YIEPI meets DNA: on vehicle datasets analysis

on-going progress, findings and record-matching strategies

YIEPI meets DNA:
on vehicle datasets analysis

let's go!

vehicle datasets analysis

we all want to identify sold vehicles...

identify

we all want to

ok, a glimpse on the data!

vehicles...

± 260K

b2b data

± 6.25Mio.

b2c data

b2b not even
5% of the volume!

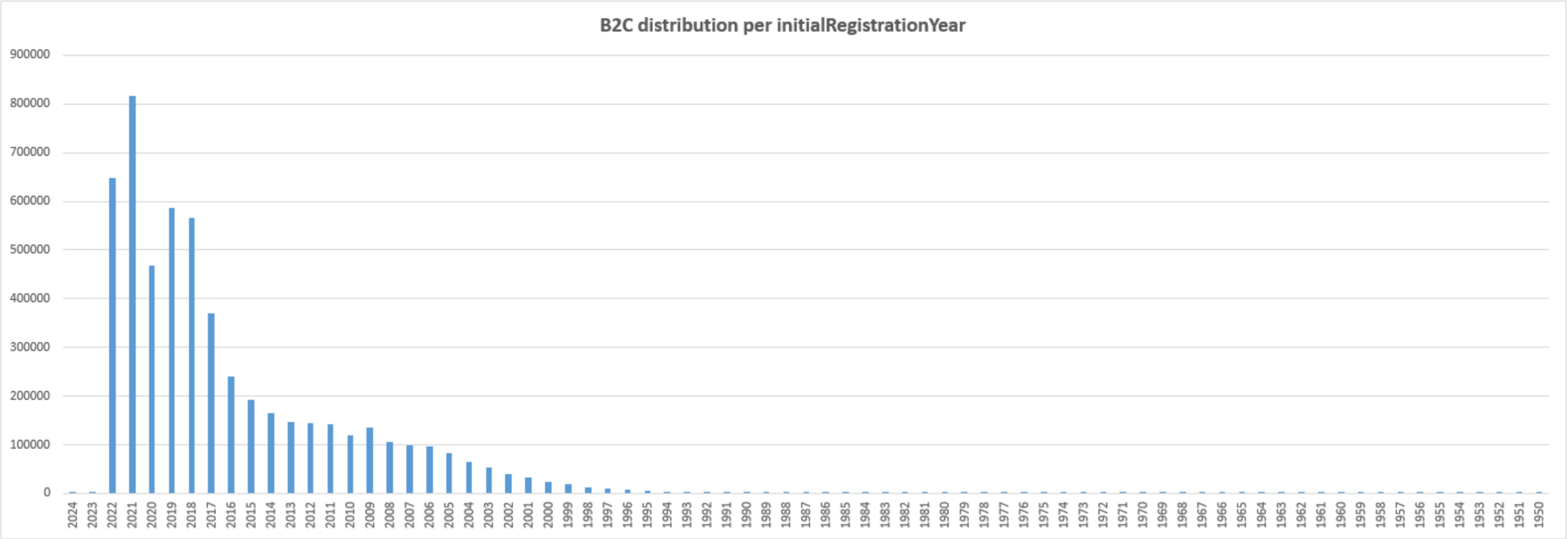
± 260K

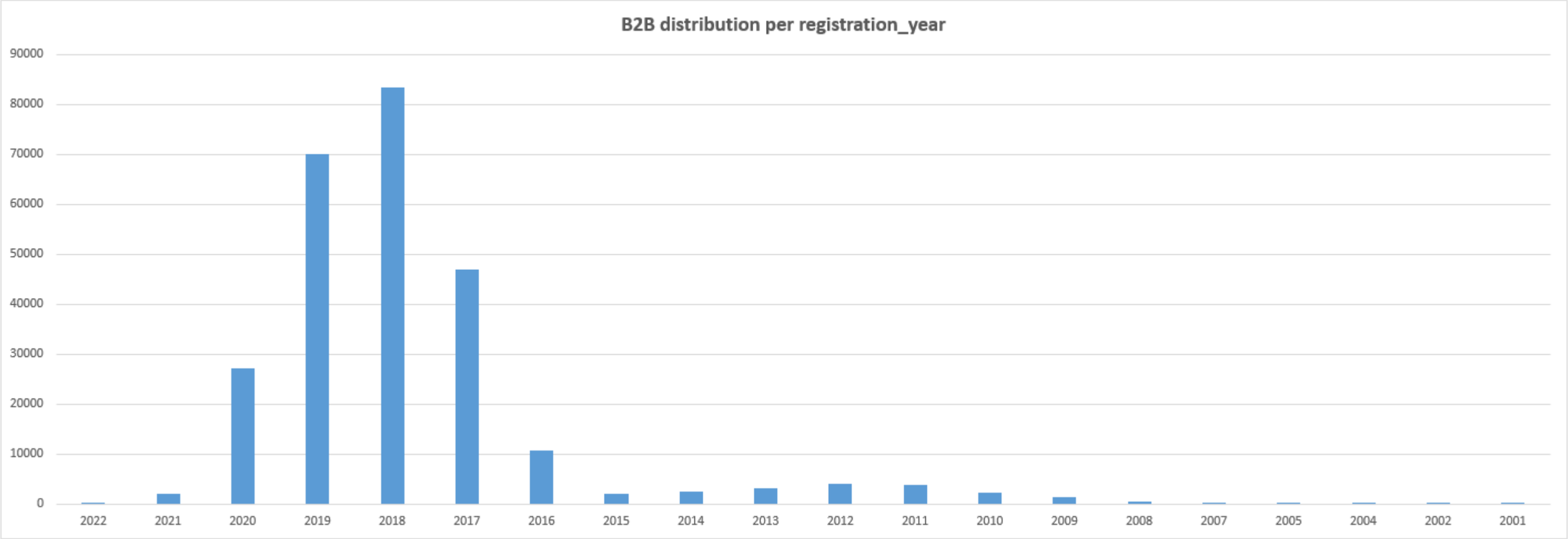
b2b data

± 6.25Mio.

b2c data

some other quick insights...


B2C distribution per initialRegistrationYear
900000
800000
700000
600000
500000
400000
300000
200000
100000
0
2024 2023 2022 2021 2020 2019 2018 2017 2016 2015 2014 2013 2012 2011 2010 2009 2008 2007 2006 2005 2004 2003 2002 2001 2000 1999 1998 1997 1996 1995 1994 1993 1992 1991 1990 1989 1988 1987 1986 1985 1984 1983 1982 1981 1980 1979 1978 1977 1976 1975 1974 1973 1972 1971 1970 1969 1968 1967 1966 1965 1964 1963 1962 1961 1960 1959 1958 1957 1956 1955 1954 1953 1952 1951 1950


B2B distribution per registration_year
90000
80000
70000
60000
50000
40000
30000
20000
10000
0
2022
2021
2020
2019
2018
2017
2016
2015
2014
2013
2012
2011
2010
2009
2008
2007
2005
2004
2002
2001

hey, what about the vin number?

hey, what about the vin number?

null in approx. 85% of b2c records!

we all want to identify sold vehicles...

we all want to

ok, so let’s match the vins!

vehicles...

± 60K

matching vins

± 260K

b2b data

± 6.25Mio.

b2c data

± 60K

matching vins

23%, and around 1.000 duplicated

± 260K

b2b data

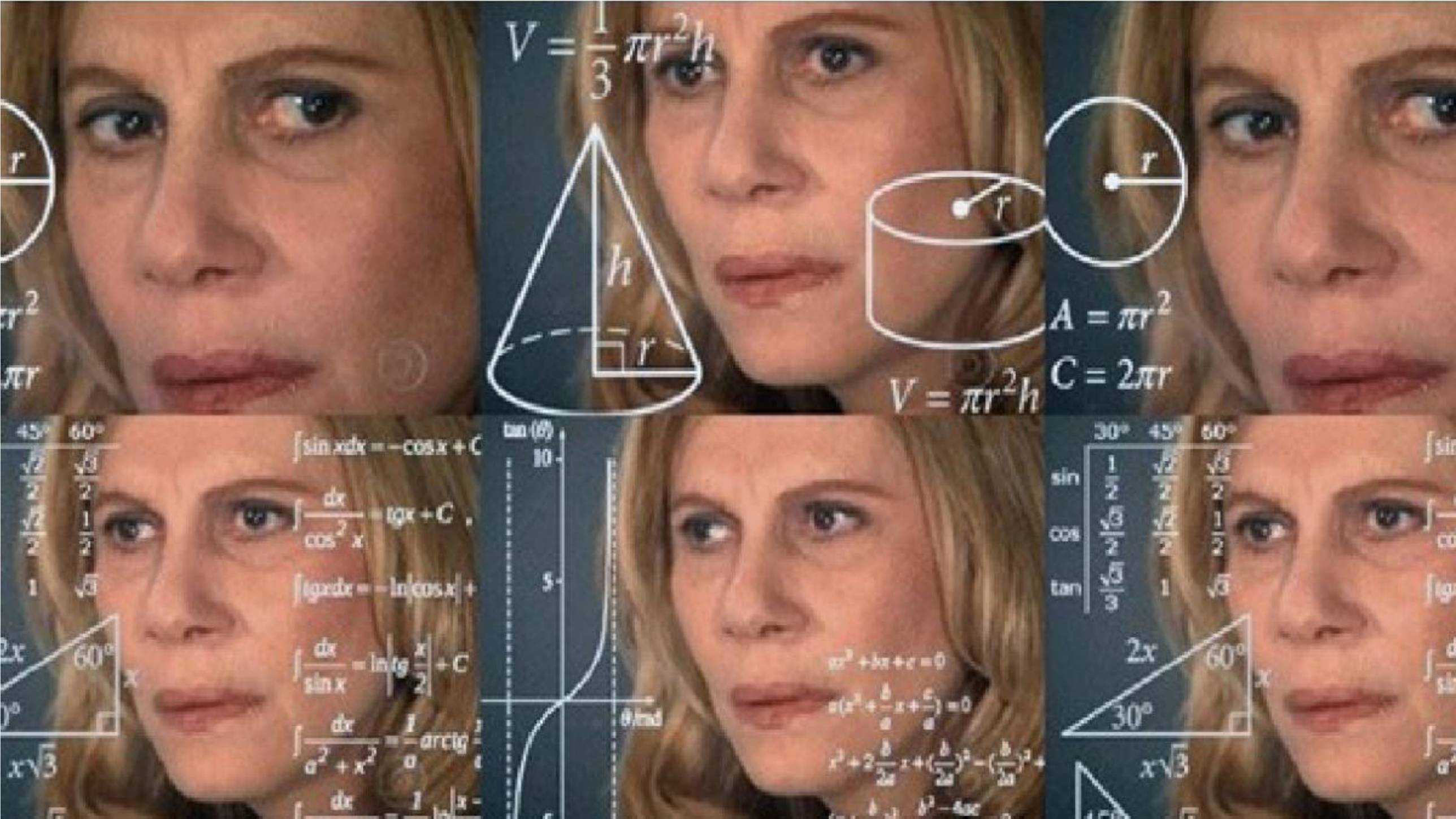
± 6.25Mio.

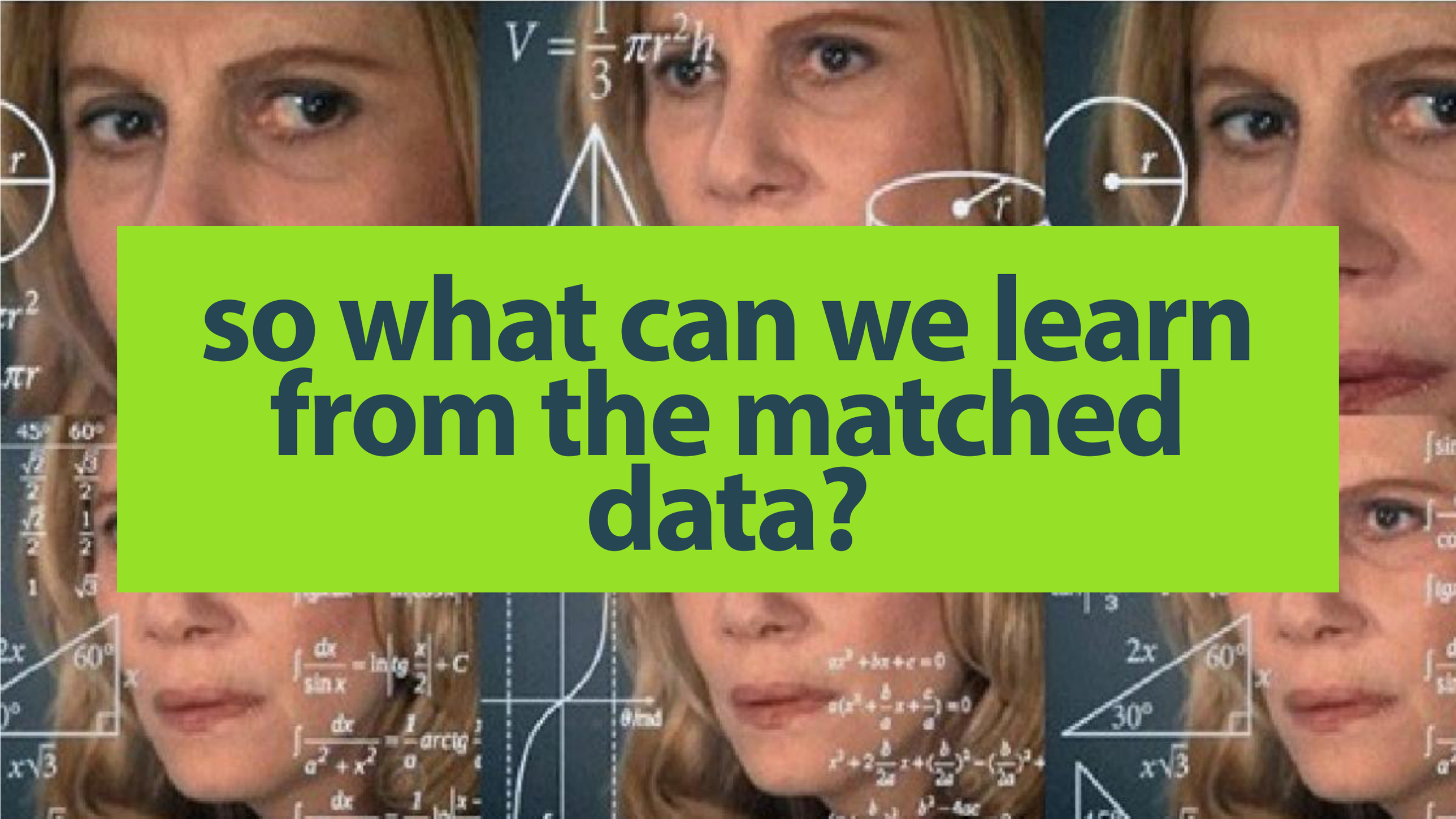
b2c data

ok, Super VIN is our first hero

ok, Super VIN is our first hero

but we still have 75% of the b2b left to tackle


$V = \frac{1}{3}\pi r^2 h$
$A = \pi r^2$
$C = 2\pi r$
$V = \pi r^2 h$
$\int \sin x dx = -\cos x + C$
$\int \frac{dx}{\cos^2 x} = tgx + C$
$\int \frac{dx}{\sin x} = \ln \left| tg \frac{x}{2} \right| + C$
$ax^2 + bx + c = 0$
$a(x^2 + \frac{b}{a}x + \frac{c}{a}) = 0$
$x^2 + 2\frac{b}{2a}x + (\frac{b}{2a})^2 - (\frac{b}{2a})^2 +$
$\frac{b^2 - 4ac}{}$
sin
cos
tan
30° 45° 60°
2x
x√3



so what can we learn from the matched data?

± 60K

matched vins

± 60K

matched vins

let's call it our "golden dataset"

quick hypotheses:

quick hypotheses:

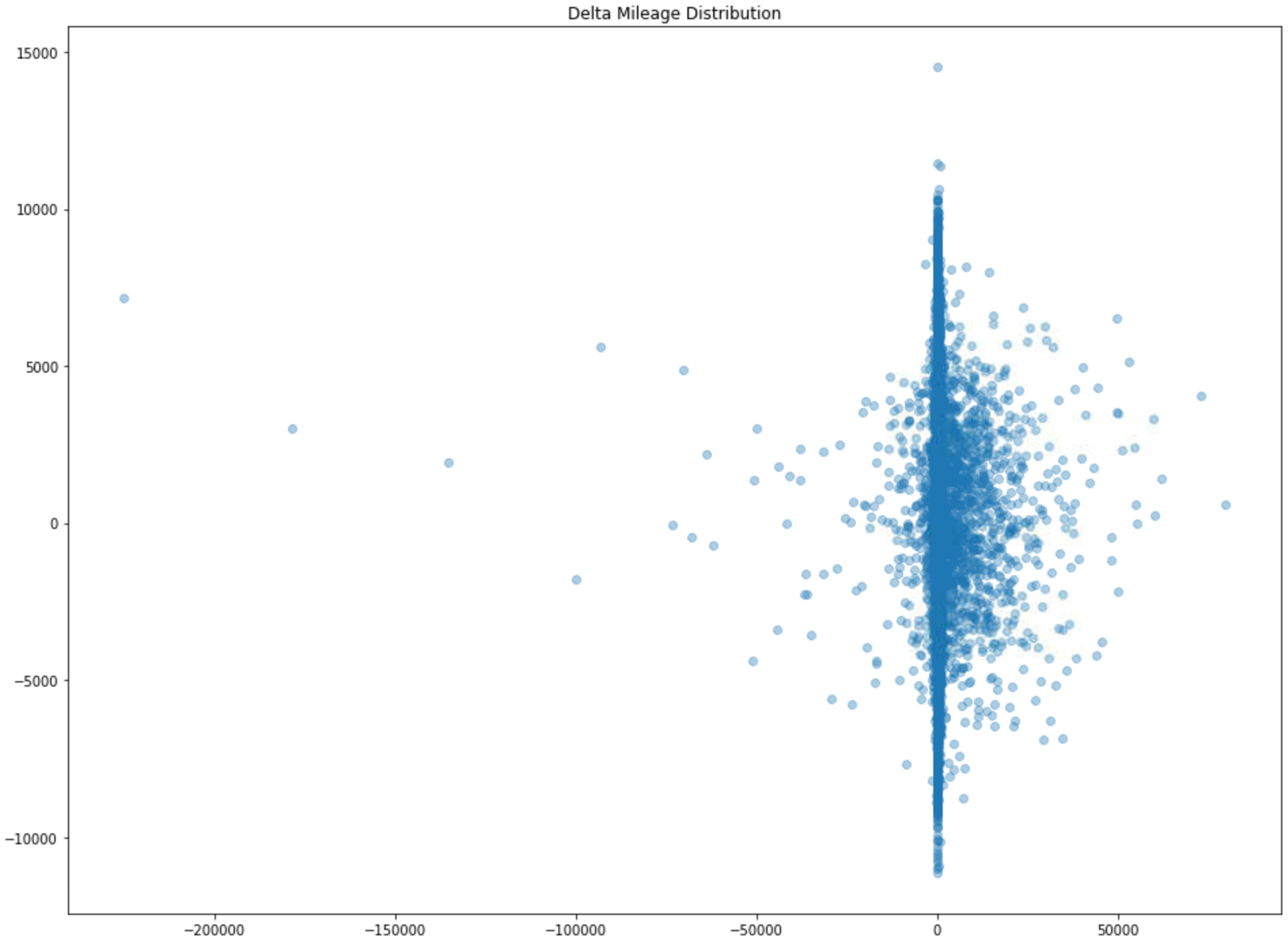
- maybe model matches;

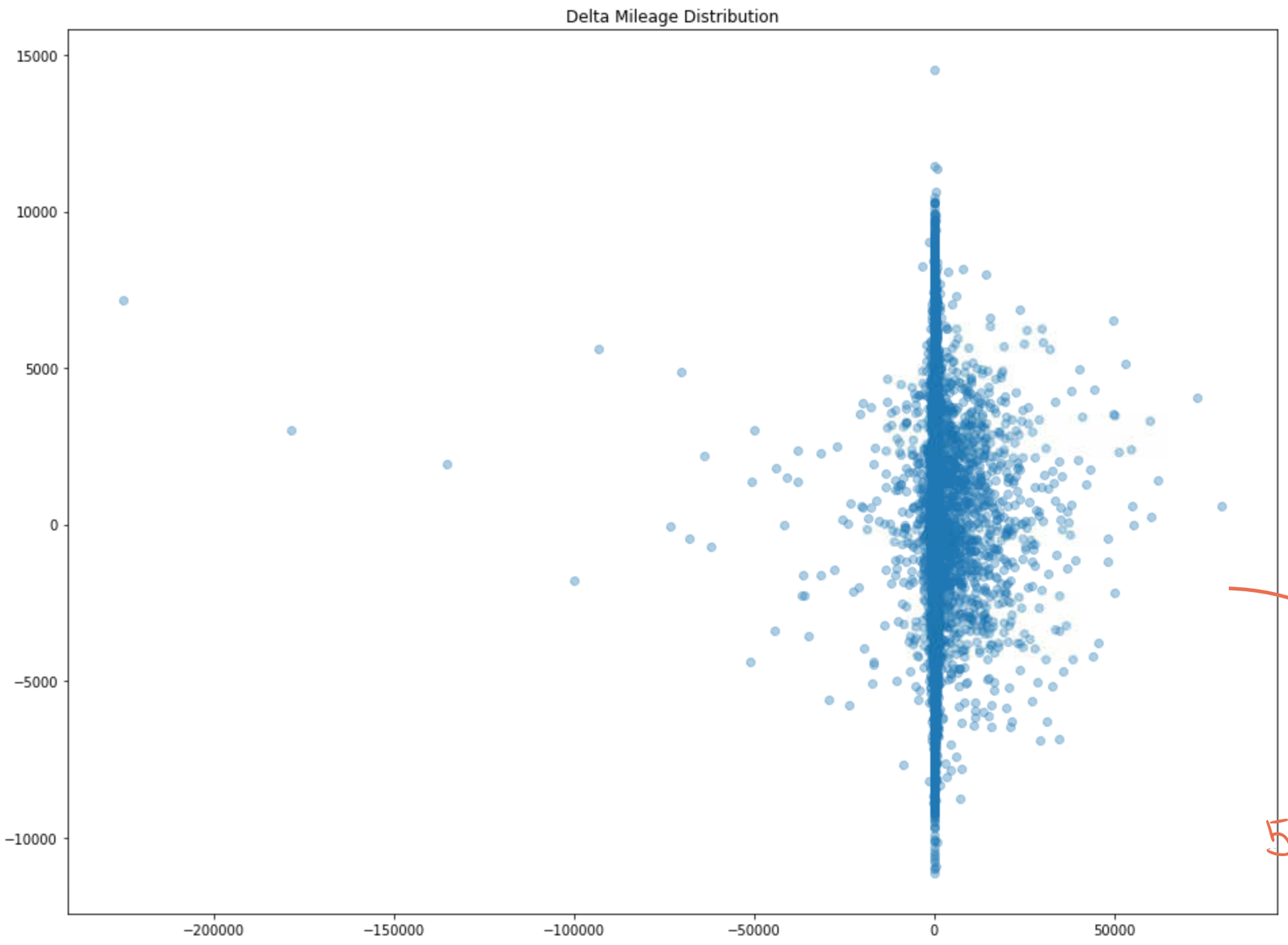
quick hypotheses:
- maybe model matches;

98.70% of the cases!

quick hypotheses:

- maybe model matches;
- maybe mileage matches;


Delta Mileage Distribution
15000
10000
5000
0
−5000
−10000
−200000
−150000
−100000
−50000
0
50000



50% have a delta of zero

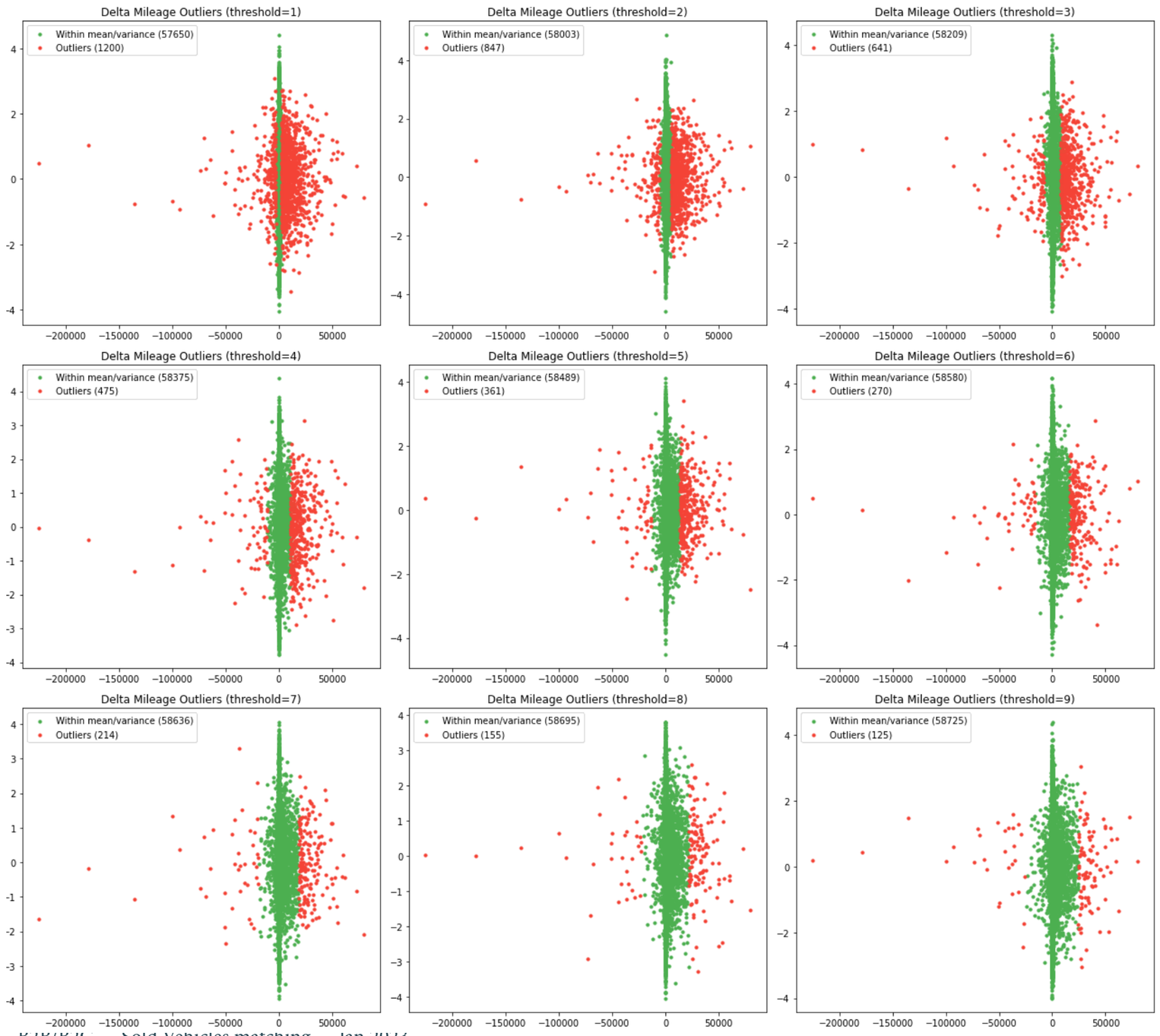
ok, let's take a deeper look here

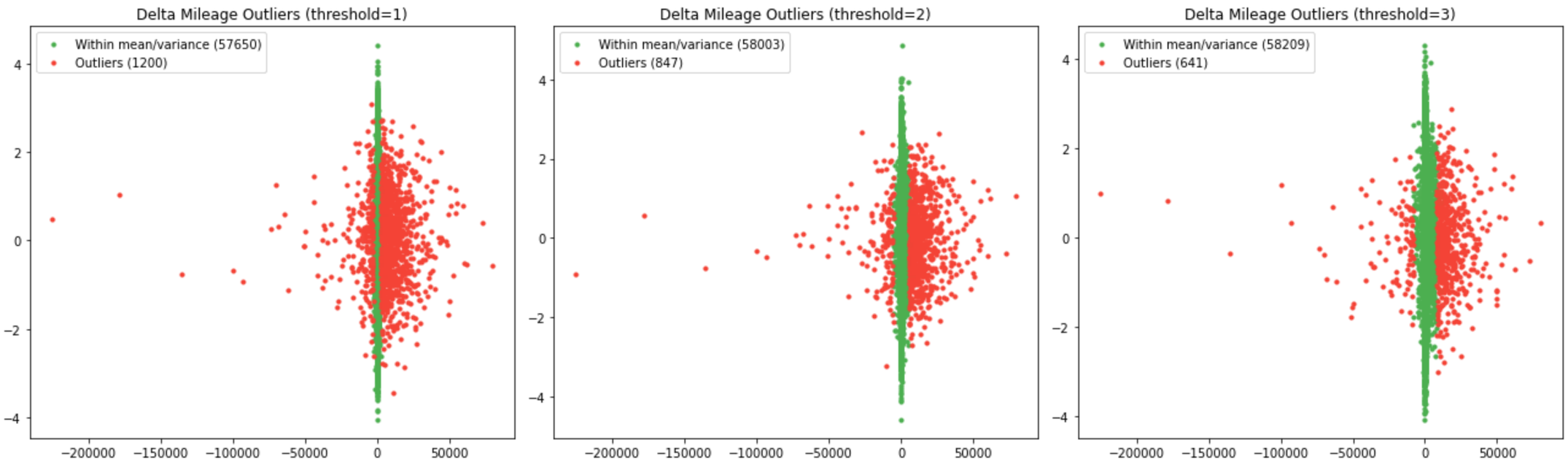

Delta Mileage Distribution
15000
10000
5000
0
−5000
−10000
−200000
−150000
−100000
−50000
0
50000
how to know when a delta is an outlier?

z-score to the rescue

z-score to the rescue

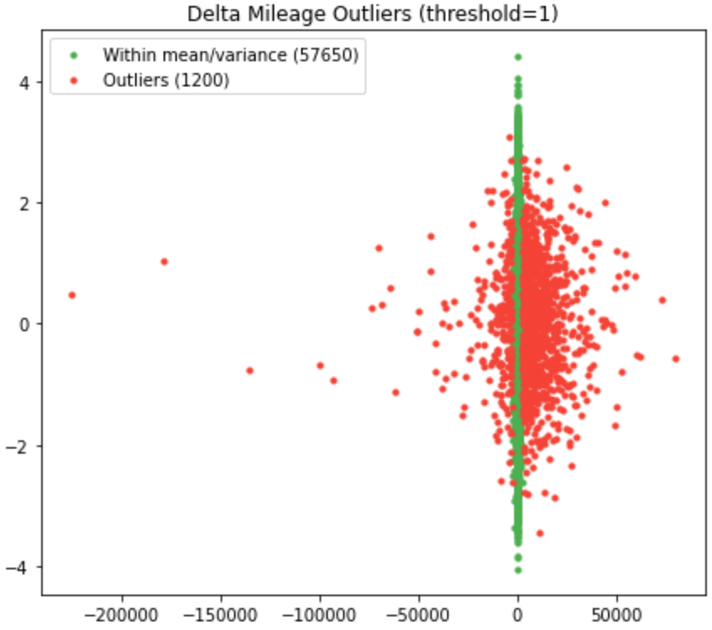
nothing fancy, just how many std deviations away from mean a point can be, before it is seen as "outlier"


Delta Mileage Outliers (threshold=1)
Within mean/variance (57650)
Outliers (1200)
Delta Mileage Outliers (threshold=2)
Within mean/variance (58003)
Outliers (847)
Delta Mileage Outliers (threshold=3)
Within mean/variance (58209)
Outliers (641)
Delta Mileage Outliers (threshold=4)
Within mean/variance (58375)
Outliers (475)
Delta Mileage Outliers (threshold=5)
Within mean/variance (58489)
Outliers (361)
Delta Mileage Outliers (threshold=6)
Within mean/variance (58580)
Outliers (270)
Delta Mileage Outliers (threshold=7)
Within mean/variance (58636)
Outliers (214)
Delta Mileage Outliers (threshold=8)
Within mean/variance (58695)
Outliers (155)
Delta Mileage Outliers (threshold=9)
Within mean/variance (58725)
Outliers (125)

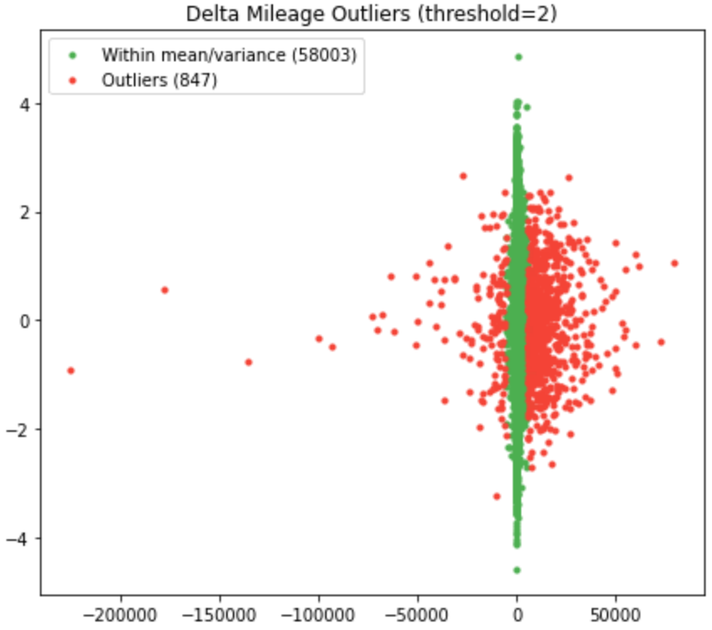


rule of thumb, z-score in [1,3]

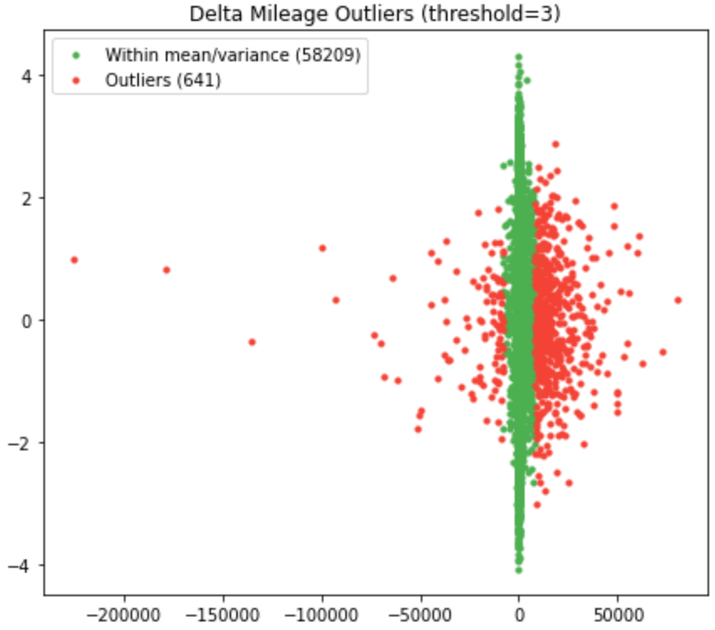
97.96%



98.56%



98.91%



rule of thumb, z-score in [1,3]

quick hypotheses:

- maybe model matches;
- maybe mileage matches;
- **maybe registration year matches;**

quick hypotheses:

- maybe model matches;
- maybe mileage matches;
- **maybe registration year matches;**

99.80% of the cases!

quick hypotheses:

- maybe model matches;
- maybe mileage matches;
- maybe registration year matches;
- maybe more features match;

quick hypotheses:

- maybe model matches;
- maybe mileage matches;
- maybe registration year matches;

correlation matrix as a rescue tool

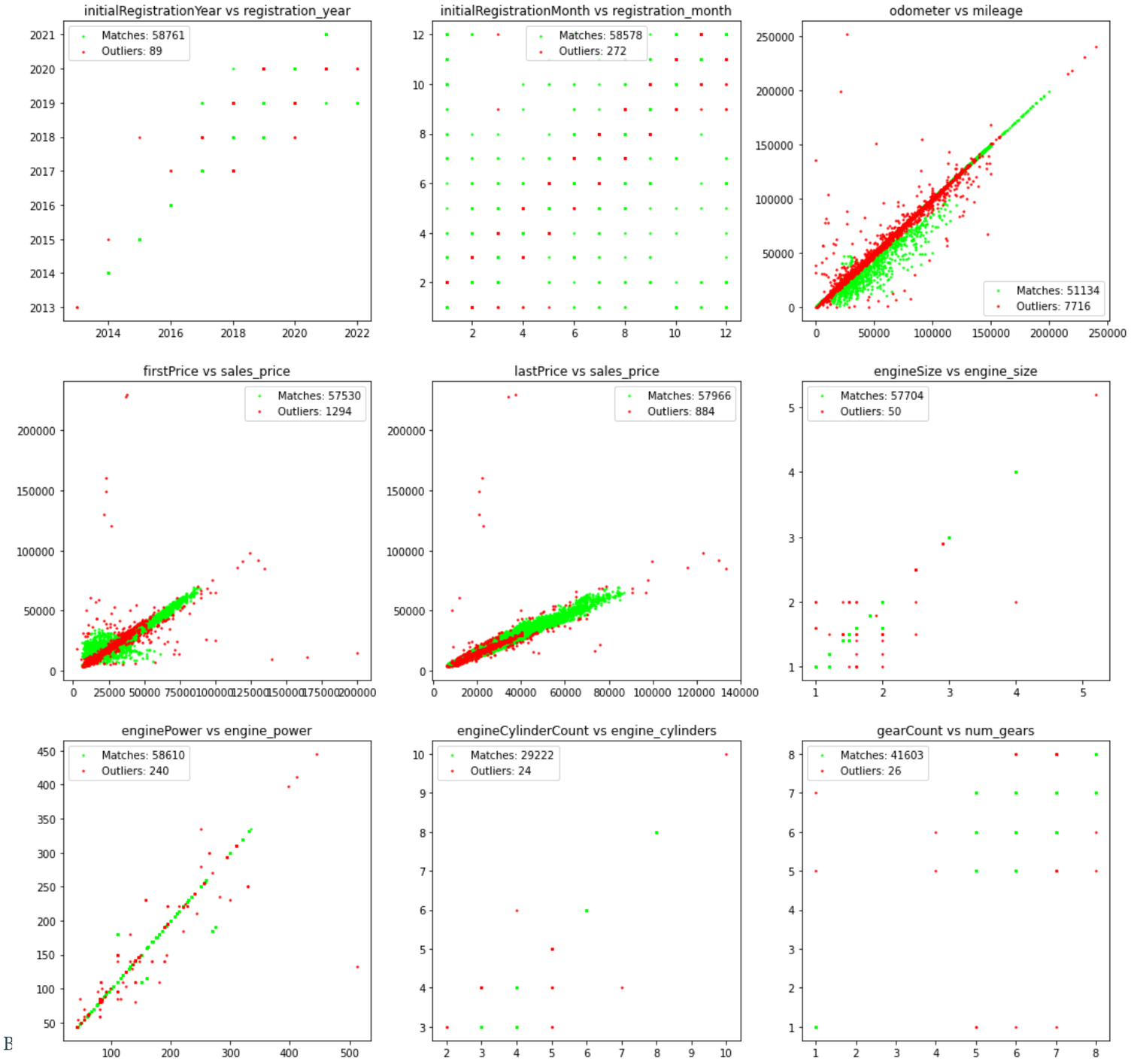
quick hypotheses:

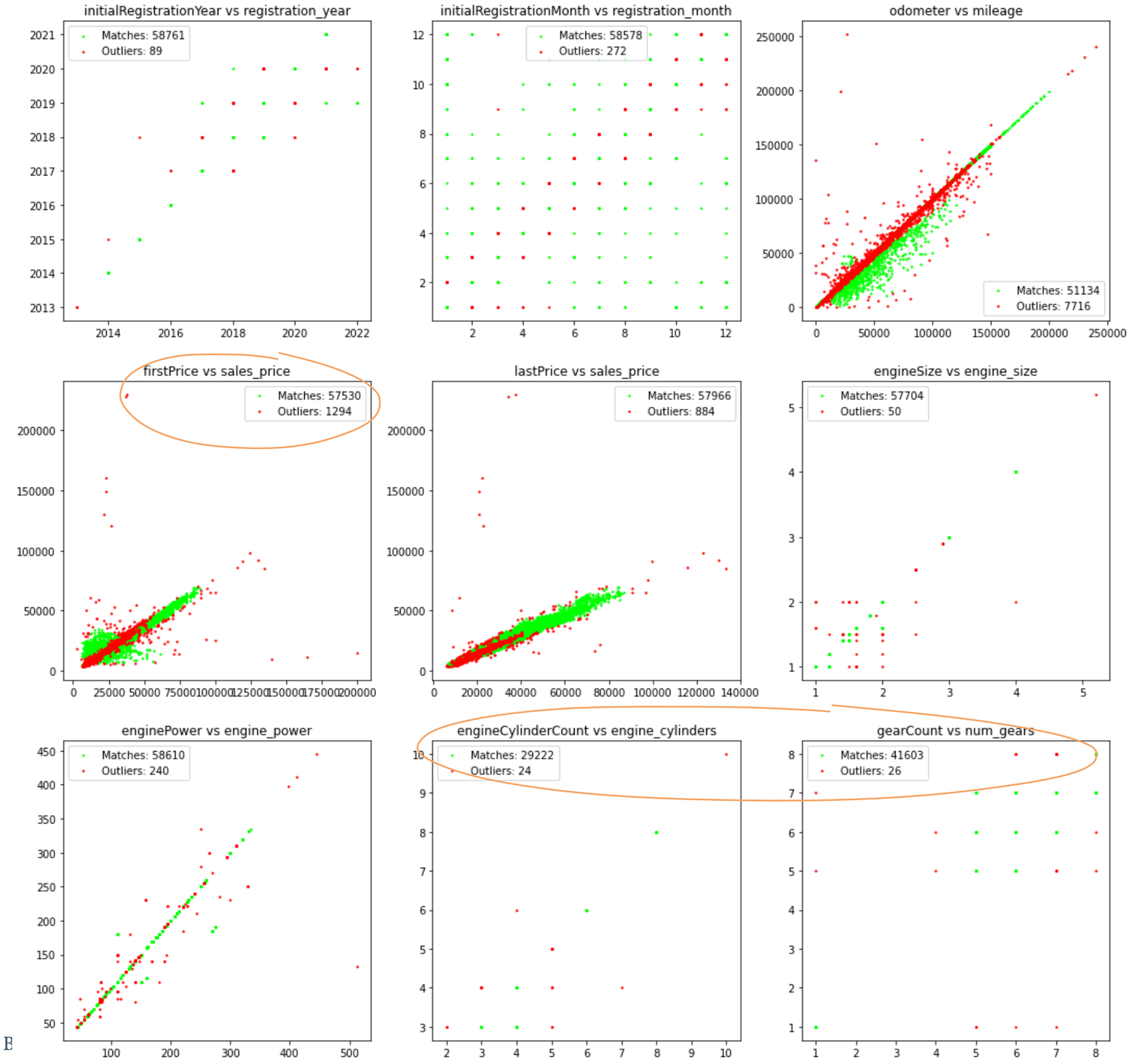
- maybe model matches;
- maybe mileage matches;
- maybe registration year matches;

correlation matrix as a rescue tool

better analyzed with Spearman method

B2C		B2B
initialRegistrationYear	99.85%	registration_year
initialRegistrationMonth	99.25%	registration_month
odometer	99.60%	mileage
firstPrice	96.98%	sales_price
lastPrice	97.86%	sales_price
engineSize	99.87%	engine_size
enginePower	99.72%	engine_power
engineCylinderCount	99.51%	engine_cylinders
gearCount	97.63%	num_gears


initialRegistrationYear vs registration_year
Matches: 58761
Outliers: 89
initialRegistrationMonth vs registration_month
Matches: 58578
Outliers: 272
odometer vs mileage
Matches: 51134
Outliers: 7716
firstPrice vs sales_price
Matches: 57530
Outliers: 1294
lastPrice vs sales_price
Matches: 57966
Outliers: 884
engineSize vs engine_size
Matches: 57704
Outliers: 50
enginePower vs engine_power
Matches: 58610
Outliers: 240
engineCylinderCount vs engine_cylinders
Matches: 29222
Outliers: 24
gearCount vs num_gears
Matches: 41603
Outliers: 26


initialRegistrationYear vs registration_year
Matches: 58761
Outliers: 89
initialRegistrationMonth vs registration_month
Matches: 58578
Outliers: 272
odometer vs mileage
Matches: 51134
Outliers: 7716
firstPrice vs sales_price
Matches: 57530
Outliers: 1294
lastPrice vs sales_price
Matches: 57966
Outliers: 884
engineSize vs engine_size
Matches: 57704
Outliers: 50
enginePower vs engine_power
Matches: 58610
Outliers: 240
engineCylinderCount vs engine_cylinders
Matches: 29222
Outliers: 24
gearCount vs num_gears
Matches: 41603
Outliers: 26

time for a research question:

RQ1: Due to the proximity of these features, can the matched 60K points be reidentified against B2C but without VIN?

RQ1: Due to the proximity of these features, can the matched 60K points be reidentified against B2C but without VIN?

after all, we still have records with no label

good idea, bad results.

euclidean distance

euclidean distance

mileage, engine_size, engine_power, sales_price, registration_year, registration_month, model*

stacked B2C and B2B, all normalized (RobustScaler)

Hard Filtered

euclidean distance

mileage, engine_size, engine_power, sales_price, registration_year, registration_month, model*

stacked B2C and B2B, all normalized (RobustScaler)

on a first attempt to rematch the “golden dataset”:

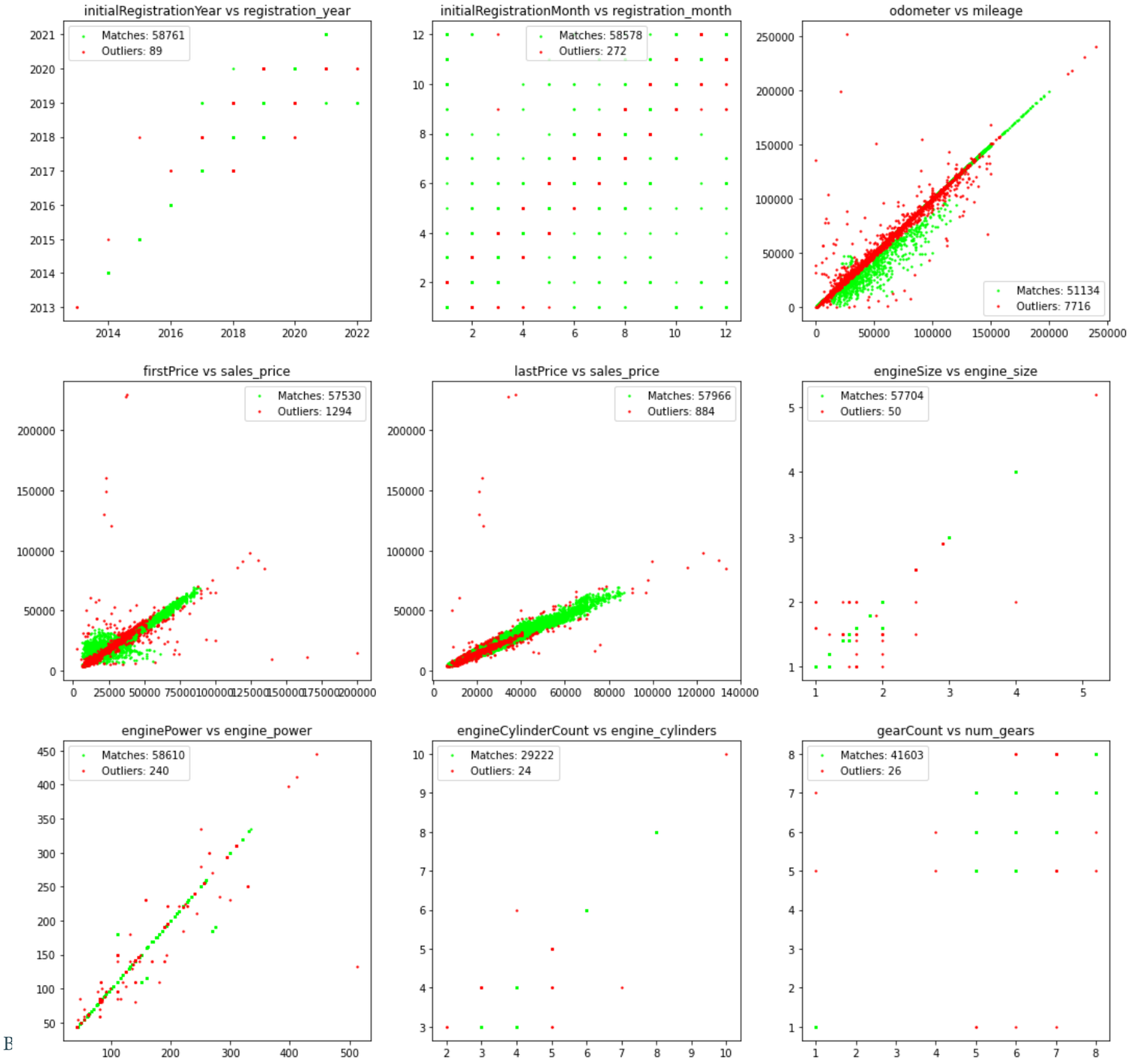
1%

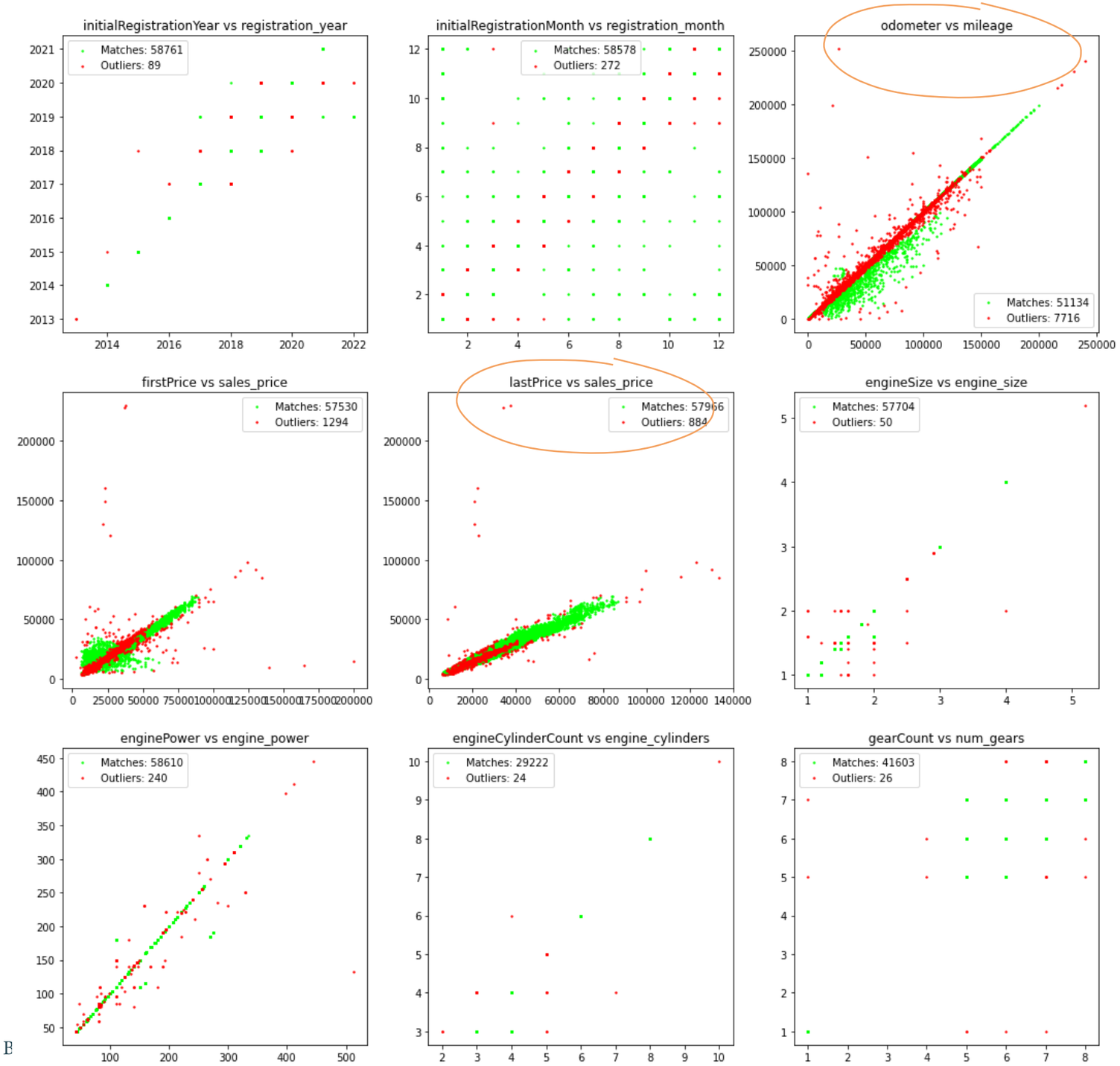
when searching for closest point against full B2C dataset

10%

when searching for closest point against 60K B2C dataset

but wait a minute!


initialRegistrationYear vs registration_year
Matches: 58761
Outliers: 89
initialRegistrationMonth vs registration_month
Matches: 58578
Outliers: 272
odometer vs mileage
Matches: 51134
Outliers: 7716
firstPrice vs sales_price
Matches: 57530
Outliers: 1294
lastPrice vs sales_price
Matches: 57966
Outliers: 884
engineSize vs engine_size
Matches: 57704
Outliers: 50
enginePower vs engine_power
Matches: 58610
Outliers: 240
engineCylinderCount vs engine_cylinders
Matches: 29222
Outliers: 24
gearCount vs num_gears
Matches: 41603
Outliers: 26



these are continuous features!!!

RQ1: Due to the proximity of these features, can the matched 60K points be reidentified against B2C but without VIN?

RQ1: Due to the proximity of these features, can the matched 60K points be reidentified against B2C but without VIN?

Answer: Apparently not. **(RQ1.1)** But what if we boost the continuous variables to distantiate from the non-matching points?

the incredible Mileage Booster

Hard Filtered

euclidean distance

mileage, engine_size, engine_power, sales_price, registration_year, registration_month, model*

stacked B2C and B2B, all normalized (RobustScaler)

Hard Filtered Boosted

euclidean distance

MILEAGE, engine_size, engine_power, sales_price, registration_year, registration_month, model*

stacked B2C and B2B, all normalized (RobustScaler)

methodology:

methodology:

- grid search;

methodology:
- grid search;

arbitrary values inside interval [2, 40.000]

methodology:

- grid search;
- matches;

methodology:

- grid search;
- **matches;**

against hardly-filtered B2C dataset! (year / month / model)

methodology:

- grid search;
- matches;
- likelihood;

methodology:

- grid search;
- matches;
- **likelihood;**

different TOP N rankings, useful to assist on labeling unknown data
(± 200K B2B records)

methodology:

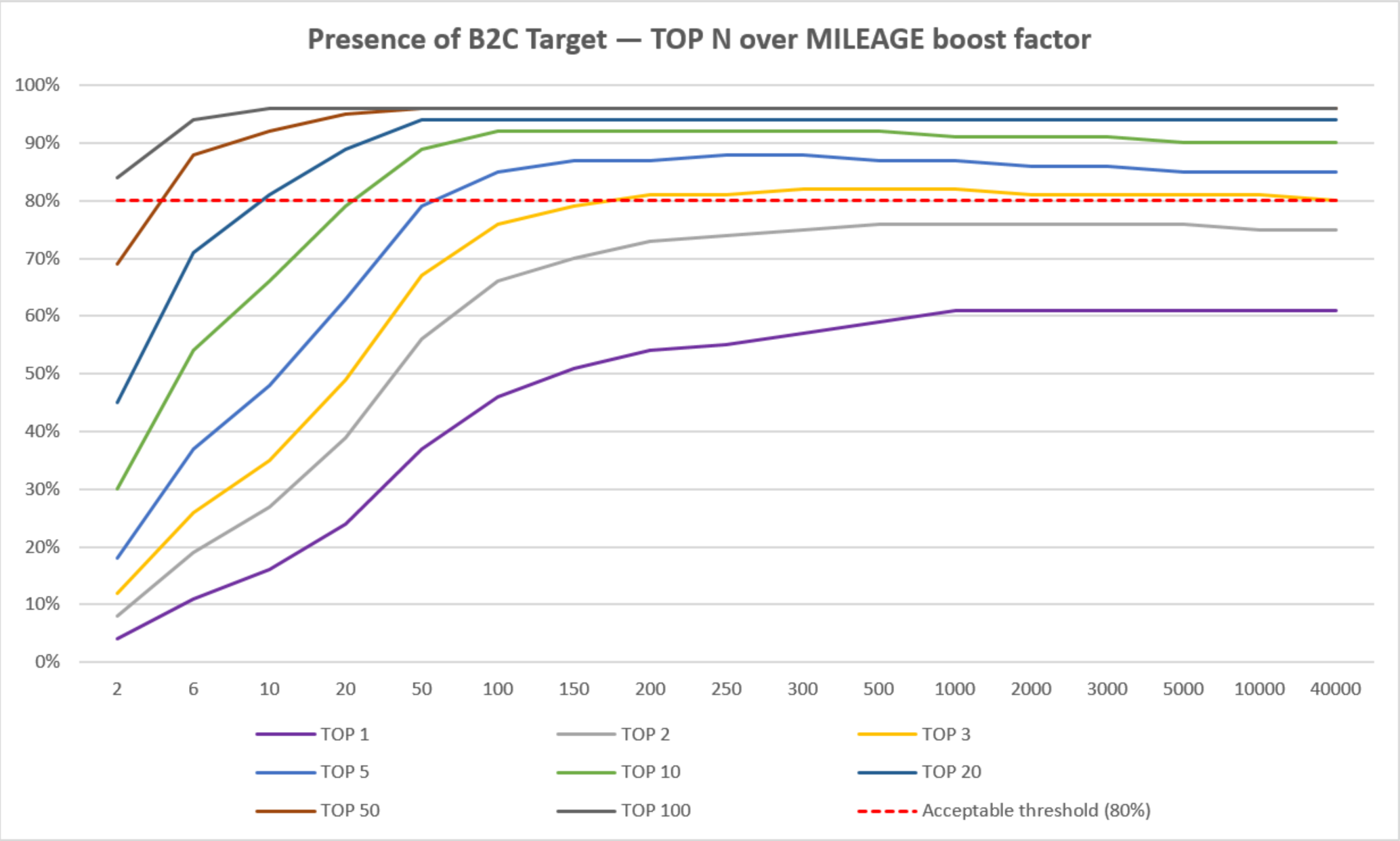
- grid search;
- matches;
- likelihood;
- acceptance threshold.

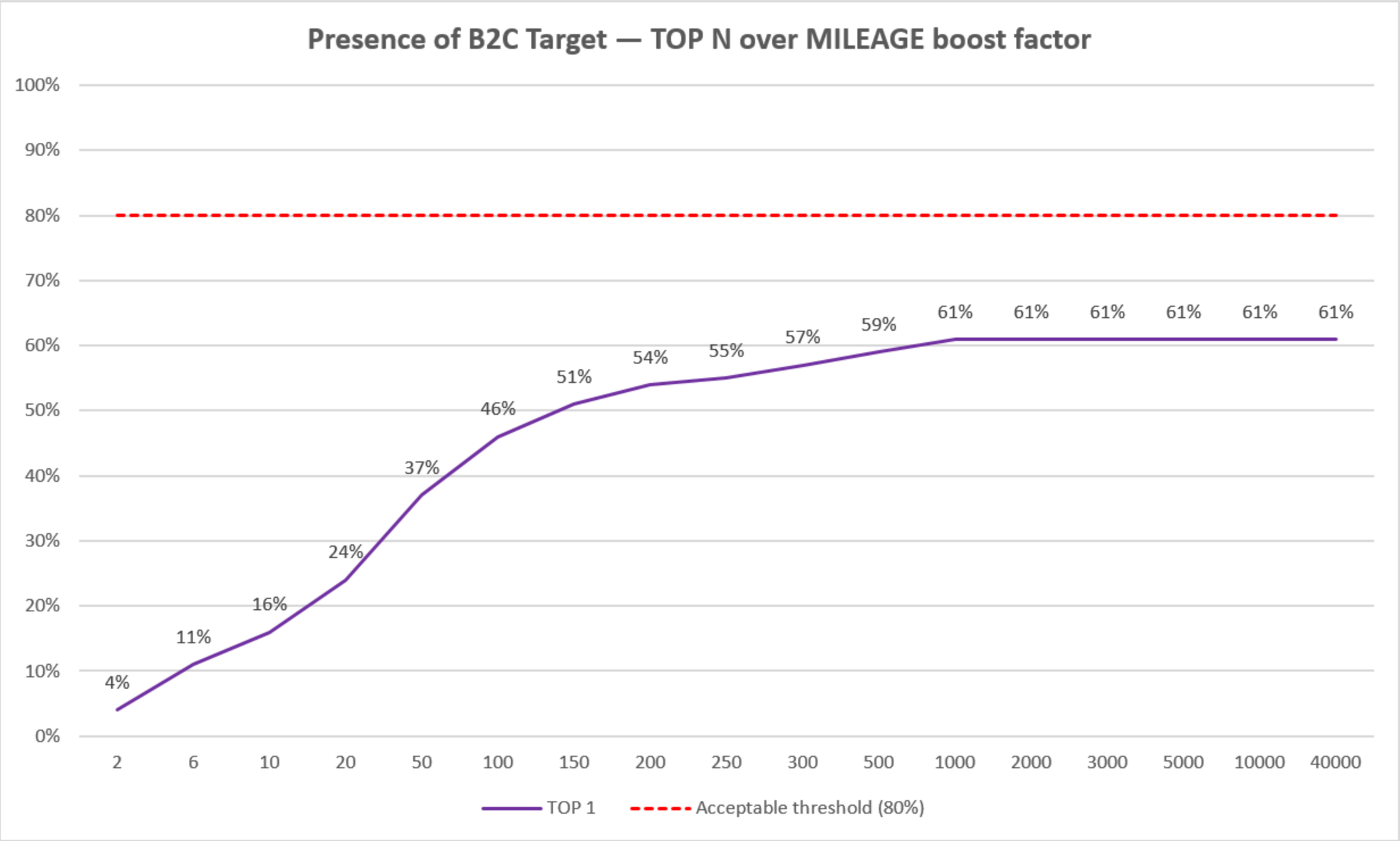
methodology:

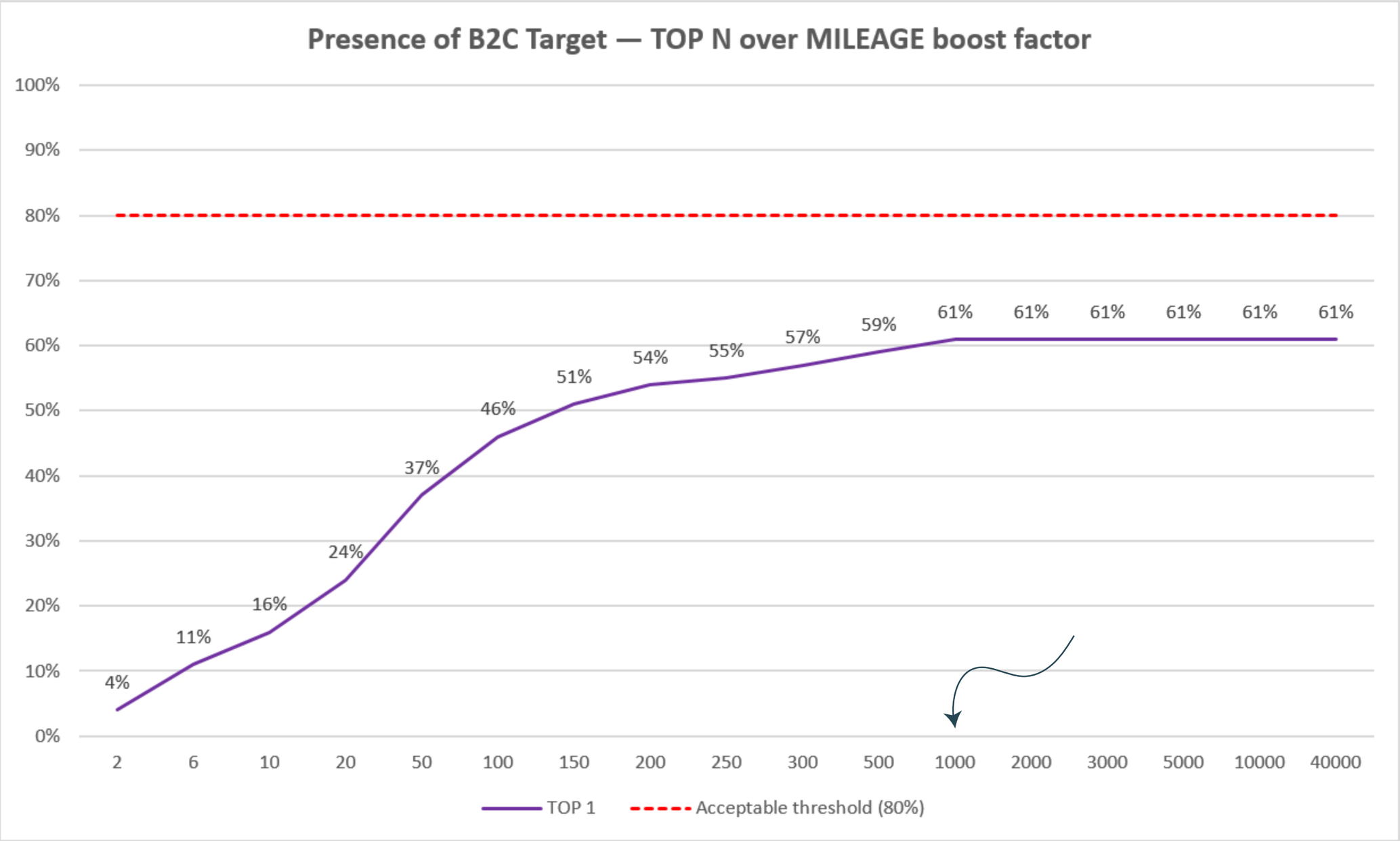
- grid search;
- matches;
- likelihood;
- **acceptance threshold.**

80% as challenge for useful results
(for exact matches, or for likelihood — target in TOP N)

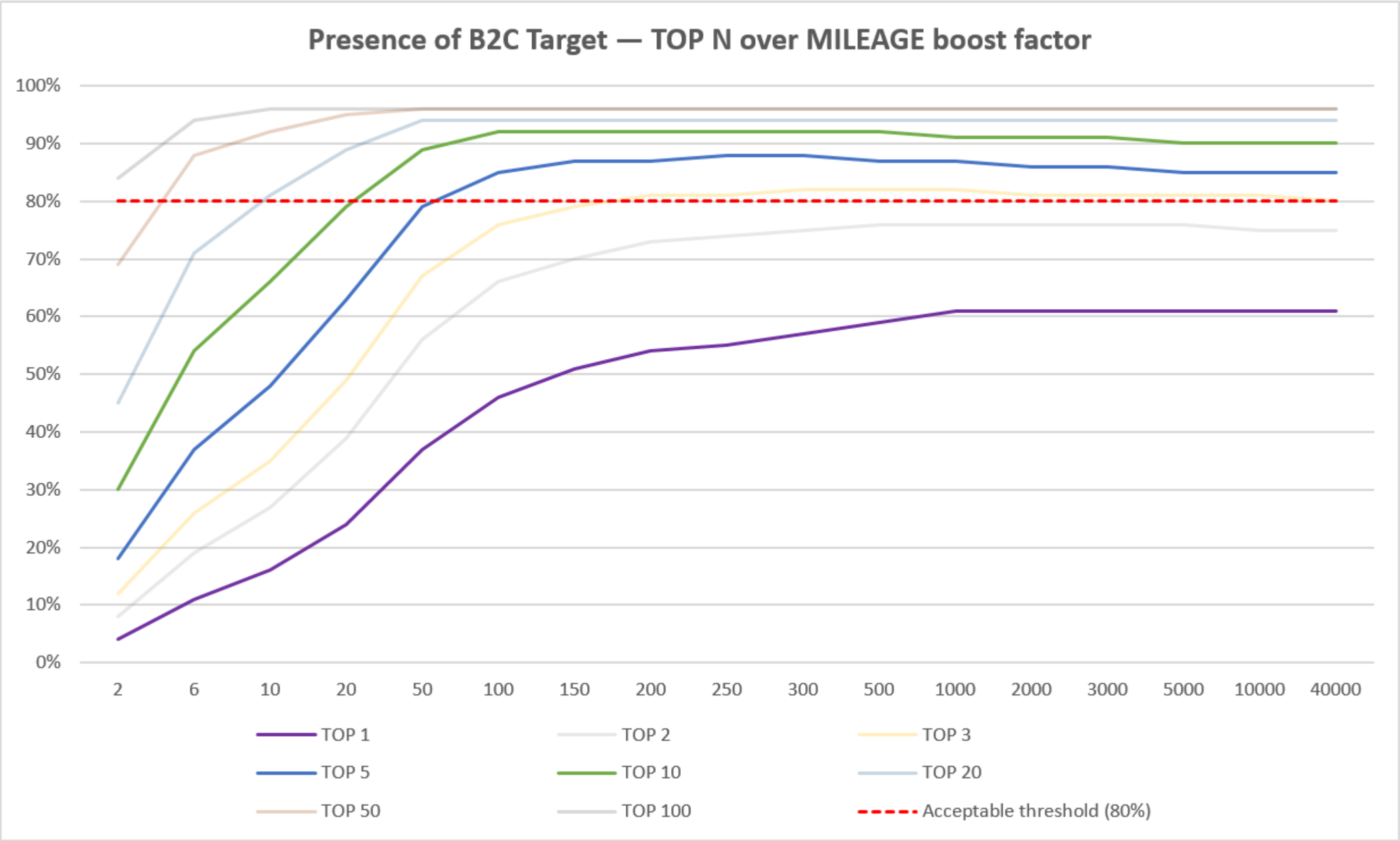
this is quite a lot to process!

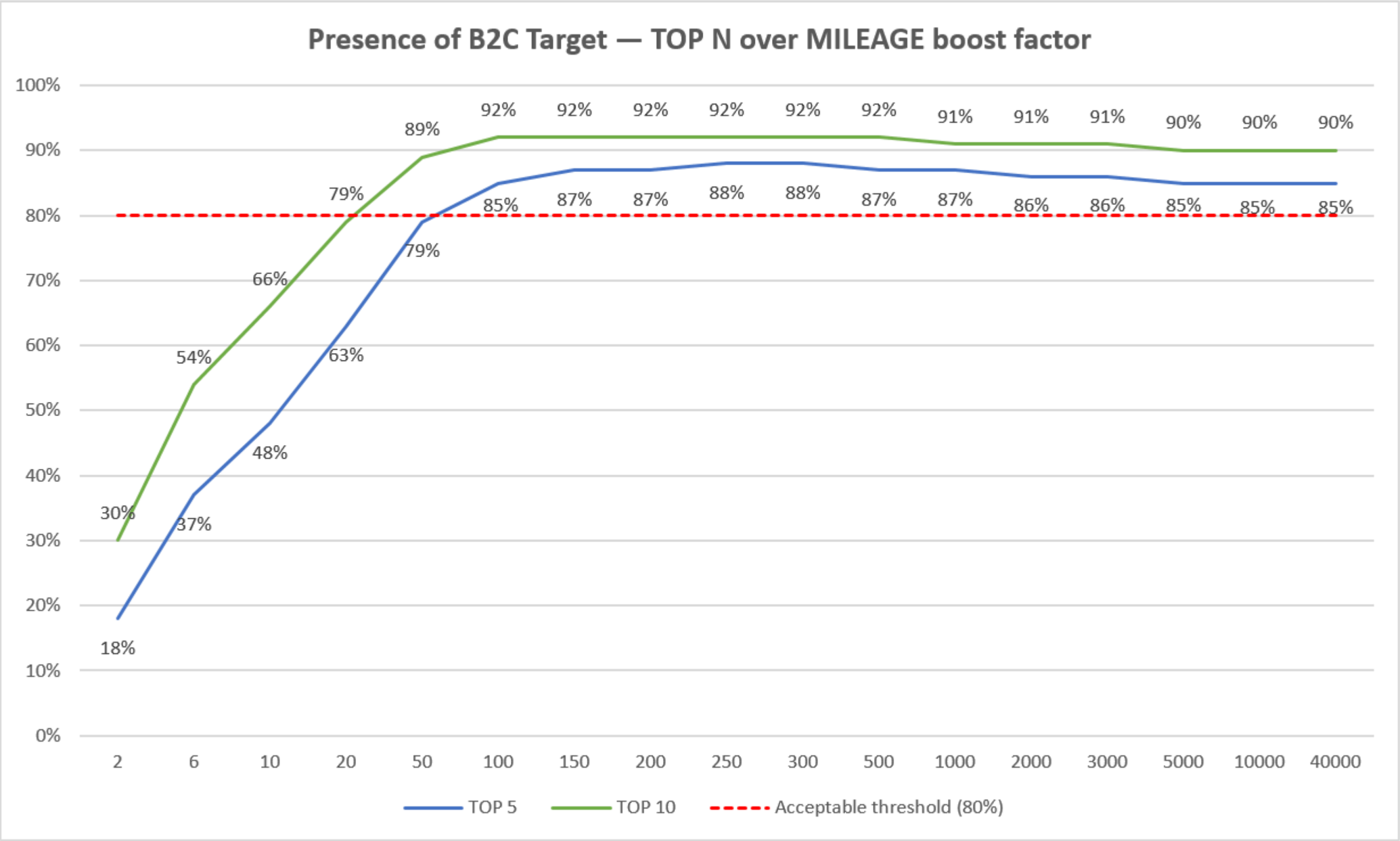

Presence of B2C Target — TOP N over MILEAGE boost factor
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
2
6
10
20
50
100
150
200
250
300
500
1000
2000
3000
5000
10000
40000
TOP 1
TOP 2
TOP 3
TOP 5
TOP 10
TOP 20
TOP 50
TOP 100
Acceptable threshold (80%)

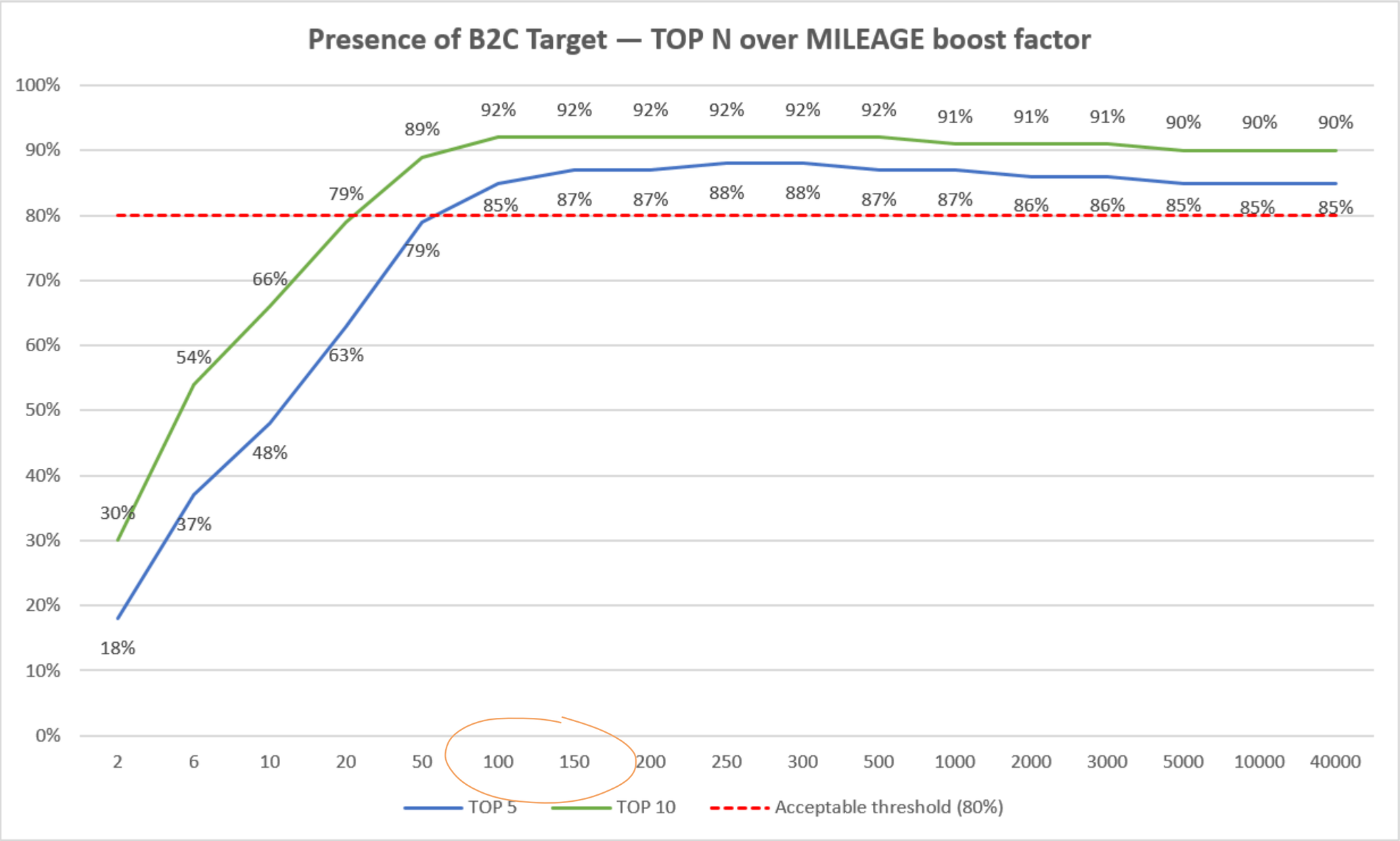

Presence of B2C Target — TOP N over MILEAGE boost factor
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
4%
11%
16%
24%
37%
46%
51%
54%
55%
57%
59%
61%
61%
61%
61%
61%
61%
2
6
10
20
50
100
150
200
250
300
500
1000
2000
3000
5000
10000
40000
TOP 1
Acceptable threshold (80%)


Presence of B2C Target — TOP N over MILEAGE boost factor
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
4%
11%
16%
24%
37%
46%
51%
54%
55%
57%
59%
61%
61%
61%
61%
61%
61%
2
6
10
20
50
100
150
200
250
300
500
1000
2000
3000
5000
10000
40000
TOP 1
Acceptable threshold (80%)

let’s not forget the amount of records with zero mileage delta!


Presence of B2C Target — TOP N over MILEAGE boost factor
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
2
6
10
20
50
100
150
200
250
300
500
1000
2000
3000
5000
10000
40000
TOP 1
TOP 2
TOP 3
TOP 5
TOP 10
TOP 20
TOP 50
TOP 100
Acceptable threshold (80%)


Presence of B2C Target — TOP N over MILEAGE boost factor
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
2
6
10
20
50
100
150
200
250
300
500
1000
2000
3000
5000
10000
40000
30%
54%
66%
79%
89%
92%
92%
92%
92%
92%
92%
91%
91%
91%
90%
90%
90%
18%
37%
48%
63%
79%
85%
87%
87%
88%
88%
87%
87%
86%
86%
85%
85%
85%
TOP 5
TOP 10
Acceptable threshold (80%)


Presence of B2C Target — TOP N over MILEAGE boost factor
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
2
6
10
20
50
100
150
200
250
300
500
1000
2000
3000
5000
10000
40000
30%
54%
66%
79%
89%
92%
92%
92%
92%
92%
92%
91%
91%
91%
90%
90%
90%
18%
37%
48%
63%
79%
85%
87%
87%
88%
88%
87%
87%
86%
86%
85%
85%
85%
TOP 5
TOP 10
Acceptable threshold (80%)

Mileage Boost	TOP 1	TOP 2	TOP 3	TOP 5	TOP 10	TOP 20	TOP 50	TOP 100
2	4%	8%	12%	18%	30%	45%	69%	84%
6	11%	19%	26%	37%	54%	71%	88%	94%
10	16%	27%	35%	48%	66%	81%	92%	96%
20	24%	39%	49%	63%	79%	89%	95%	96%
50	37%	56%	67%	79%	89%	94%	96%	96%
100	46%	66%	76%	85%	92%	94%	96%	96%
150	51%	70%	79%	87%	92%	94%	96%	96%
200	54%	73%	81%	87%	92%	94%	96%	96%
250	55%	74%	81%	88%	92%	94%	96%	96%
300	57%	75%	82%	88%	92%	94%	96%	96%
500	59%	76%	82%	87%	92%	94%	96%	96%
1000	61%	76%	82%	87%	91%	94%	96%	96%
2000	61%	76%	81%	86%	91%	94%	96%	96%
3000	61%	76%	81%	86%	91%	94%	96%	96%
5000	61%	76%	81%	85%	90%	94%	96%	96%
10000	61%	75%	81%	85%	90%	94%	96%	96%
40000	61%	75%	80%	85%	90%	94%	96%	96%

Mileage Boost	TOP 1	TOP 2	TOP 3	TOP 5	TOP 10	TOP 20	TOP 50	TOP 100
2	4%	8%	12%	18%	30%	45%	69%	84%
6	11%	19%	26%	37%	54%	71%	88%	94%
10	16%	27%	35%	48%	66%	81%	92%	96%
3000	61%	76%	81%	86%	91%	94%	96%	96%
5000	61%	76%	81%	85%	90%	94%	96%	96%
10000	61%	75%	81%	85%	90%	94%	96%	96%
40000	61%	75%	80%	85%	90%	94%	96%	96%

short info: boosting sales_price did not prove successful :-(

Mileage Boost	TOP 1	TOP 2	TOP 3	TOP 5	TOP 10	TOP 20	TOP 50	TOP 100
2	4%	8%	12%	18%	30%	45%	69%	84%
6	11%	19%	26%	37%	54%	71%	88%	94%
10	16%	27%	35%	48%	66%	81%	92%	96%
40000	61%	75%	80%	85%	90%	94%	96%	96%

short info: boosting sales_price did not prove successful :-(

what if we “degraded” this feature, then? (idea)

pause for a quick breath...

can we sketch a pipeline?

Pipeline

Pipeline

- TOP 1 == VIN;

Pipeline

- TOP 1 == VIN;

- Reduce dataset for non-matching VINS;

# of course, the set of features can be revisited!

Pipeline

- TOP 1 == VIN;
- Reduce dataset for non-matching VINS;
- **Use boosted distance to output likelihood — TOP N;**

Pipeline

- TOP 1 == VIN;
- Reduce dataset for non-matching VINS;
- **Use boosted distance to output likelihood — TOP N;**

focusing on statistical approach, no complex n.n.

Pipeline

- TOP 1 == VIN;
- Reduce dataset for non-matching VINS;
- Use boosted distance to output likelihood — TOP N;
- **Input likelihood on a second (or more) model(s): best probable B2C match.**

Pipeline

- TOP 1 == VIN;
- Reduce dataset for non-matching VINS;
- Use boosted distance to output likelihood — TOP N;
- **Input likelihood on a second (or more) model(s): best probable B2C match.**

maybe filtering mileage? or extra parameters for other features

Pipeline

- TOP 1 == VIN;
- Reduce dataset for non-matching VINS;
- Use boosted distance to output likelihood — TOP N;
- **Input likelihood on a second (or more) model(s): best probable B2C match.**

also, dealer info can be handy

to be continued...

Further superheroes to be unlocked...

~~to be continued...~~