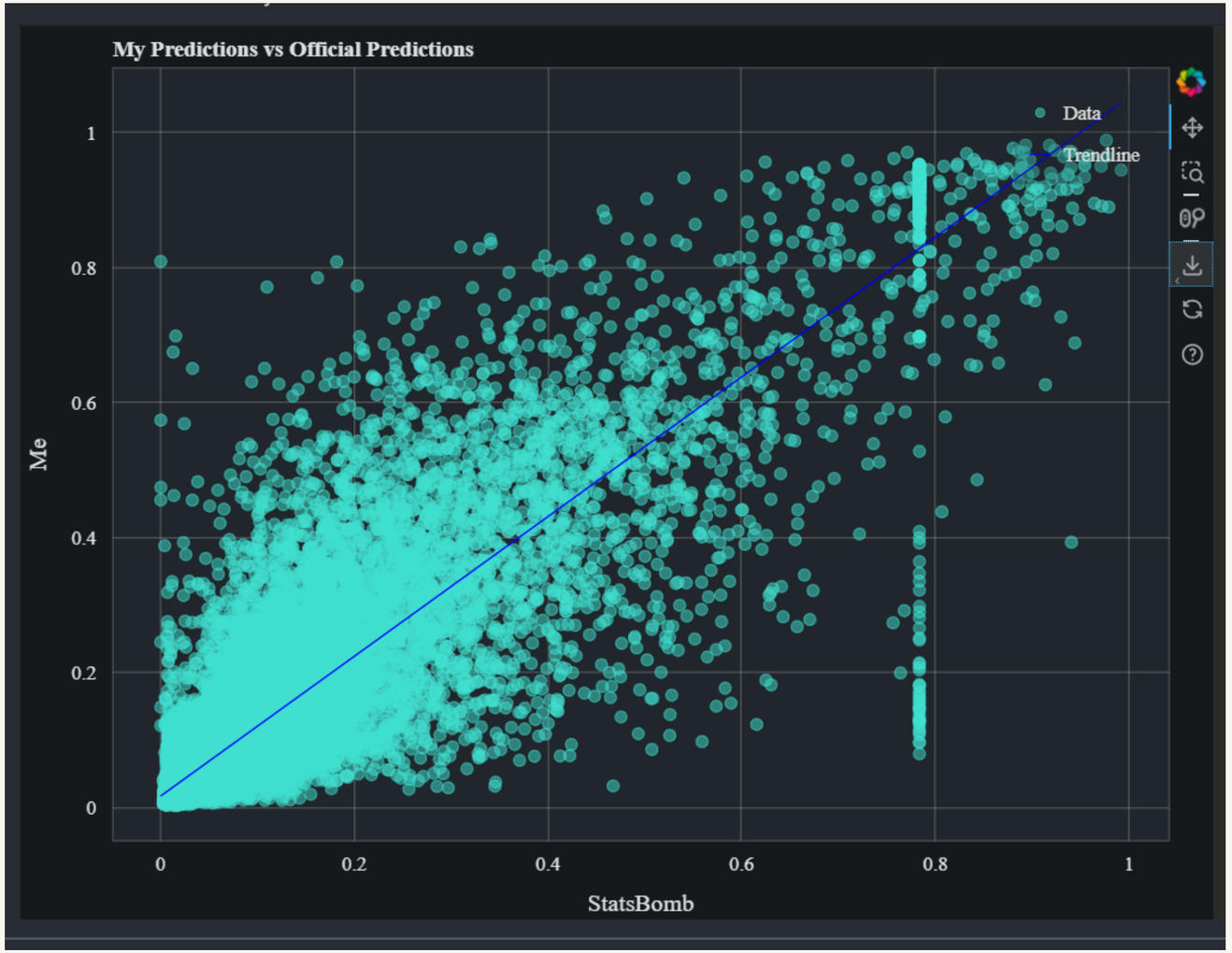
A trendline is worth a thousand rows



Visualizations have “impact”

```
In 30   1  pprint(list(zip(sb, [x[1] for x in gbt_xg])))

    (0.030884951, 0.028960792654741923),
    (0.16717885, 0.19885242023419034),
    (0.1706786, 0.09184453972297935),
    (0.06590014, 0.05933262335893341),
    (0.03487592, 0.06192678025923776),
    (0.006050177, 0.007101748932612316),
    (0.21281049, 0.49333134004161283),
    (0.018406134, 0.03293154851152055),
    (0.047982257, 0.04285772772047425),
    (0.1429732, 0.18915507075235233),
    (0.049008705, 0.08878594024612965),
    (0.0278987, 0.05769674039826831),
    (0.075503685, 0.10227964638959555),
    (0.040092945, 0.04169406181737197),
    (0.023332302, 0.01914528426543526),
    (0.03465536, 0.03079233267349761),
    (0.22949053, 0.44328667053484927),
    (0.006160782, 0.005327651859718585),
    (0.022147518, 0.025348764101981735),
    (0.00825277, 0.0647523159619699),
    (0.14014418, 0.10667029543535332),
    (0.016255273, 0.014603446834778858),
```



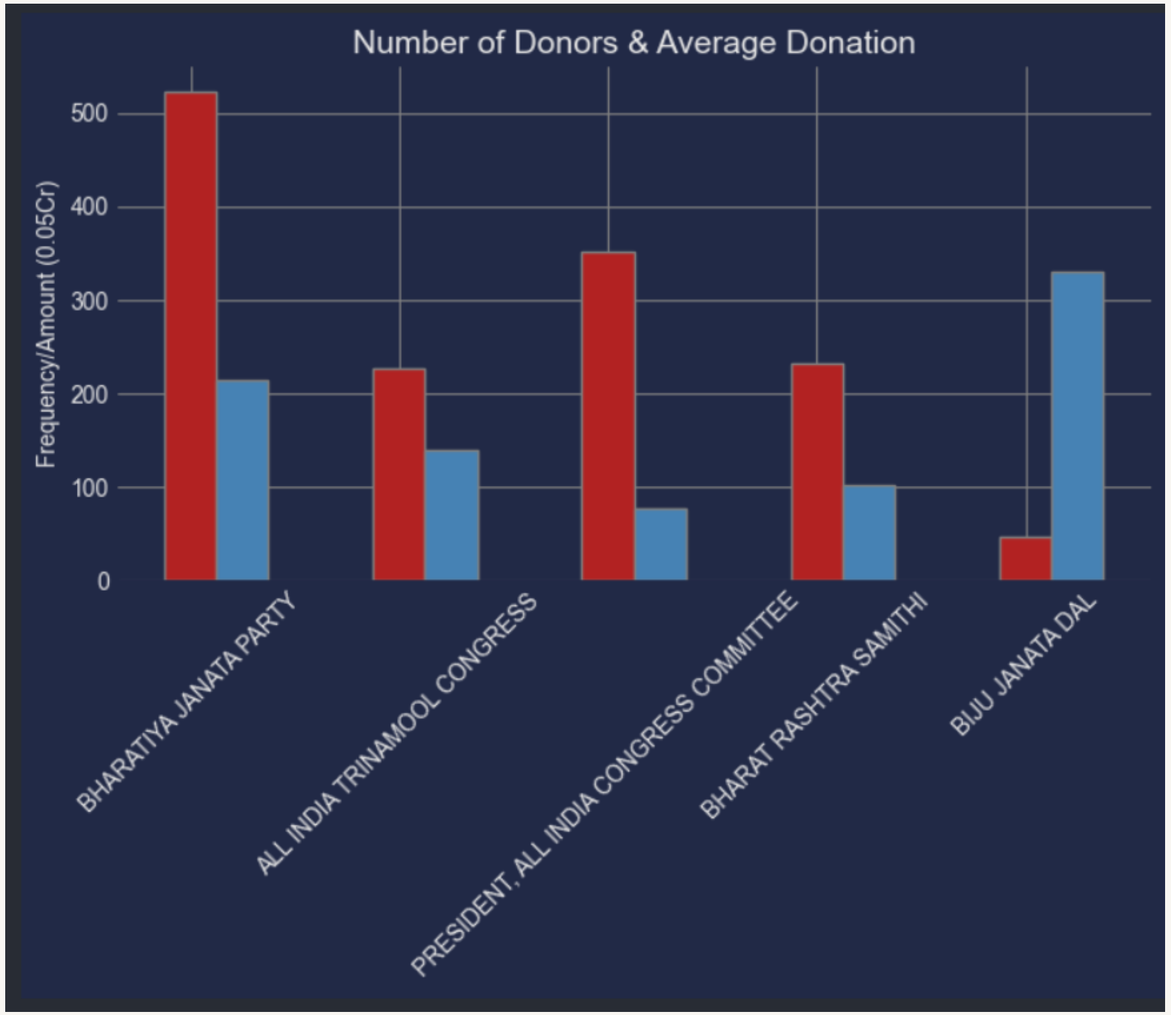
Visualizations are information-dense

```
In 27   1  parties_agg[["Buyer Count", "Buyer Average"]]
```

Out 27

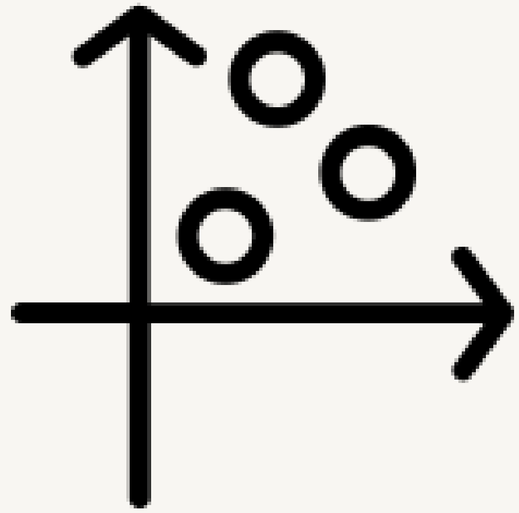
	Buyer Count	Buyer Average
0	523	10.696369
1	227	7.015513
2	351	3.849272
3	233	5.112274
4	47	16.500000
5	15	42.133333
6	52	6.322115
7	37	5.718378
8	35	4.355754
9	24	3.020833
10	55	1.186364
11	8	5.125000
12	6	6.083333
13	13	2.192308
14	5	4.200000
15	3	4.403333
16	7	1.785714
17	3	4.000000
18	7	1.037143

23 rows × 2 columns Open in new tab

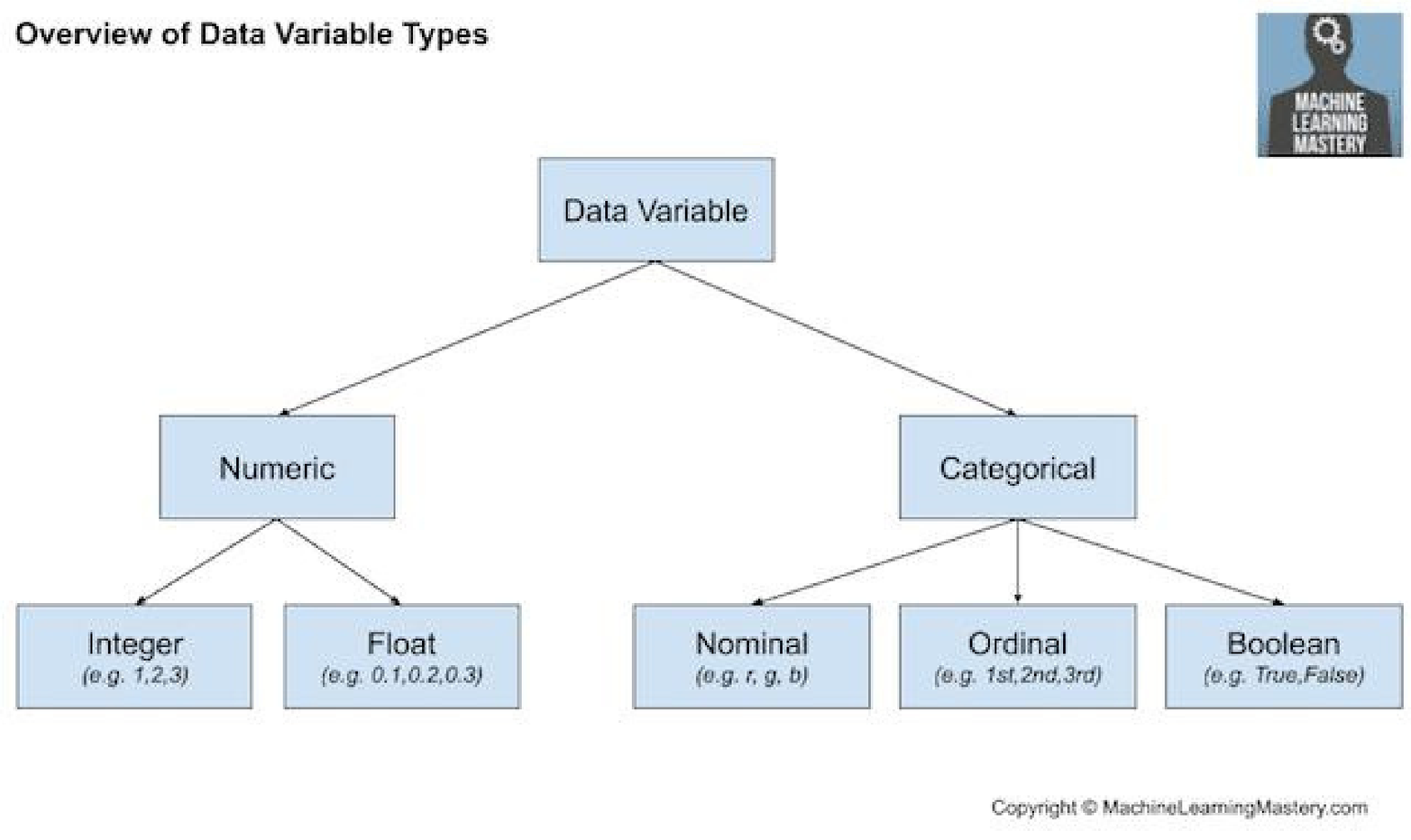


Part II (what)

Bread and butter



Data is either binary or not binary



Choice shapes choice

	A	B	C	D	E	F	G	H	I
1	Student Mess Meals								
2	Name	Date	Time	Meal	TKS	Batch	Major	Price	Count
3	Karnav Popat	26-03-2024	03:26:54	Lunch	FALSE	UG24	CS	80	58
4	Ayushman Roy	25-03-2024	03:31:40	Lunch	TRUE	UG27	Economics	120	23
5	Arushi Jain	26-03-2024	20:14:07	Dinner	FALSE	UG24	Psychology	80	61
6	Yash Gautam	22-03-2024	21:01:06	Dinner	TRUE	UG25	Economics	80	8
7	Jagrit Khatri	26-03-2024	22:14:20	Dinner	TRUE	UG25	English	120	95
8	Garbage	Numeric	Numeric	Categorica	Categorica	Categorica	Categorica	Numeric	Numeric

Staples of visualization

Impact

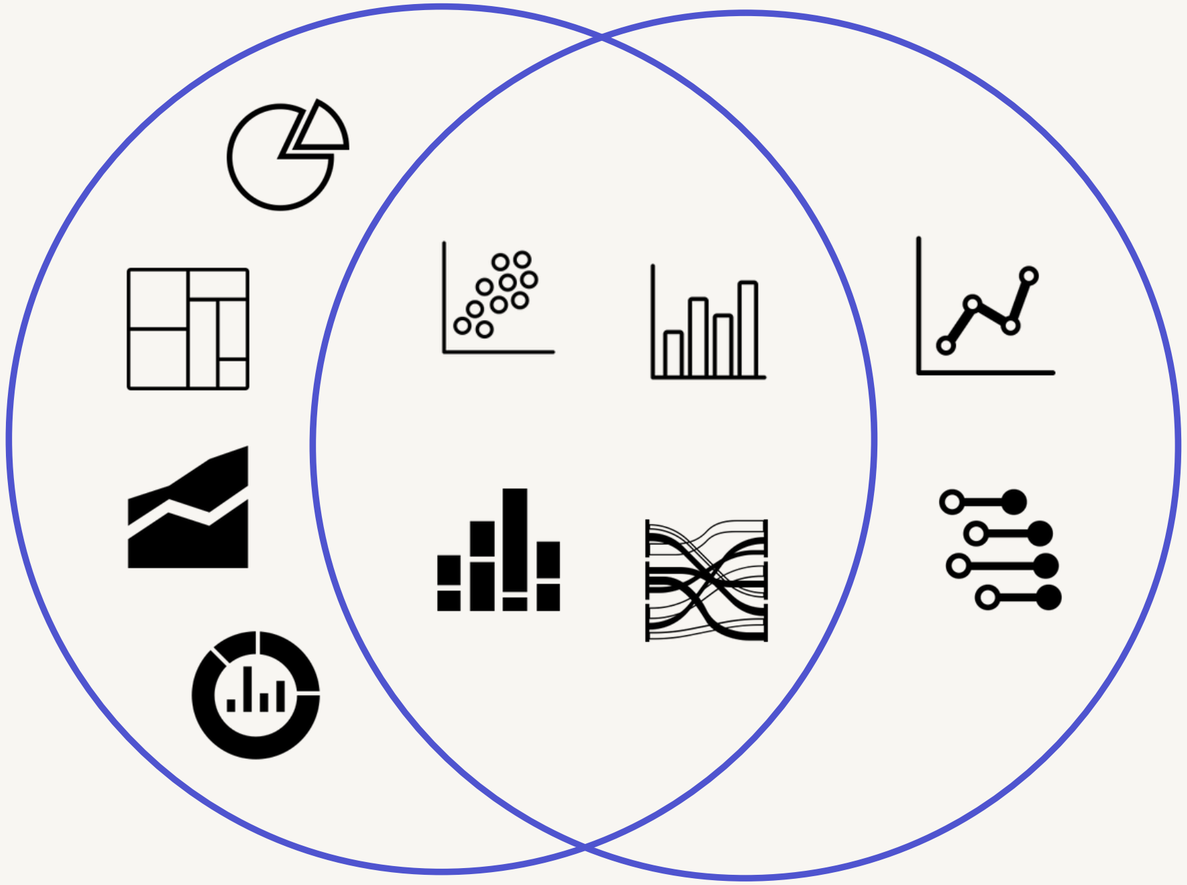
Information

Impact-Information Tradeoff

	Impact	Information	Feature Count	Feature Type
	High	Low	1	Numerical (Share%)
	High	High	2-4	Numerical + Categorical
	High	High	1+	Numerical
	Low	High	3+	Numerical + Categorical

Mix and match

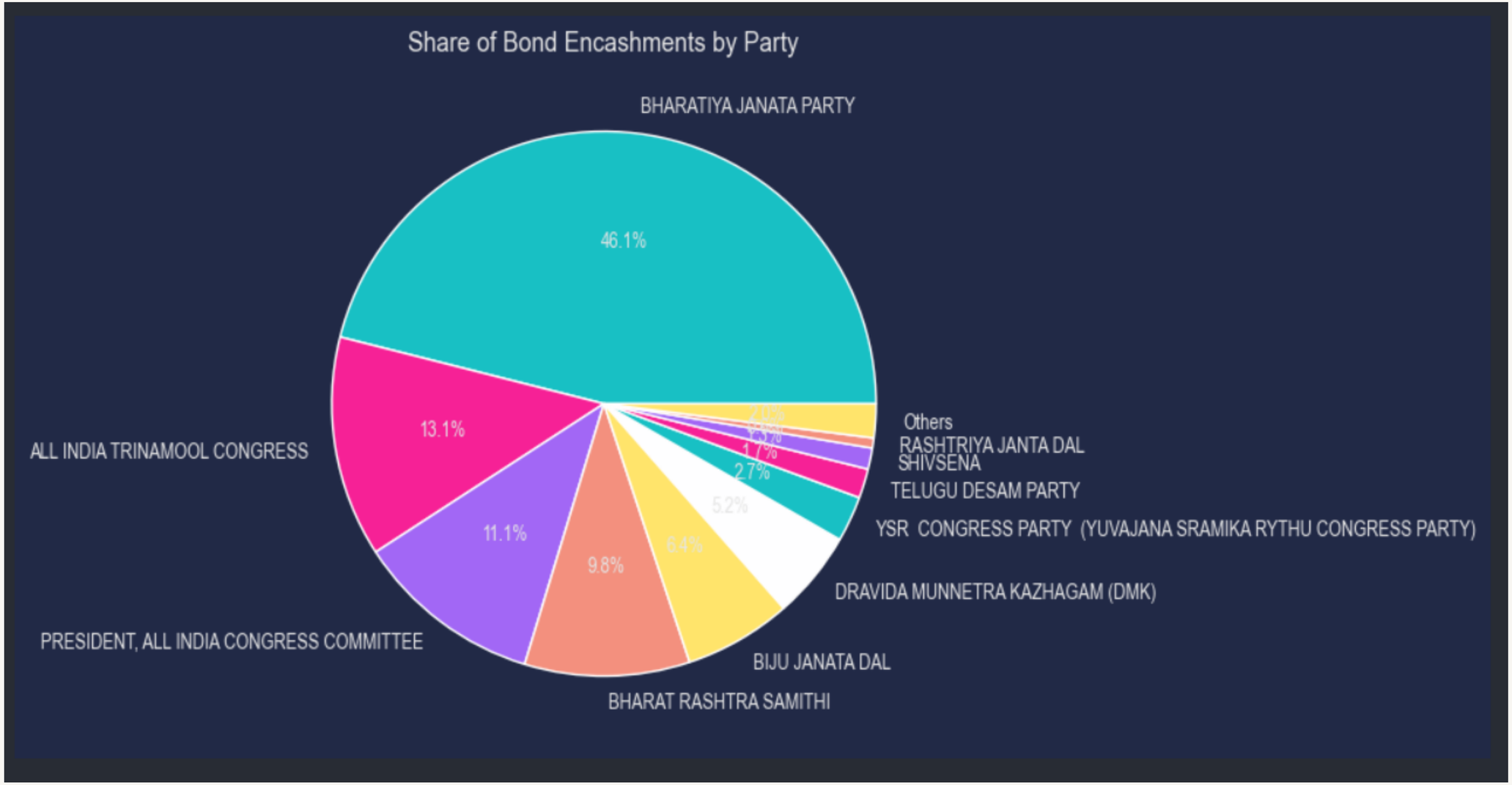
Impact

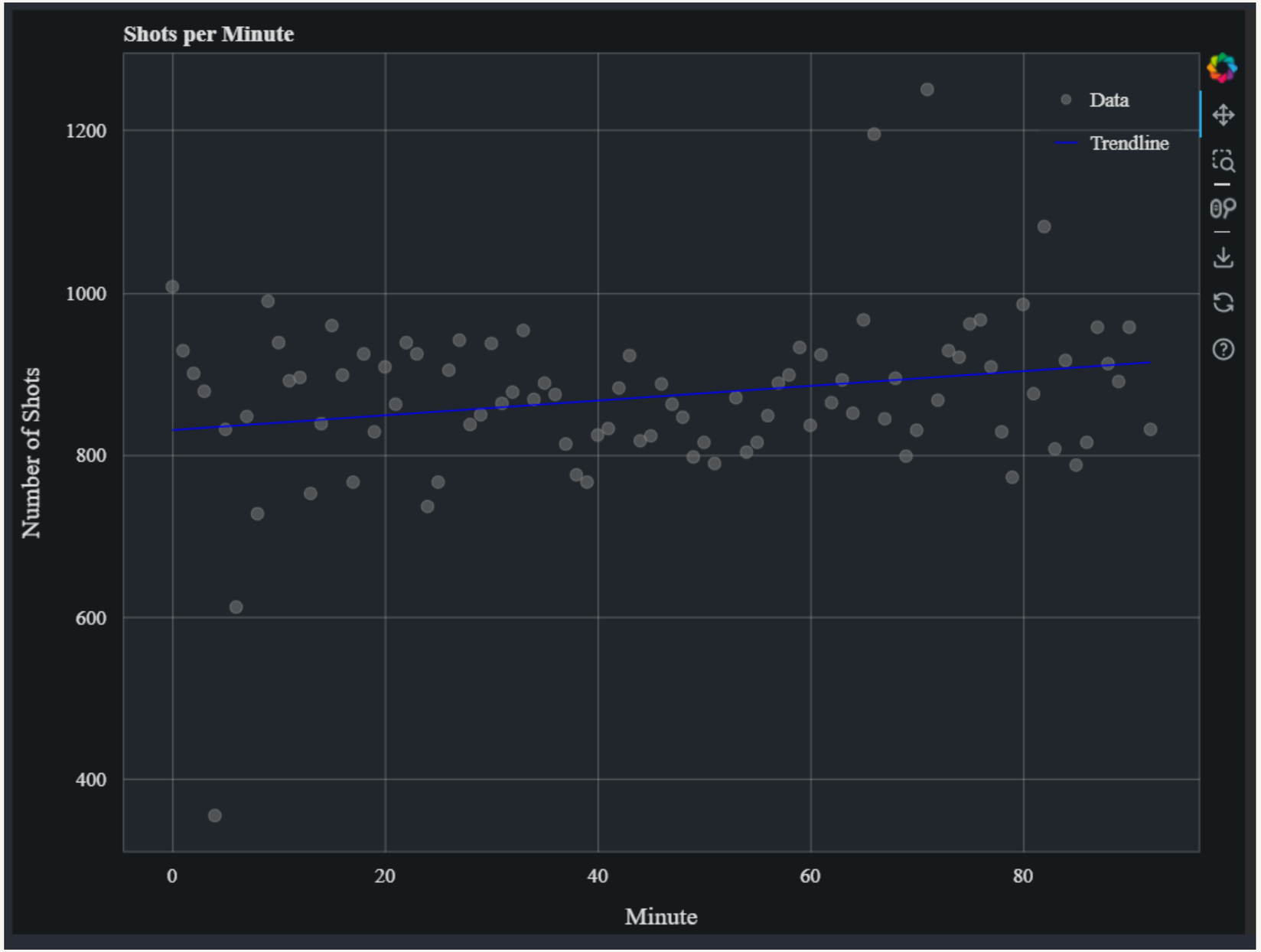


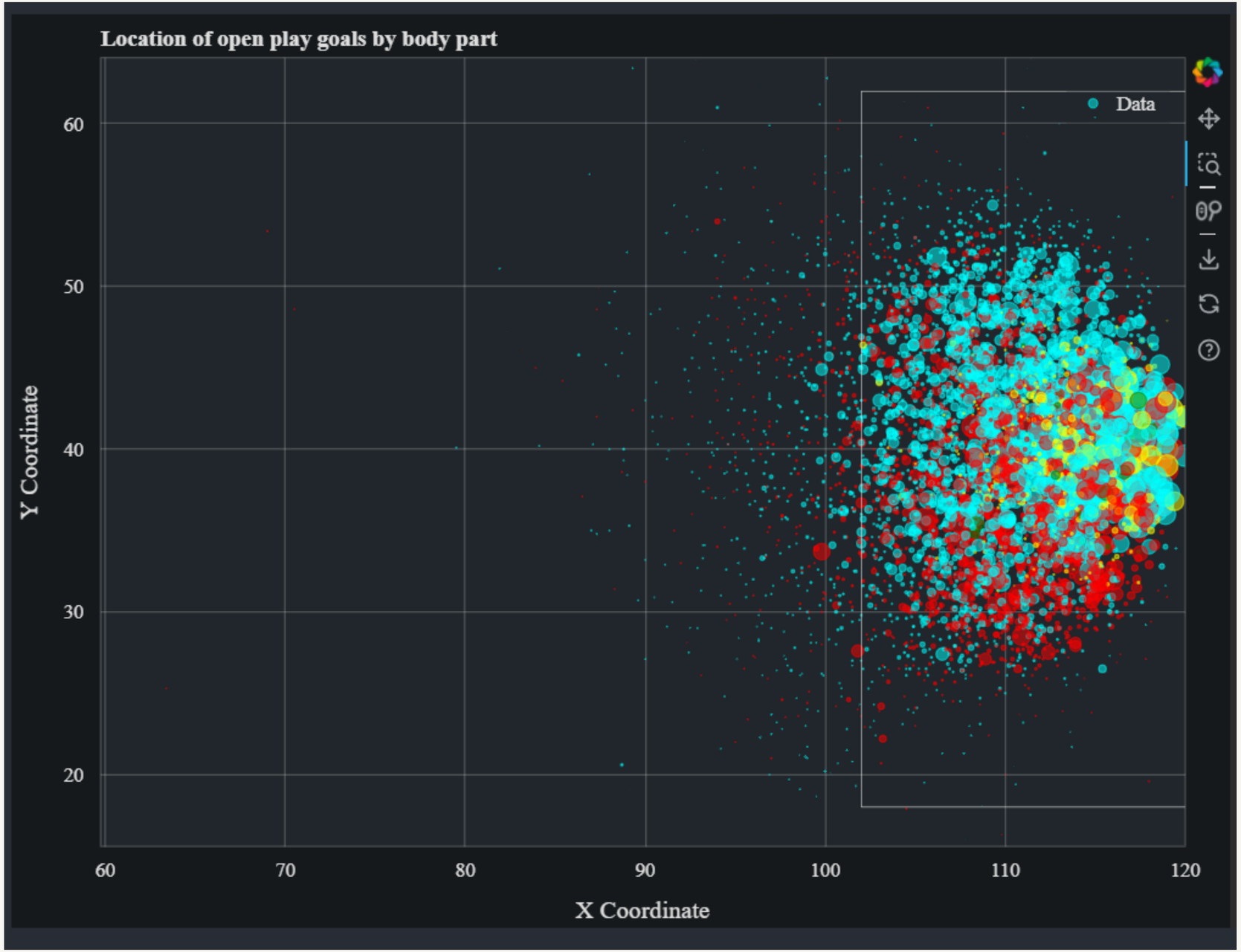
Information

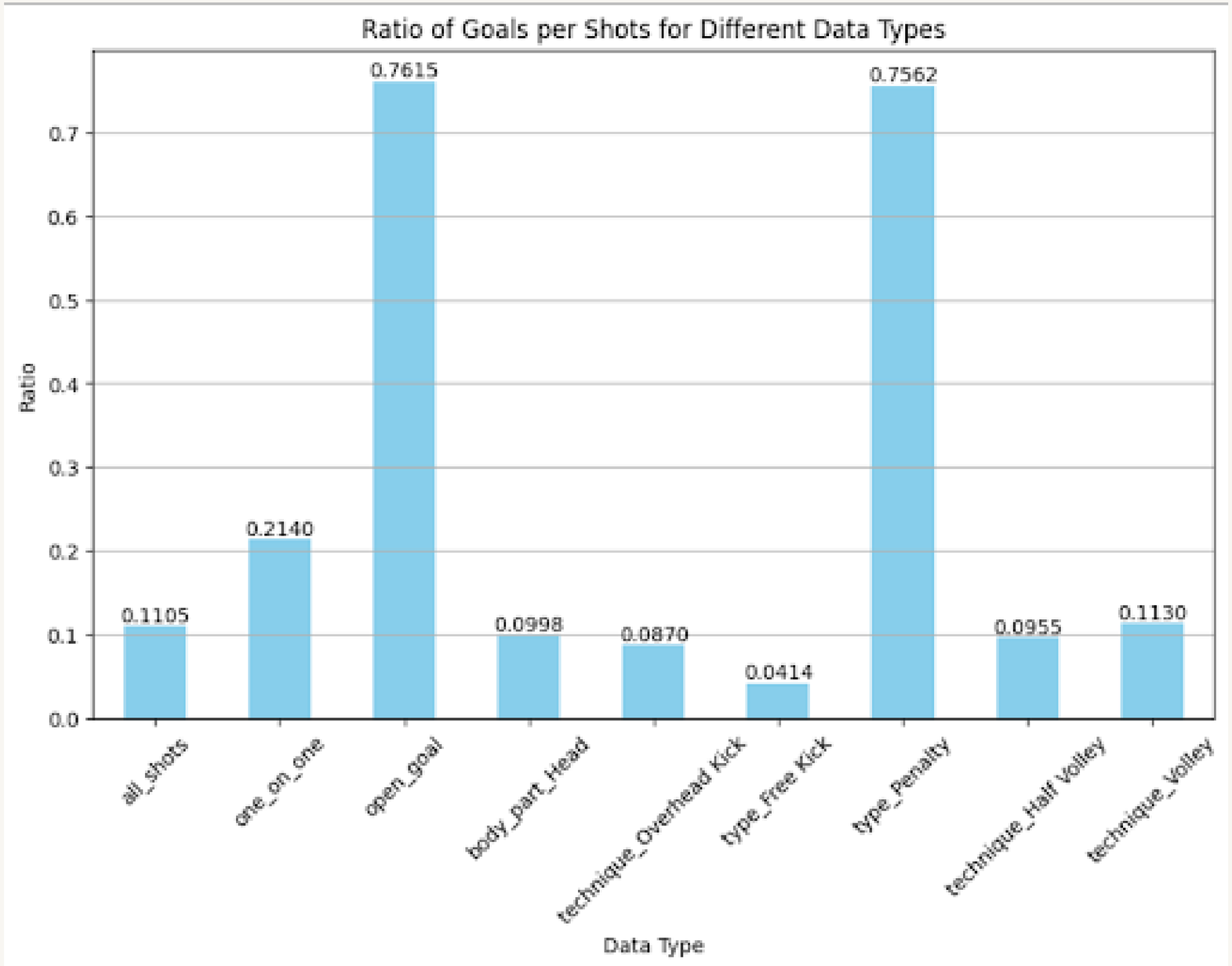
Part III

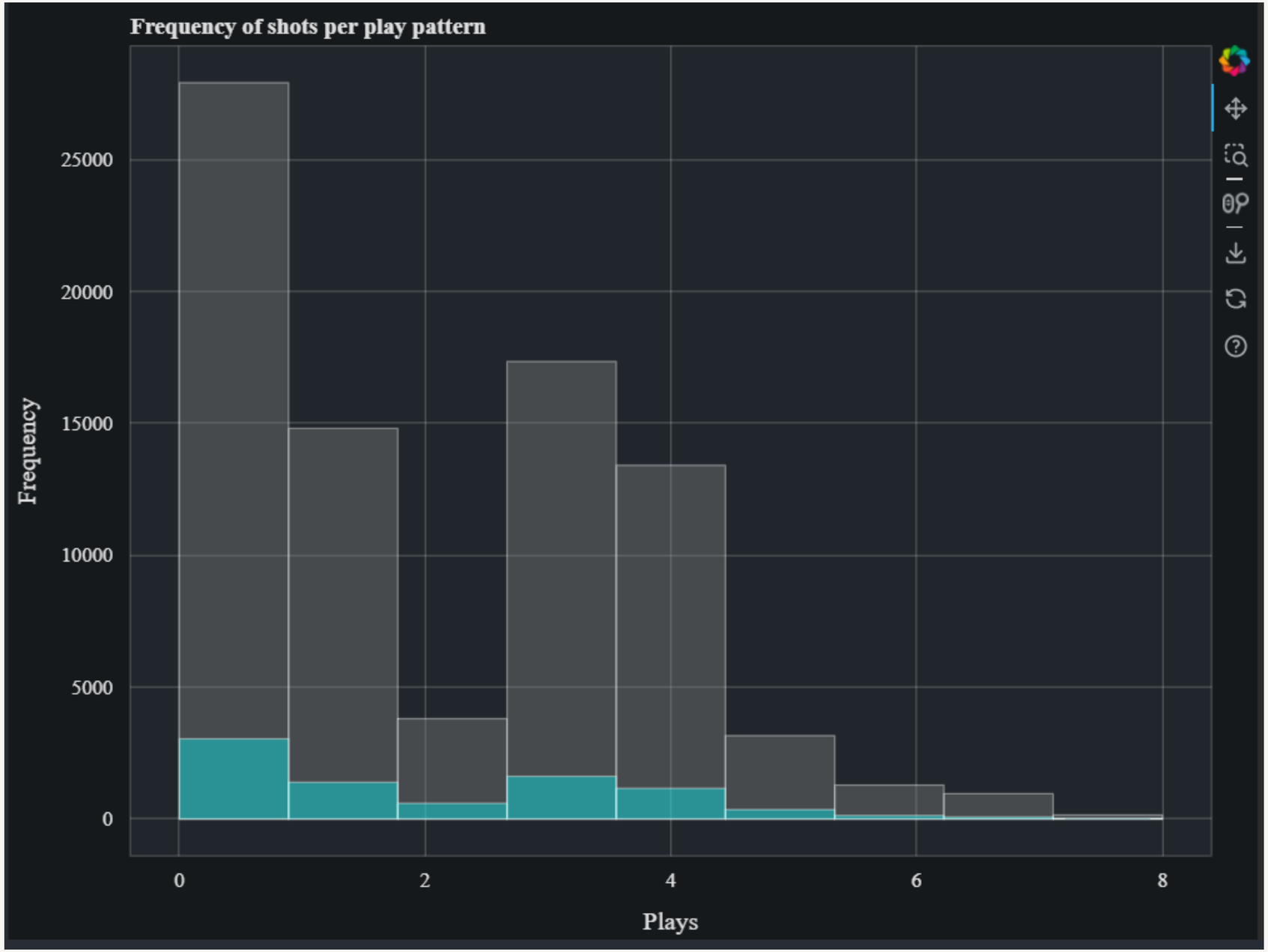
Bad graphs and good graphs

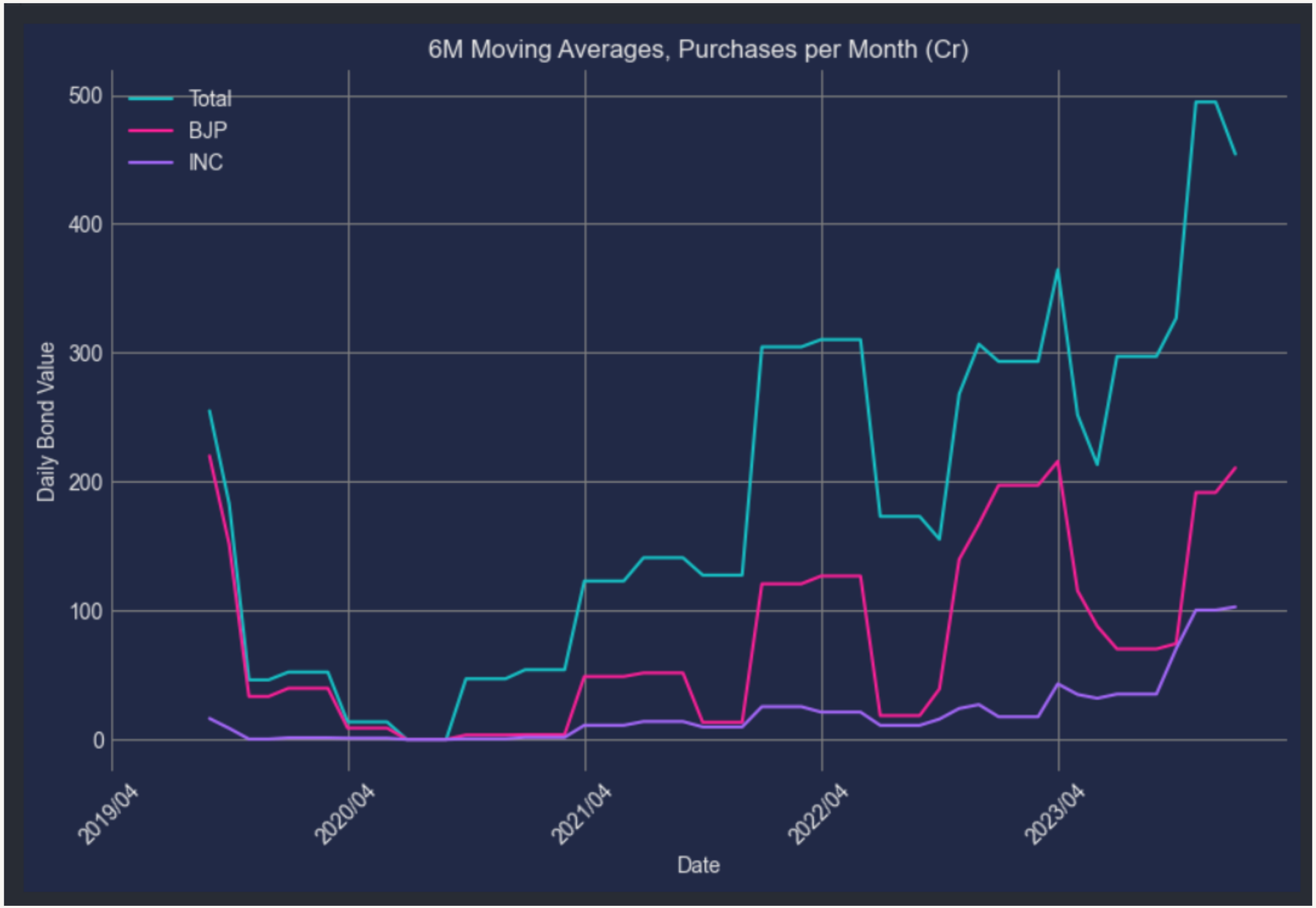

Share of Bond Encashments by Party
BHARATIYA JANATA PARTY
46.1%
ALL INDIA TRINAMOOL CONGRESS
13.1%
PRESIDENT, ALL INDIA CONGRESS COMMITTEE
11.1%
BHARAT RASHTRA SAMITHI
9.8%
BIJU JANATA DAL
6.4%
DRAVIDA MUNNETRA KAZHAGAM (DMK)
5.2%
YSR CONGRESS PARTY (YUVAJANA SRAMIKA RYTHU CONGRESS PARTY)
2.7%
TELUGU DESAM PARTY
SHIVSENA
RASHTRIYA JANTA DAL
Others


Shots per Minute
Data
Trendline
1200
1000
800
600
400
Number of Shots
0
20
40
60
80
Minute


Location of open play goals by body part
Data
Y Coordinate
60
50
40
30
20
60
70
80
90
100
110
120
X Coordinate


Ratio of Goals per Shots for Different Data Types
Ratio
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.1105
0.2140
0.7615
0.0998
0.0870
0.0414
0.7562
0.0955
0.1130
all_shots
one_on_one
open_goal
body_part_Head
technique_Overhead Kick
type_Free Kick
type_Penalty
technique_Half Volley
technique_Volley
Data Type


Frequency of shots per play pattern
25000
20000
15000
10000
5000
0
Frequency
0
2
4
6
8
Plays

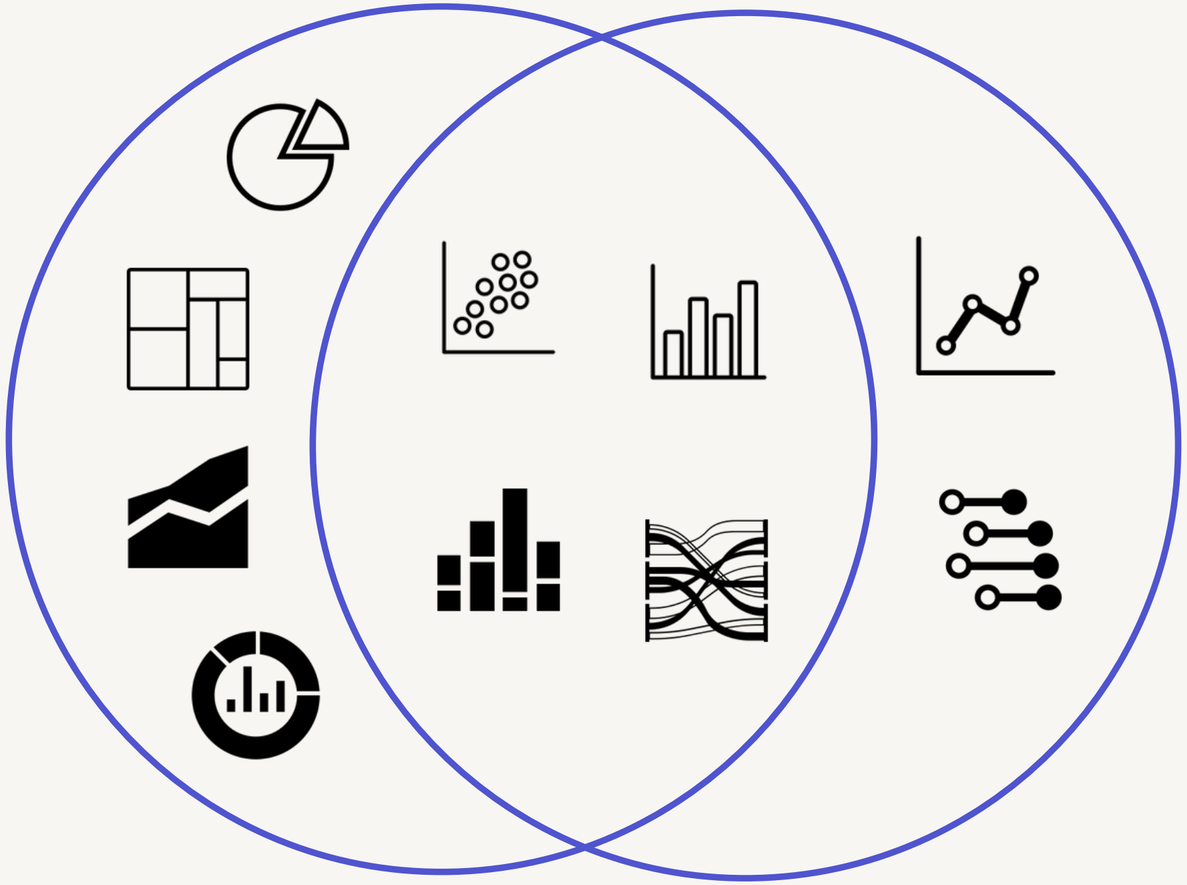

6M Moving Averages, Purchases per Month (Cr)
Total
BJP
INC
500
400
300
200
100
0
Daily Bond Value
2019/04
2020/04
2021/04
2022/04
2023/04
Date

Are all graphs numbers?

	Impact	Information	Feature Count	**Feature Type**
	High	Low	1	**Numerical** (Share%)
	High	High	2-4	**Numerical**, Categorical
	High	High	1+	**Numerical**
	Low	High	3+	**Numerical**, Categorical

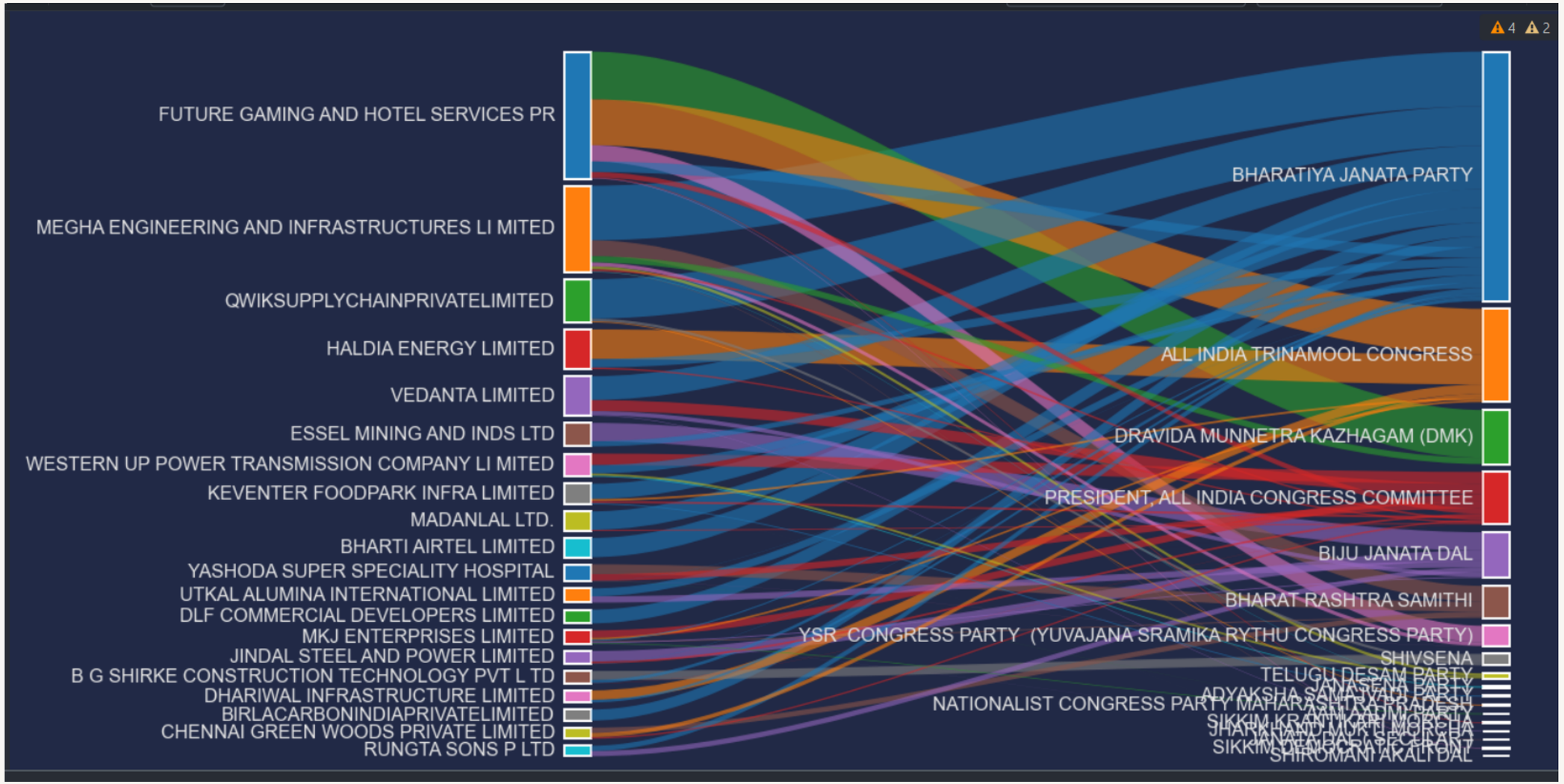
Mapping categories to categories

Impact

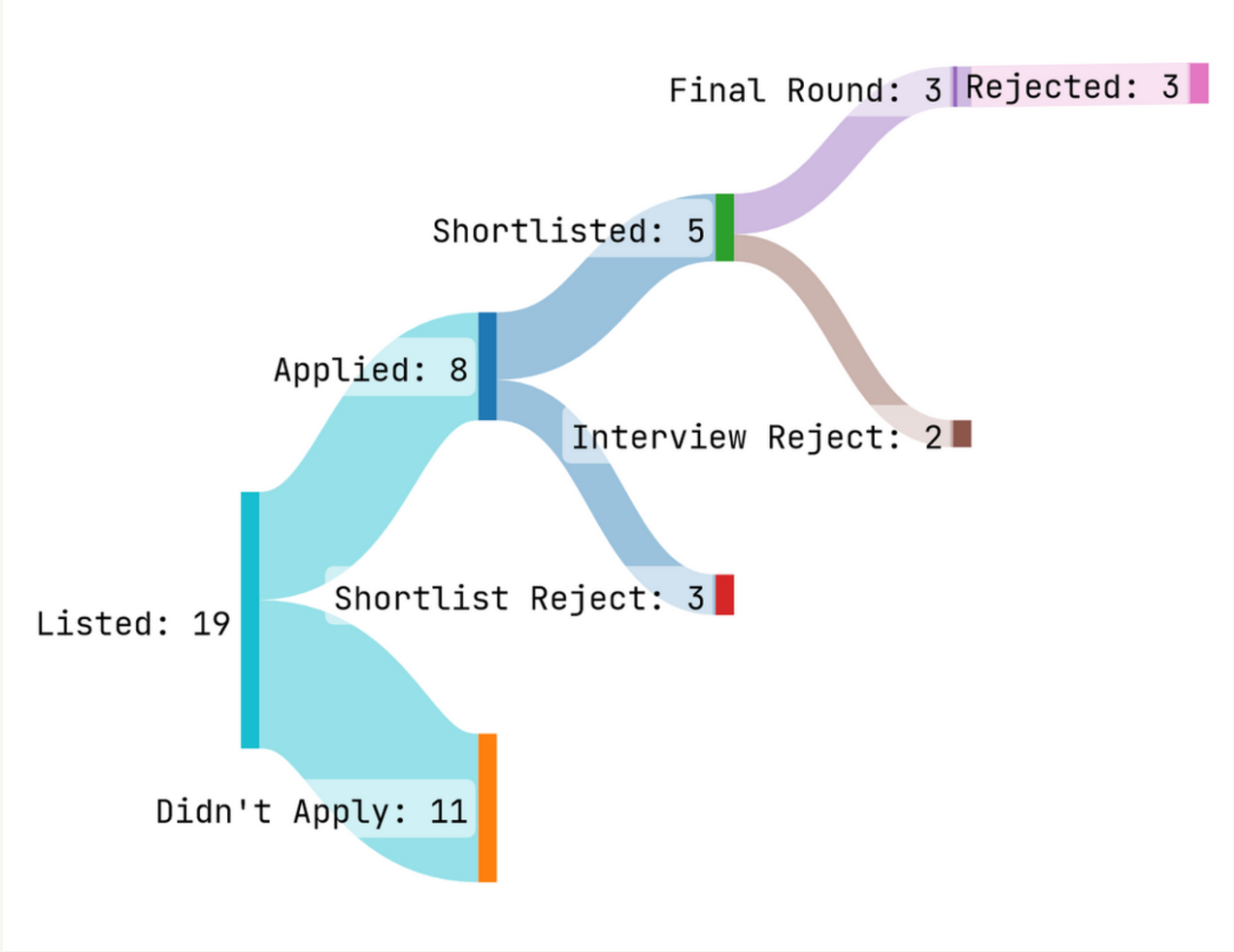


Information

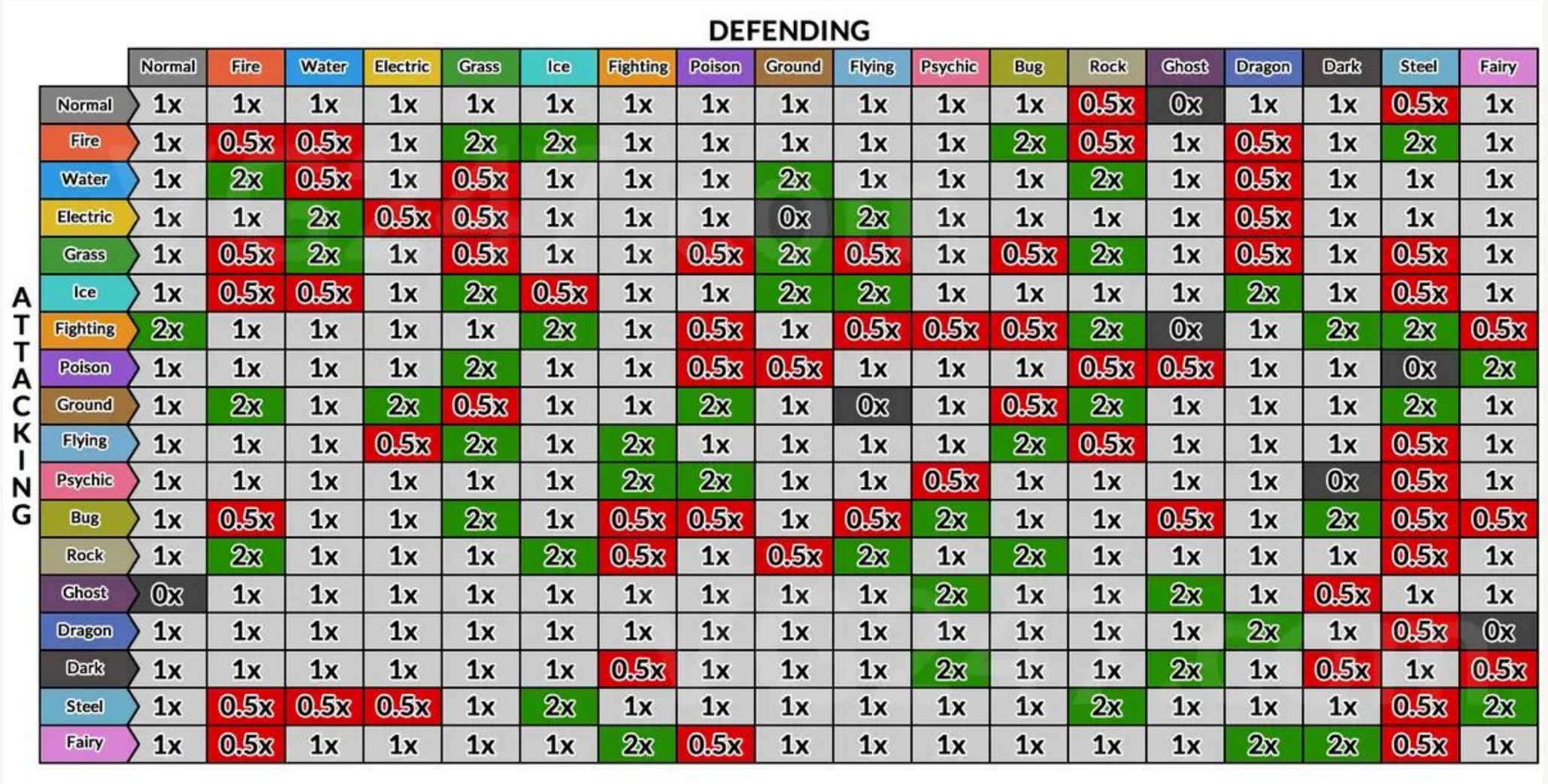
Mapping categories to categories



Mapping categories to categories



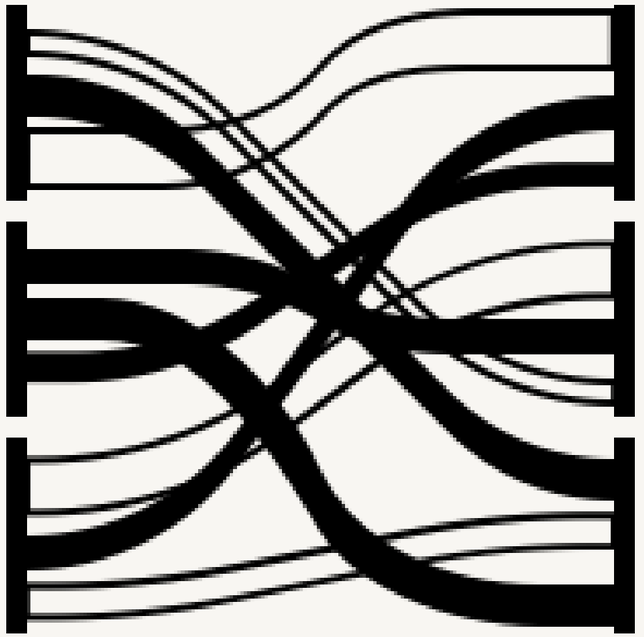
Mapping categories to categories



ATTACKING \ DEFENDING	Normal	Fire	Water	Electric	Grass	Ice	Fighting	Poison	Ground	Flying	Psychic	Bug	Rock	Ghost	Dragon	Dark	Steel	Fairy
Normal	1x	1x	1x	1x	1x	1x	1x	1x	1x	1x	1x	1x	0.5x	0x	1x	1x	0.5x	1x
Fire	1x	0.5x	0.5x	1x	2x	2x	1x	1x	1x	1x	1x	2x	0.5x	1x	0.5x	1x	2x	1x
Water	1x	2x	0.5x	1x	0.5x	1x	1x	1x	2x	1x	1x	1x	2x	1x	0.5x	1x	1x	1x
Electric	1x	1x	2x	0.5x	0.5x	1x	1x	1x	0x	2x	1x	1x	1x	1x	0.5x	1x	1x	1x
Grass	1x	0.5x	2x	1x	0.5x	1x	1x	0.5x	2x	0.5x	1x	0.5x	2x	1x	0.5x	1x	0.5x	1x
Ice	1x	0.5x	0.5x	1x	2x	0.5x	1x	1x	2x	2x	1x	1x	1x	1x	2x	1x	0.5x	1x
Fighting	2x	1x	1x	1x	1x	2x	1x	0.5x	1x	0.5x	0.5x	0.5x	2x	0x	1x	2x	2x	0.5x
Poison	1x	1x	1x	1x	2x	1x	1x	0.5x	0.5x	1x	1x	1x	0.5x	0.5x	1x	1x	0x	2x
Ground	1x	2x	1x	2x	0.5x	1x	1x	2x	1x	0x	1x	0.5x	2x	1x	1x	1x	2x	1x
Flying	1x	1x	1x	0.5x	2x	1x	2x	1x	1x	1x	1x	2x	0.5x	1x	1x	1x	0.5x	1x
Psychic	1x	1x	1x	1x	1x	1x	2x	2x	1x	1x	0.5x	1x	1x	1x	1x	0x	0.5x	1x
Bug	1x	0.5x	1x	1x	2x	1x	0.5x	0.5x	1x	0.5x	2x	1x	1x	0.5x	1x	2x	0.5x	0.5x
Rock	1x	2x	1x	1x	1x	2x	0.5x	1x	0.5x	2x	1x	2x	1x	1x	1x	1x	0.5x	1x
Ghost	0x	1x	1x	1x	1x	1x	1x	1x	1x	1x	2x	1x	1x	2x	1x	0.5x	1x	1x
Dragon	1x	1x	1x	1x	1x	1x	1x	1x	1x	1x	1x	1x	1x	1x	2x	1x	0.5x	0x
Dark	1x	1x	1x	1x	1x	1x	0.5x	1x	1x	1x	2x	1x	1x	2x	1x	0.5x	1x	0.5x
Steel	1x	0.5x	0.5x	0.5x	1x	2x	1x	1x	1x	1x	1x	1x	2x	1x	1x	1x	0.5x	2x
Fairy	1x	0.5x	1x	1x	1x	1x	2x	0.5x	1x	1x	1x	1x	1x	1x	2x	2x	0.5x	1x

Part IV

Messy dirty spaghetti code



Libraries

- **Matplotlib.PyPlot** (Python default)
- **Seaborn** (Matplotlib styling)
- **Bokeh** (Interactive and more customizations)
- **SankeyFlow** (Sankey diagrams in Python)
- **ggplot2** (R and Python)

The staples (dataviz.py)

```
43
44  plt.scatter(football_data["location_x_distance"], football_data["location_y_distance"])
45  plt.show()
46
47  plt.pie(parties["Denominations"], labels=parties["Party"])
48  plt.scatter(football_data["location_x"], football_data["location_y"])
49  football_data["statsbomb_xg"].hist(bins=10, xlabelsize=10, ylabelsize=6, color="cyan")
50  plt.plot(copy["minute"], copy["possession"])
51
```

Structure

Thank you!

Get in touch @ karnav.popat_ug24@ashoka.edu.in