Overview

Dataset statistics

Number of variables	6
Number of observations	1035
Missing cells	792
Missing cells (%)	12.8%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	48.6 KiB
Average record size in memory	48.1 B

Variable types

Categorical	1
Numeric	5

Alerts

`orbital_period` is highly correlated with `mass`	High correlation
`mass` is highly correlated with `orbital_period`	High correlation
`method` is highly correlated with `distance` and 1 other fields (distance, year)	High correlation
`distance` is highly correlated with `method`	High correlation
`year` is highly correlated with `method`	High correlation
`orbital_period` has 43 (4.2%) missing values	Missing
`mass` has 522 (50.4%) missing values	Missing
`distance` has 227 (21.9%) missing values	Missing
`orbital_period` is highly skewed (γ1 = 24.27306658)	Skewed

Reproduction

Analysis started	2022-09-25 10:41:45.282736
Analysis finished	2022-09-25 10:41:54.251714
Duration	8.97 seconds
Software version	pandas-profiling v3.1.0 (https://github.com/pandas-profiling/pandas-profiling)
Download configuration	config.json (data:text/plain;charset=utf-8,%7B%22title%22%3A%20%22Pandas%20Profiling%20Report%22%2C%20%22dataset%22%3A%20%7B%22description%22%3A%20%22%22%2C%2

Variables

method

Categorical

HIGH CORRELATION (This variable has a high correlation with 2 fields: distance, year)

Distinct	10
Distinct (%)	1.0%
Missing	0
Missing (%)	0.0%
Memory size	8.2 KiB

Length

Max length	29
Median length	15
Mean length	11.7294686
Min length	7

Characters and Unicode

Total characters	0	
Distinct characters	0	
Distinct categories	0 (https://en.wikipedia.org/wiki/Unicode_character_property#General_Category)	?
Distinct scripts	0 (https://en.wikipedia.org/wiki/Script_(Unicode)#List_of_scripts_in_Unicode)	?
Distinct blocks	0 (https://en.wikipedia.org/wiki/Unicode_block)	?

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

Unique	1	?
Unique (%)	0.1%	

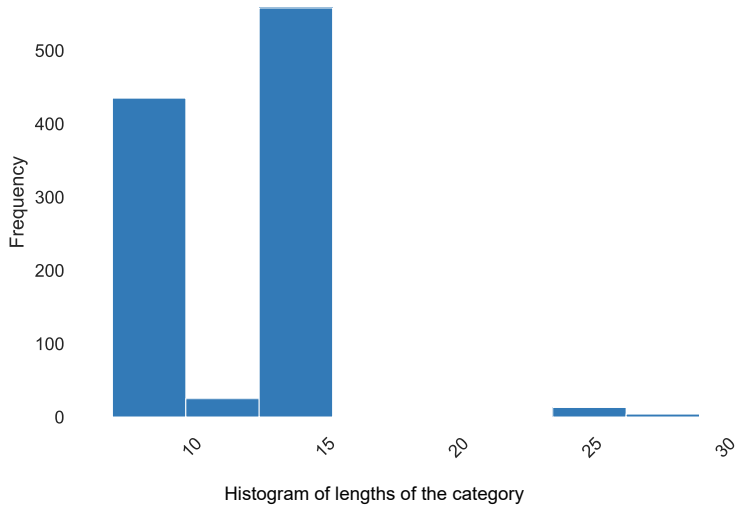
Sample

1st row	Radial Velocity
2nd row	Radial Velocity
3rd row	Radial Velocity
4th row	Radial Velocity
5th row	Radial Velocity

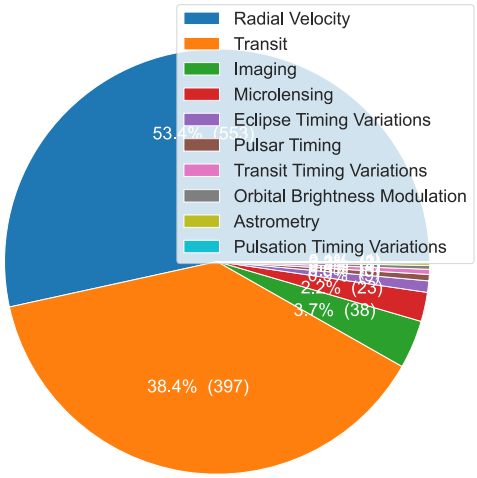
Common Values

Value	Count	Frequency (%)
Radial Velocity	553	53.4%
Transit	397	38.4%
Imaging	38	3.7%
Microlensing	23	2.2%
Eclipse Timing Variations	9	0.9%
Pulsar Timing	5	0.5%
Transit Timing Variations	4	0.4%
Orbital Brightness Modulation	3	0.3%
Astrometry	2	0.2%
Pulsation Timing Variations	1	0.1%

Length



Pie chart


Radial Velocity
Transit
Imaging
Microlensing
Eclipse Timing Variations
Pulsar Timing
Transit Timing Variations
Orbital Brightness Modulation
Astrometry
Pulsation Timing Variations
38.4% (397)
3.7% (38)
2.2% (23)

Value	Count	Frequency (%)
radial	553	34.0%
velocity	553	34.0%
transit	401	24.6%
imaging	38	2.3%
microlensing	23	1.4%
timing	19	1.2%
variations	14	0.9%
eclipse	9	0.6%
pulsar	5	0.3%
orbital	3	0.2%
Other values (4)	9	0.6%

Most occurring characters

Value	Count	Frequency (%)

No values found.

Most occurring categories

Value	Count	Frequency (%)
No values found.		

Most frequent character per category

Most occurring scripts

Value	Count	Frequency (%)
No values found.		

Most frequent character per script

Most occurring blocks

Value	Count	Frequency (%)
No values found.		

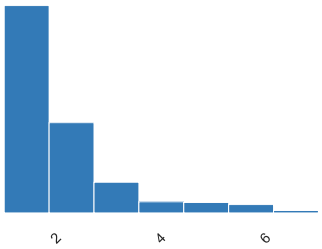
Most frequent character per block

number

Real number ($\mathbb{R}_{\geq 0}$)

Distinct	7
Distinct (%)	0.7%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	1.785507246

Minimum	1
Maximum	7
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	8.2 KiB



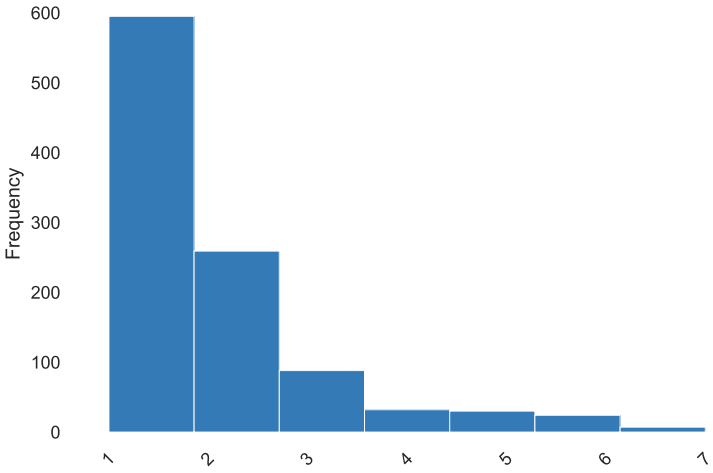
Quantile statistics

Minimum	1
5-th percentile	1
Q1	1
median	1
Q3	2
95-th percentile	5
Maximum	7
Range	6
Interquartile range (IQR)	1

Descriptive statistics

Standard deviation	1.240976174
Coefficient of variation (CV)	0.6950272405
Kurtosis	4.020604399
Mean	1.785507246
Median Absolute Deviation (MAD)	0
Skewness	2.030571847
Sum	1848

Variance	1.540021865
Monotonicity	Not monotonic



Histogram with fixed size bins (bins=7)

Value	Count	Frequency (%)
1	595	57.5%
2	259	25.0%
3	88	8.5%
4	32	3.1%
5	30	2.9%
6	24	2.3%
7	7	0.7%

Value	Count	Frequency (%)
1	595	57.5%
2	259	25.0%
3	88	8.5%
4	32	3.1%
5	30	2.9%
6	24	2.3%
7	7	0.7%

Value	Count	Frequency (%)
7	7	0.7%
6	24	2.3%
5	30	2.9%
4	32	3.1%
3	88	8.5%
2	259	25.0%
1	595	57.5%

orbital_period

Real number ($\mathbb{R}_{\geq 0}$)

HIGH

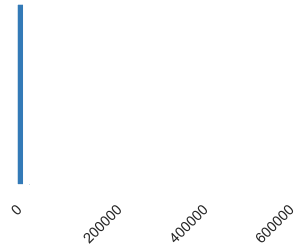
CORRELATION (This variable has a high correlation with 1 fields: mass)

MISSING

SKEWED

Distinct	988
Distinct (%)	99.6%
Missing	43
Missing (%)	4.2%
Infinite	0
Infinite (%)	0.0%
Mean	2002.917596

Minimum	0.09070629
Maximum	730000
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	8.2 KiB



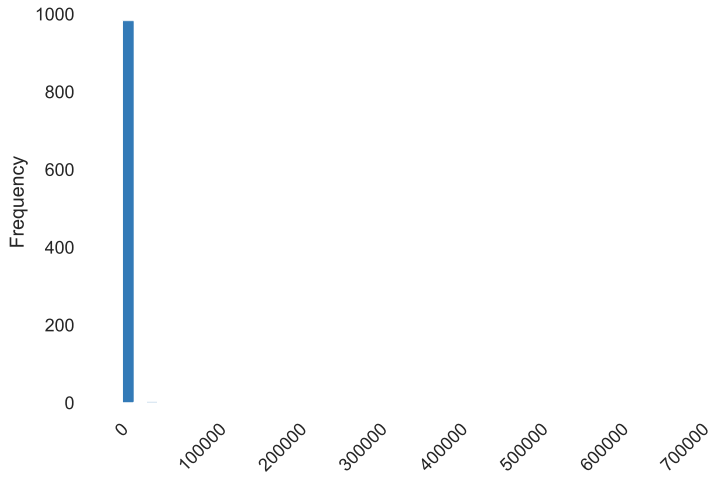
Quantile statistics

Minimum	0.09070629
5-th percentile	1.884364165
Q1	5.4425405
median	39.9795
Q3	526.005
95-th percentile	3360.45
Maximum	730000
Range	729999.9093
Interquartile range (IQR)	520.5624595

Descriptive statistics

Standard deviation	26014.7283
Coefficient of variation (CV)	12.98841668
Kurtosis	643.5748366
Mean	2002.917596
Median Absolute Deviation (MAD)	37.81445684

Skewness	24.27306658
Sum	1986894.255
Variance	676766088.7
Monotonicity	Not monotonic



Histogram with fixed size bins (bins=50)

Value	Count	Frequency (%)
1845	2	0.2%
883	2	0.2%
428.5	2	0.2%
123	2	0.2%
19.57706	1	0.1%
6.098493	1	0.1%
77.61184	1	0.1%
10.854092	1	0.1%
160	1	0.1%
3.6961219	1	0.1%
Other values (978)	978	94.5%
(Missing)	43	4.2%

Value	Count	Frequency (%)
0.09070629	1	0.1%
0.240104	1	0.1%
0.342887	1	0.1%
0.355	1	0.1%
0.45328509	1	0.1%
0.66931	1	0.1%
0.73654	1	0.1%
0.74296	1	0.1%
0.78884	1	0.1%
0.813475	1	0.1%

Value	Count	Frequency (%)
730000	1	0.1%
318280	1	0.1%
170000	1	0.1%
69000	1	0.1%
40000	1	0.1%
37000	1	0.1%
36525	1	0.1%
18000	1	0.1%
17337.5	1	0.1%
14002	1	0.1%

mass

Real number ($\mathbb{R}_{\geq 0}$)

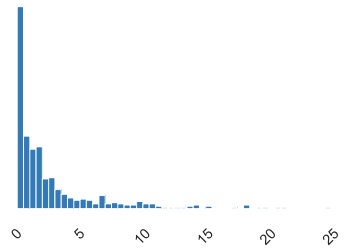
HIGH

CORRELATION (This variable has a high correlation with 1 fields: orbital period)

MISSING

Distinct	381
Distinct (%)	74.3%
Missing	522
Missing (%)	50.4%
Infinite	0
Infinite (%)	0.0%
Mean	2.638160585

Minimum	0.0036
Maximum	25
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	8.2 KiB



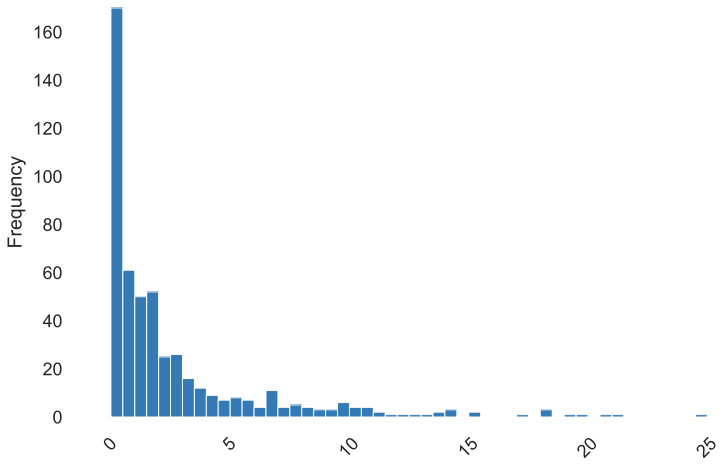
Quantile statistics

Minimum	0.0036
5-th percentile	0.01836
Q1	0.229
median	1.26
Q3	3.04
95-th percentile	10.528
Maximum	25
Range	24.9964
Interquartile range (IQR)	2.811

Descriptive statistics

Standard deviation	3.818616651
Coefficient of variation (CV)	1.447454212
Kurtosis	7.550015591
Mean	2.638160585
Median Absolute Deviation (MAD)	1.181
Skewness	2.550824321

Sum	1353.37638
Variance	14.58183313
Monotonicity	Not monotonic



Histogram with fixed size bins (bins=50)

Value	Count	Frequency (%)
1.9	6	0.6%
1.8	6	0.6%
2.3	5	0.5%
2.6	5	0.5%
2.7	4	0.4%
1.6	4	0.4%
0.82	4	0.4%
0.48	4	0.4%
1.7	4	0.4%
2.1	4	0.4%
Other values (371)	467	45.1%
(Missing)	522	50.4%

Value	Count	Frequency (%)
0.0036	1	0.1%
0.006	1	0.1%
0.00755	1	0.1%
0.008	2	0.2%
0.0085	1	0.1%
0.0087	1	0.1%
0.0097	1	0.1%
0.011	1	0.1%
0.01133	1	0.1%
0.0117	1	0.1%

Value	Count	Frequency (%)
25	1	0.1%
21.42	1	0.1%
20.6	1	0.1%
19.8	1	0.1%
19.4	1	0.1%
18.37	1	0.1%
18.15	1	0.1%
18.1	1	0.1%
17.4	1	0.1%
15.5	1	0.1%

distance

Real number ($\mathbb{R}_{\geq 0}$)

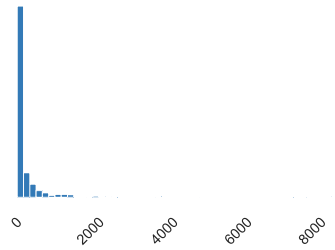
HIGH

CORRELATION (This variable has a high correlation with 1 fields: method)

MISSING

Distinct	552
Distinct (%)	68.3%
Missing	227
Missing (%)	21.9%
Infinite	0
Infinite (%)	0.0%
Mean	264.0692822

Minimum	1.35
Maximum	8500
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	8.2 KiB



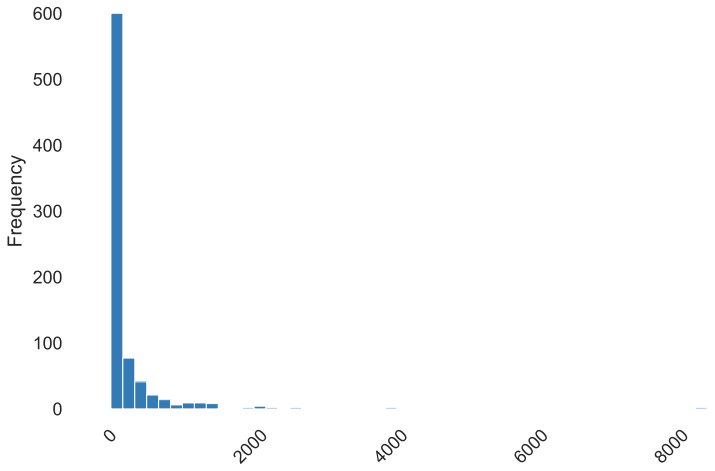
Quantile statistics

Minimum	1.35
5-th percentile	11.6715
Q1	32.56
median	55.25
Q3	178.5
95-th percentile	1180
Maximum	8500
Range	8498.65
Interquartile range (IQR)	145.94

Descriptive statistics

Standard deviation	733.1164929
Coefficient of variation (CV)	2.77622784
Kurtosis	68.32362922
Mean	264.0692822
Median Absolute Deviation (MAD)	37.28
Skewness	7.362539121

Sum	213367.98
Variance	537459.7922
Monotonicity	Not monotonic



Histogram with fixed size bins (bins=50)

Value	Count	Frequency (%)
780	7	0.7%
613	6	0.6%
12.83	6	0.6%
303	6	0.6%
6.8	6	0.6%
39.39	6	0.6%
368	5	0.5%
290	5	0.5%
1200	5	0.5%
1400	5	0.5%
Other values (542)	751	72.6%
(Missing)	227	21.9%

Value	Count	Frequency (%)
1.35	1	0.1%
3.22	1	0.1%
4.54	1	0.1%
4.7	4	0.4%
4.94	1	0.1%
6.06	3	0.3%
6.27	4	0.4%
6.8	6	0.6%
7.01	1	0.1%
7.69	1	0.1%

Value	Count	Frequency (%)
8500	2	0.2%
7720	1	0.1%
7560	1	0.1%
4970	1	0.1%
4080	2	0.2%
3600	1	0.1%
3200	1	0.1%
2800	1	0.1%
2700	1	0.1%
2570	1	0.1%

year

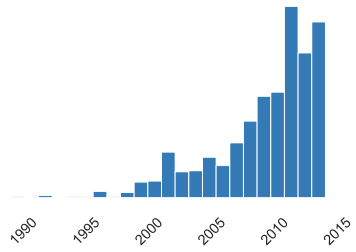
Real number ($\mathbb{R}_{\geq 0}$)

HIGH

CORRELATION (This variable has a high correlation with 1 fields: method)

Distinct	23
Distinct (%)	2.2%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	2009.070531

Minimum	1989
Maximum	2014
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	8.2 KiB



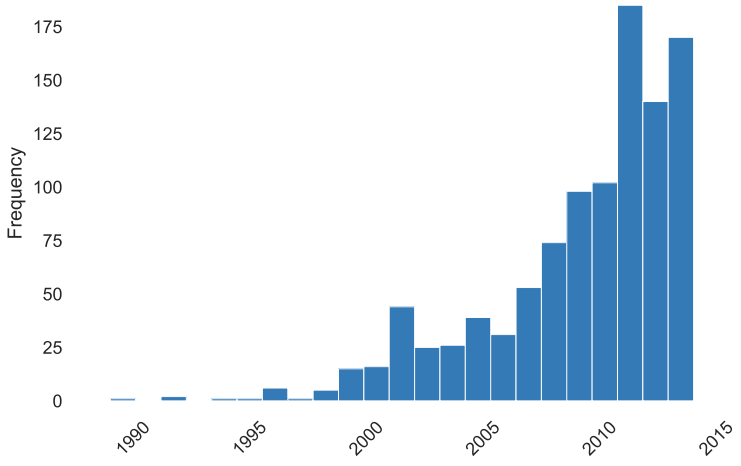
Quantile statistics

Minimum	1989
5-th percentile	2001
Q1	2007
median	2010
Q3	2012
95-th percentile	2013.3
Maximum	2014
Range	25
Interquartile range (IQR)	5

Descriptive statistics

Standard deviation	3.972566857
Coefficient of variation (CV)	0.001977315776
Kurtosis	1.776910442
Mean	2009.070531
Median Absolute Deviation (MAD)	2
Skewness	-1.326369808

Sum	2079388
Variance	15.78128743
Monotonicity	Not monotonic



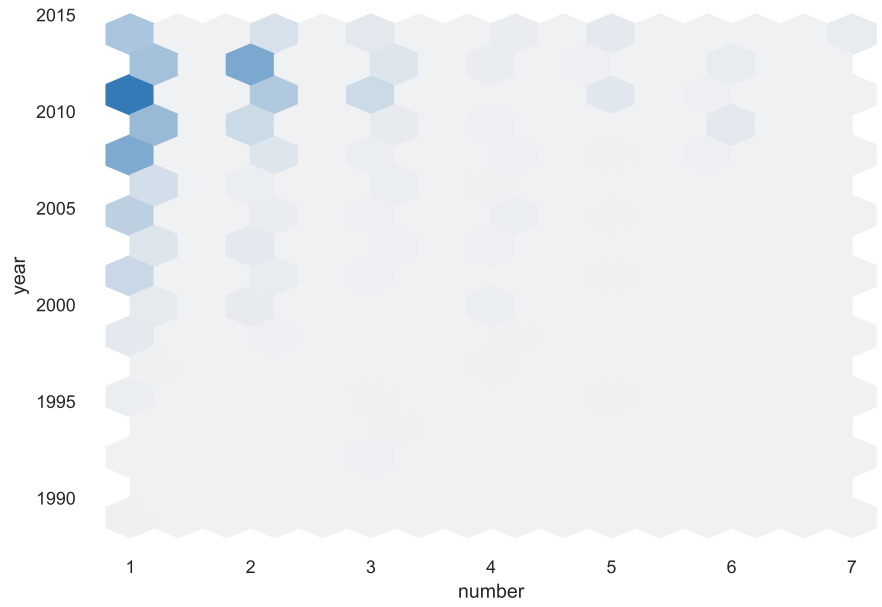
Histogram with fixed size bins (bins=23)

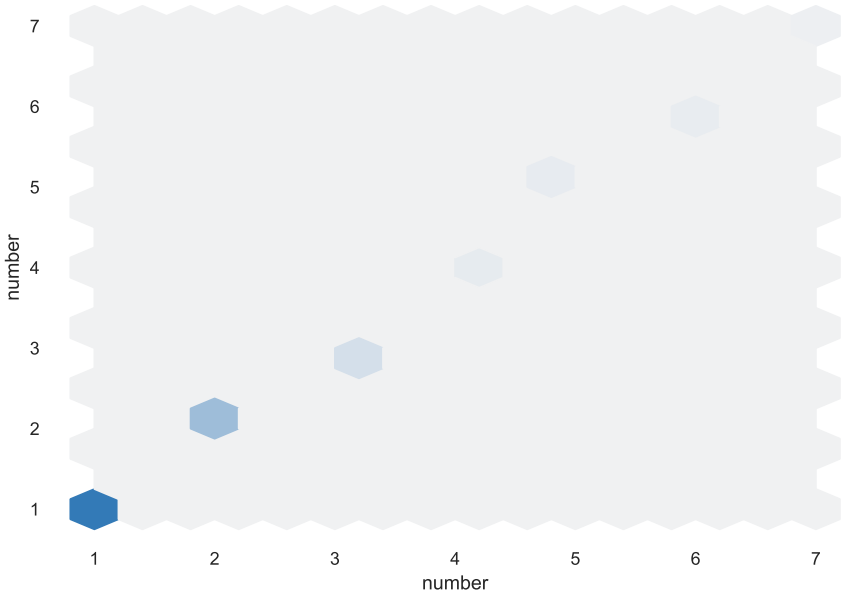
Value	Count	Frequency (%)
2011	185	17.9%
2012	140	13.5%
2013	118	11.4%
2010	102	9.9%
2009	98	9.5%
2008	74	7.1%
2007	53	5.1%
2014	52	5.0%
2005	39	3.8%
2002	32	3.1%
Other values (13)	142	13.7%

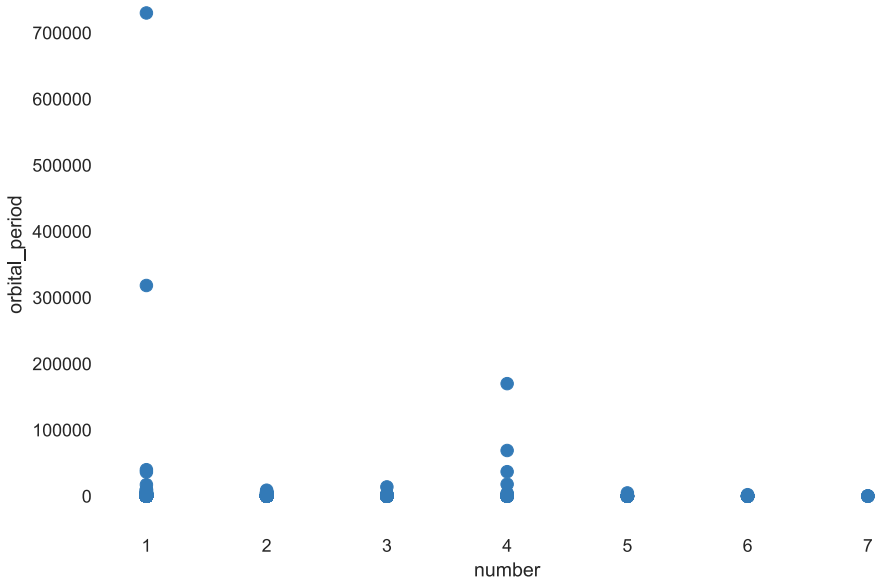
Value	Count	Frequency (%)
1989	1	0.1%
1992	2	0.2%
1994	1	0.1%
1995	1	0.1%
1996	6	0.6%
1997	1	0.1%
1998	5	0.5%
1999	15	1.4%
2000	16	1.5%
2001	12	1.2%

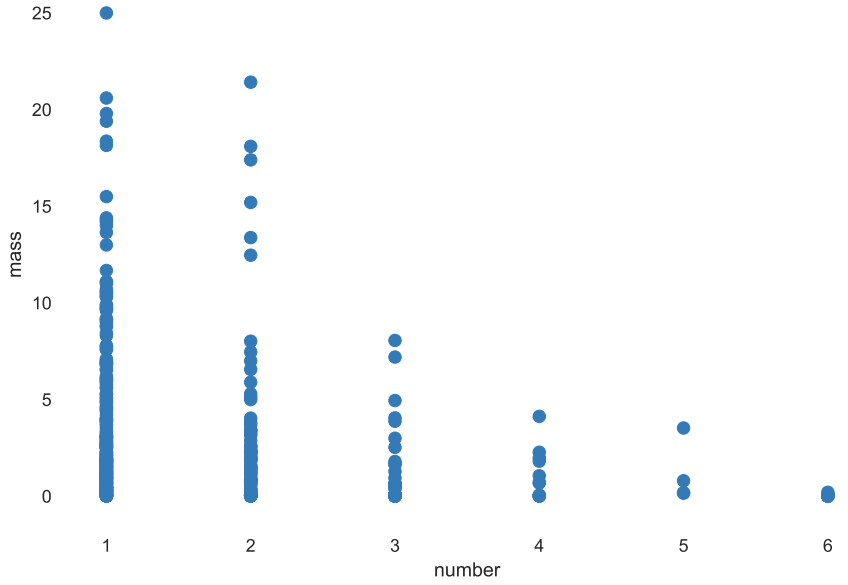
Value	Count	Frequency (%)
2014	52	5.0%
2013	118	11.4%
2012	140	13.5%
2011	185	17.9%
2010	102	9.9%
2009	98	9.5%
2008	74	7.1%
2007	53	5.1%
2006	31	3.0%
2005	39	3.8%

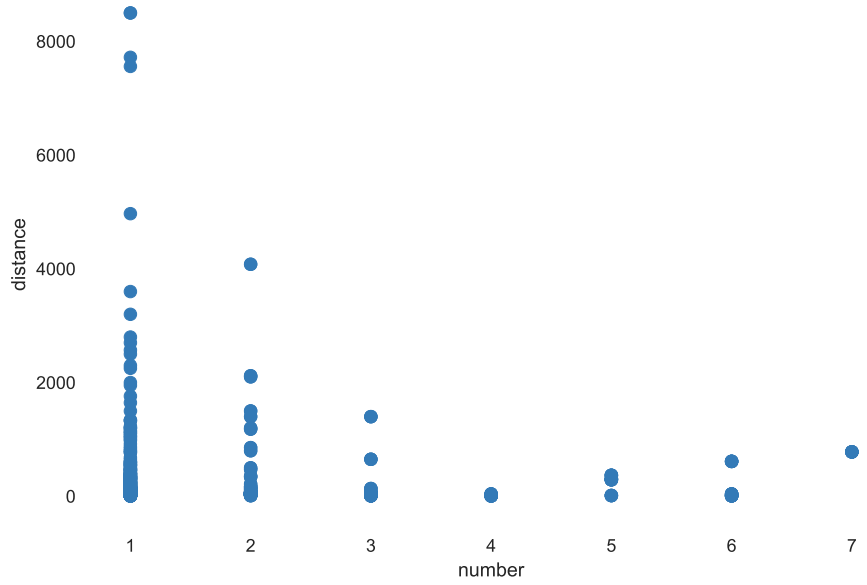
Interactions

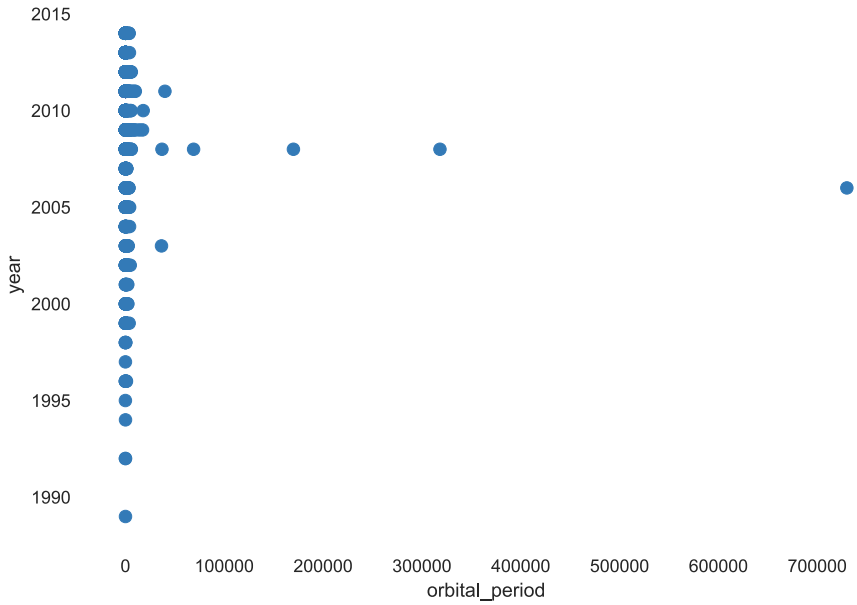


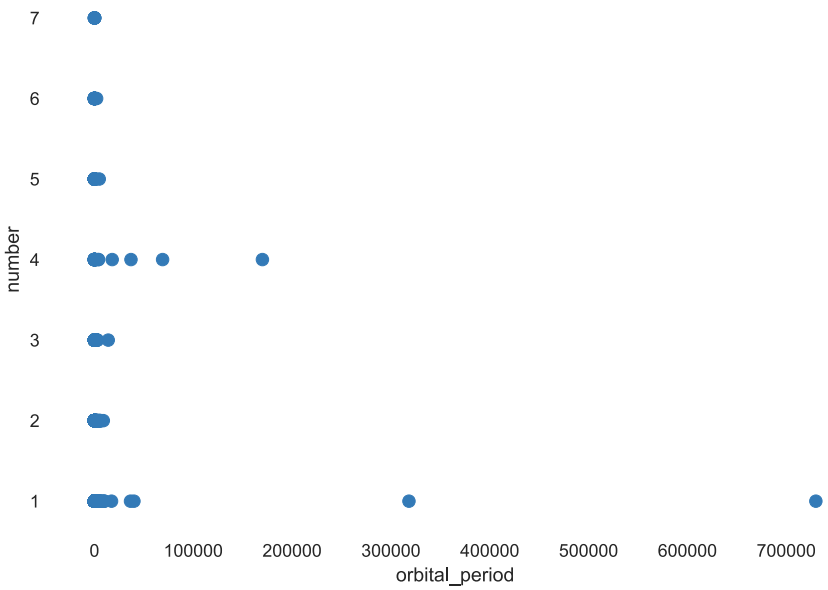

7
6
5
4
3
2
1
number
1
2
3
4
5
6
7
number

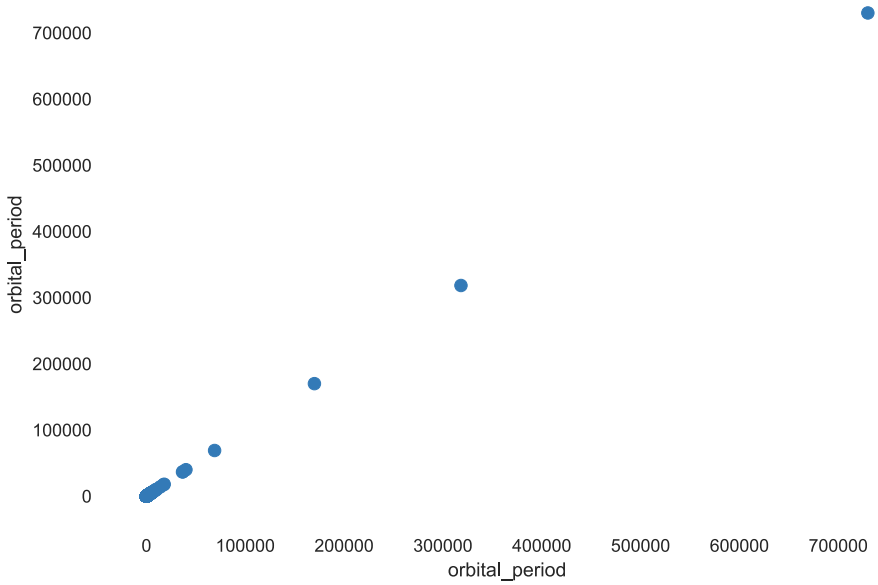

700000
600000
500000
400000
300000
200000
100000
0
orbital_period
1
2
3
4
5
6
7
number

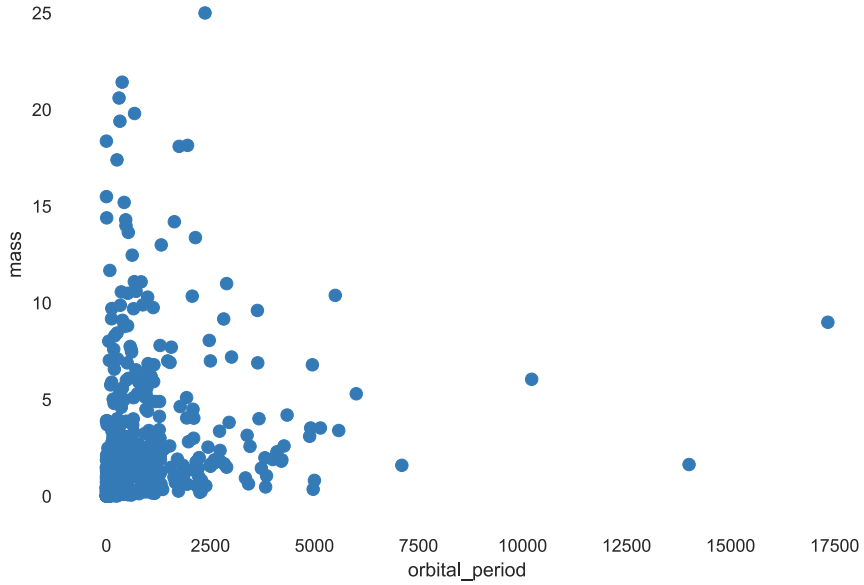

25
20
15
10
5
0
mass
1
2
3
4
5
6
number

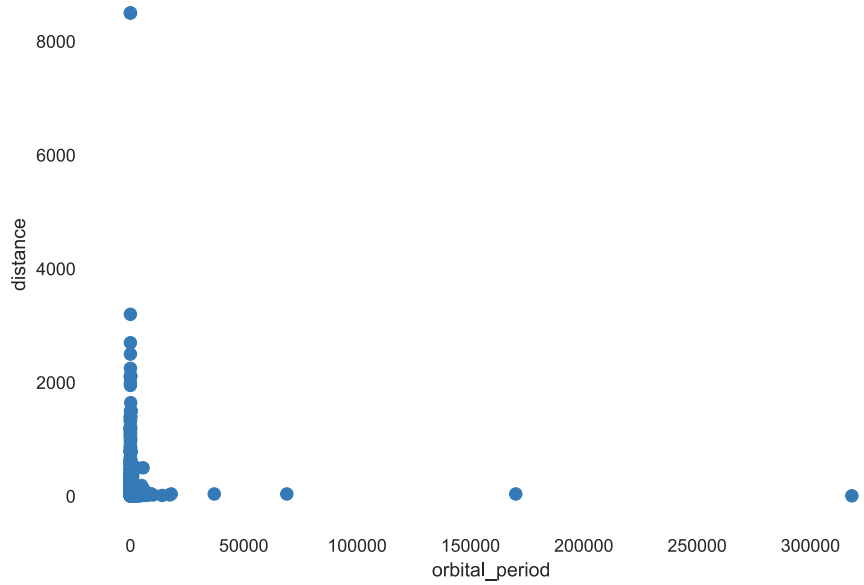

8000
6000
4000
2000
0
distance
1
2
3
4
5
6
7
number

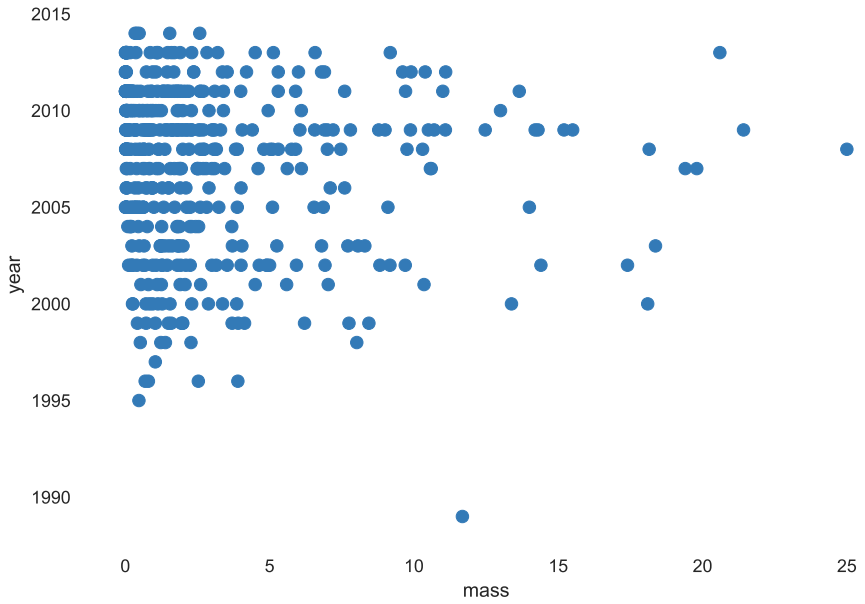

2015
2010
2005
2000
1995
1990
year
0
100000
200000
300000
400000
500000
600000
700000
orbital_period

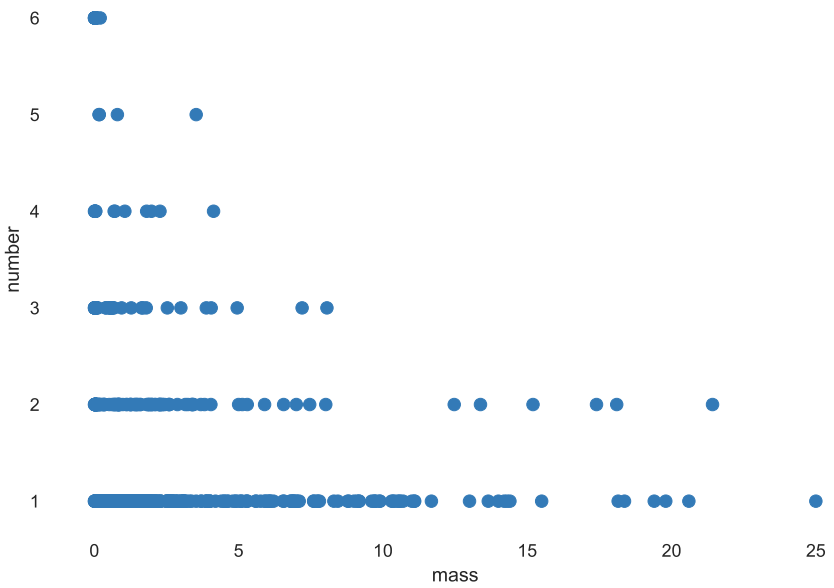

7
6
5
number
4
3
2
1
0
100000
200000
300000
400000
500000
600000
700000
orbital_period

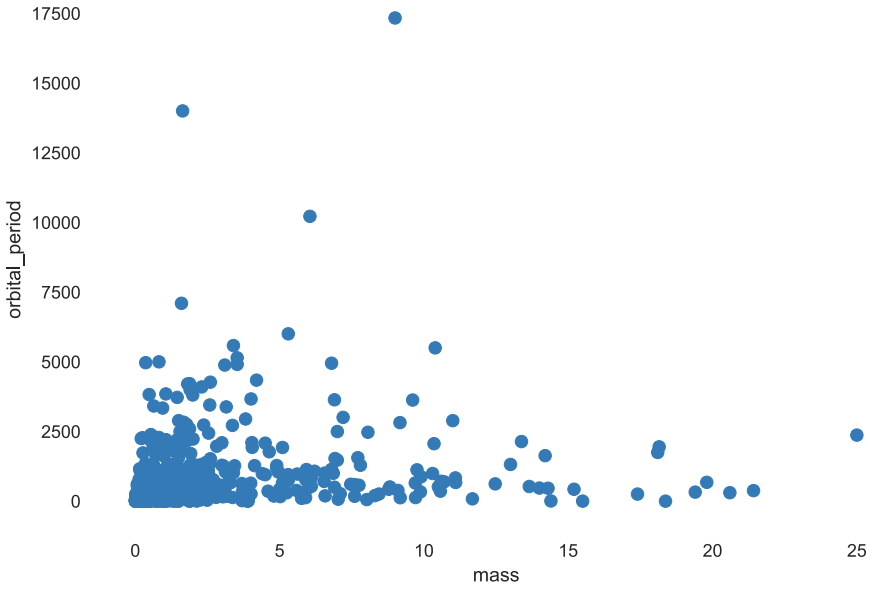

700000
600000
500000
400000
300000
200000
100000
0
orbital_period
0
100000
200000
300000
400000
500000
600000
700000
orbital_period

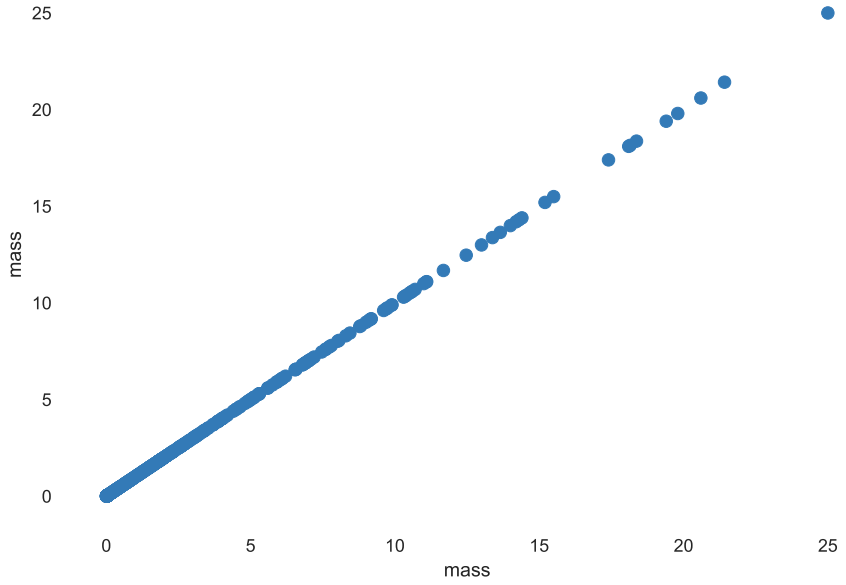

25
20
15
10
5
0
mass
0
2500
5000
7500
10000
12500
15000
17500
orbital_period

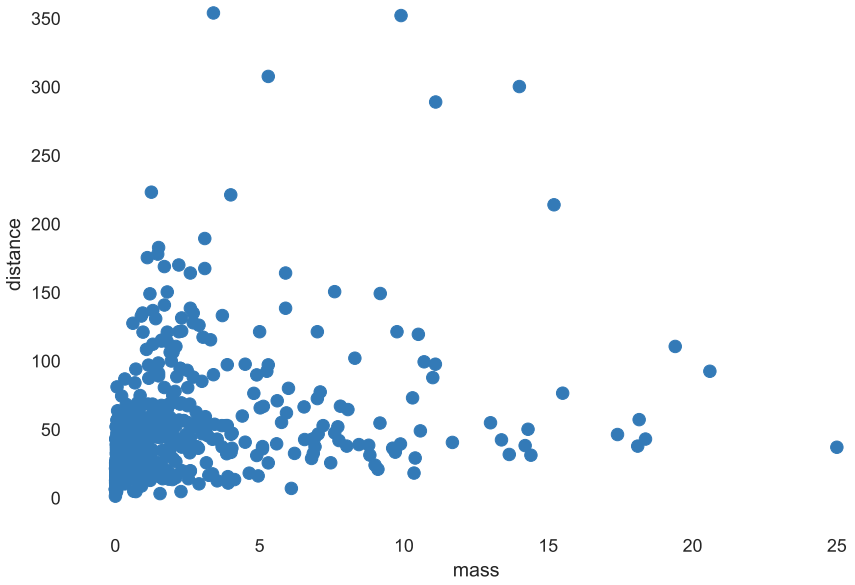

8000
6000
4000
2000
0
distance
0
50000
100000
150000
200000
250000
300000
orbital_period

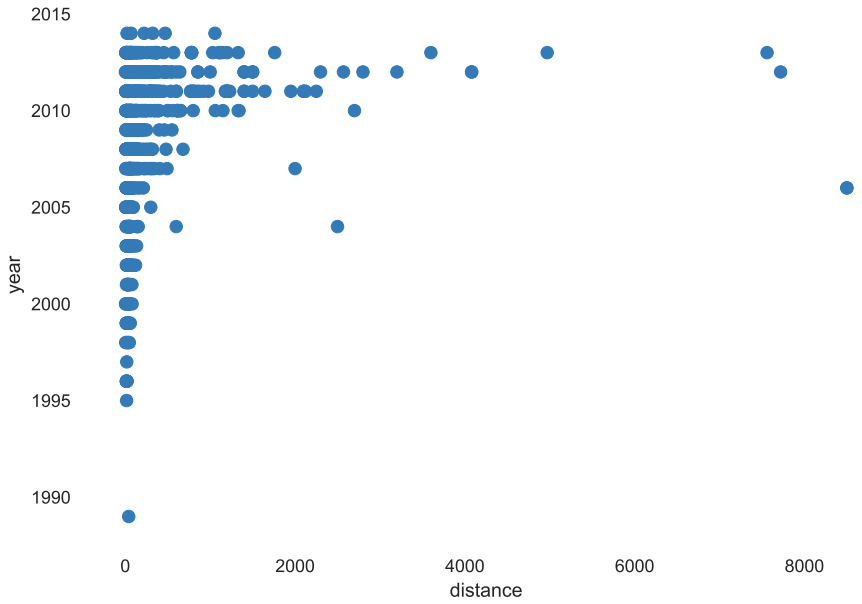

2015
2010
2005
2000
1995
1990
year
0
5
10
15
20
25
mass

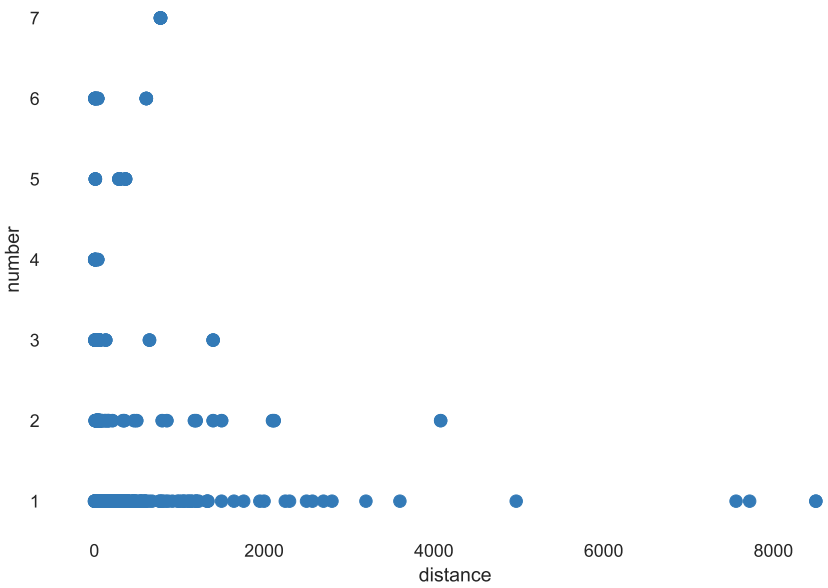

6
5
4
3
2
1
number
0
5
10
15
20
25
mass

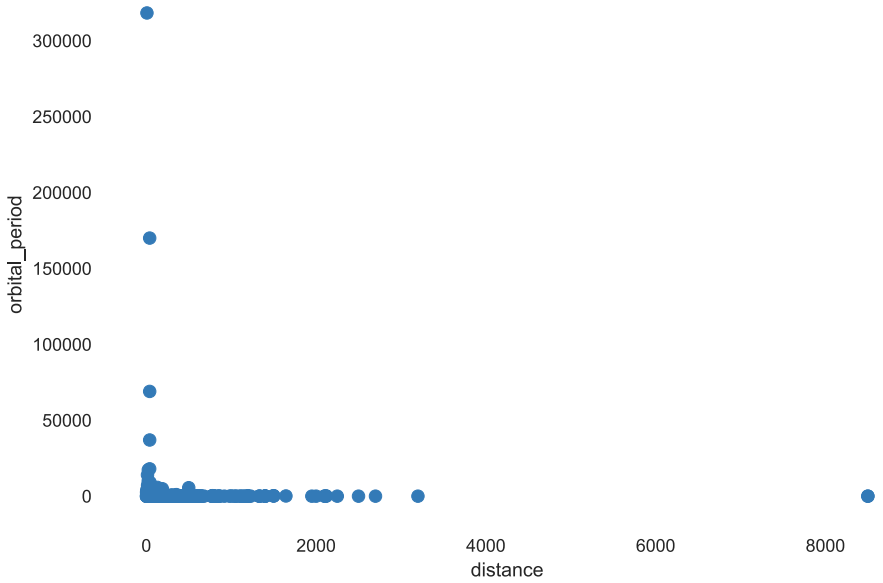

17500
15000
12500
10000
7500
5000
2500
0
orbital_period
0
5
10
15
20
25
mass

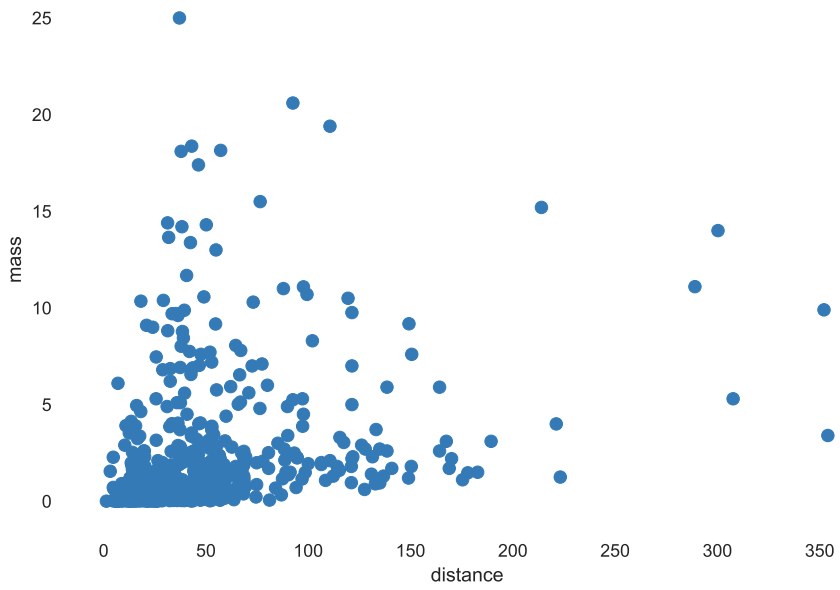

25
20
15
10
5
0
mass
0
5
10
15
20
25
mass

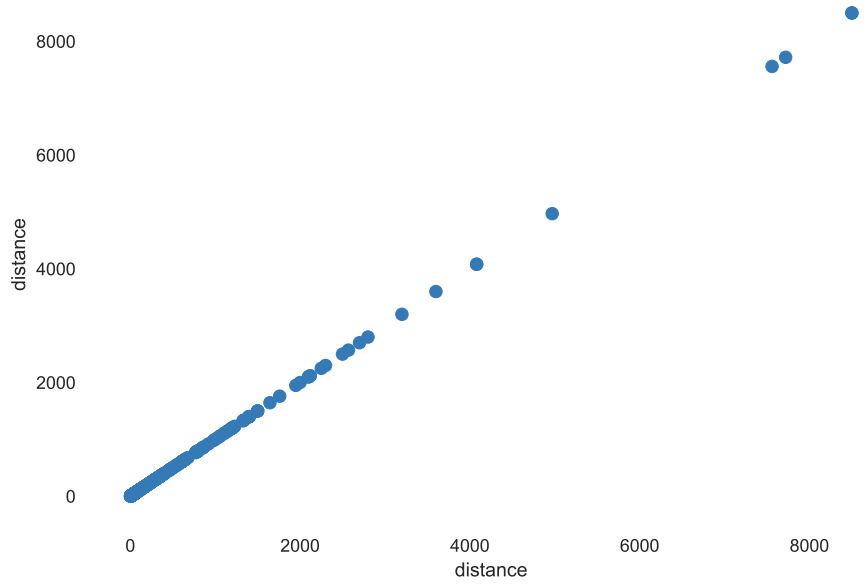

350
300
250
200
150
100
50
0
distance
0
5
10
15
20
25
mass

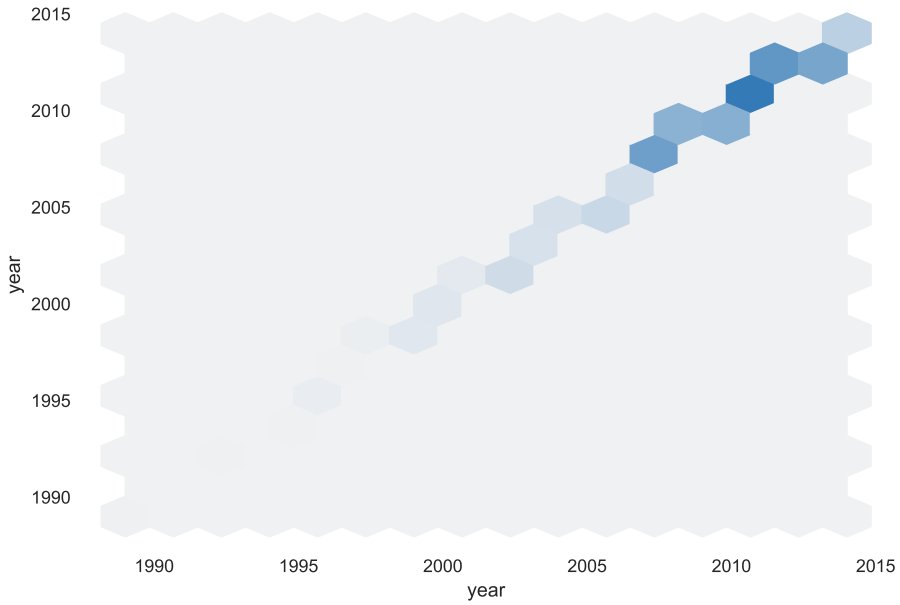

2015
2010
2005
2000
1995
1990
year
0
2000
4000
6000
8000
distance

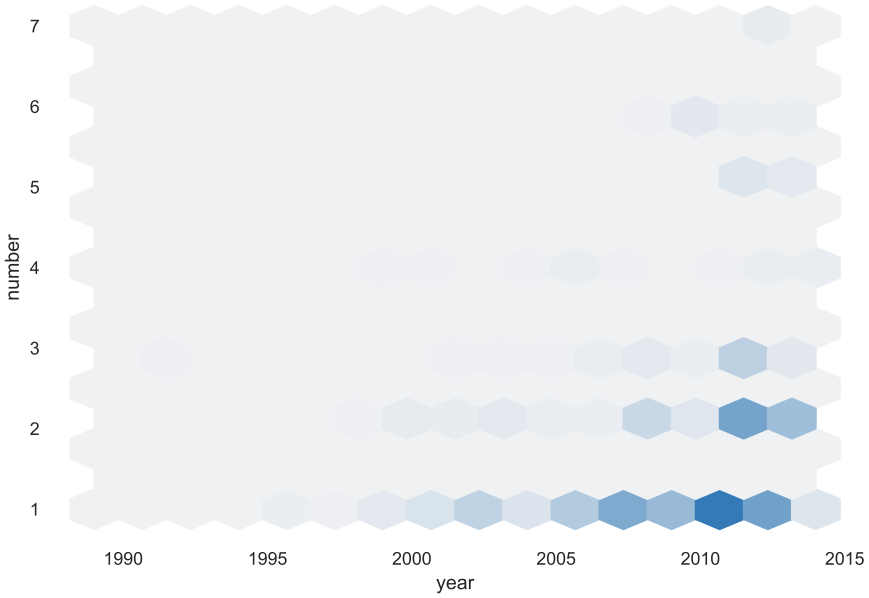

7
6
5
4
3
2
1
number
0
2000
4000
6000
8000
distance

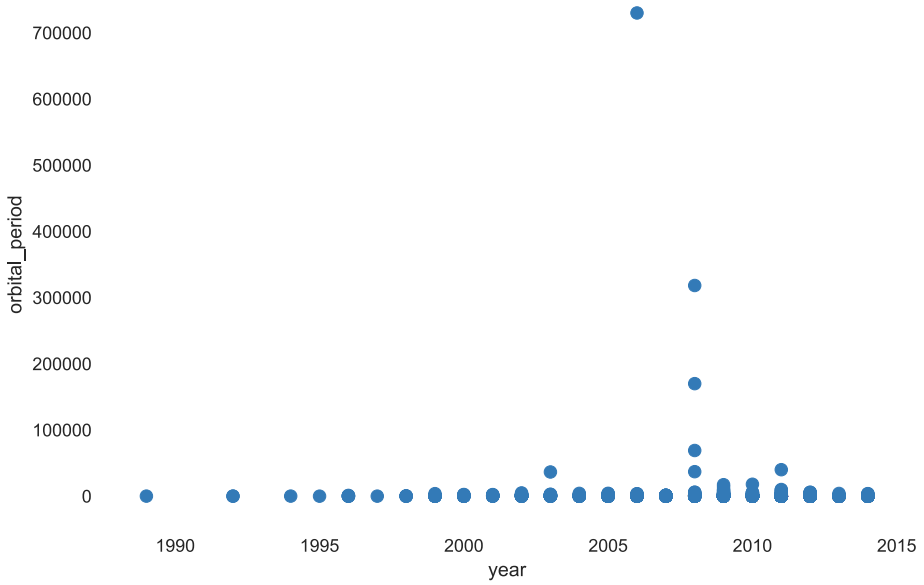

300000
250000
200000
150000
100000
50000
0
orbital_period
0
2000
4000
6000
8000
distance

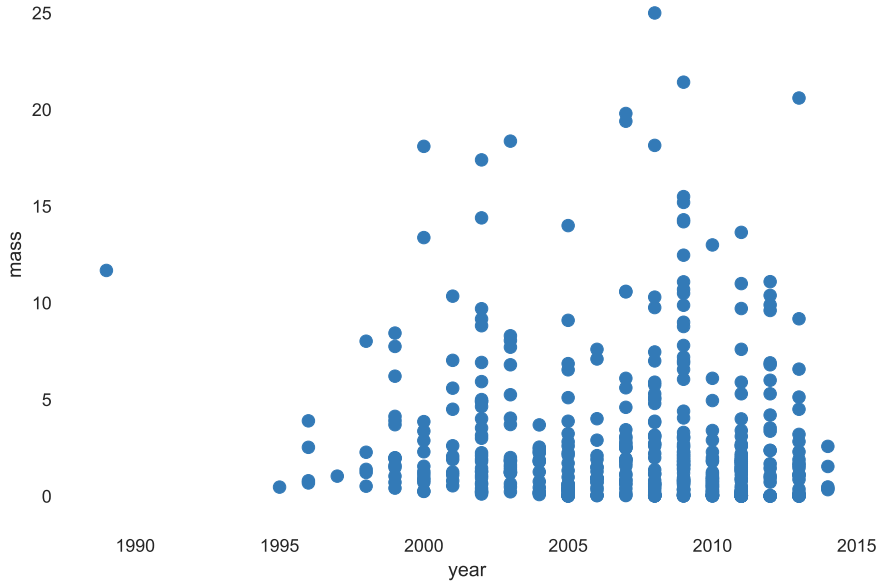

25
20
15
10
5
0
mass
0
50
100
150
200
250
300
350
distance

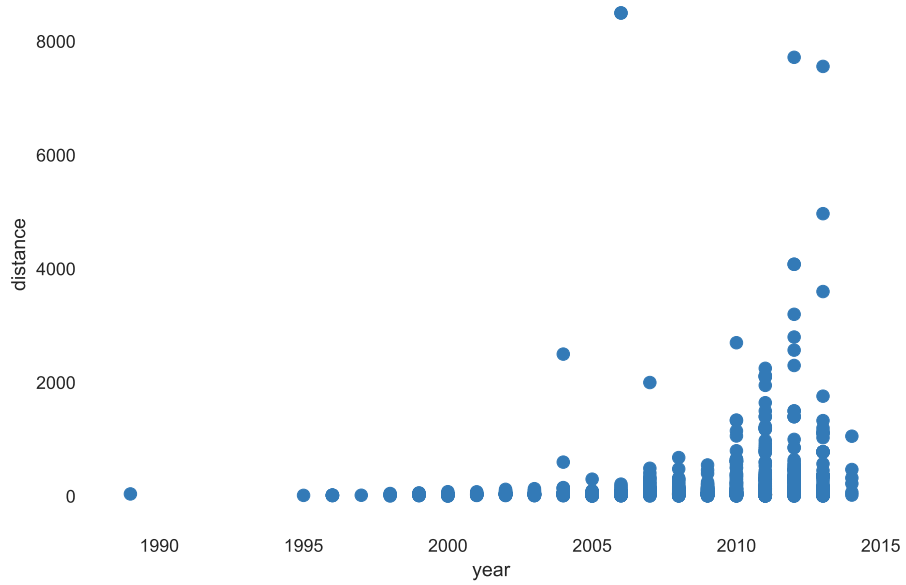

8000
6000
4000
2000
0
distance
0
2000
4000
6000
8000
distance


2015
2010
2005
2000
1995
1990
year
1990
1995
2000
2005
2010
2015
year

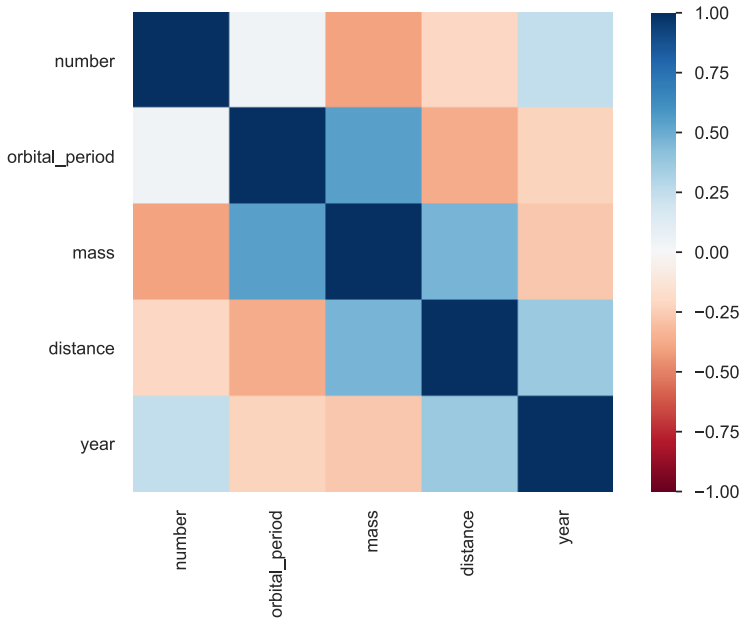

7
6
5
4
3
2
1
number
1990
1995
2000
2005
2010
2015
year

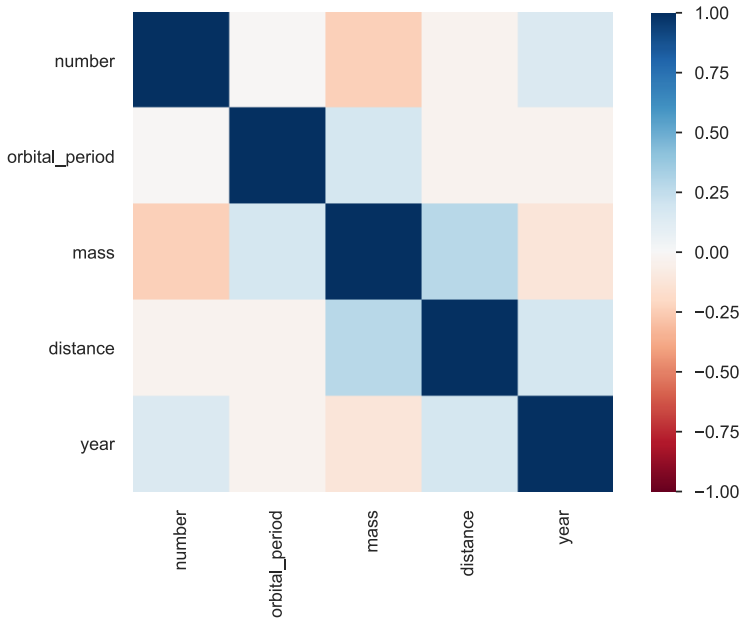

700000
600000
500000
400000
300000
200000
100000
0
orbital_period
1990
1995
2000
2005
2010
2015
year

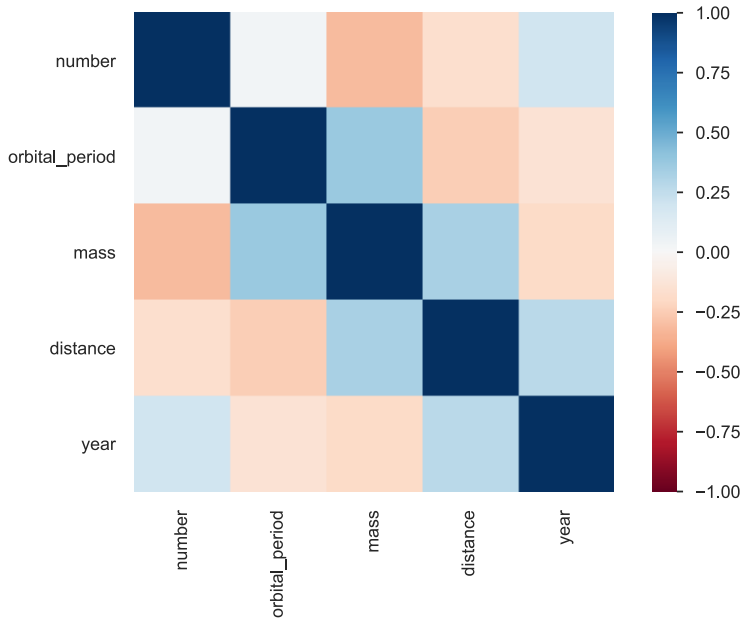

25
20
15
10
5
0
mass
1990
1995
2000
2005
2010
2015
year

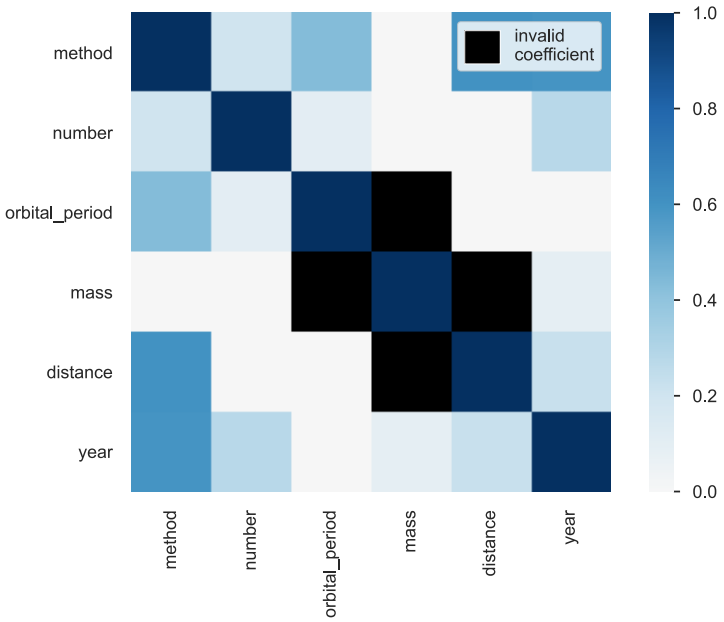

8000
6000
4000
2000
0
distance
1990
1995
2000
2005
2010
2015
year

Correlations

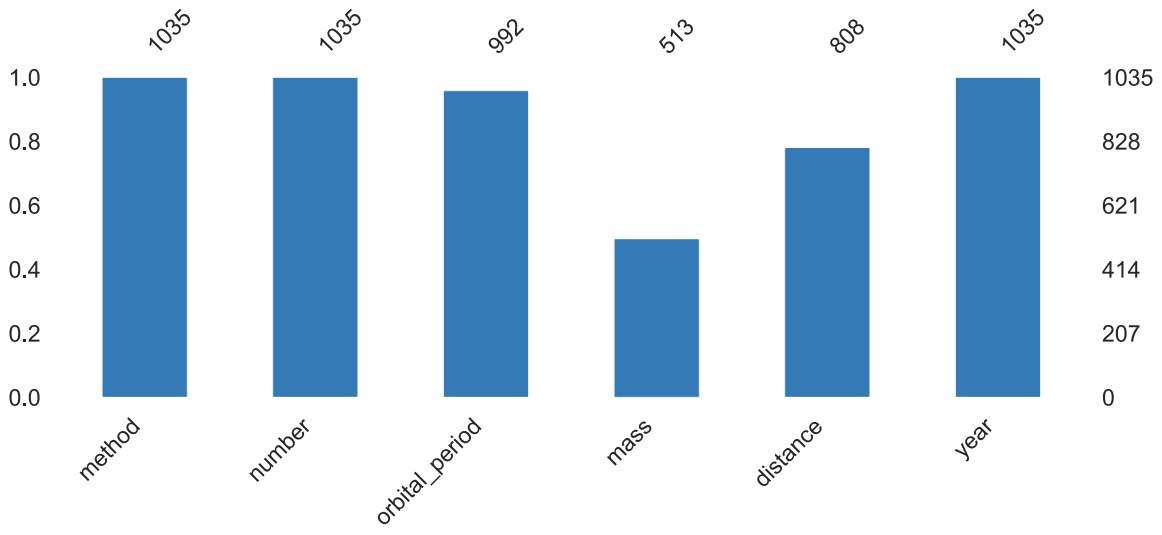



number
orbital_period
mass
distance
year
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

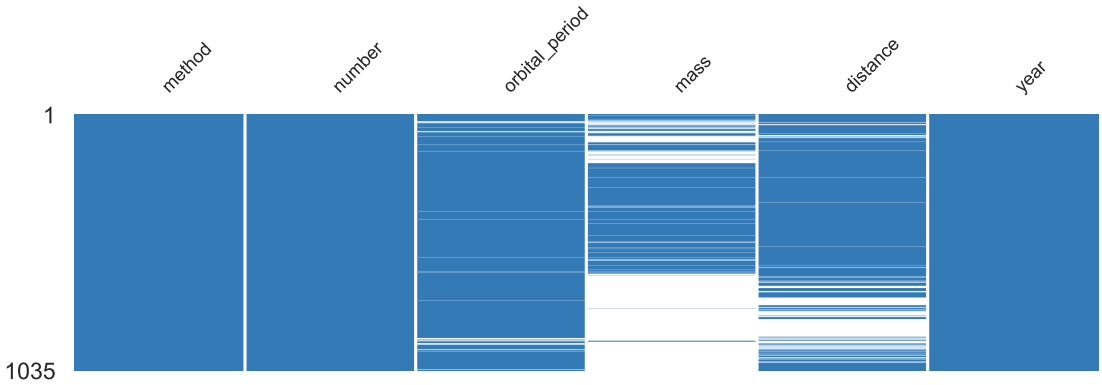

number
orbital_period
mass
distance
year
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
number
orbital_period
mass
distance
year


invalid
coefficient
method
number
orbital_period
mass
distance
year
1.0
0.8
0.6
0.4
0.2
0.0

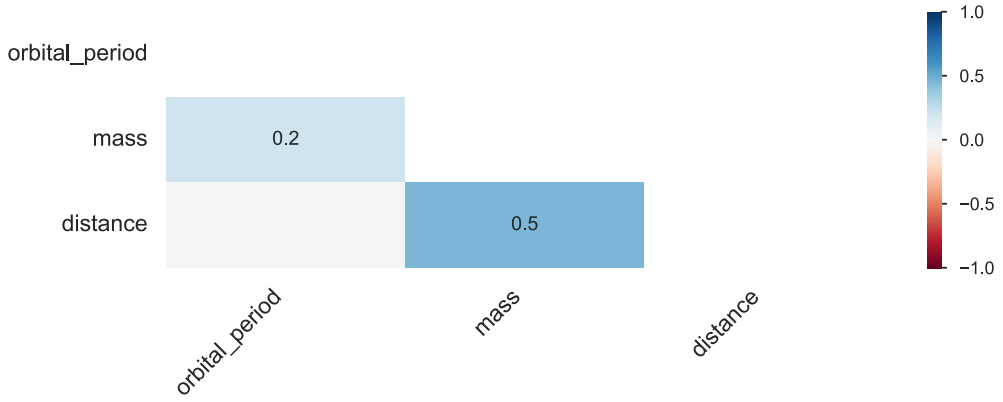
Missing values



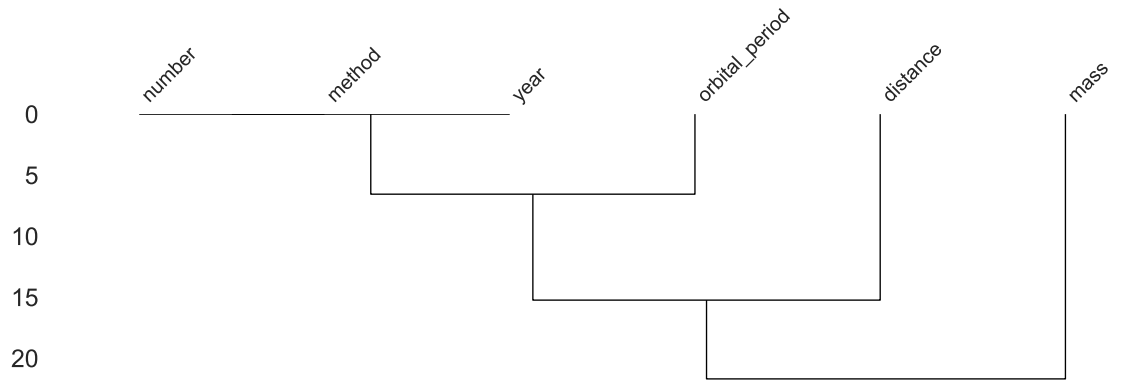
A simple visualization of nullity by column.



Nullity matrix is a data-dense display which lets you quickly visually pick out patterns in data completion.



The correlation heatmap measures nullity correlation: how strongly the presence or absence of one variable affects the presence of another.



The dendrogram allows you to more fully correlate variable completion, revealing trends deeper than the pairwise ones visible in the correlation heatmap.

Sample

First rows

	method	number	orbital_period	mass	distance	year
0	Radial Velocity	1	269.300	7.10	77.40	2006
1	Radial Velocity	1	874.774	2.21	56.95	2008
2	Radial Velocity	1	763.000	2.60	19.84	2011
3	Radial Velocity	1	326.030	19.40	110.62	2007
4	Radial Velocity	1	516.220	10.50	119.47	2009
5	Radial Velocity	1	185.840	4.80	76.39	2008
6	Radial Velocity	1	1773.400	4.64	18.15	2002
7	Radial Velocity	1	798.500	NaN	21.41	1996
8	Radial Velocity	1	993.300	10.30	73.10	2008
9	Radial Velocity	2	452.800	1.99	74.79	2010

Last rows

	method	number	orbital_period	mass	distance	year
1025	Transit	1	3.067850	NaN	60.0	2012
1026	Transit	1	0.925542	NaN	470.0	2014
1027	Imaging	1	NaN	NaN	19.2	2011
1028	Transit	1	3.352057	NaN	3200.0	2012
1029	Imaging	1	NaN	NaN	10.1	2012
1030	Transit	1	3.941507	NaN	172.0	2006
1031	Transit	1	2.615864	NaN	148.0	2007
1032	Transit	1	3.191524	NaN	174.0	2007
1033	Transit	1	4.125083	NaN	293.0	2008
1034	Transit	1	4.187757	NaN	260.0	2008

Report generated with pandas-profiling (https://github.com/pandas-profiling/pandas-profiling).