Exploratory Data Analysis (EDA)

Statsomat.com

19 April 2021

Basic Information

Automatic statistics for the file:

File
Baseball_small.csv

Your selection for the encoding: Auto
Your selection for the decimal character: Auto
Observations (rows with at least one non-missing value): 18
Variables (columns with at least one non-missing value): 23
Variables considered continuous: 16

Variables considered continuous
assist86
atbat
atbat86
error86
hits
hits86
homer86
homeruns
outs86
rbi
rbi86
runs
runs86
V1
walks
walks86

Variables considered categorical: 7

Variables considered categorical
div86
league86
name1
name2
posit86
team86
years

Results for Numerical Variables

Descriptive Statistics

Variables are sorted alphabetically. Missings are omitted in stats. CV only for positive variables.

	N Obs	N Missing	N Valid	% Complete	N Unique	Mean	SD	Median	MAD	Min	Max	Skewness	Kurtosis	CV
assist86	18	0	18	100	16	107.78	151.52	23.5	28.91	1	444	1.44	0.75	1.41
atbat	18	0	18	100	18	1872.39	1964.87	1160.5	825.07	84	6840	1.55	1.45	1.05
atbat86	18	0	18	100	18	316.83	156.39	268.5	136.4	22	589	0.23	-0.51	0.49
error86	18	0	18	100	12	7.28	4.98	5.5	4.45	1	20	1.18	1.26	0.68
hits	18	0	18	100	18	504.56	556.62	307.5	248.34	26	1910	1.59	1.53	1.1
hits86	18	0	18	100	18	82.33	41.53	72.5	37.07	10	157	0.28	-0.77	0.5
homer86	18	0	18	100	14	8.06	6.55	5.5	5.19	1	21	0.9	-0.49	0.81
homeruns	18	0	18	100	17	42.17	61.25	24	25.95	2	259	2.93	9.74	1.45
outs86	18	0	18	100	17	241.89	168	214.5	101.56	59	812	2.42	7.81	0.69
rbi	18	0	18	100	18	222.61	269.25	119.5	120.83	9	1067	2.12	4.94	1.21
rbi86	18	0	18	100	16	37.83	23.48	28.5	21.5	2	86	0.71	-0.33	0.62
runs	18	0	18	100	17	248.33	274.21	144.5	122.31	9	915	1.46	0.97	1.1
runs86	18	0	18	100	16	42.56	23.5	41	19.27	4	95	0.84	0.74	0.55
V1	18	0	18	100	18	132.94	76.2	114.5	97.11	22	273	0.19	-1.16	0.57
walks	18	0	18	100	18	154.83	173.99	98	92.66	3	576	1.6	1.74	1.12
walks86	18	0	18	100	16	27.22	18.58	21.5	13.34	1	64	1.04	0.26	0.68

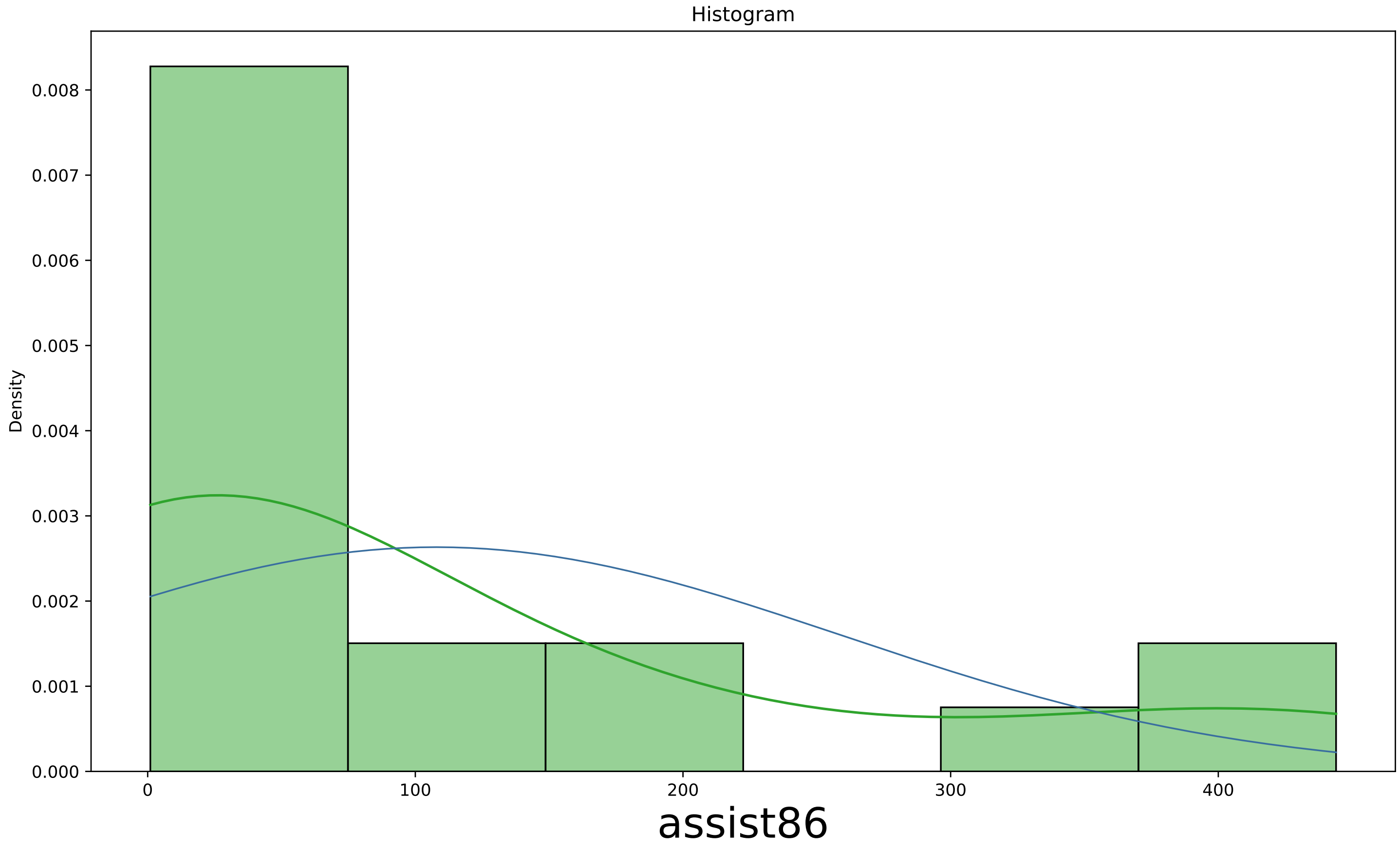
Graphics

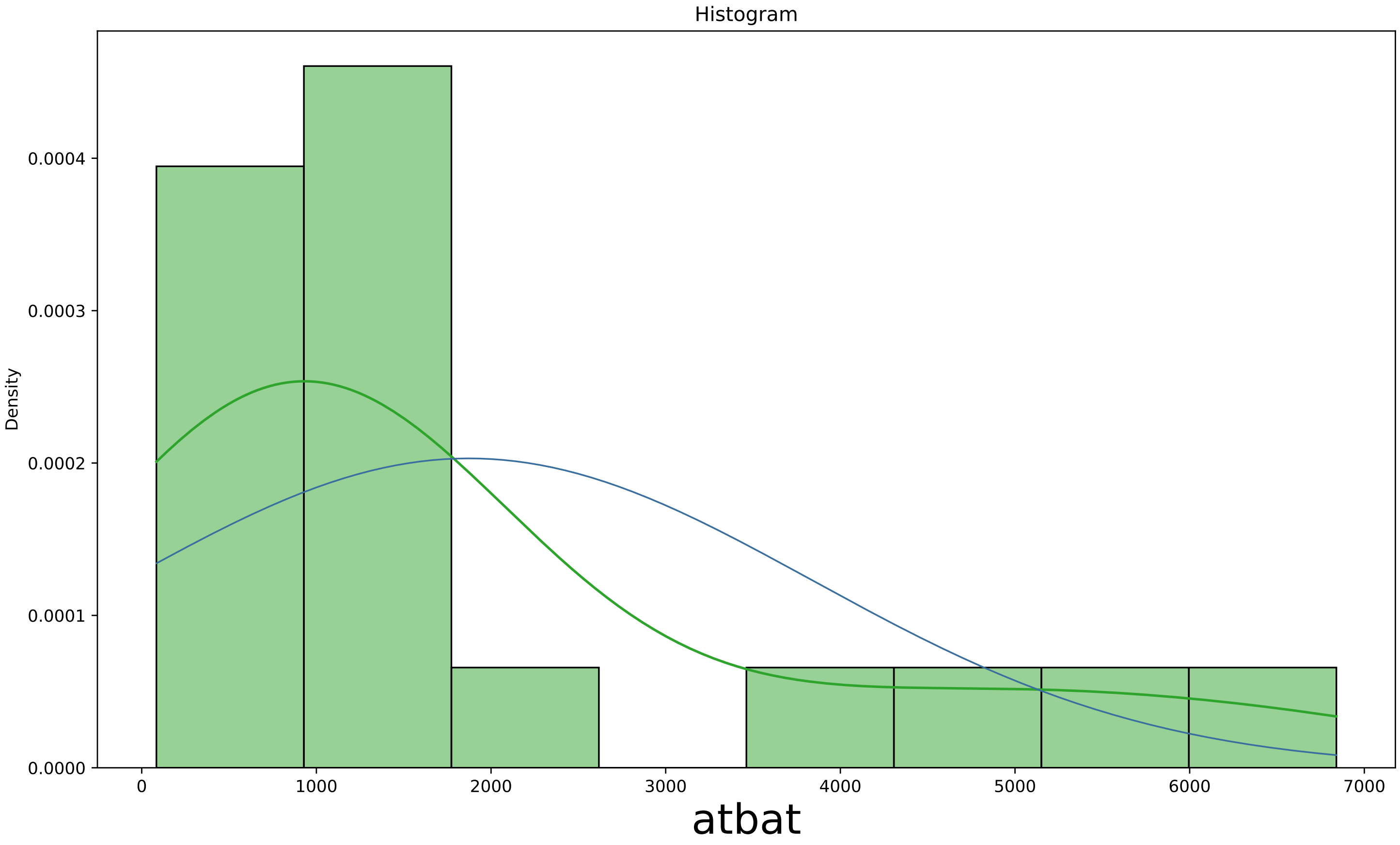
Histograms

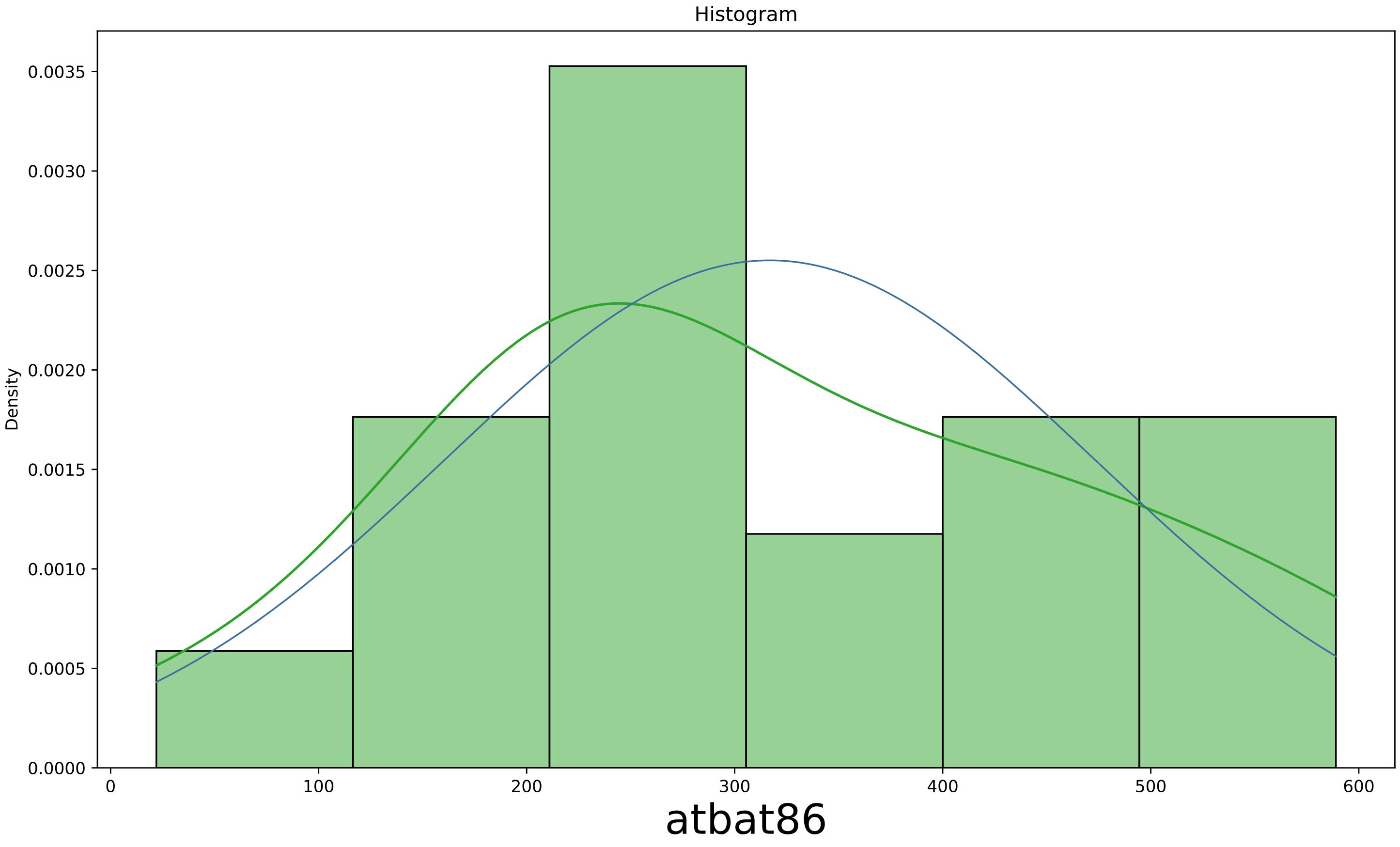
Details: Density Histograms. One large figure per page for each variable, sorted alphabetically. The blue line represents the normal density approximation. The green line represents a special kernel density approximation.

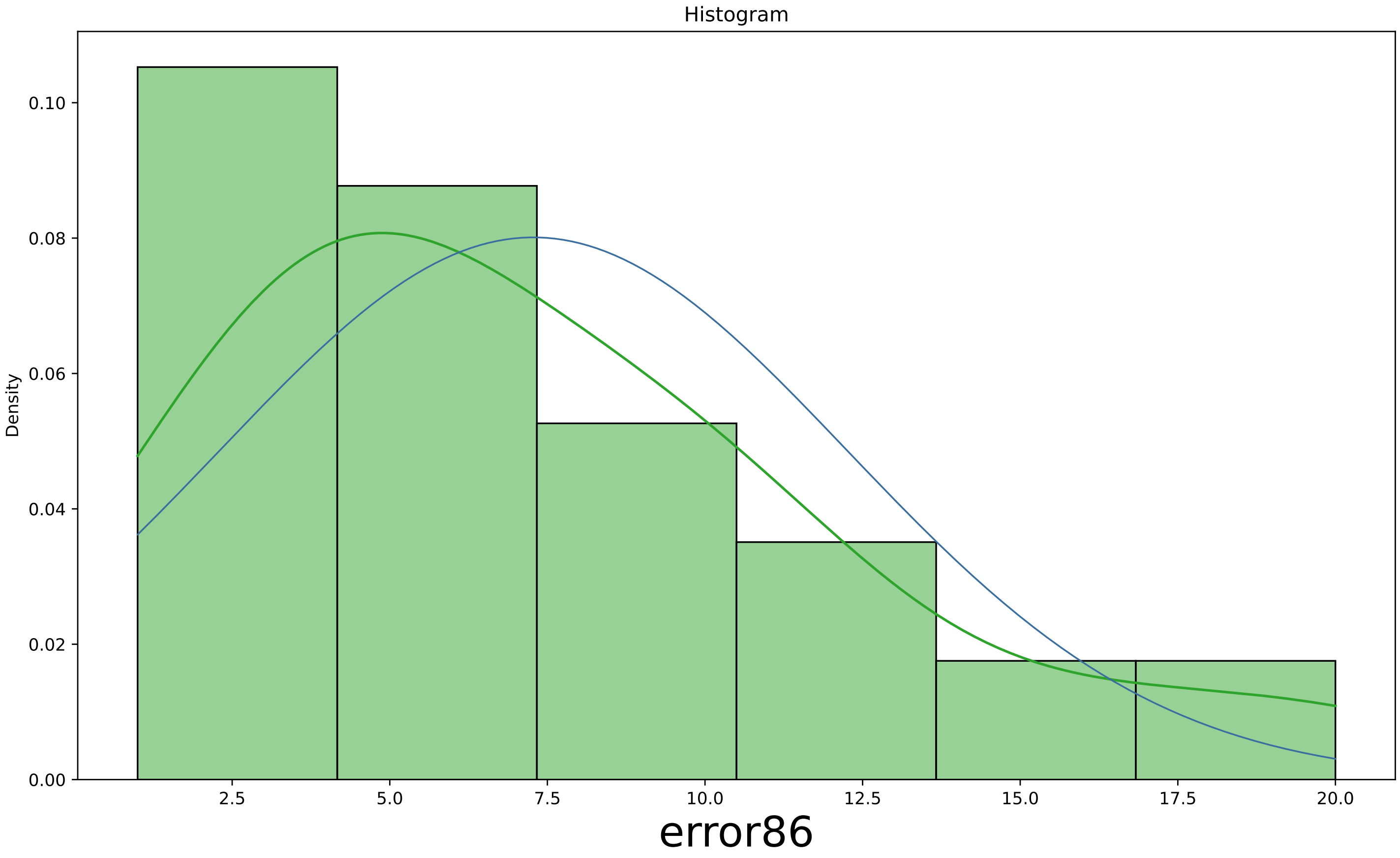
See figures on next page.

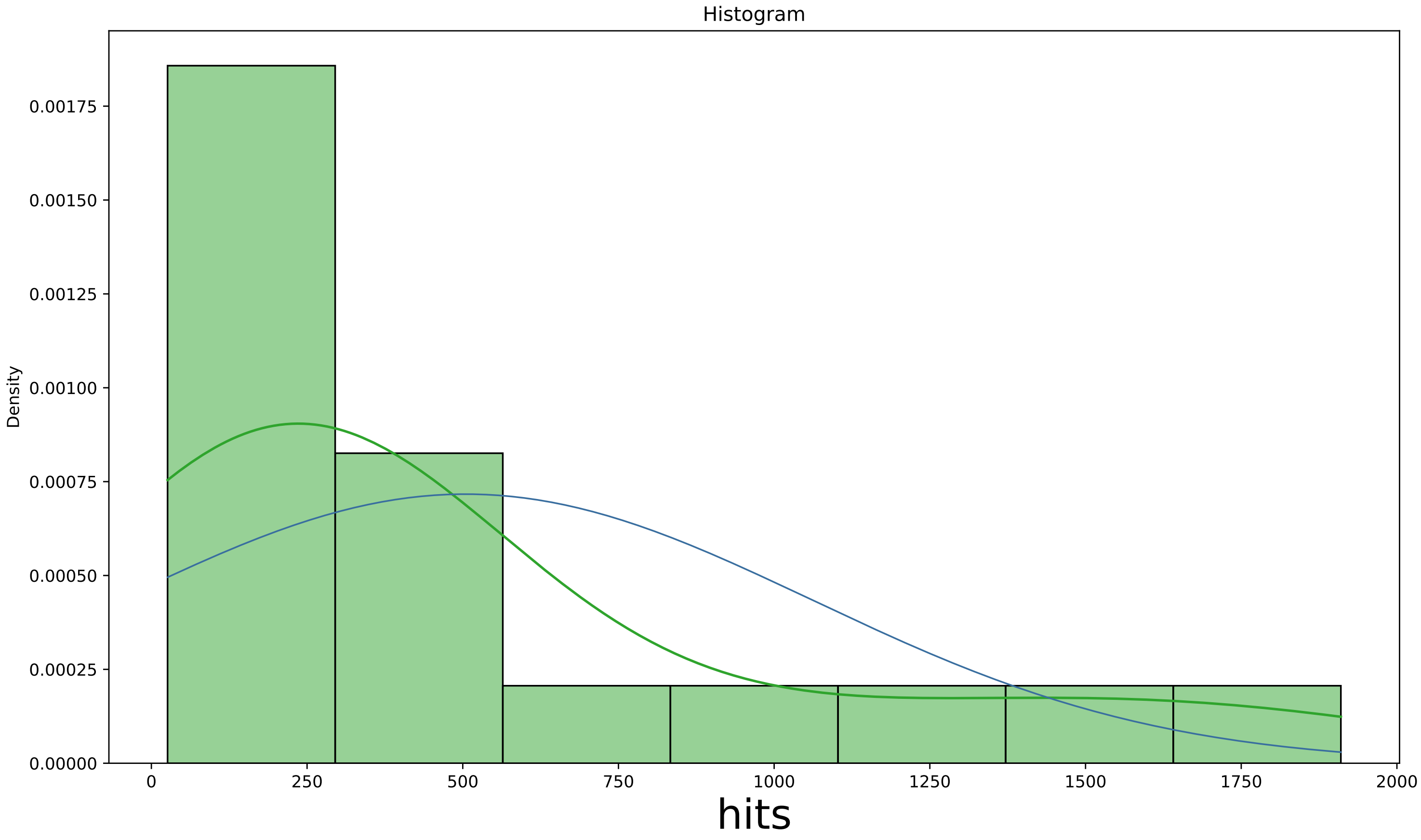
```
[]
```

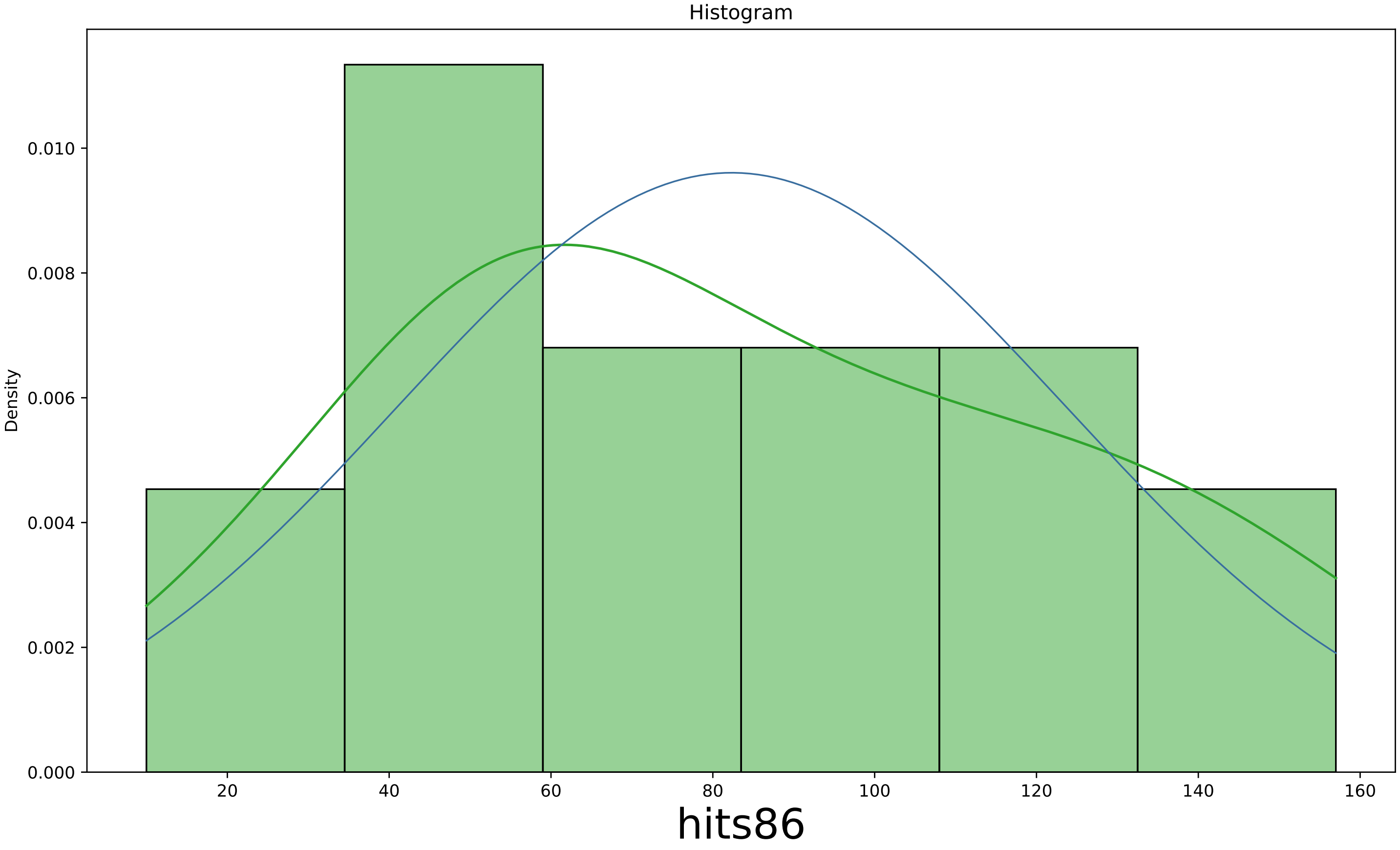

Histogram
Density
0.000
0.001
0.002
0.003
0.004
0.005
0.006
0.007
0.008
0
100
200
300
400
assist86

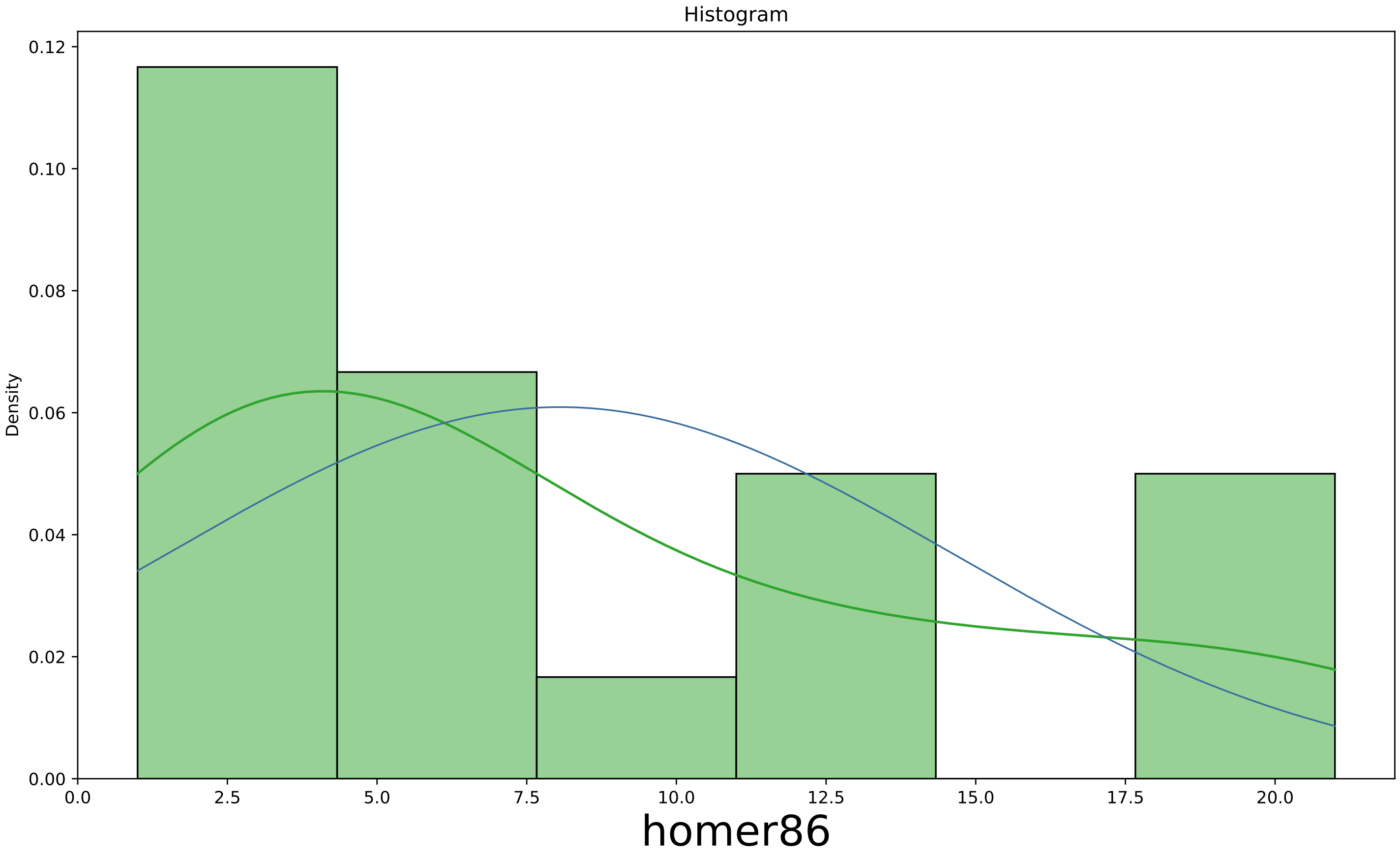

Histogram
Density
0.0000
0.0001
0.0002
0.0003
0.0004
0
1000
2000
3000
4000
5000
6000
7000
atbat

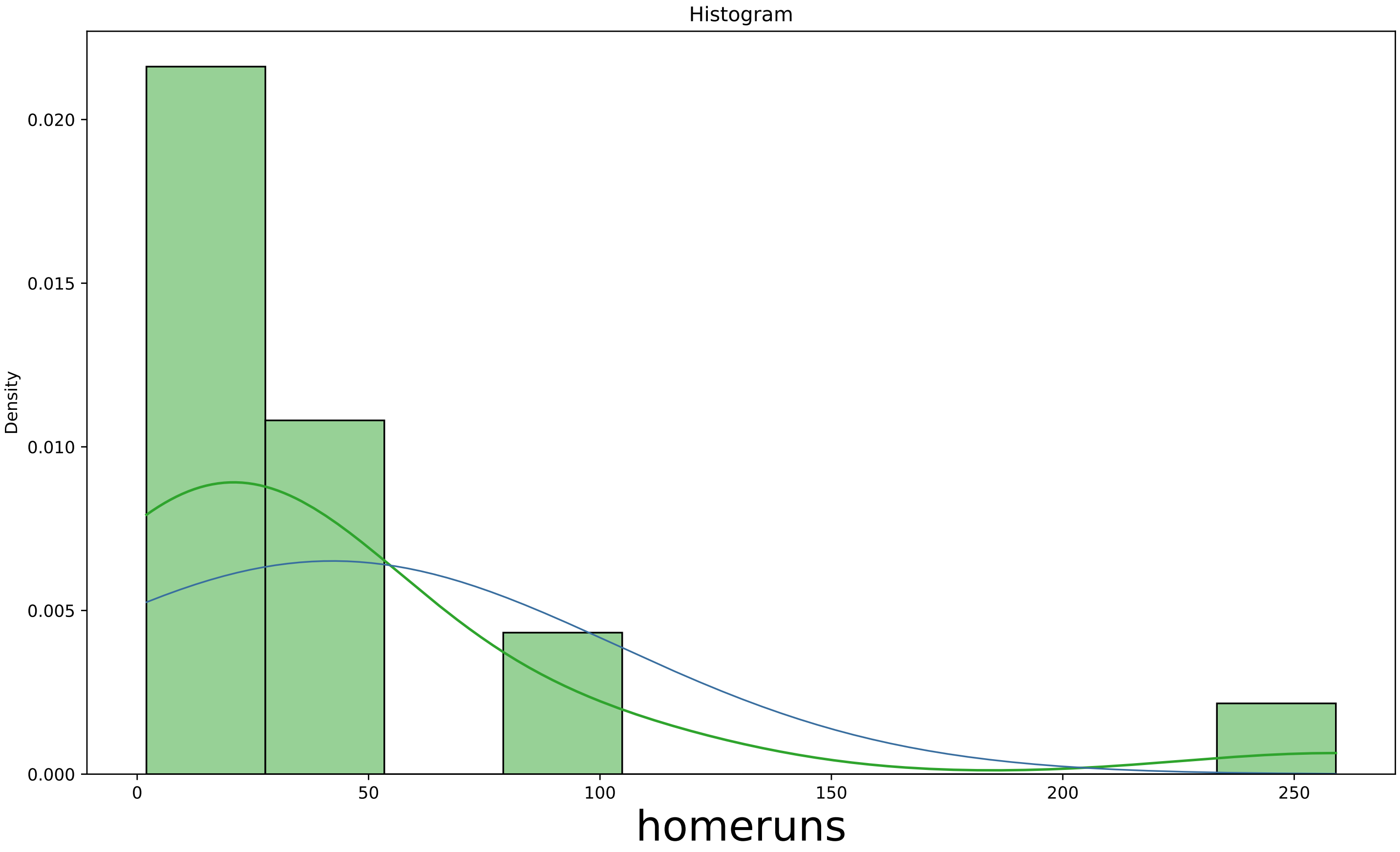

Histogram
Density
0.0000
0.0005
0.0010
0.0015
0.0020
0.0025
0.0030
0.0035
0
100
200
300
400
500
600
atbat86

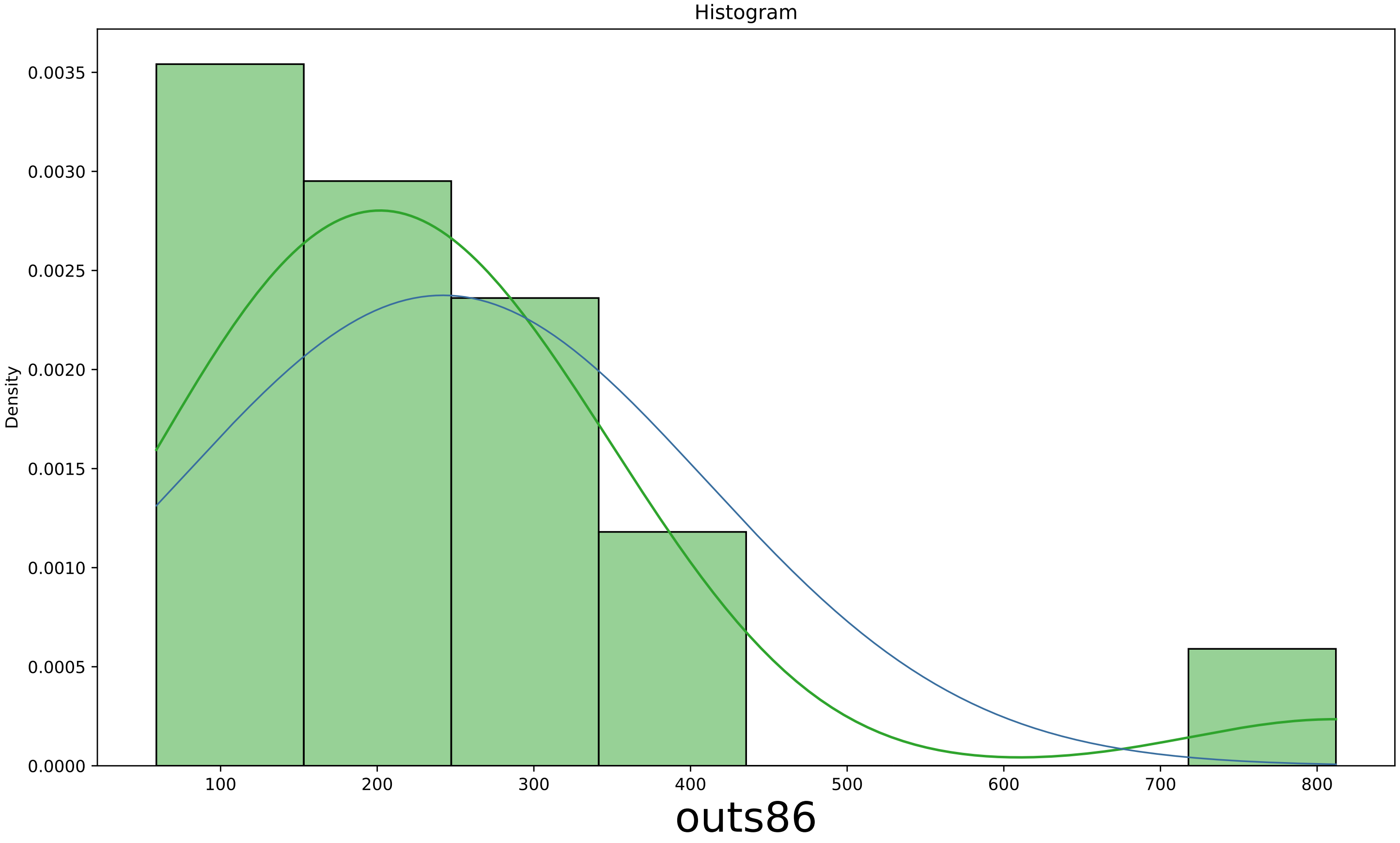

Histogram
Density
0.00
0.02
0.04
0.06
0.08
0.10
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
error86

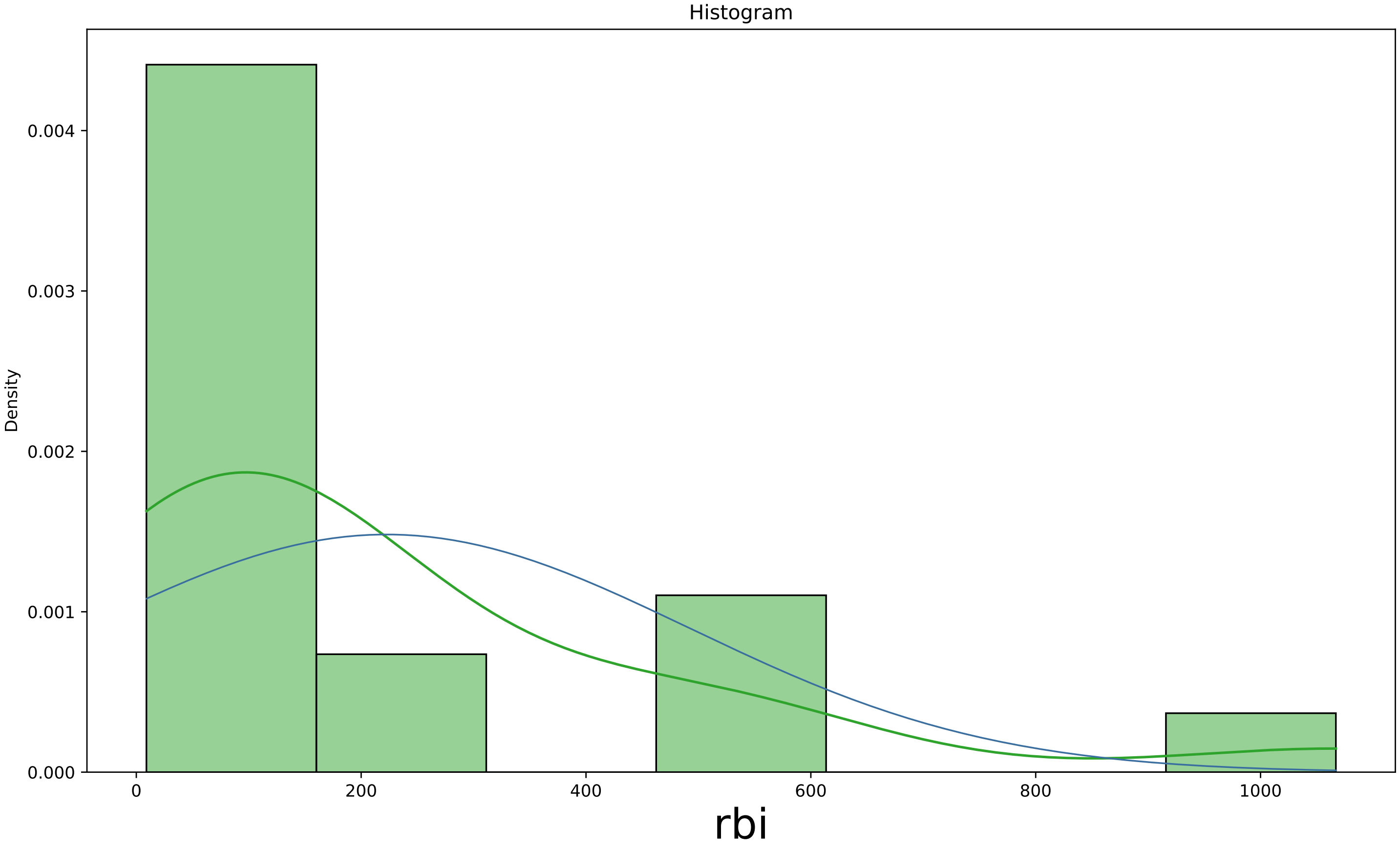

Histogram
Density
0.00000
0.00025
0.00050
0.00075
0.00100
0.00125
0.00150
0.00175
0
250
500
750
1000
1250
1500
1750
2000
hits

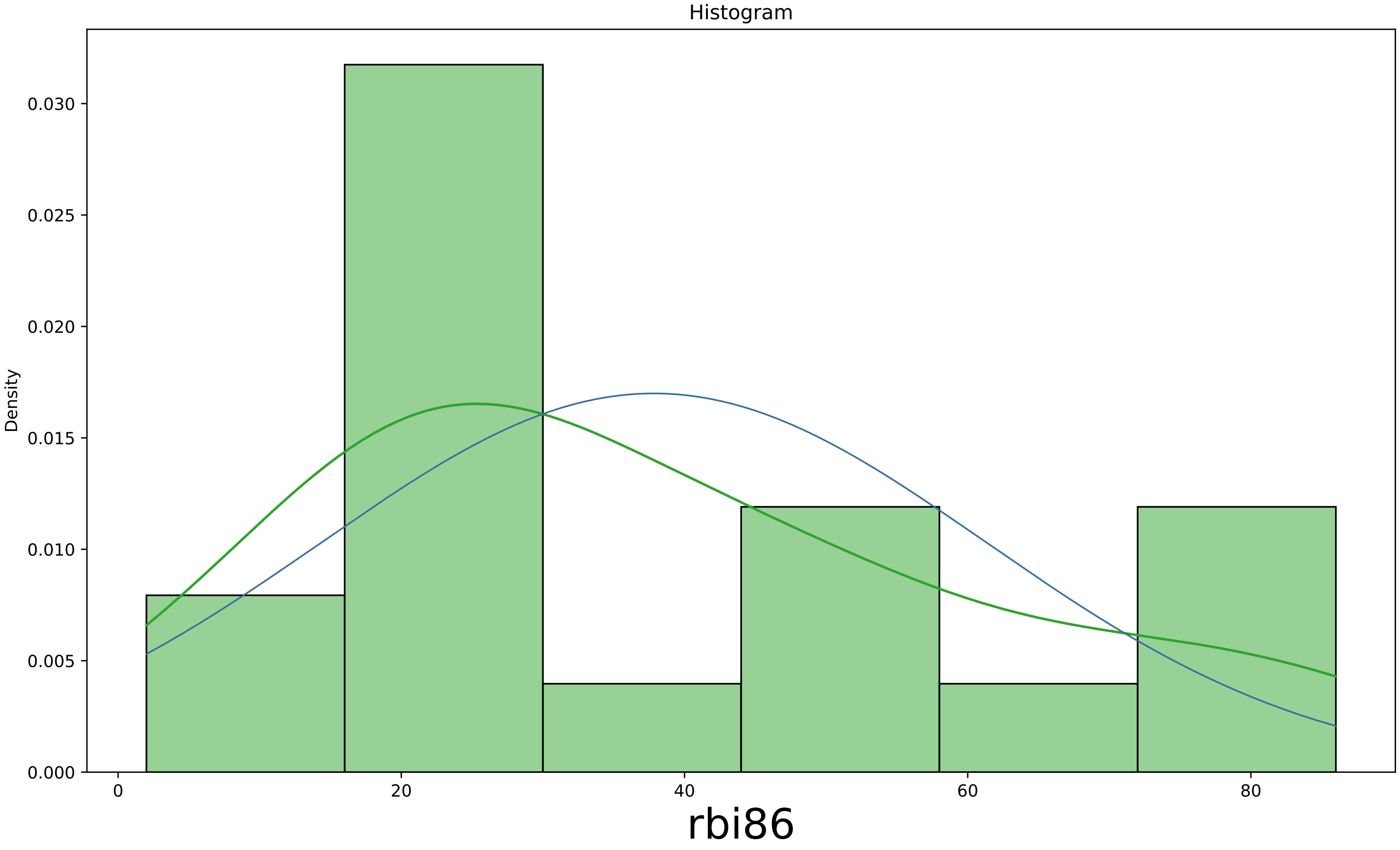

Histogram
Density
0.000
0.002
0.004
0.006
0.008
0.010
20
40
60
80
100
120
140
160
hits86

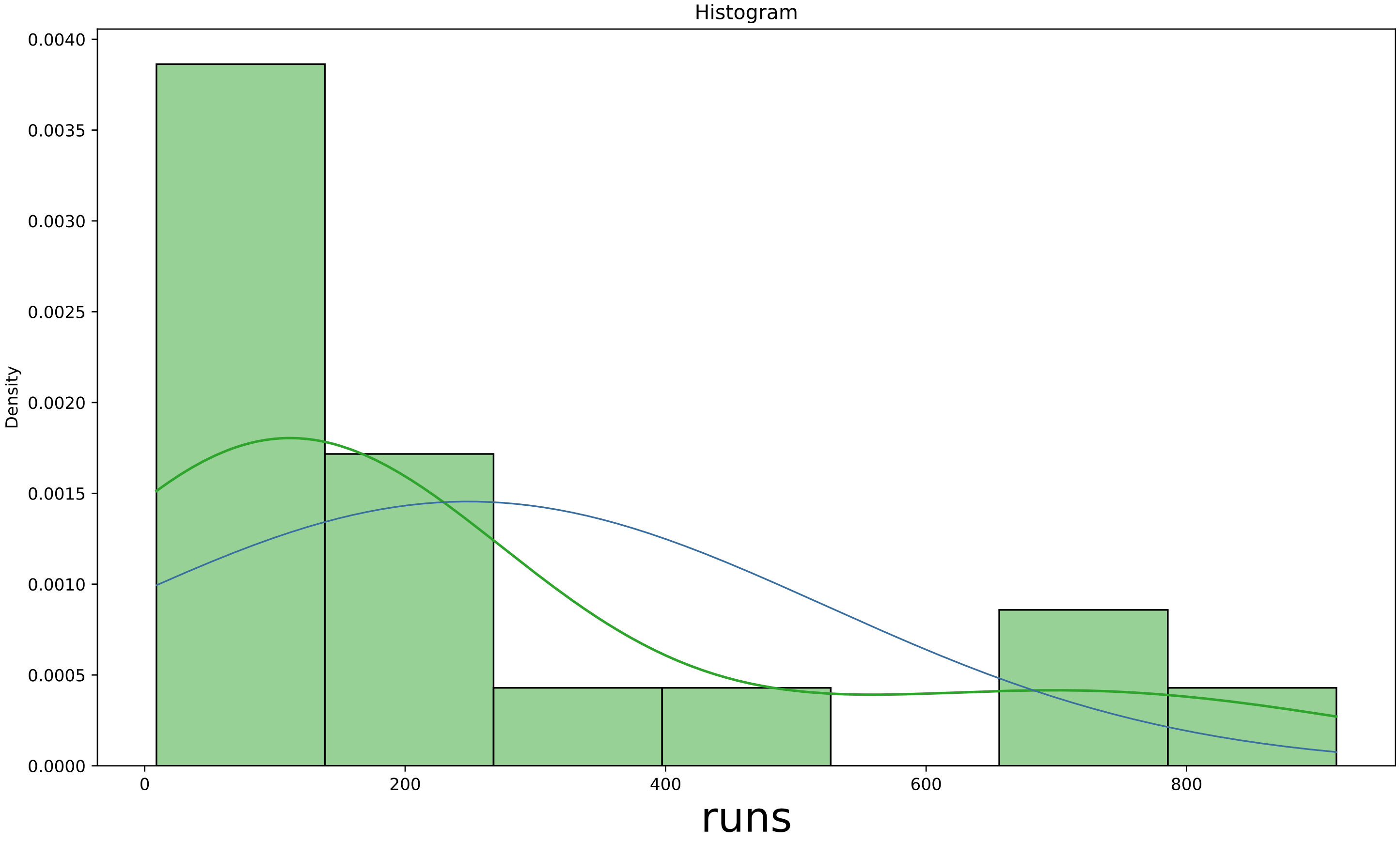

Histogram
Density
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
homer86

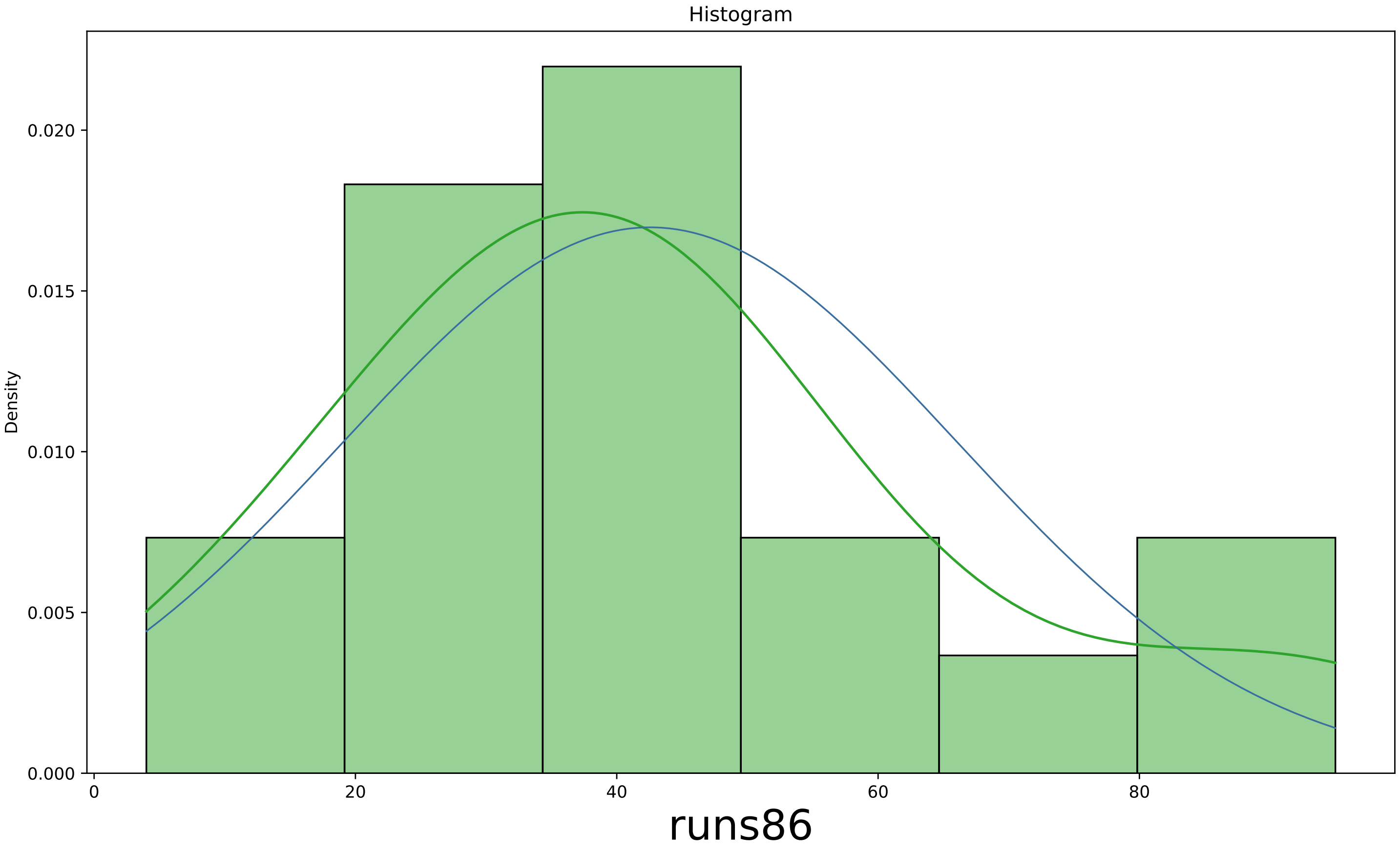

Histogram
Density
0.000
0.005
0.010
0.015
0.020
0
50
100
150
200
250
homeruns

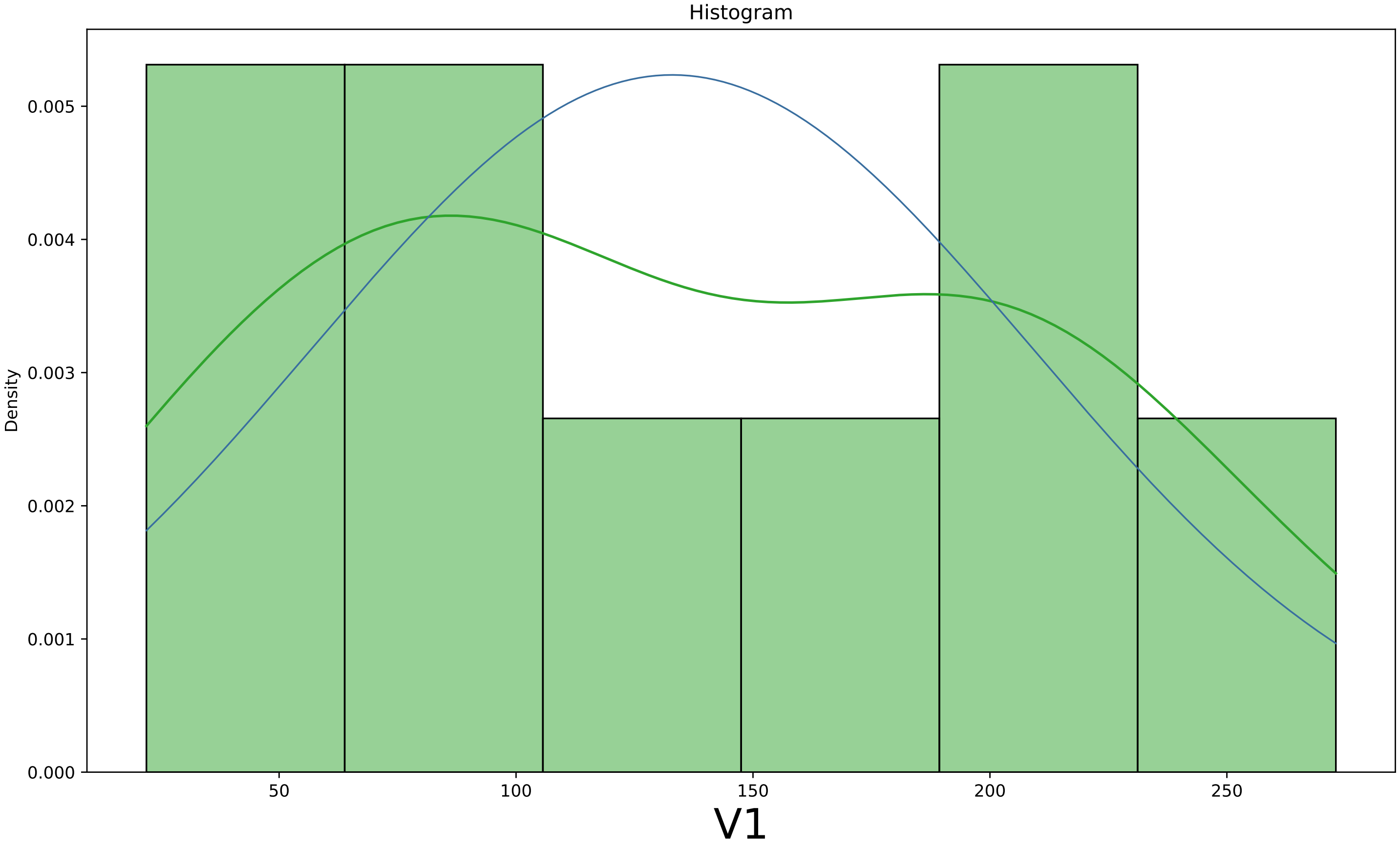

Histogram
Density
0.0035
0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
100
200
300
400
500
600
700
800
outs86

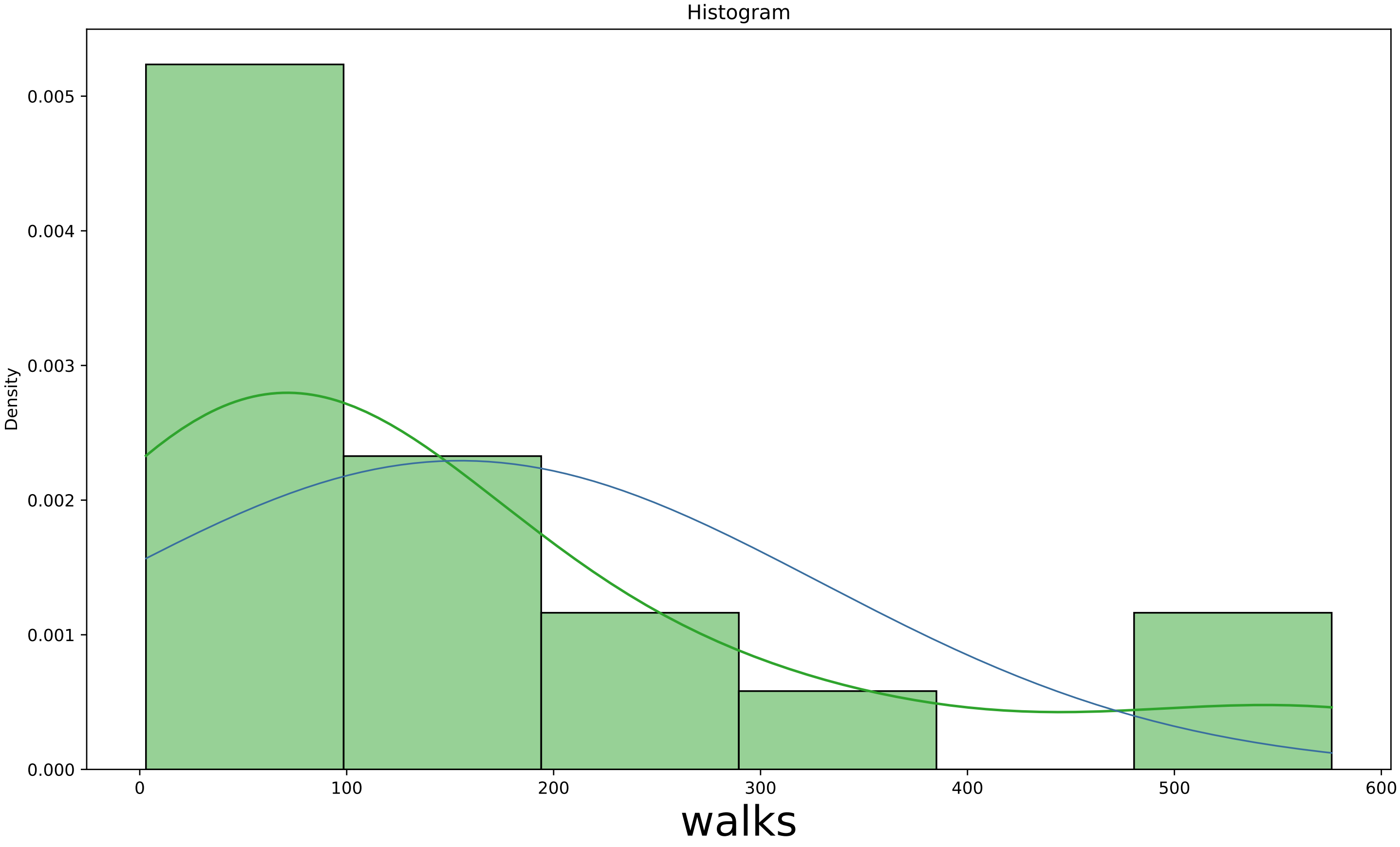

Histogram
Density
0.000
0.001
0.002
0.003
0.004
0
200
400
600
800
1000
rbi

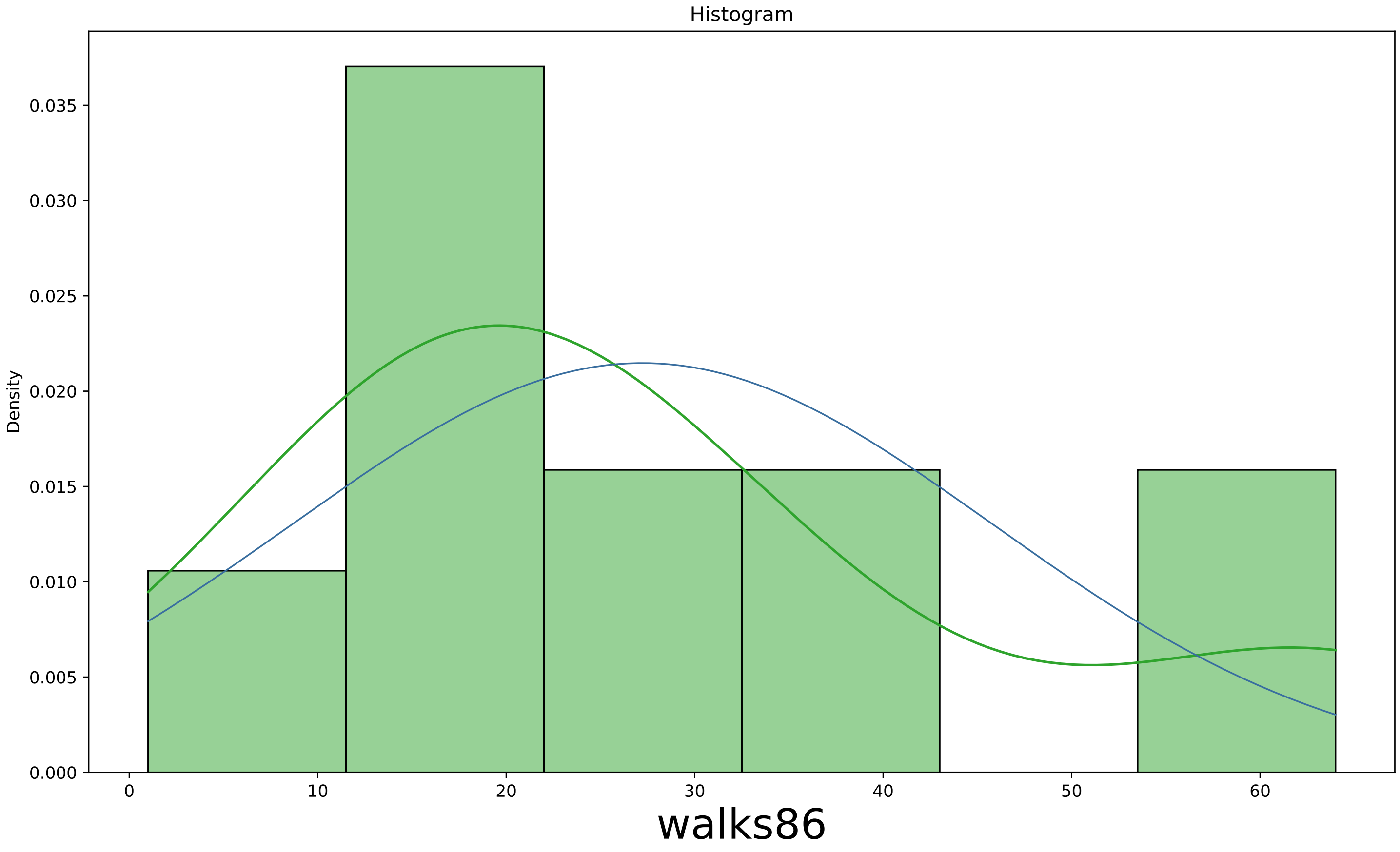

Histogram
Density
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0
20
40
60
80
rbi86


Histogram
Density
0.0000
0.0005
0.0010
0.0015
0.0020
0.0025
0.0030
0.0035
0.0040
0
200
400
600
800
runs


Histogram
Density
0.000
0.005
0.010
0.015
0.020
0
20
40
60
80
runs86

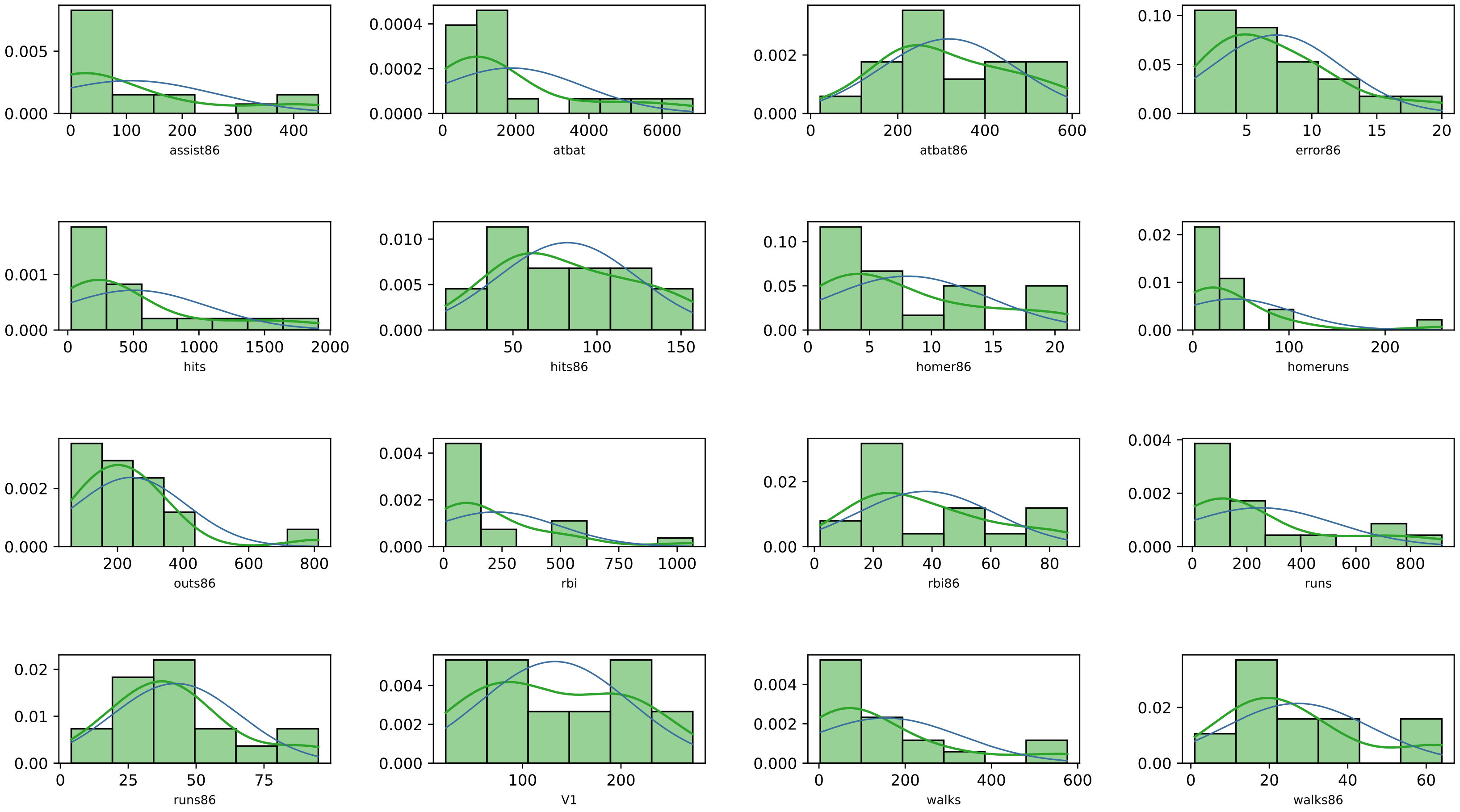

Histogram
Density
0.000
0.001
0.002
0.003
0.004
0.005
50
100
150
200
250
V1


Histogram
Density
0.000
0.001
0.002
0.003
0.004
0.005
0
100
200
300
400
500
600
walks


Histogram
Density
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
0
10
20
30
40
50
60
walks86

Histograms Summary

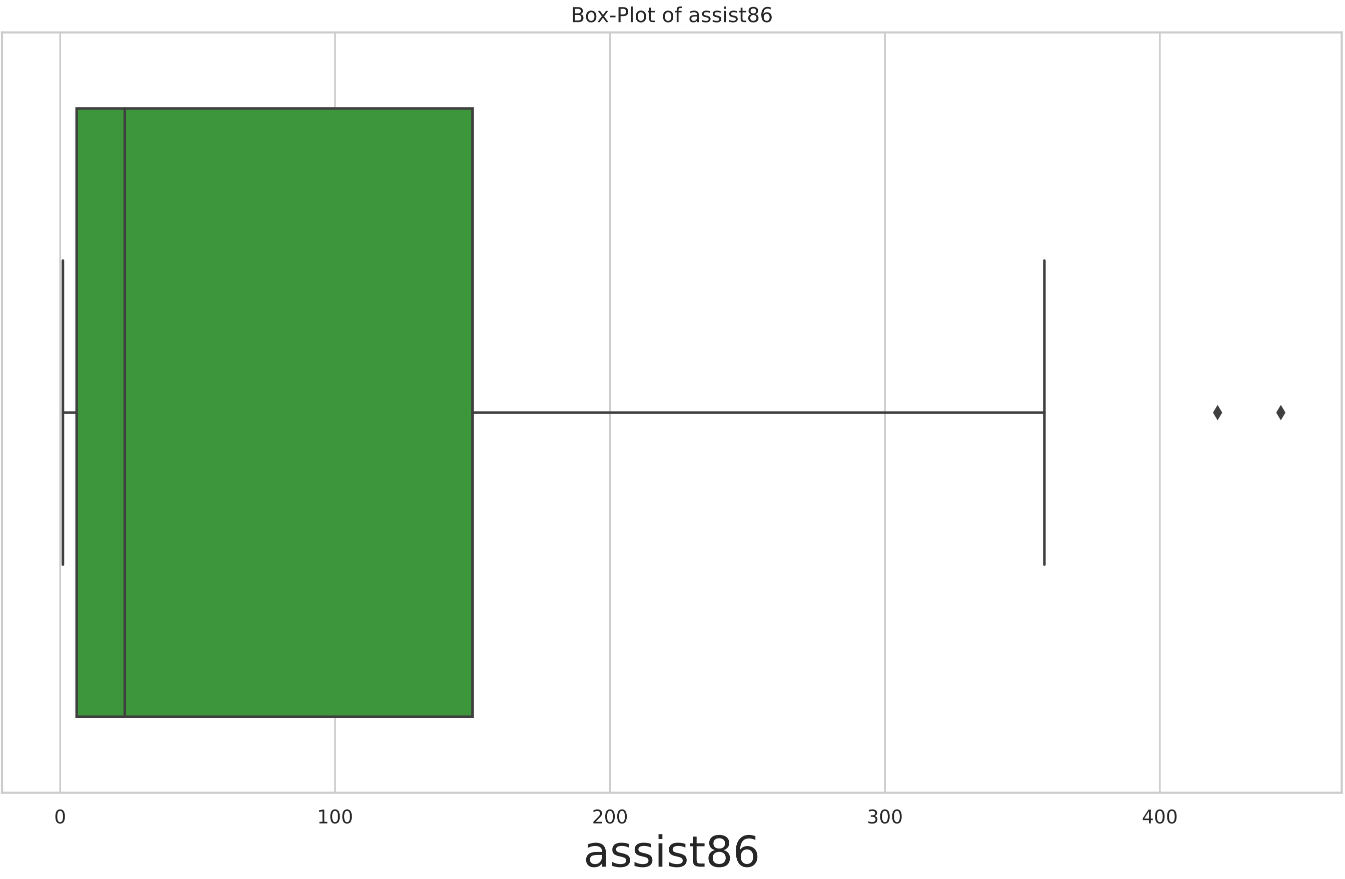
Multiple Relative Frequency Histogram in one figure. Variables are sorted alphabetically. The blue line represents the normal density approximation. The green line represents a special kernel density approximation.

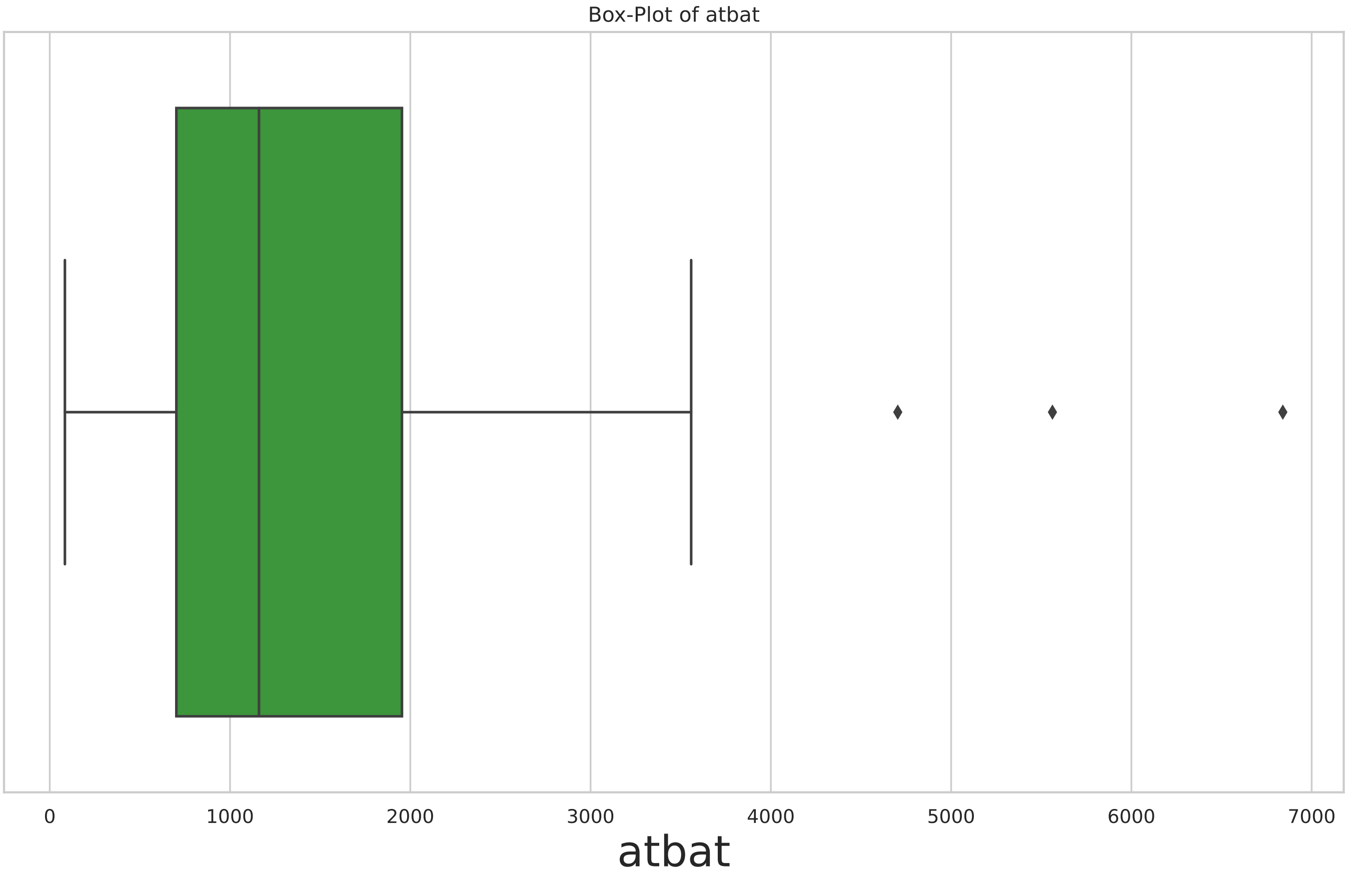


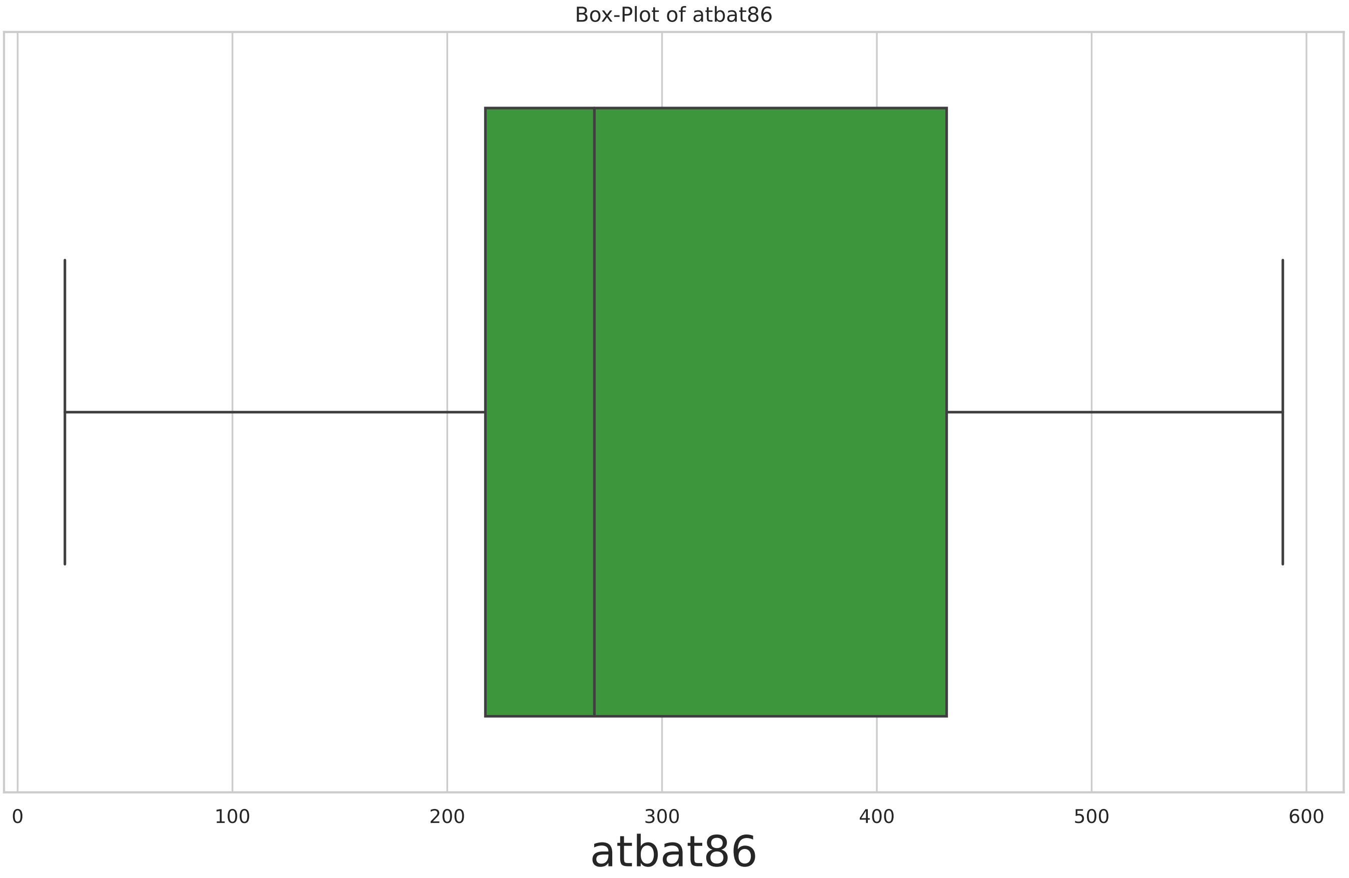
Box-Plots

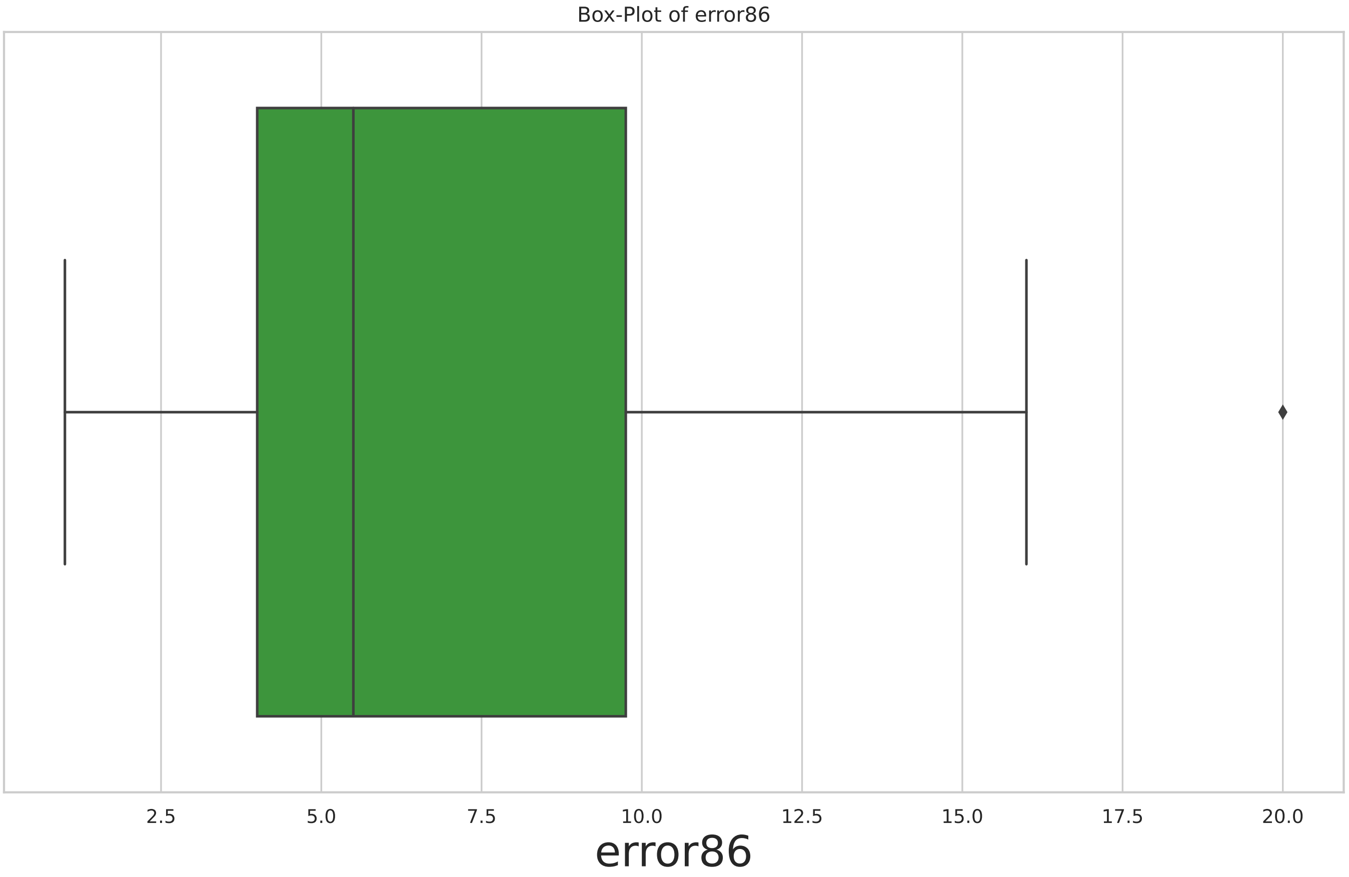
One Box-Plot per page for each variable. Variables are sorted alphabetically.

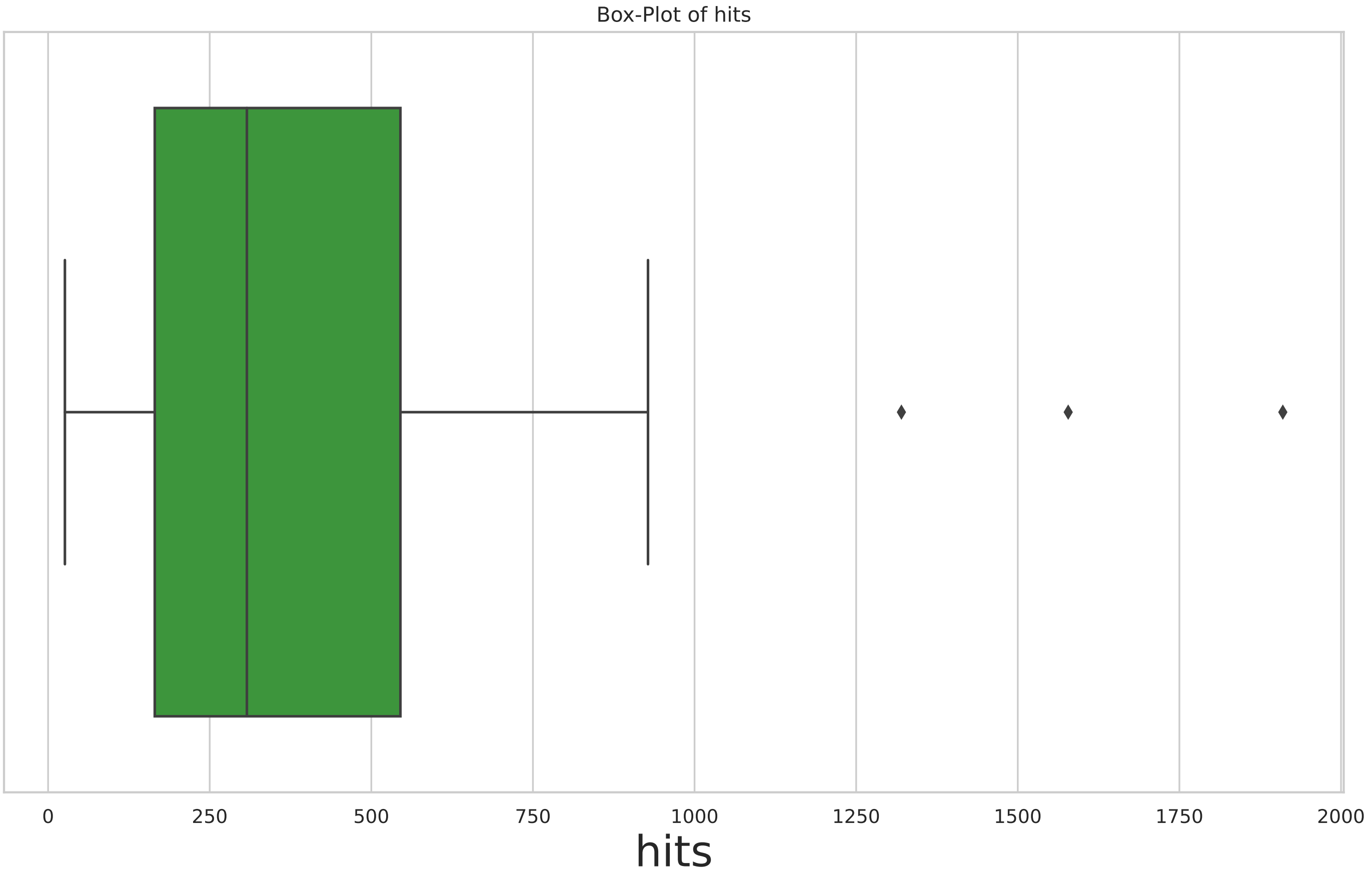
[]

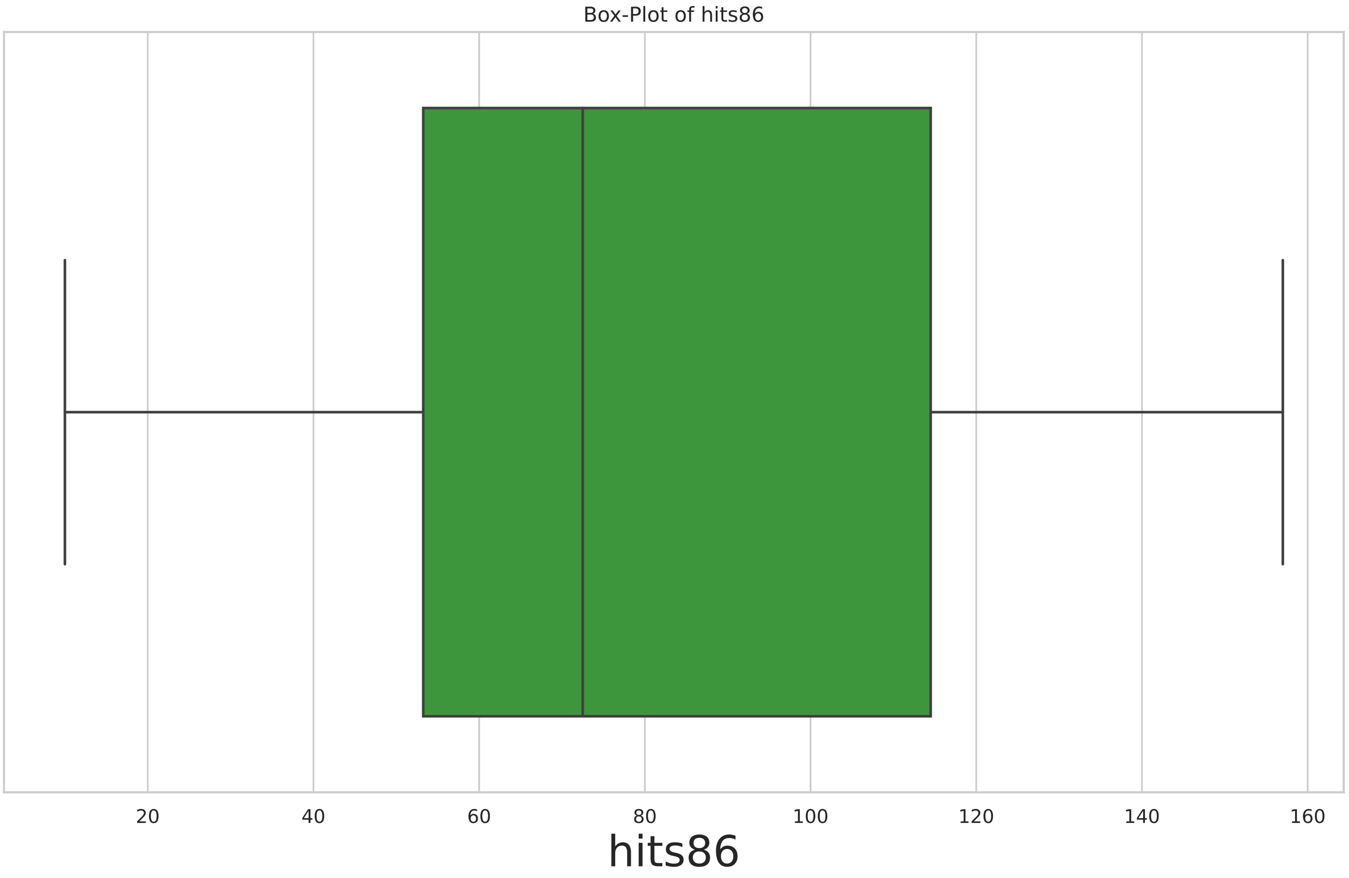


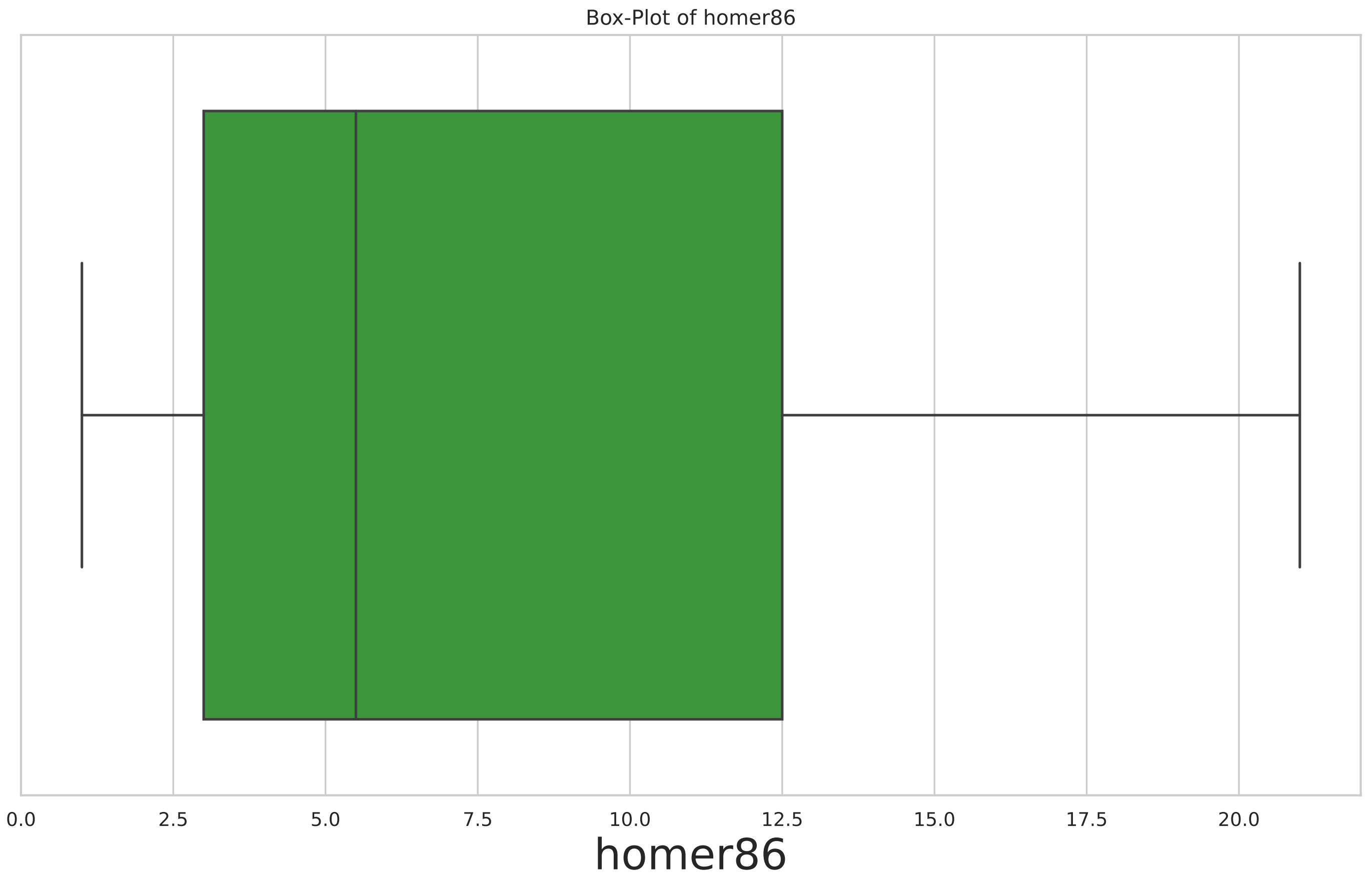

Box-Plot of atbat
0
1000
2000
3000
4000
5000
6000
7000
atbat

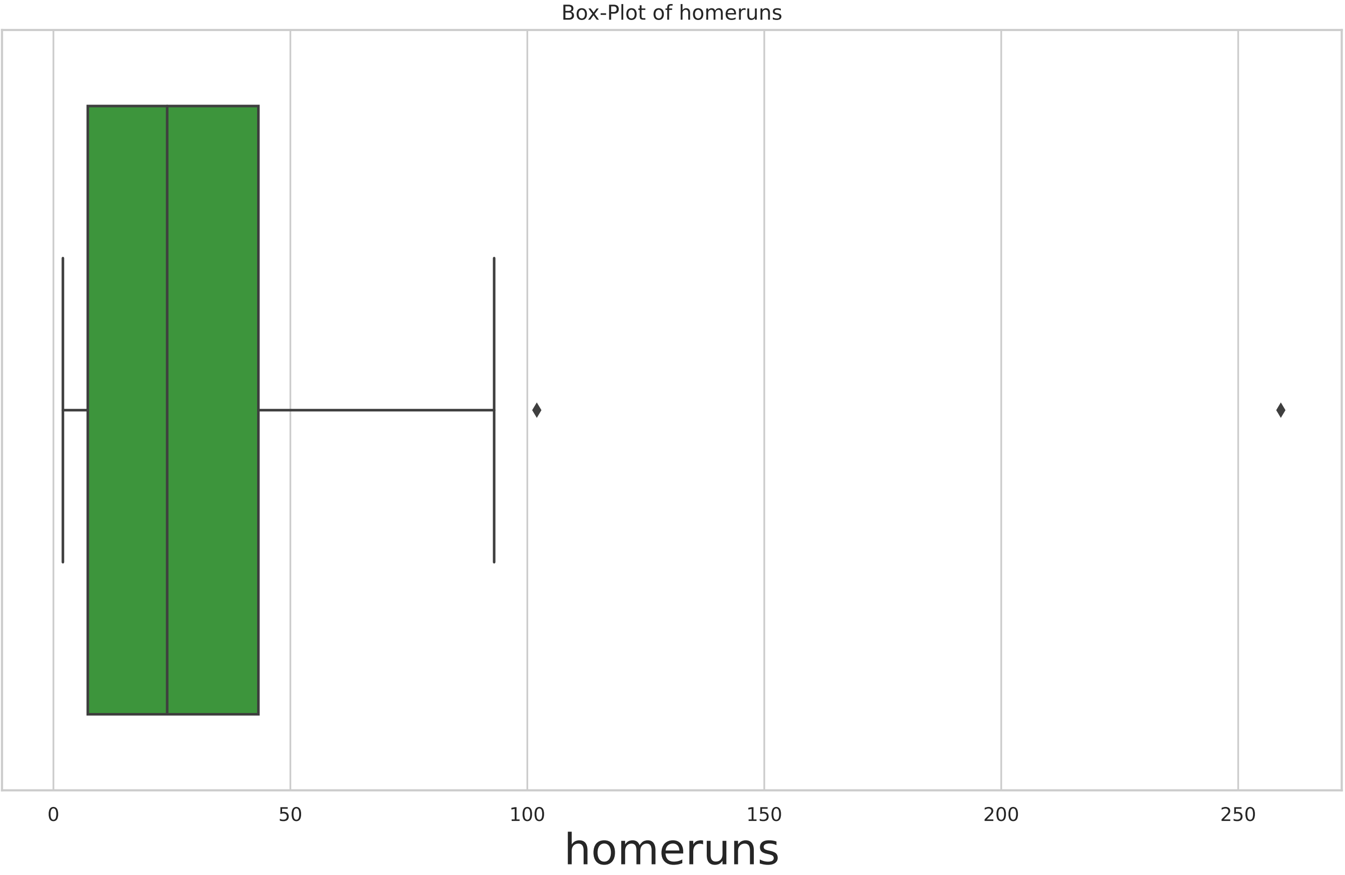

Box-Plot of atbat86
0
100
200
300
400
500
600
atbat86

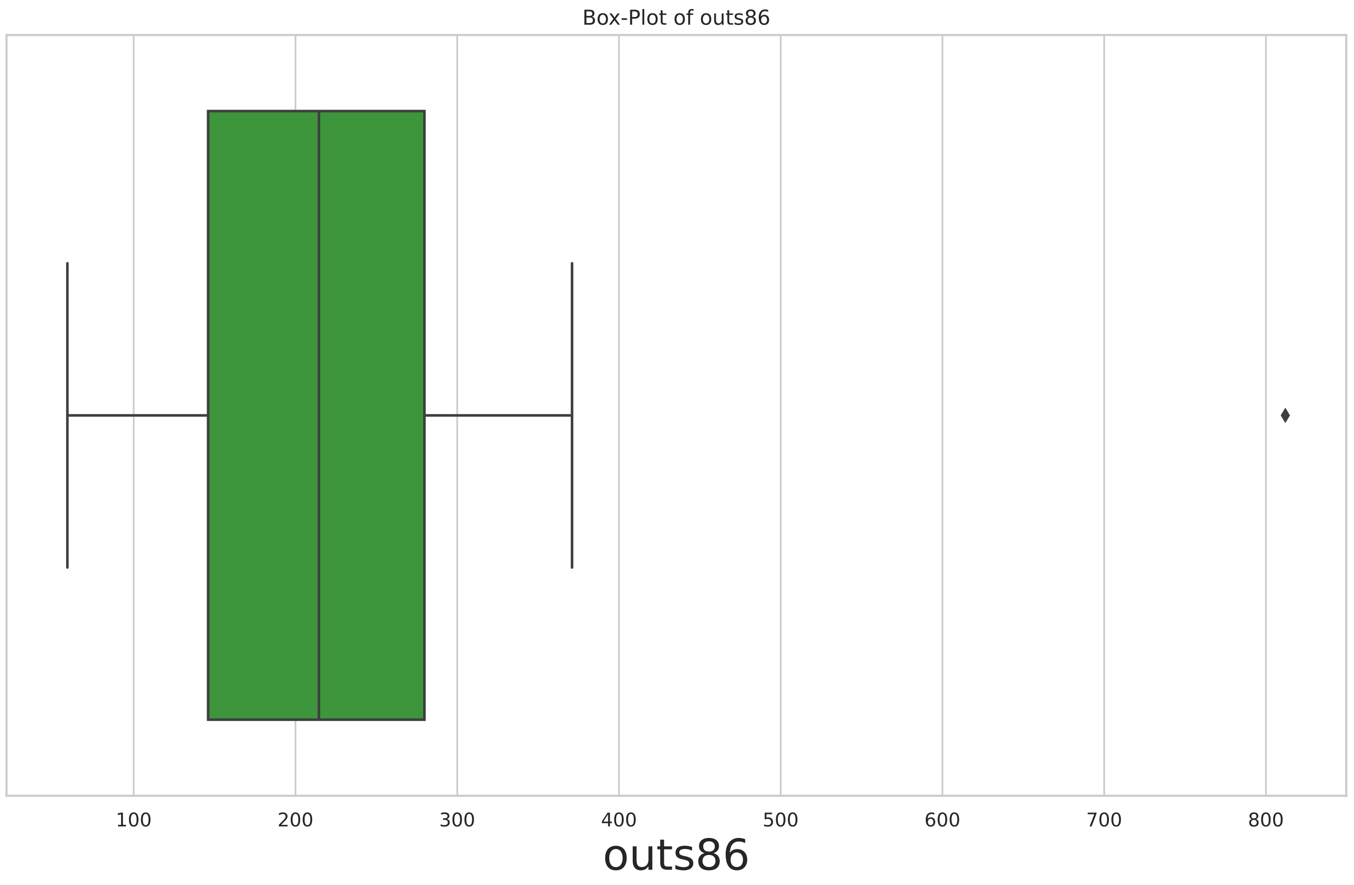

Box-Plot of error86
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
error86

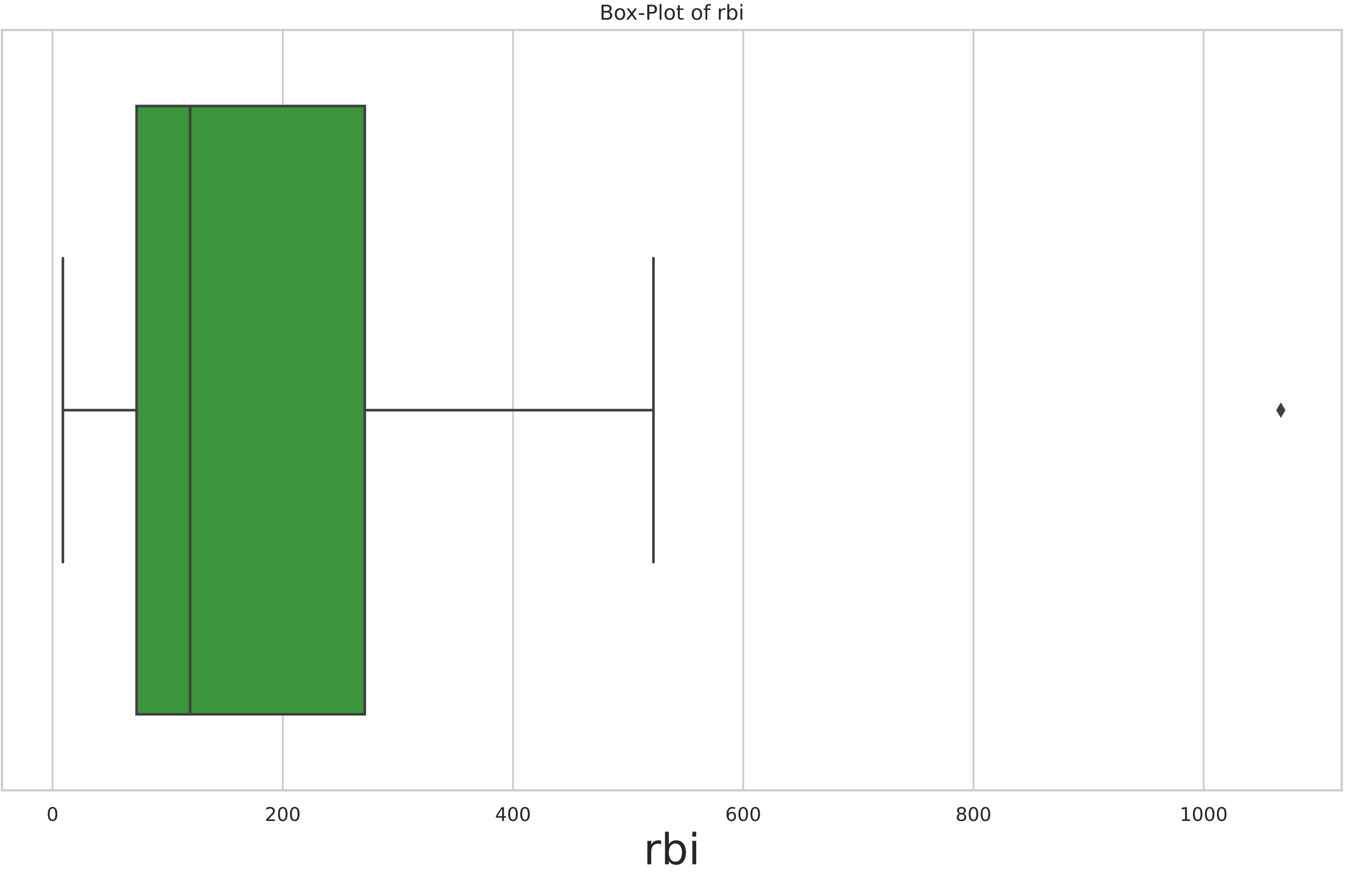

Box-Plot of hits
0
250
500
750
1000
1250
1500
1750
2000
hits

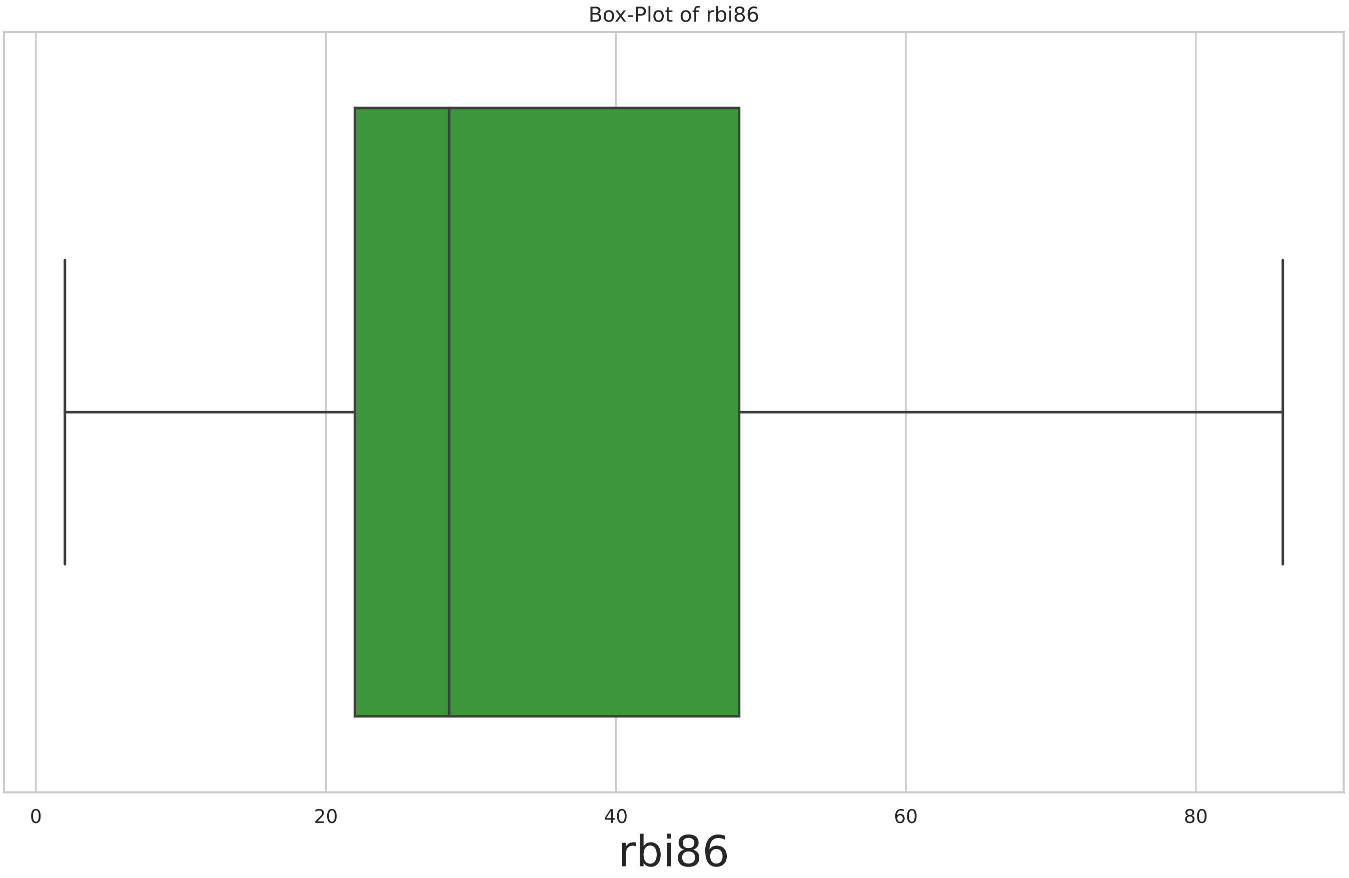

Box-Plot of hits86
20
40
60
80
100
120
140
160
hits86

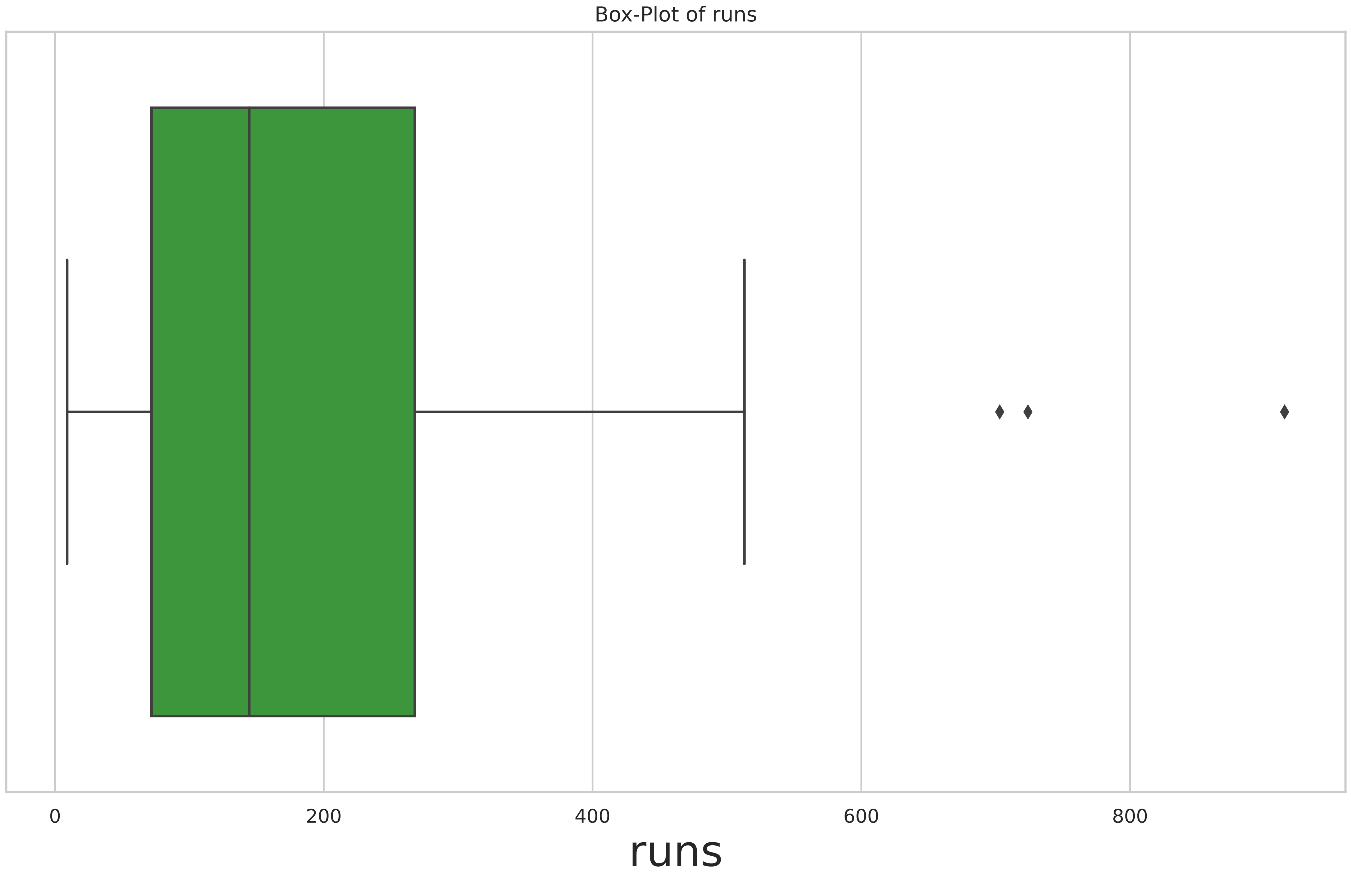

Box-Plot of homer86
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
homer86


Box-Plot of homeruns
0
50
100
150
200
250
homeruns

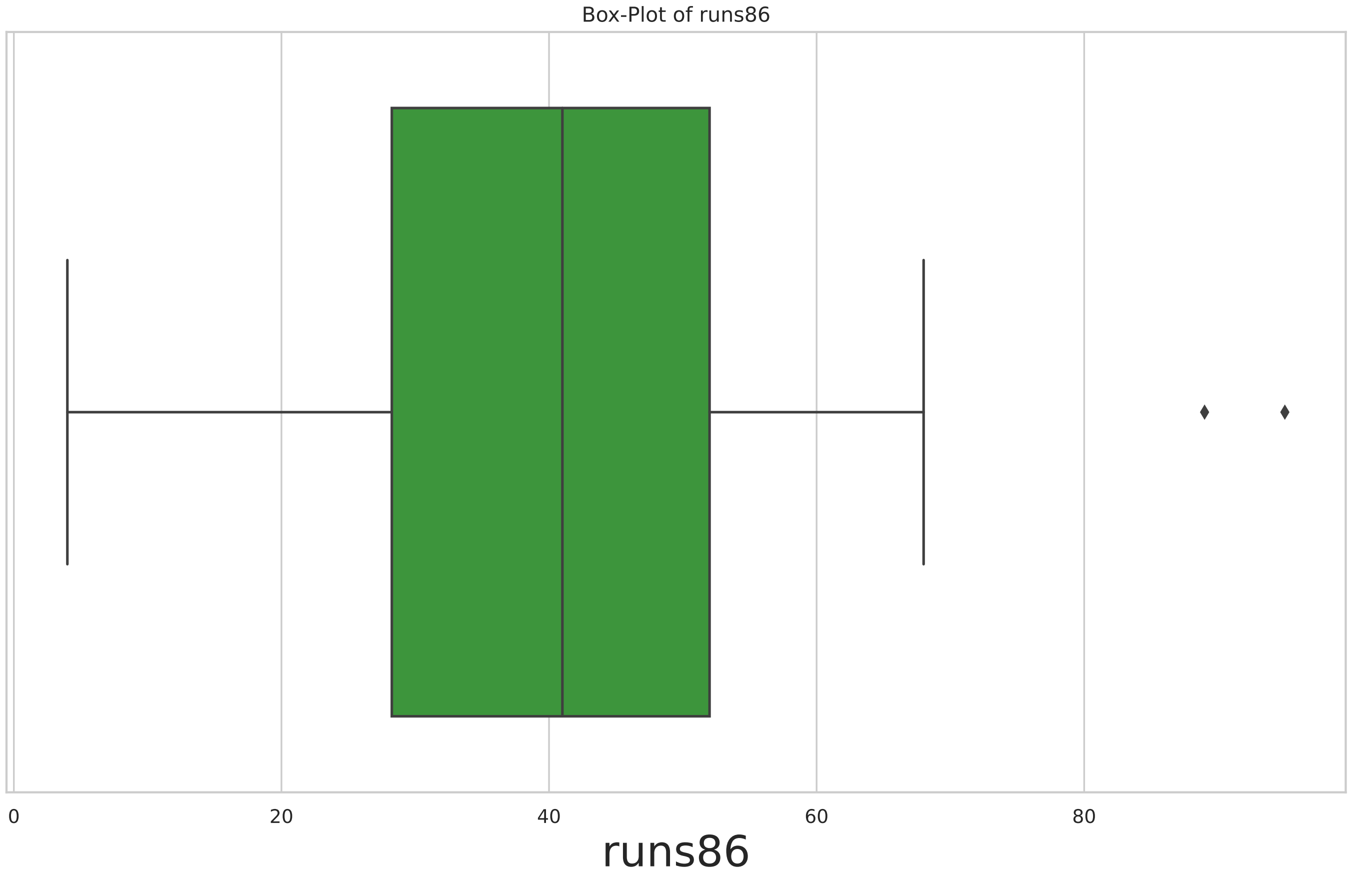

Box-Plot of outs86
100
200
300
400
500
600
700
800
outs86

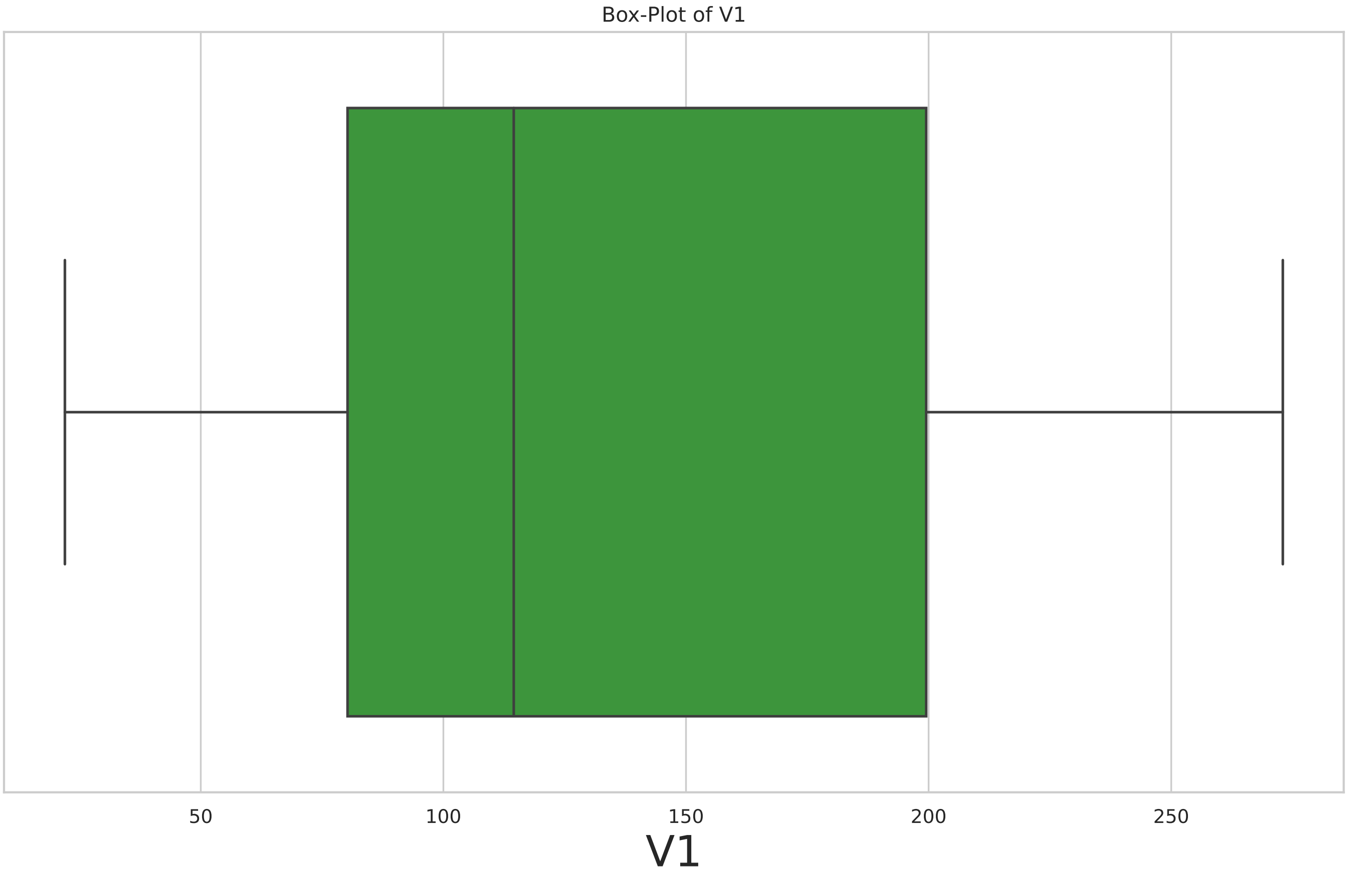

Box-Plot of rbi
0
200
400
600
800
1000
rbi


Box-Plot of rbi86
0
20
40
60
80
rbi86

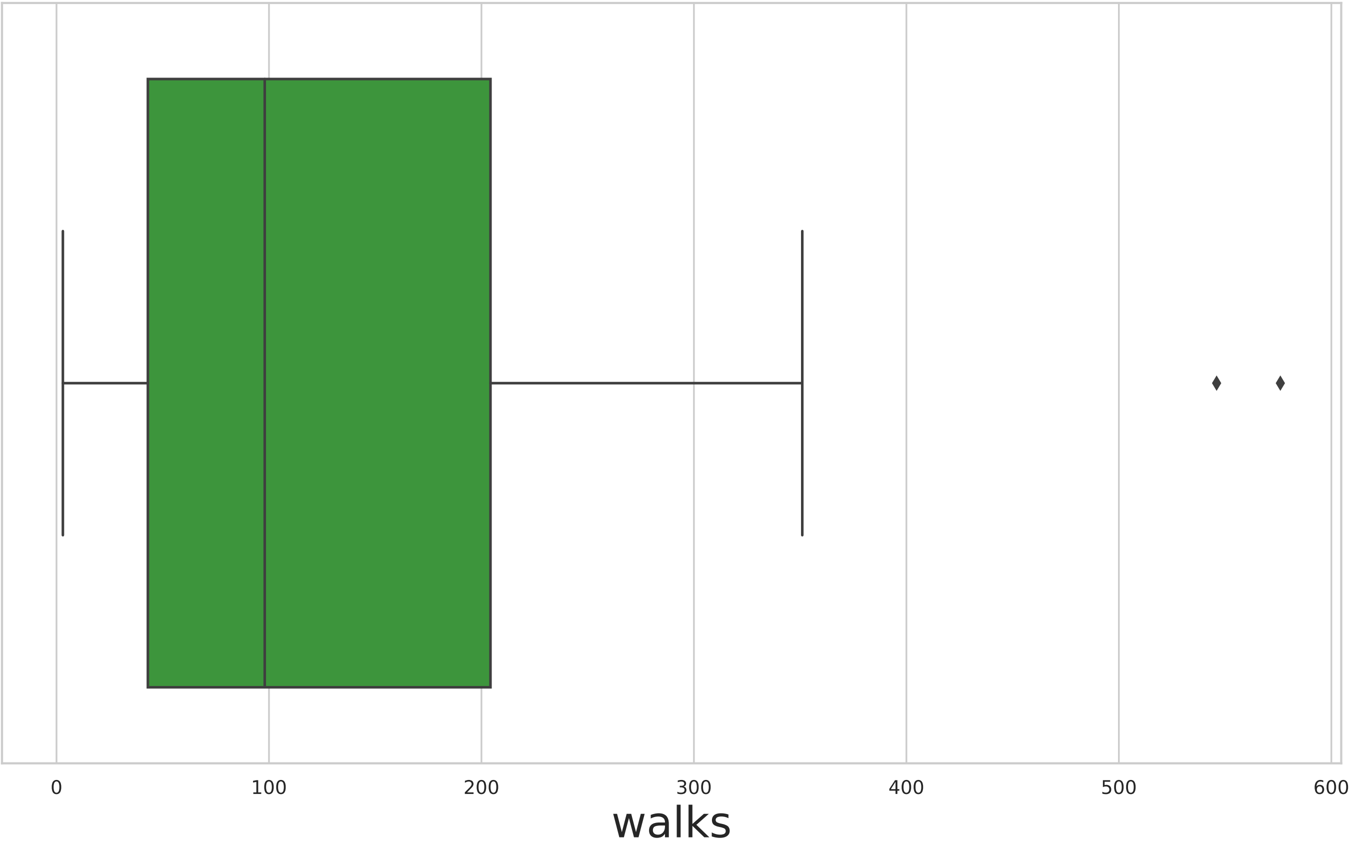


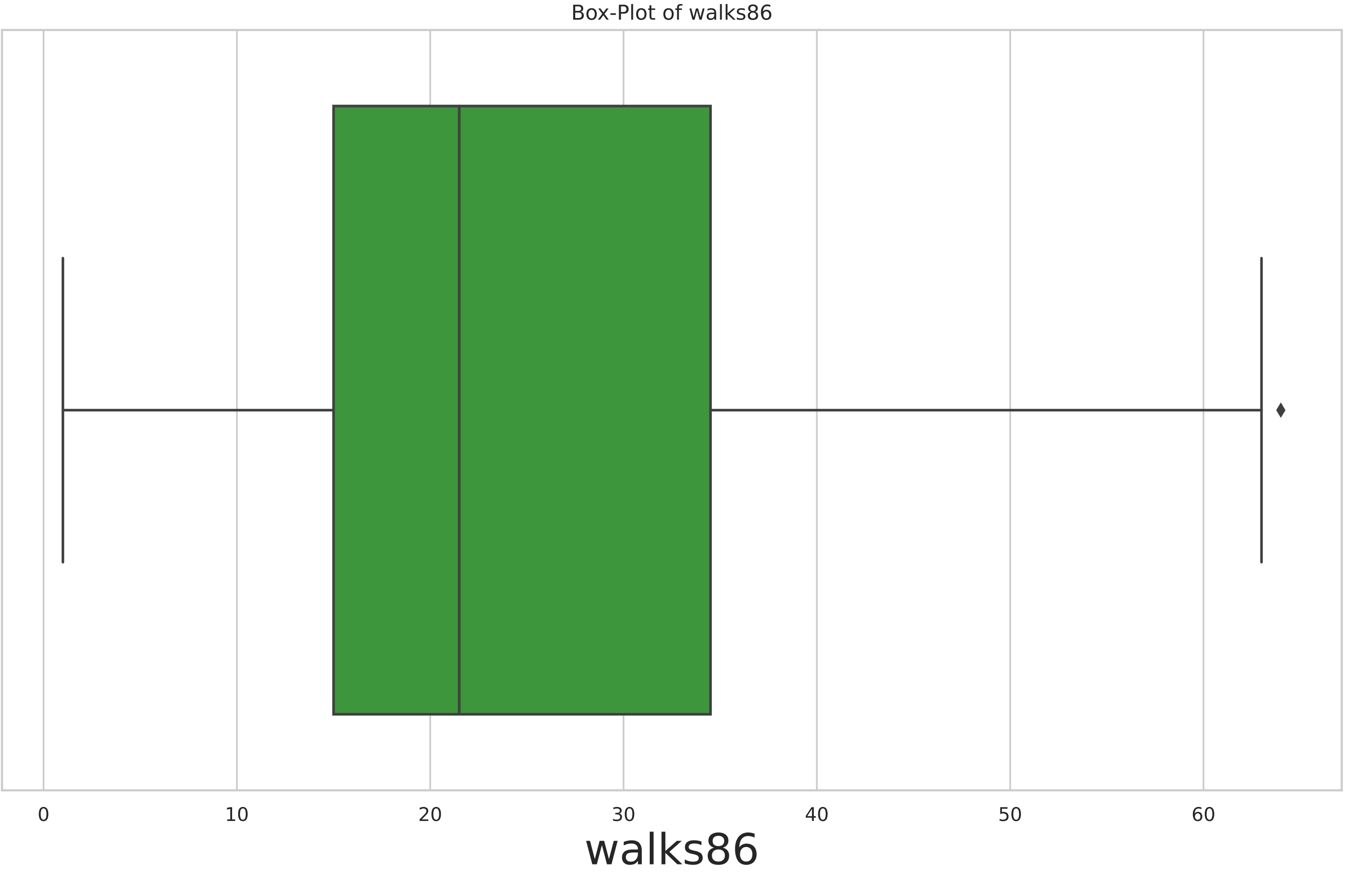
Box-Plot of runs


Box-Plot of runs86
0
20
40
60
80
runs86


Box-Plot of V1
50
100
150
200
250
V1

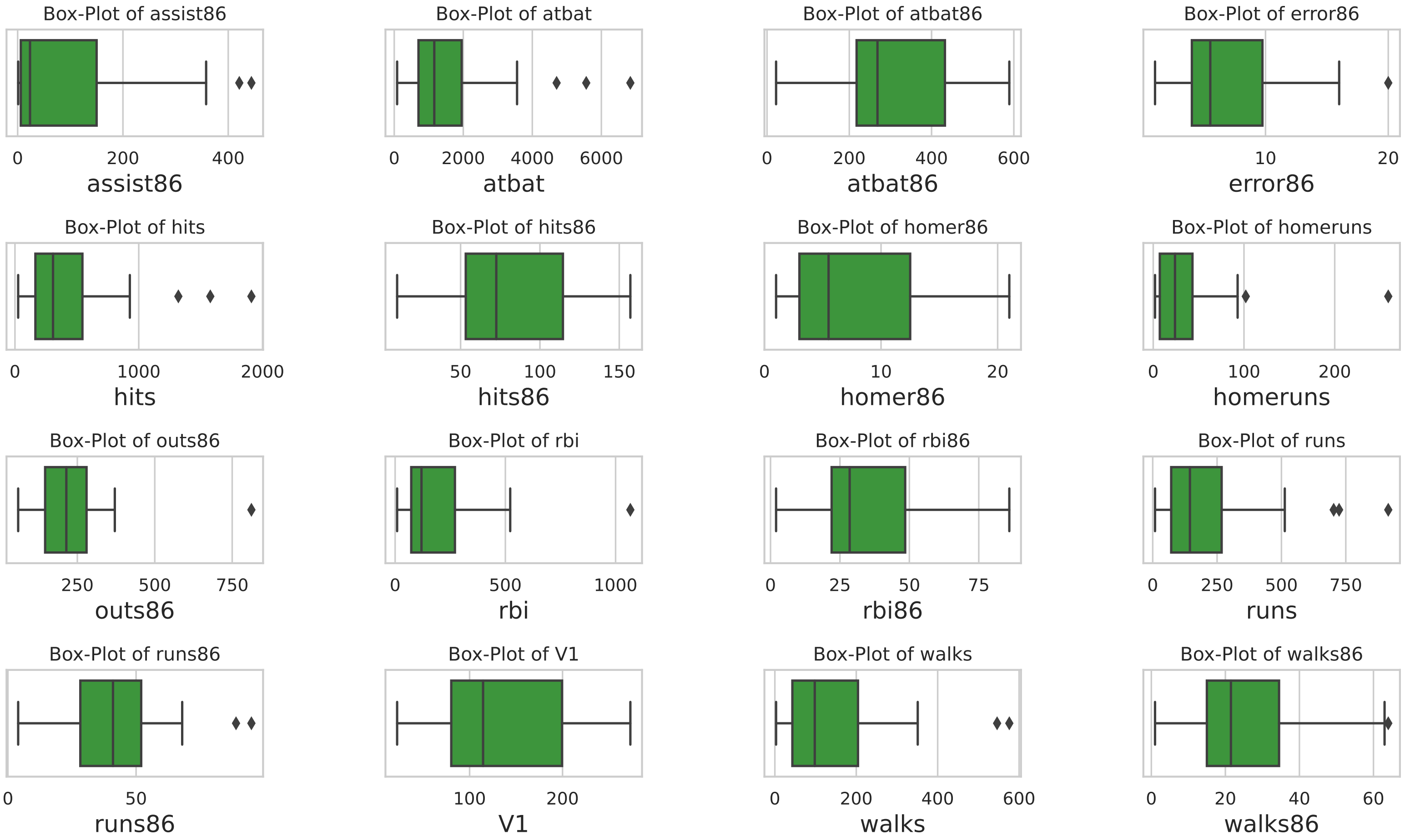
Box-Plot of walks




Box-Plot of walks86
0
10
20
30
40
50
60
walks86

Box-Plots Summary

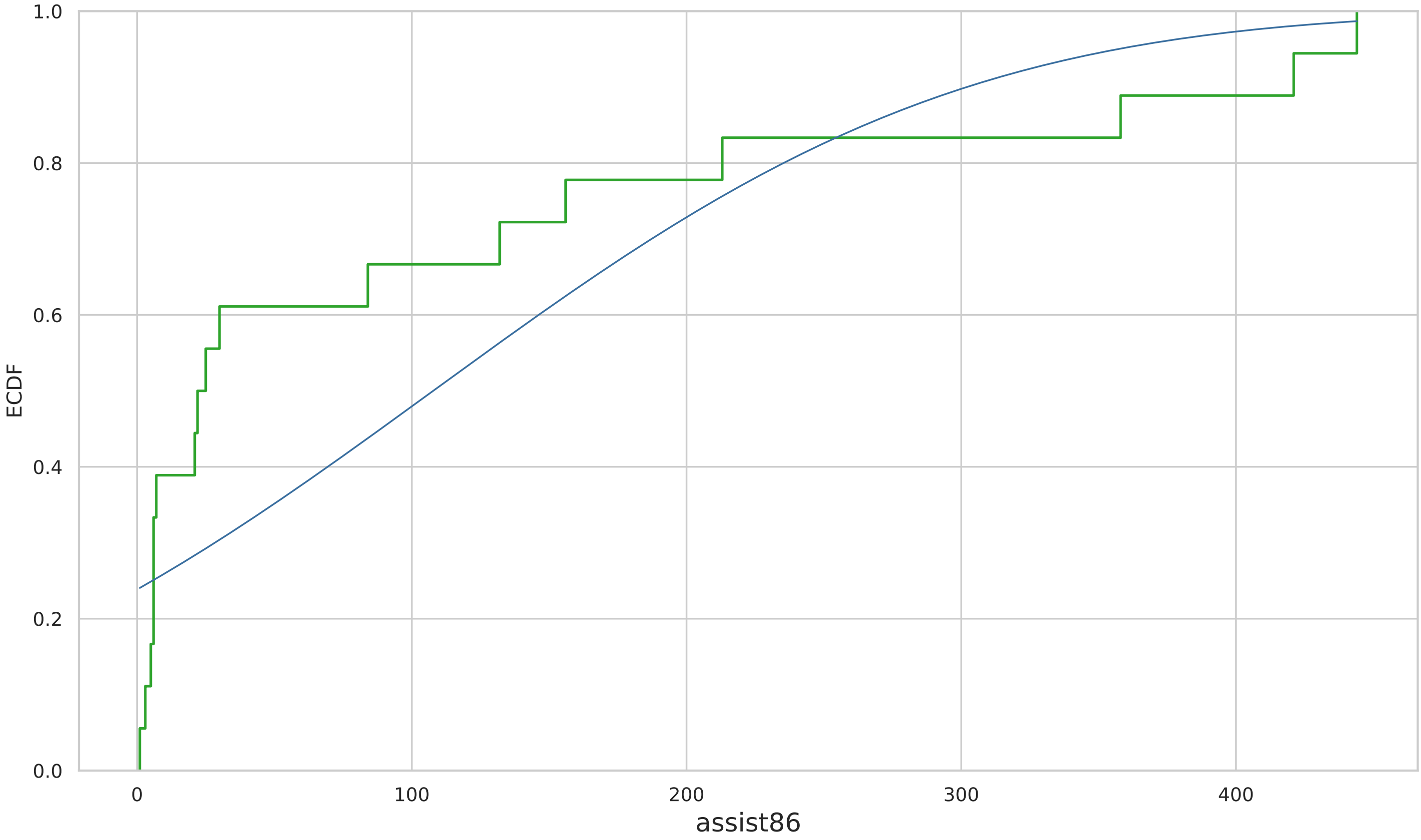
Multiple Box-Plots of variables in one figure. Variables are sorted alphabetically.

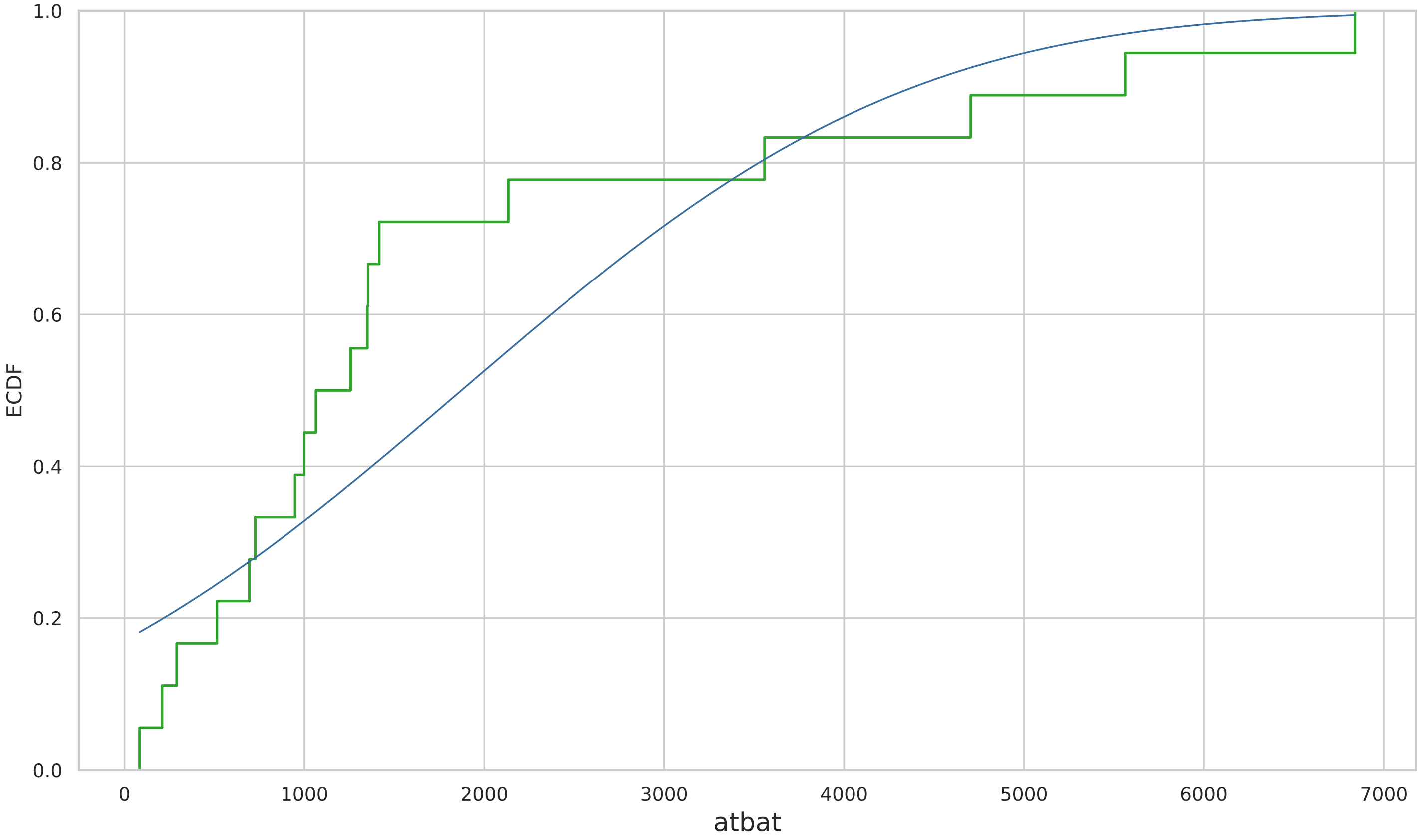


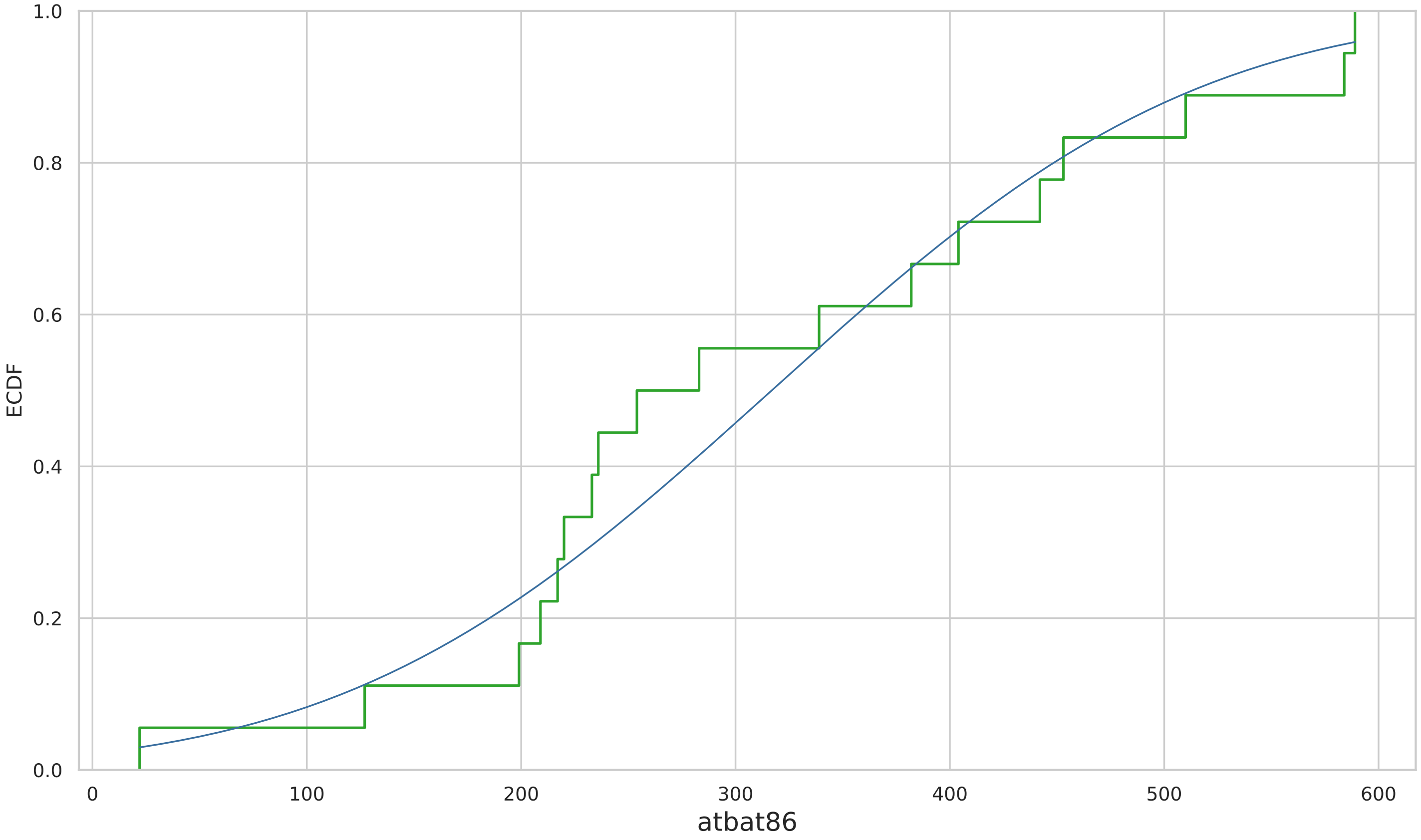
ECDF Plots

One ECDF (Empirical Cumulative Distribution Function) Plot per page for each variable. Variables are sorted alphabetically. The blue line represents the CDF of a normal distribution. If the variable is normally distributed, the blue line approximates well the ECDF.

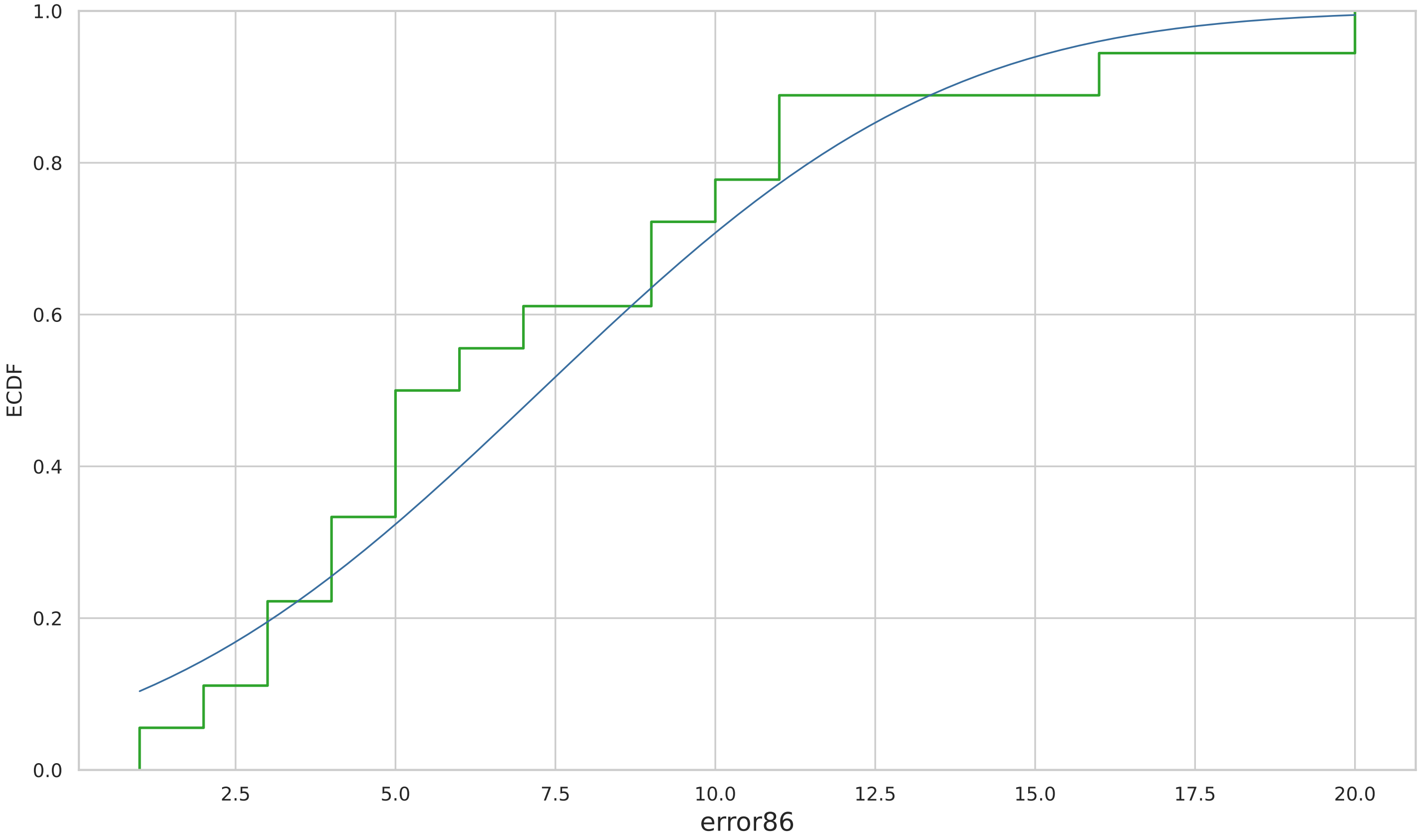
[]

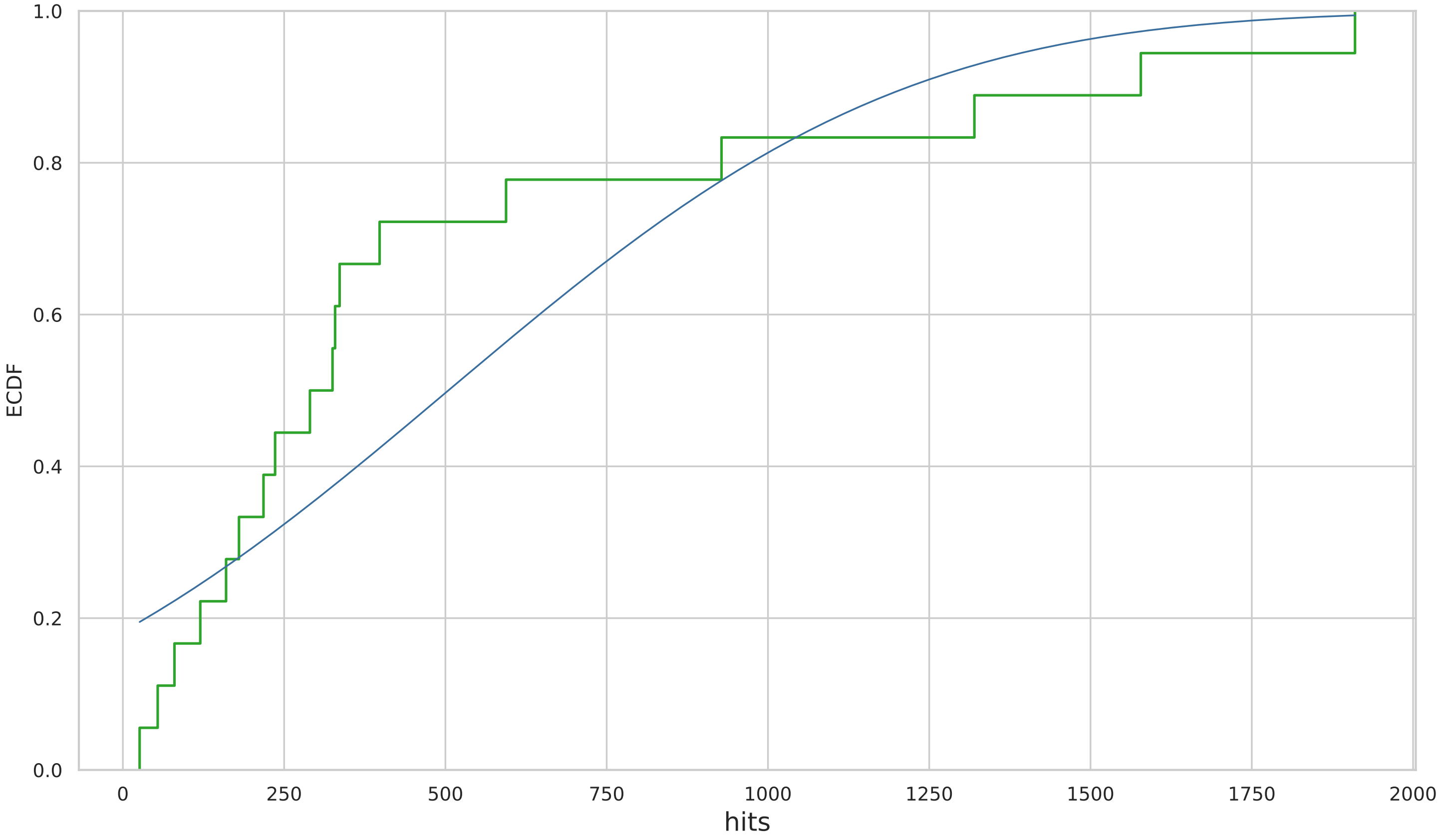


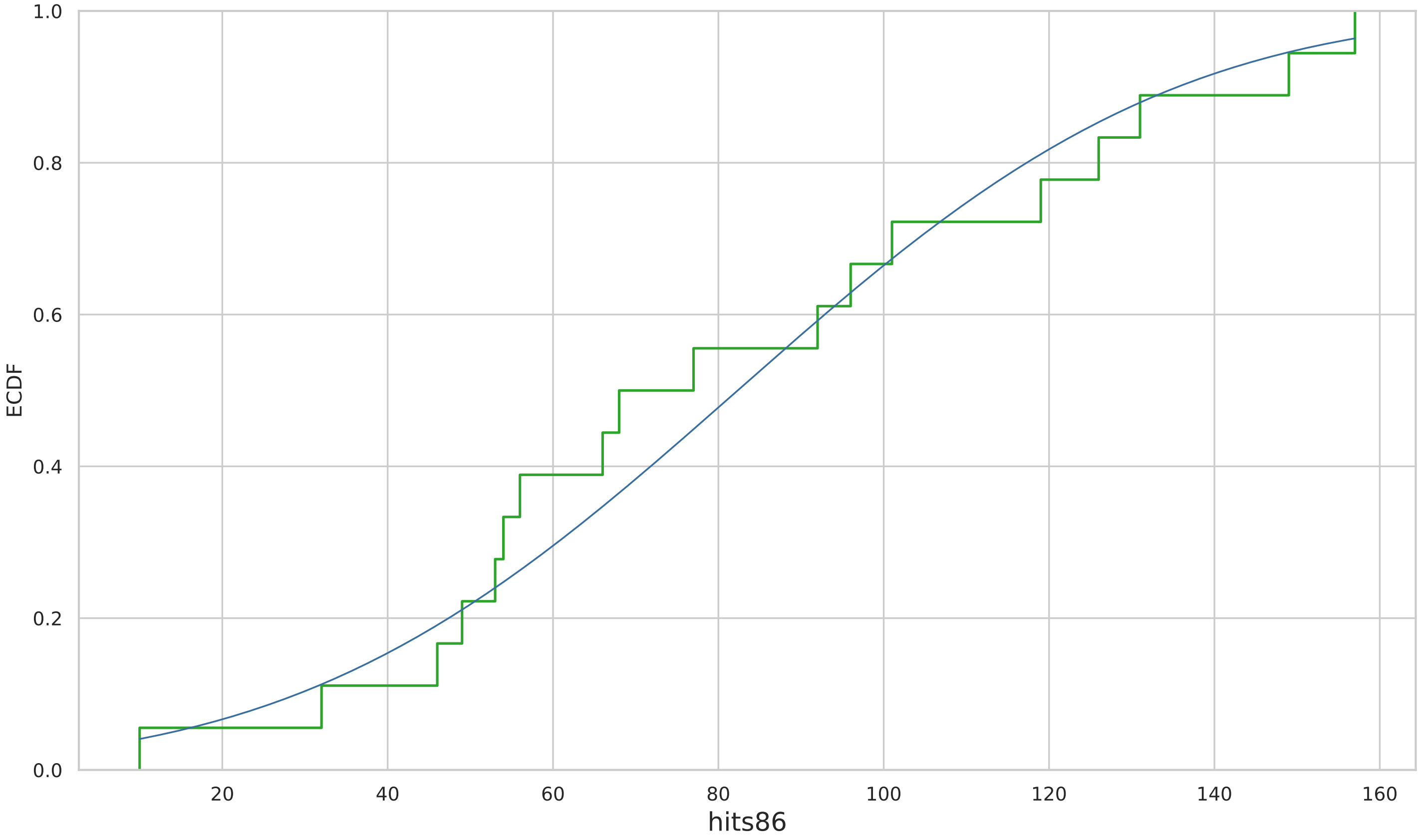

1.0
0.8
0.6
0.4
0.2
0.0
ECDF
0
1000
2000
3000
4000
5000
6000
7000
atbat

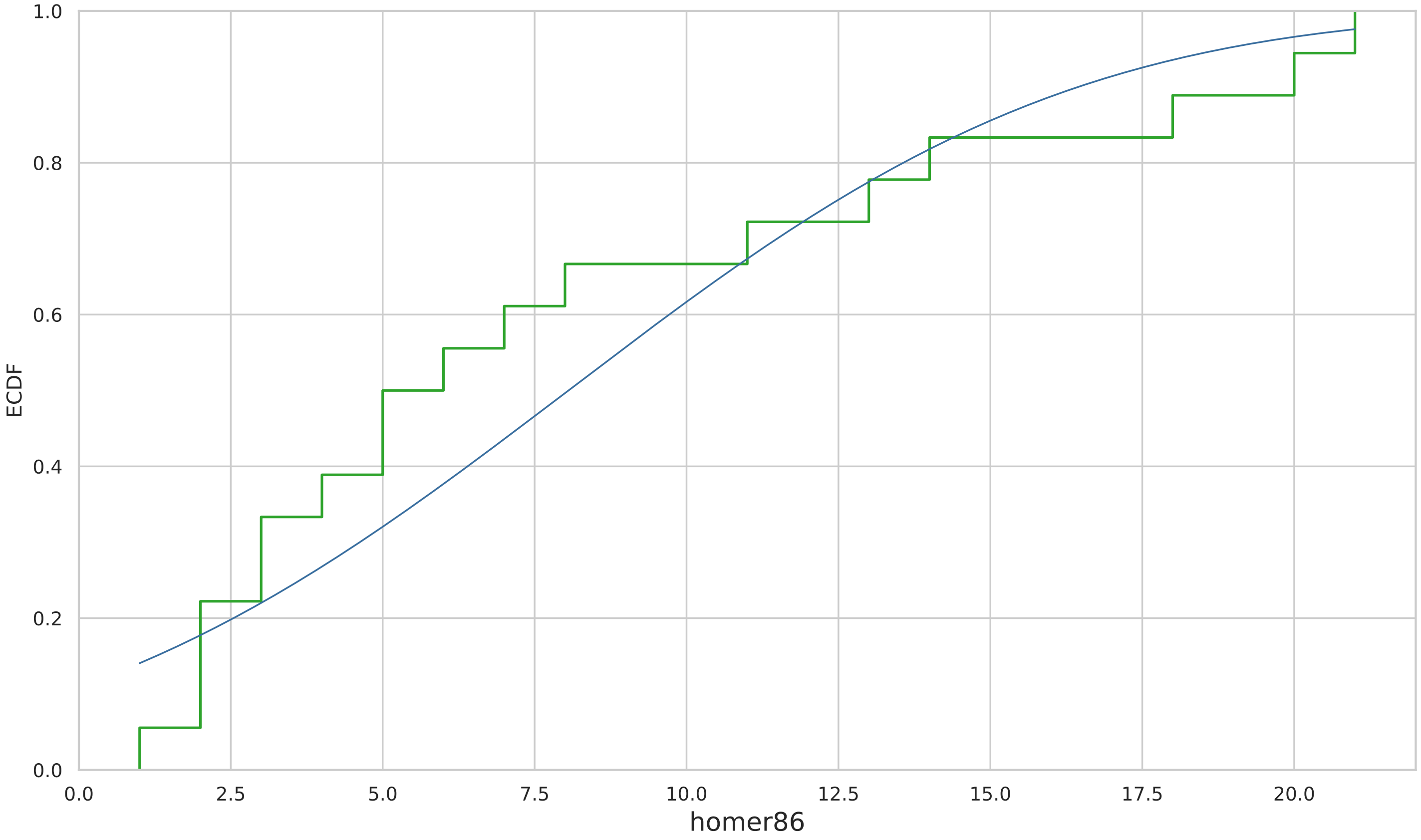


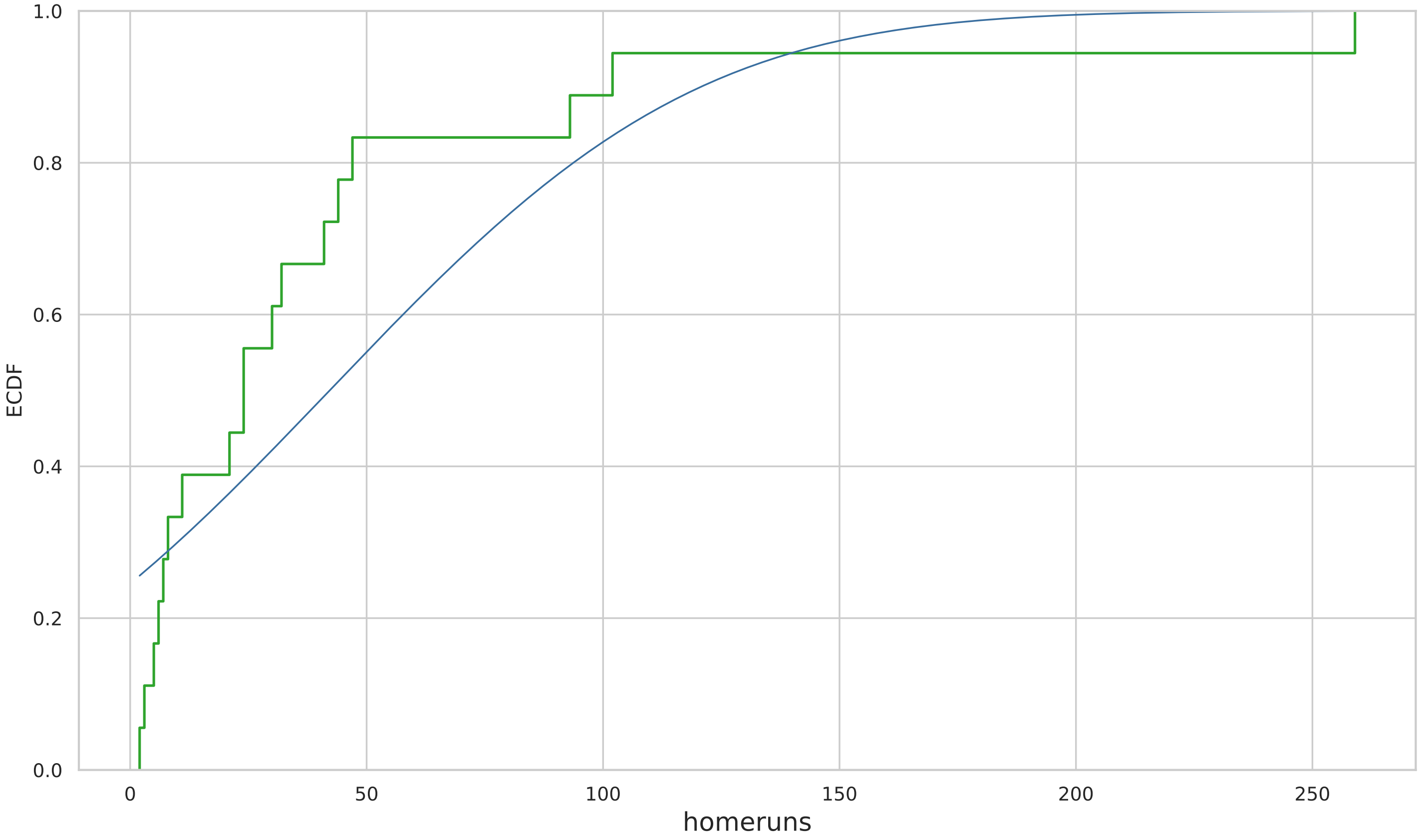
1.0
0.8
0.6
0.4
0.2
0.0
ECDF
0
100
200
300
400
500
600
atbat86

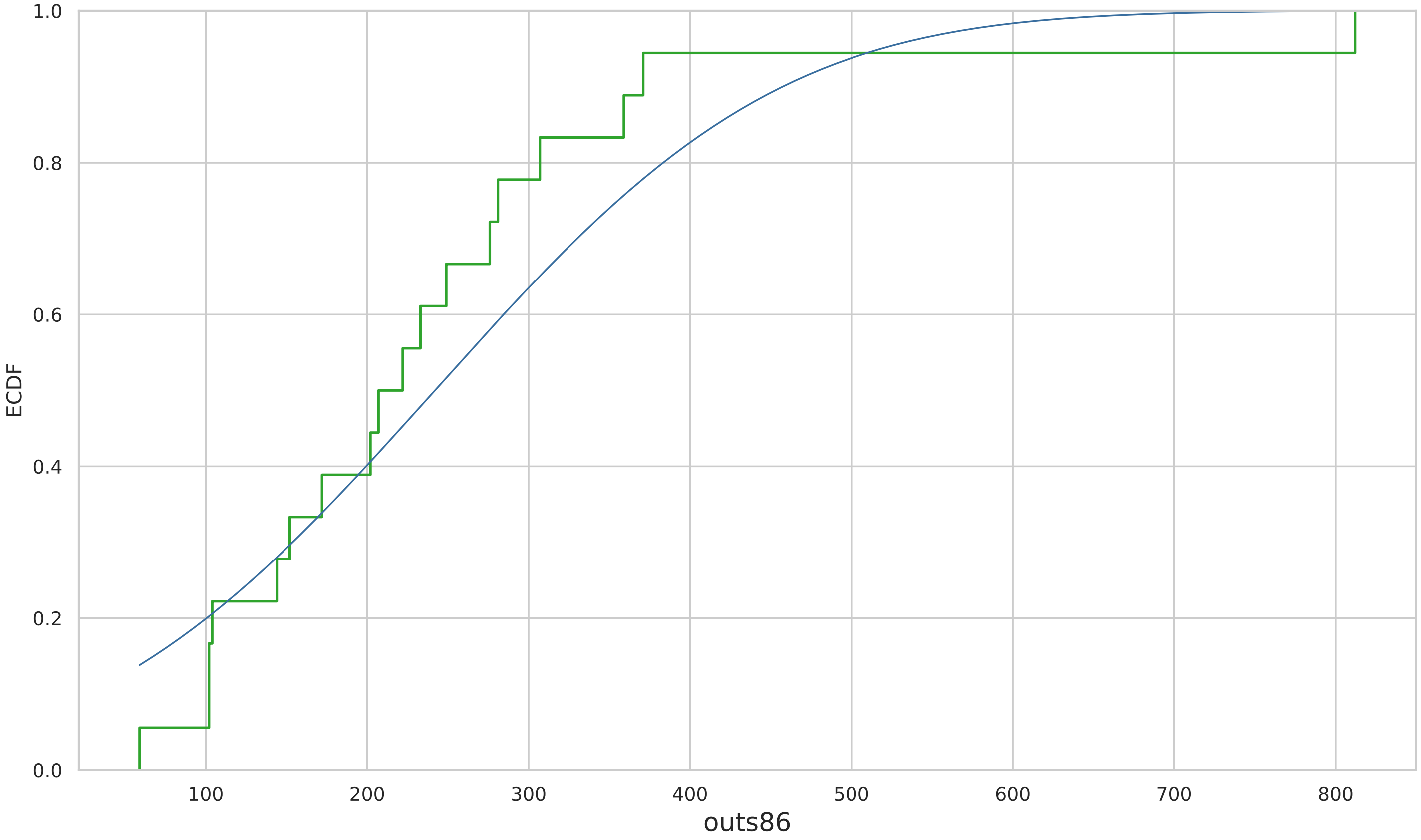

1.0
0.8
0.6
0.4
0.2
0.0
ECDF
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
error86

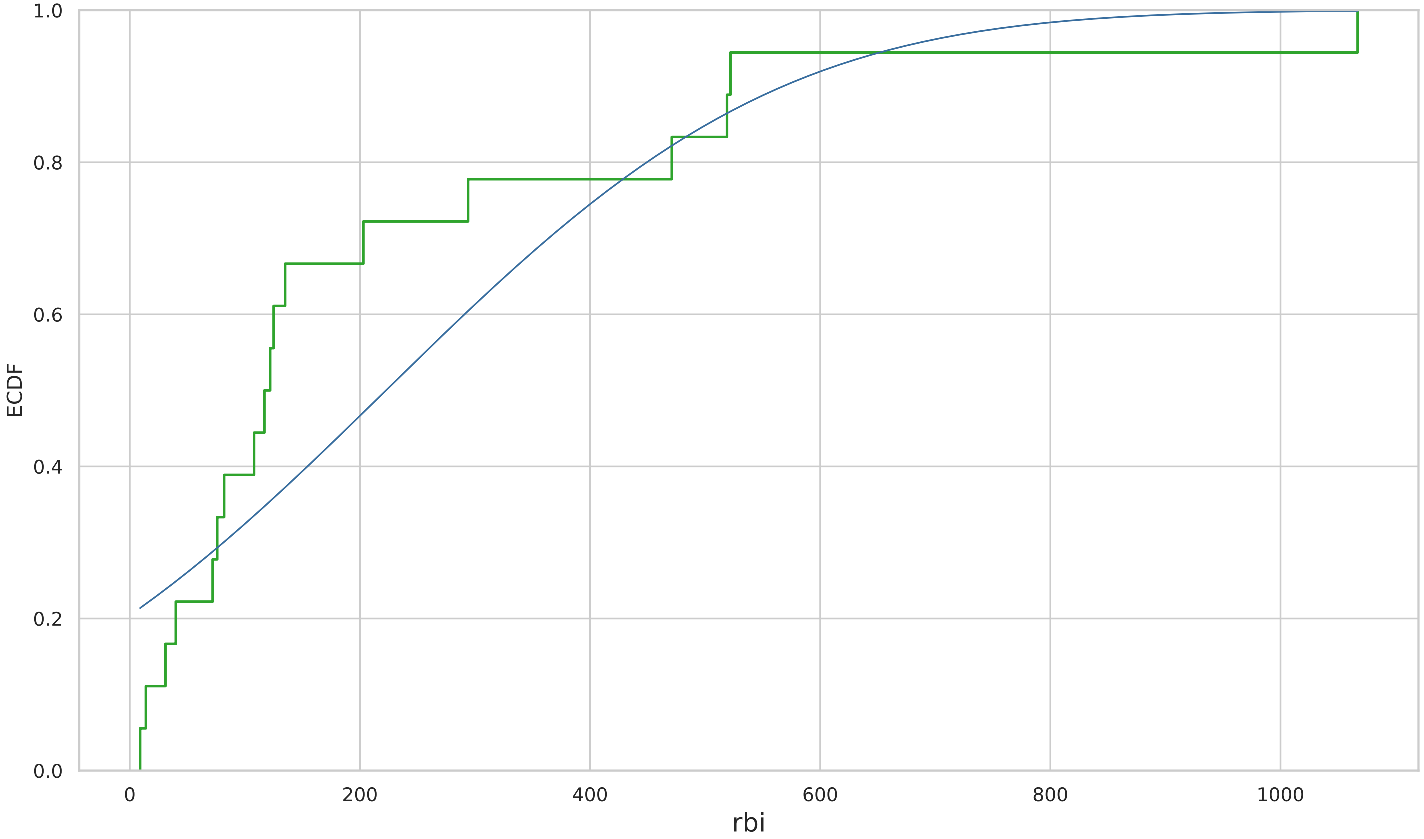

1.0
0.8
0.6
0.4
0.2
0.0
ECDF
0
250
500
750
1000
1250
1500
1750
2000
hits

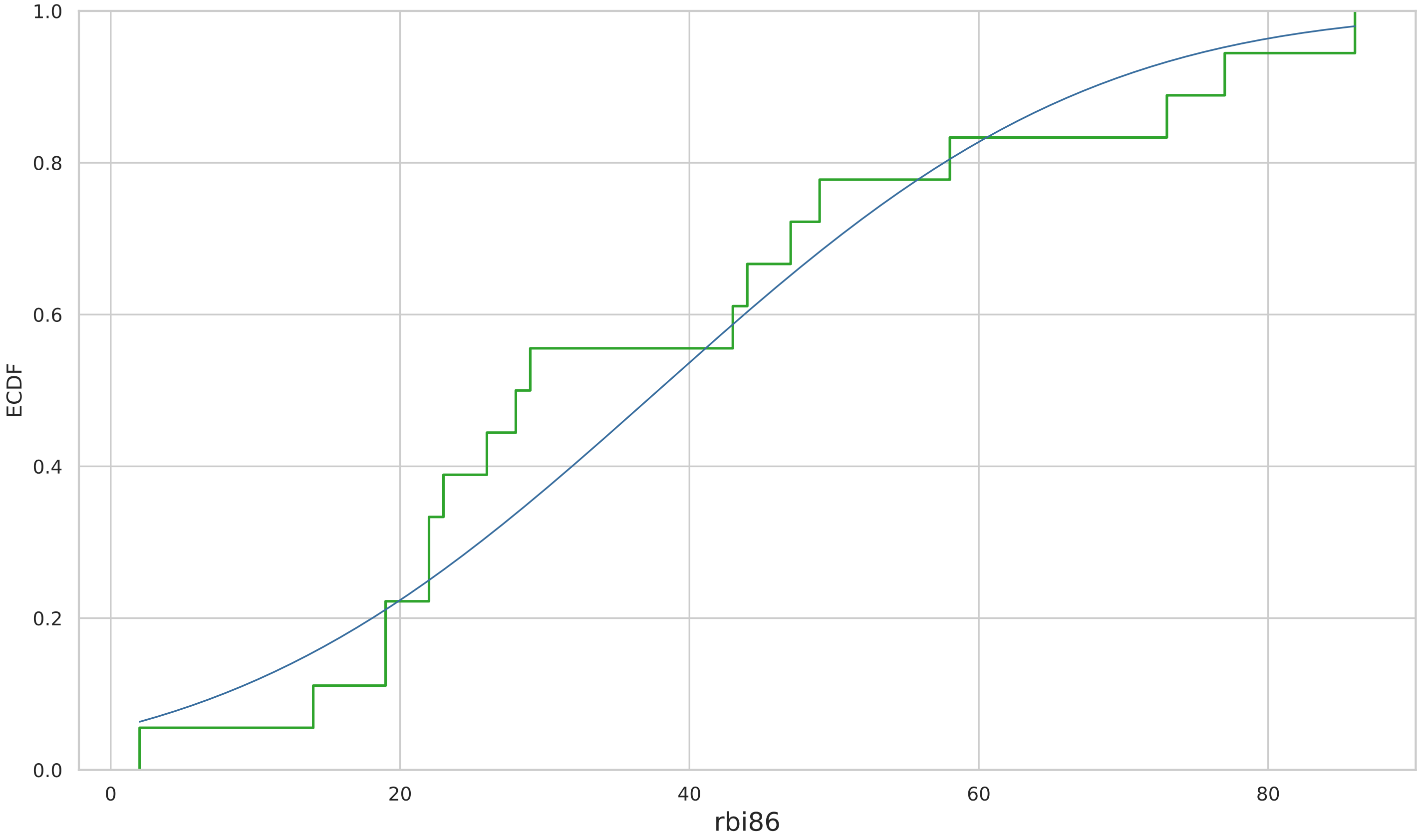

1.0
0.8
0.6
0.4
0.2
0.0
ECDF
20
40
60
80
100
120
140
160
hits86

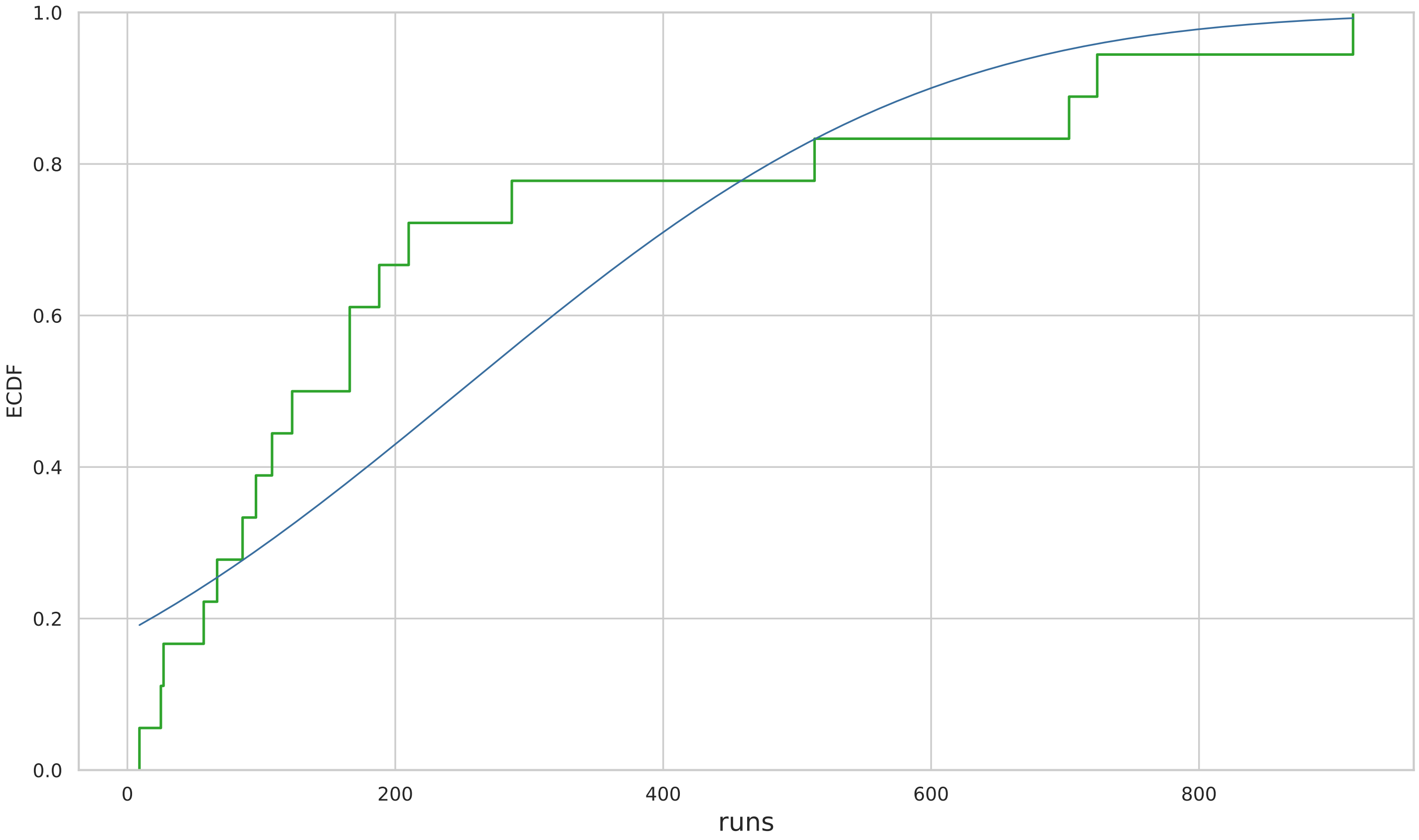

1.0
0.8
0.6
0.4
0.2
0.0
ECDF
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
homer86

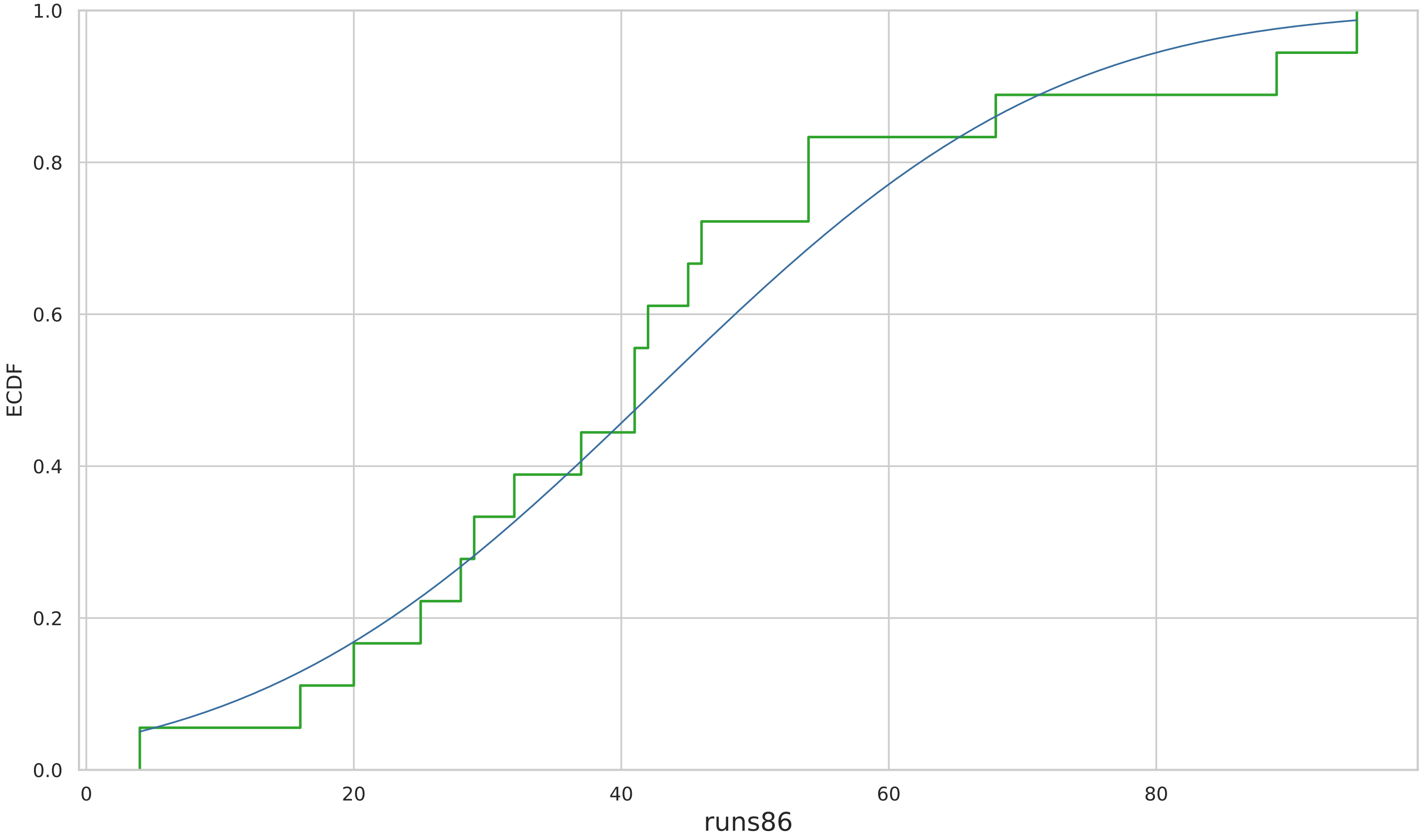

1.0
0.8
0.6
0.4
0.2
0.0
ECDF
0
50
100
150
200
250
homeruns

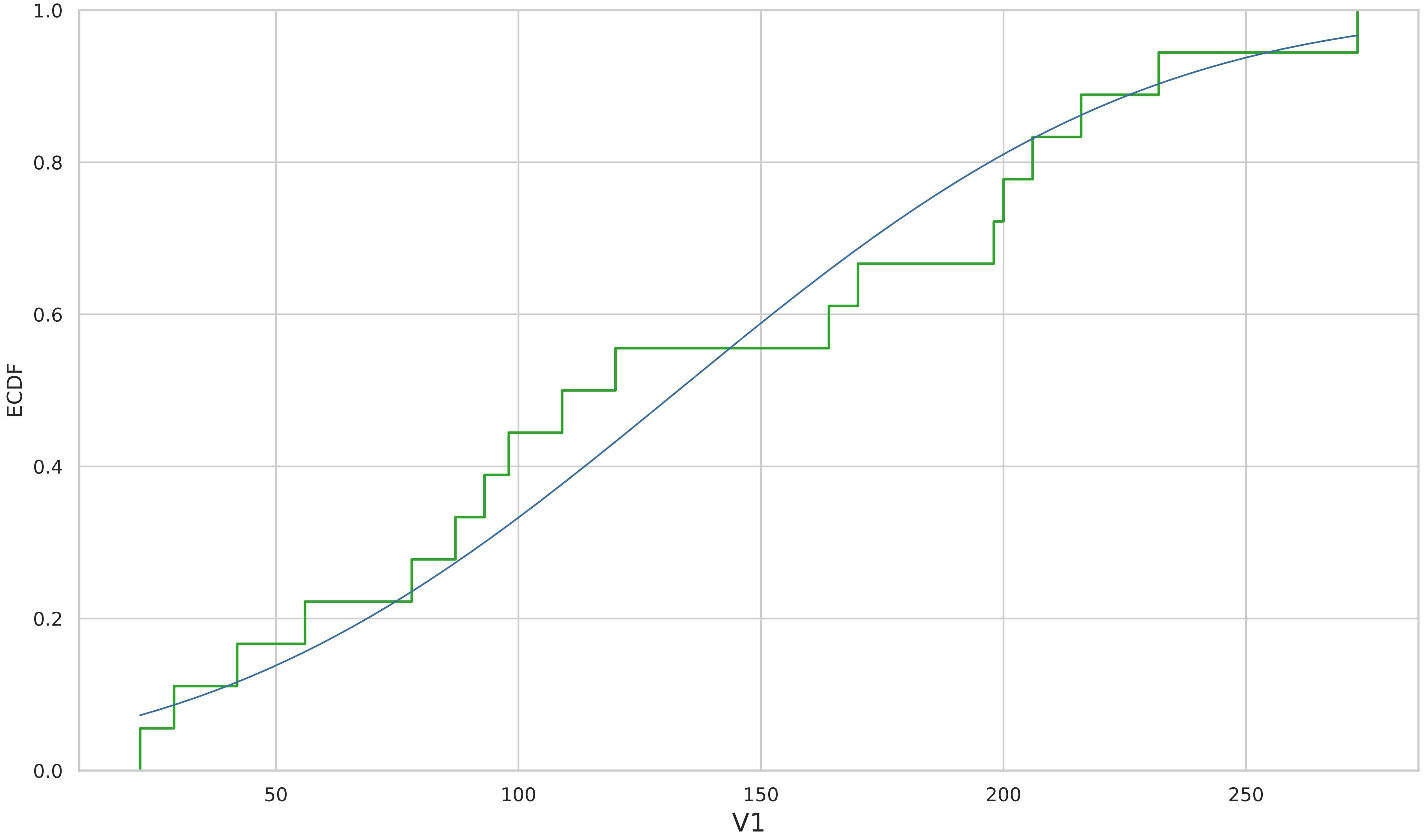

1.0
0.8
0.6
0.4
0.2
0.0
ECDF
100
200
300
400
500
600
700
800
outs86

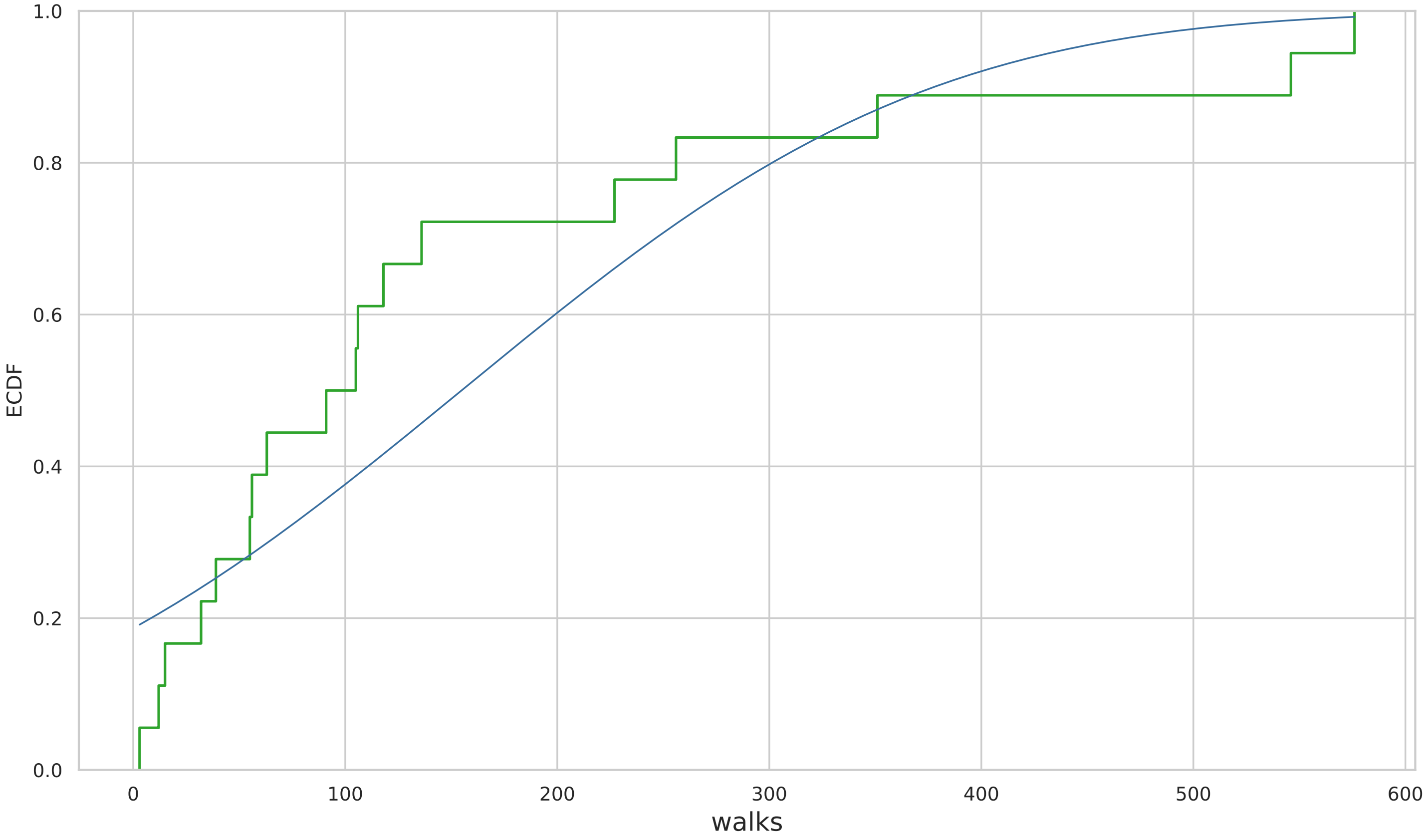

1.0
0.8
0.6
0.4
0.2
0.0
ECDF
0
200
400
600
800
1000
rbi

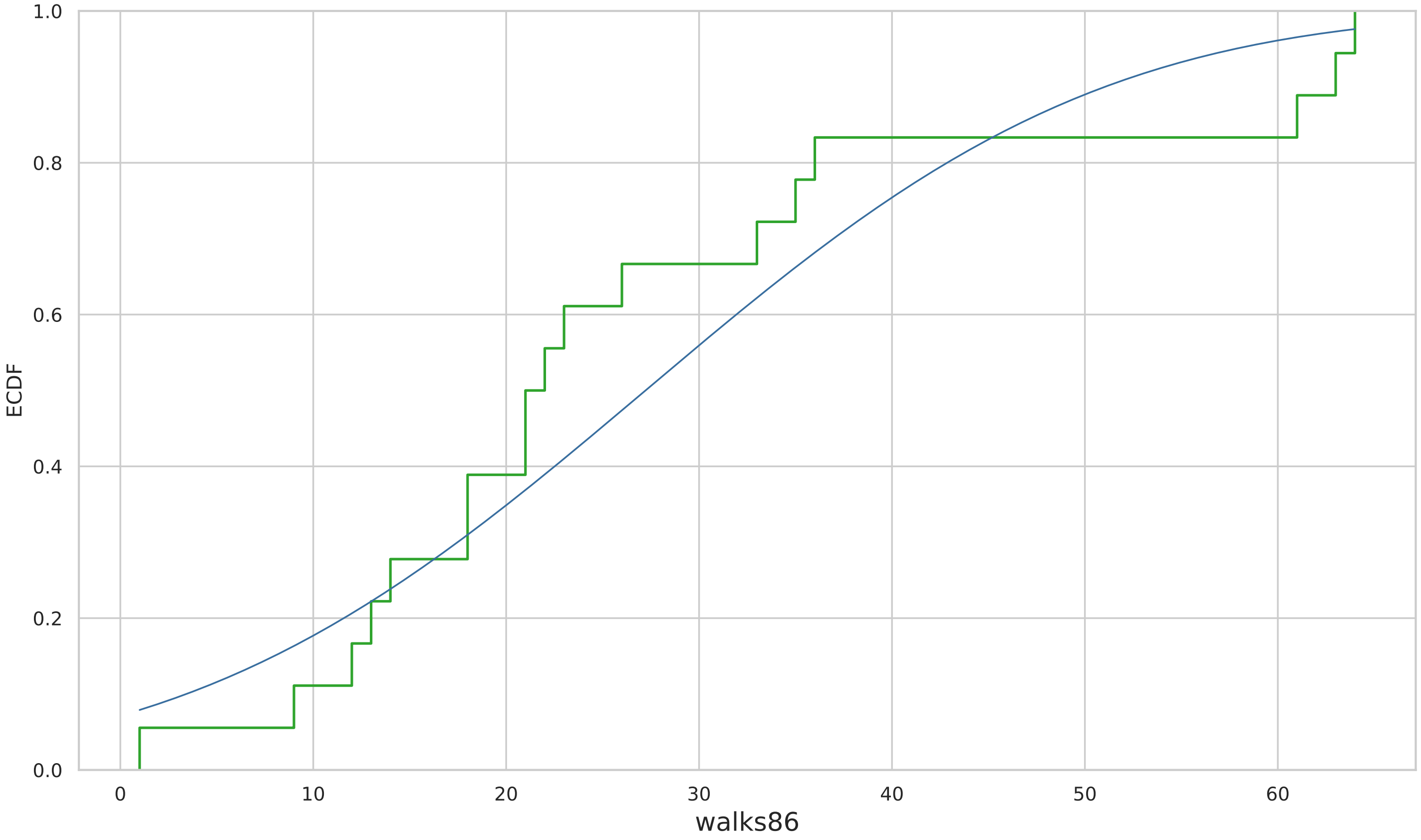

1.0
0.8
0.6
0.4
0.2
0.0
ECDF
0
20
40
60
80
rbi86


1.0
0.8
0.6
0.4
0.2
0.0
ECDF
0
200
400
600
800
runs


1.0
0.8
0.6
0.4
0.2
0.0
ECDF
0
20
40
60
80
runs86


1.0
0.8
0.6
0.4
0.2
0.0
ECDF
50
100
150
200
250
V1

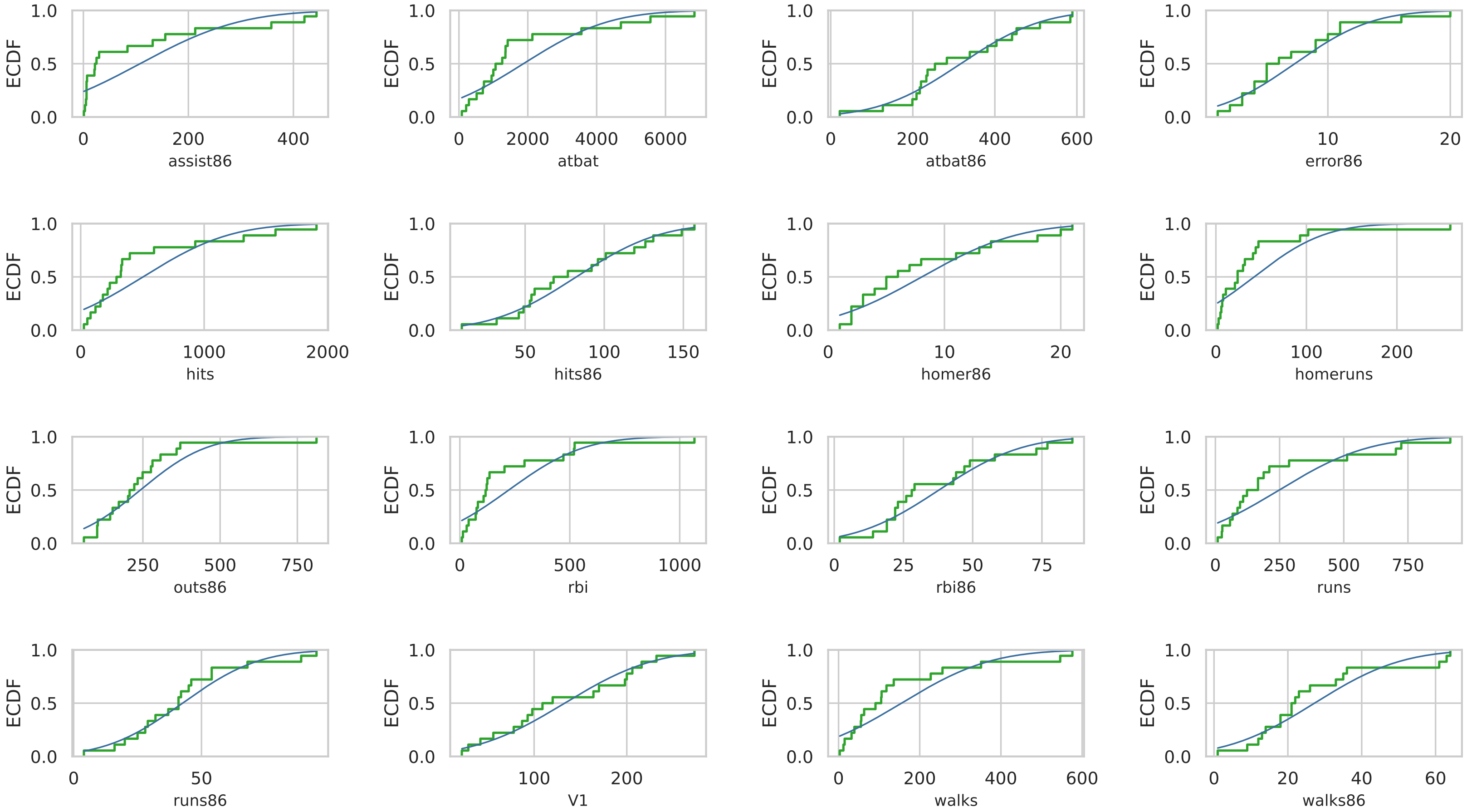

1.0
0.8
0.6
0.4
0.2
0.0
ECDF
0
100
200
300
400
500
600
walks



1.0
0.8
0.6
0.4
0.2
0.0
ECDF
0
10
20
30
40
50
60
walks86

ECDF Plots Summary

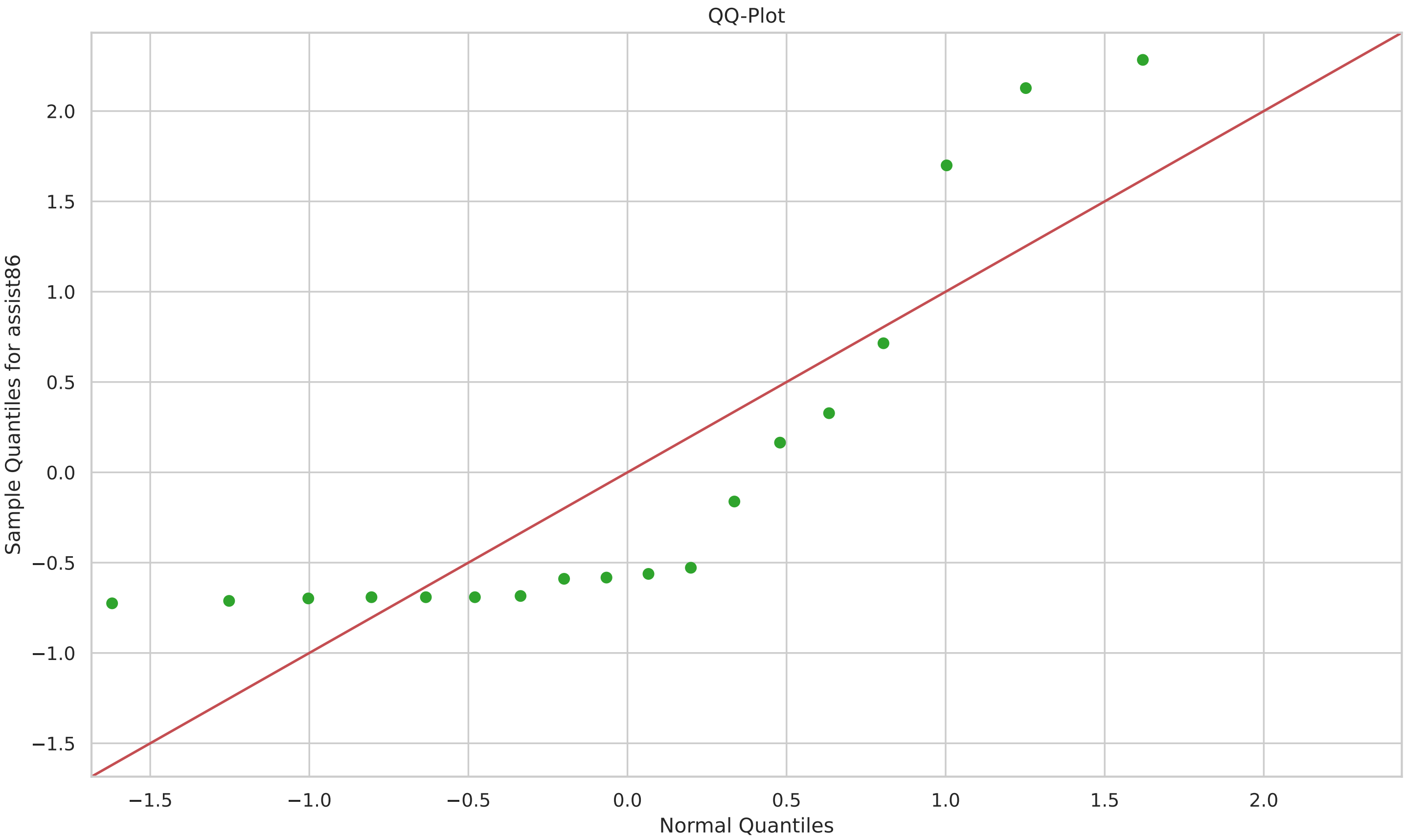
Multiple ECDF Plots of variables in one figure. Variables are sorted alphabetically.

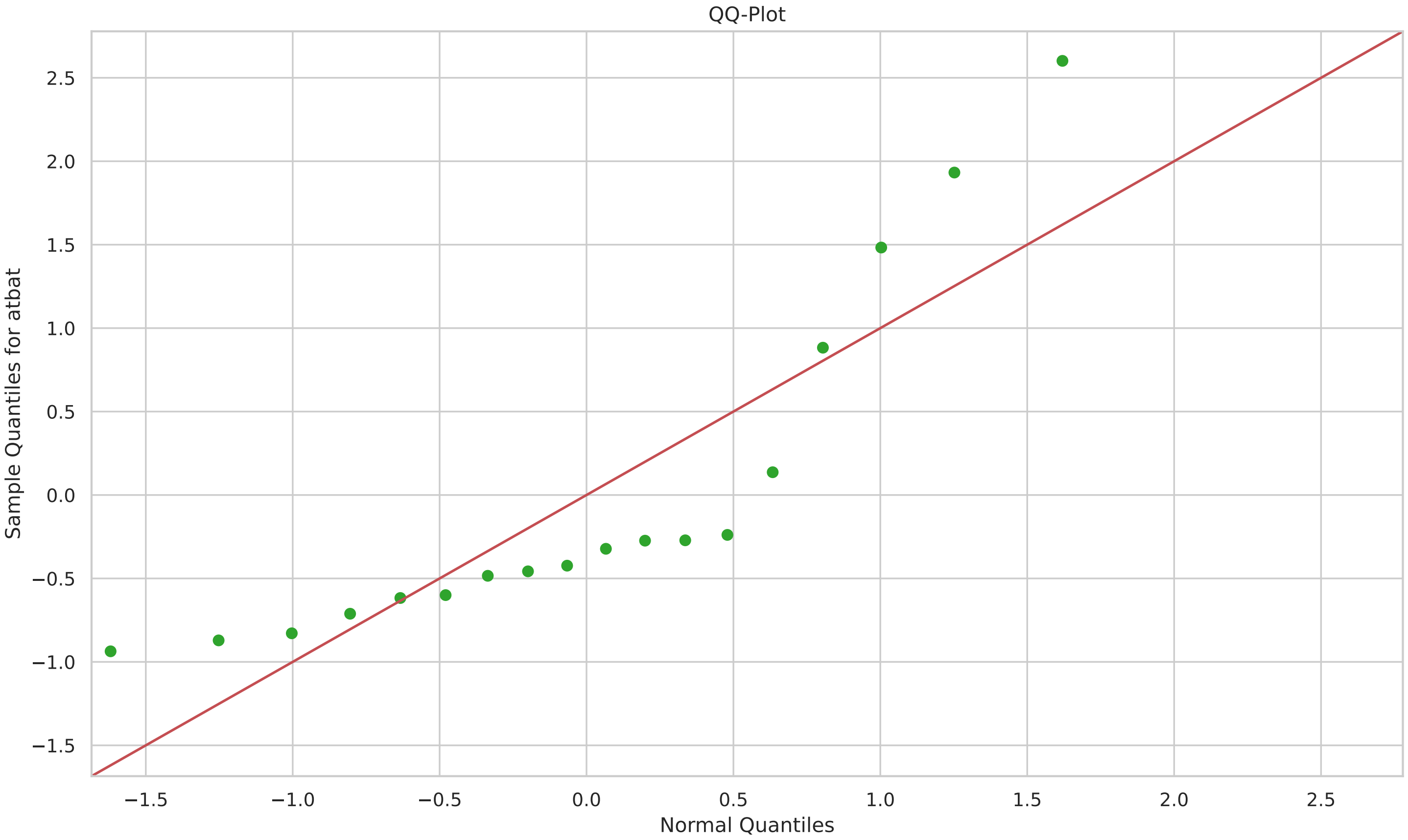


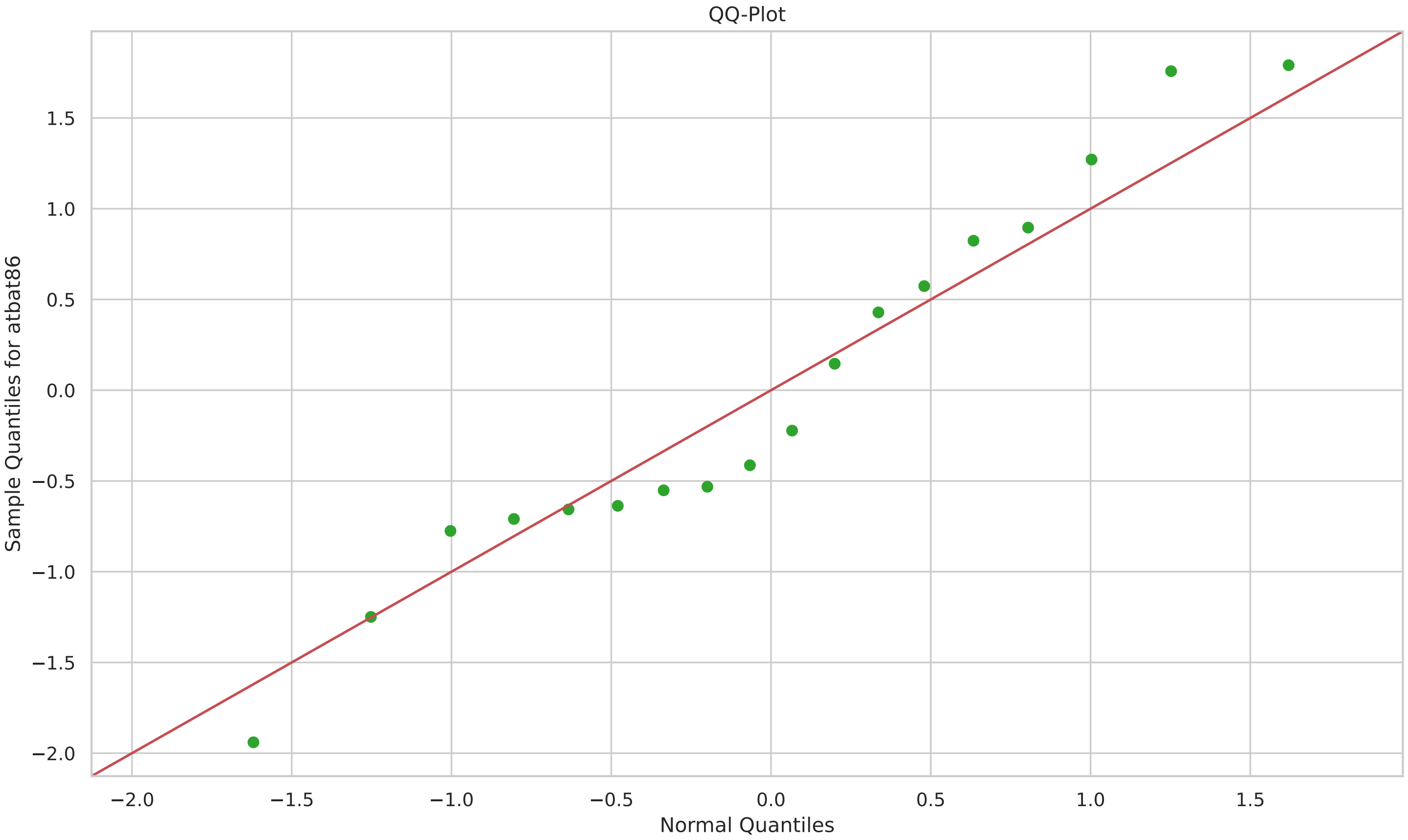
QQ-Plots

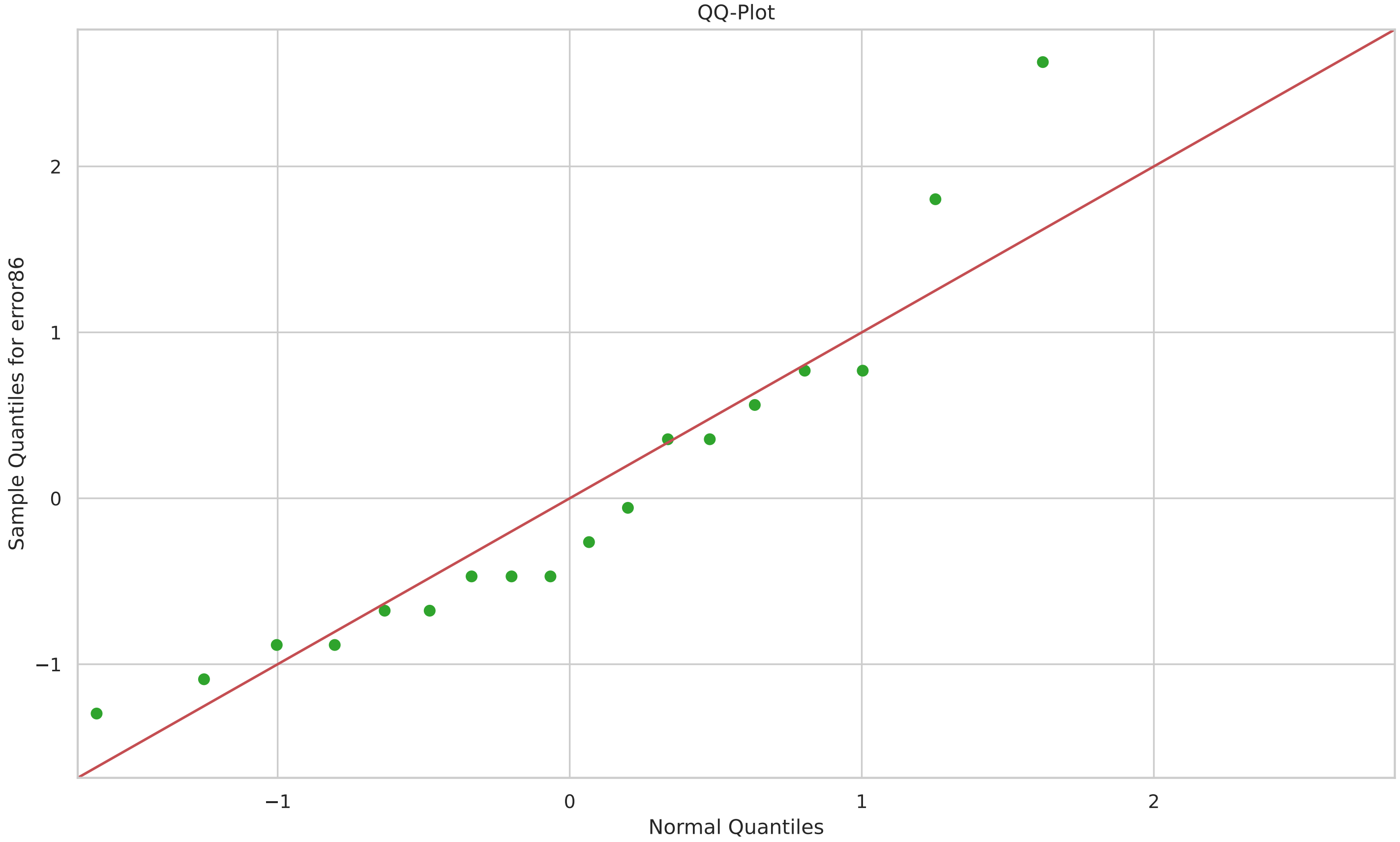
One QQ-Plot per page for each variable. Variables are sorted alphabetically.

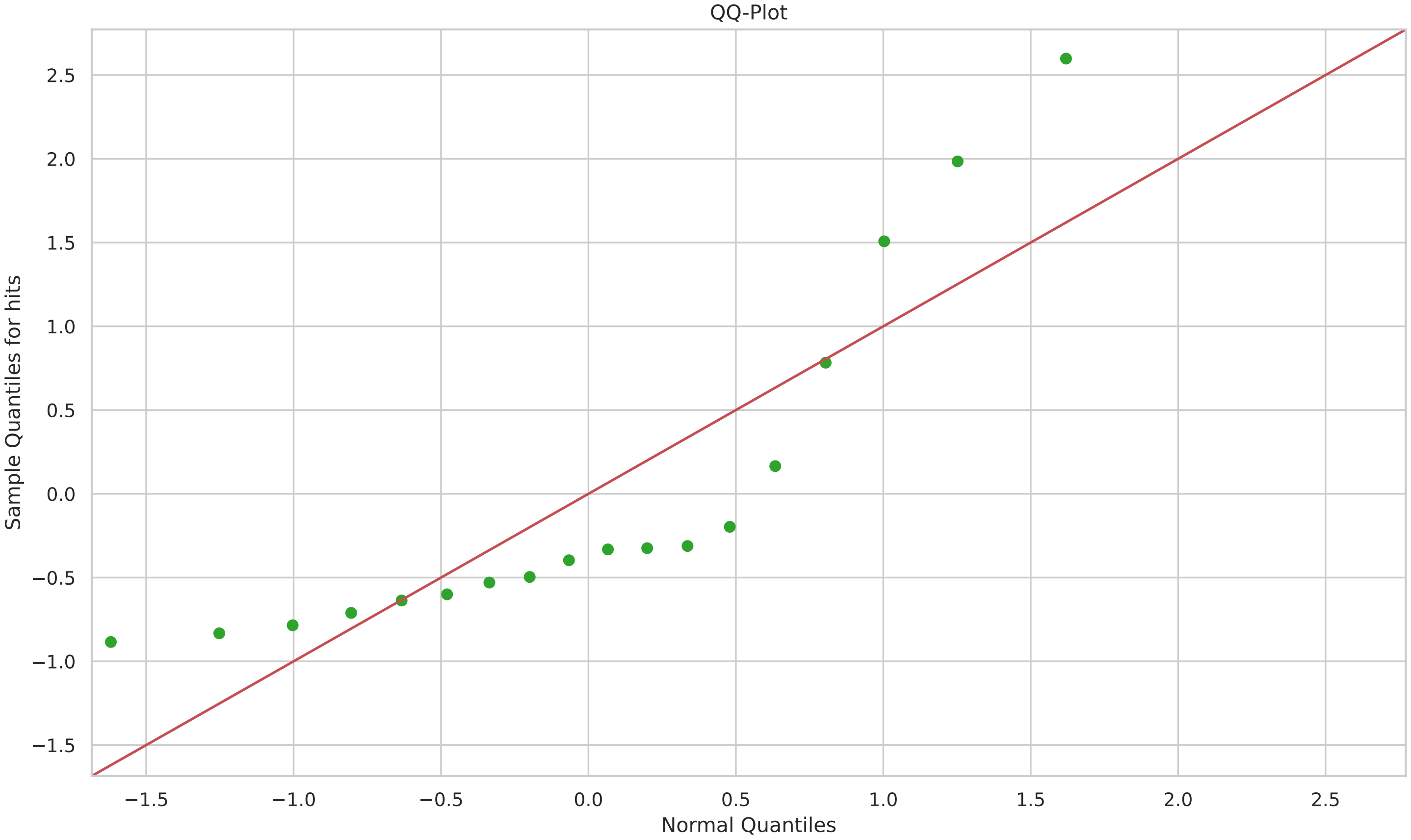
[]

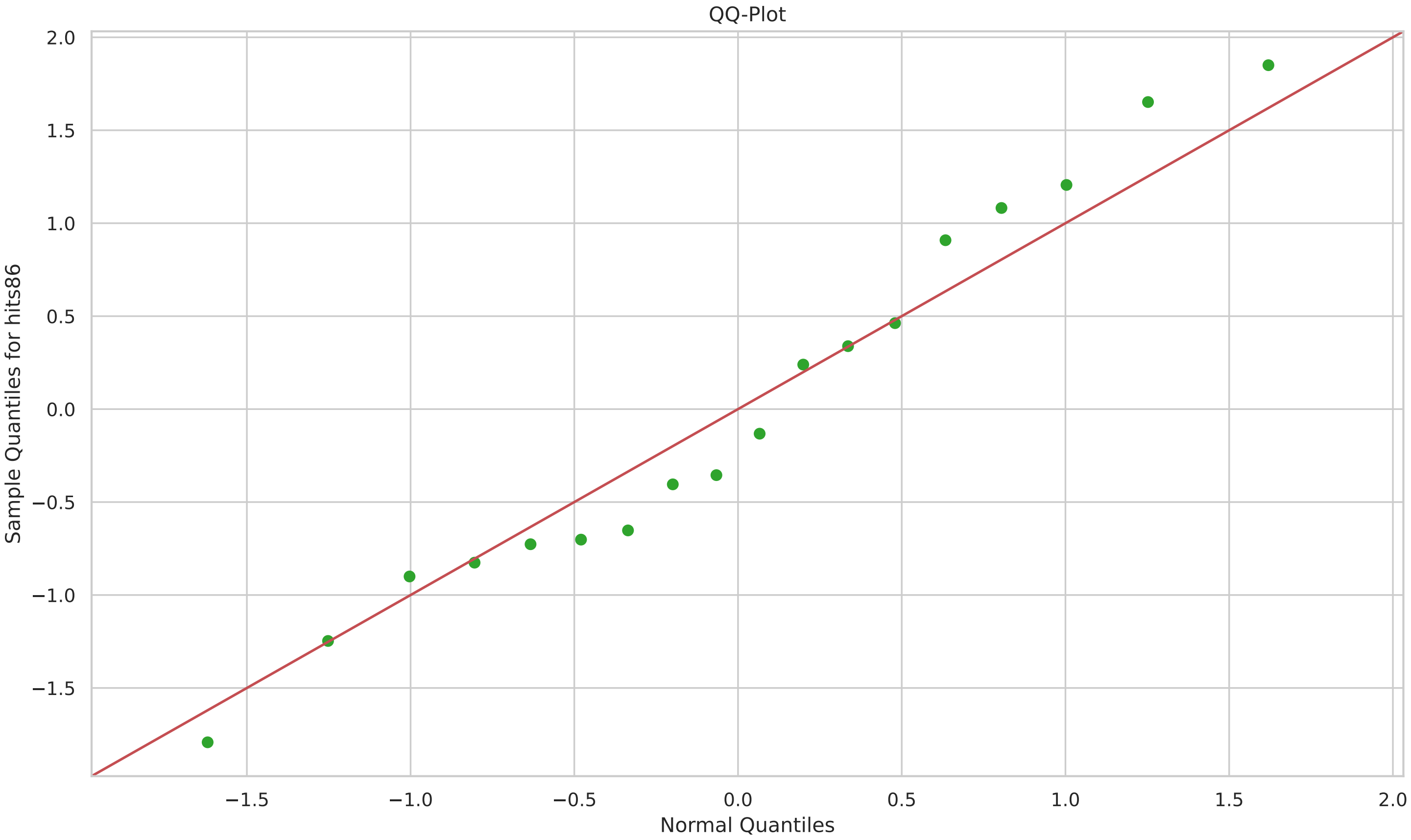


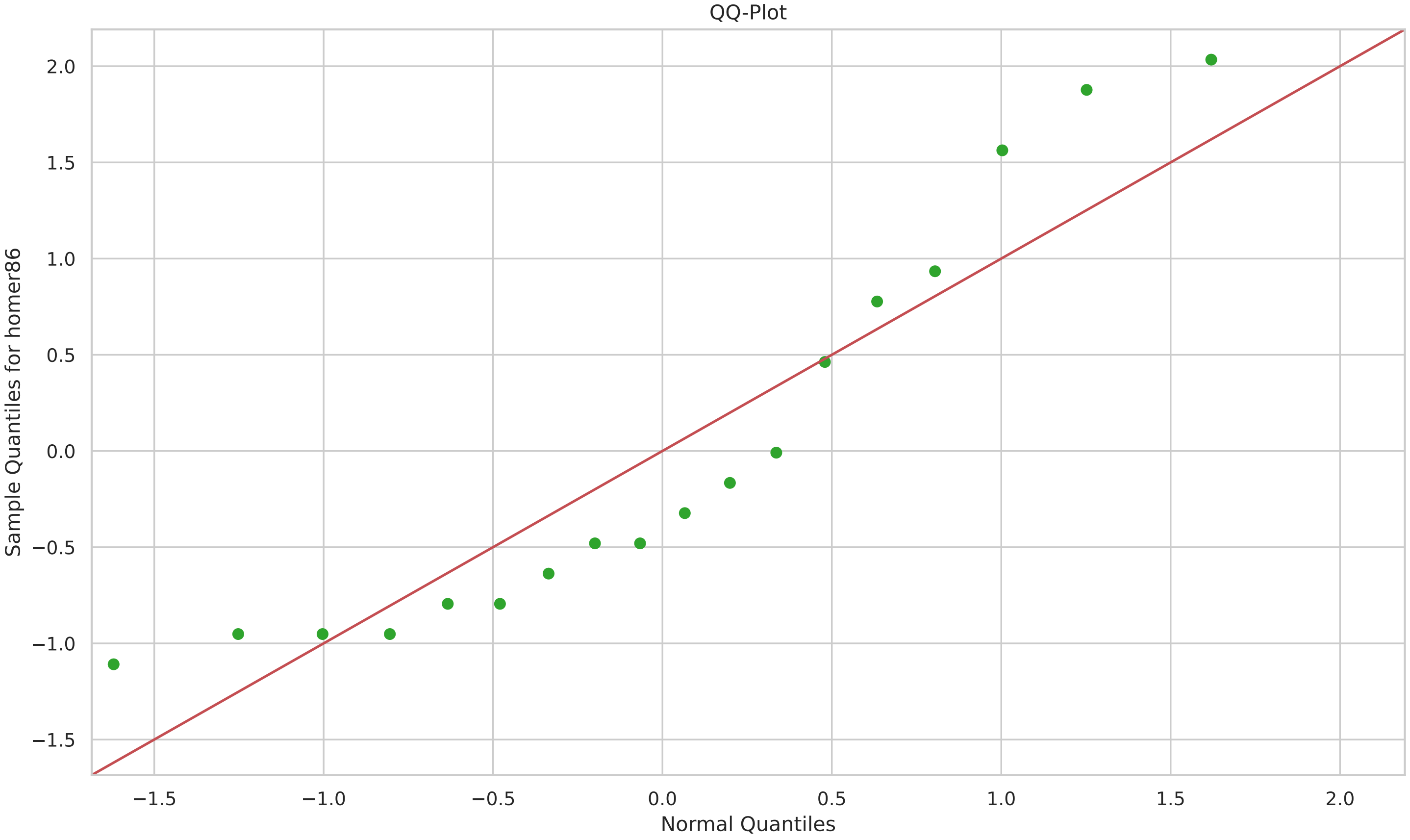

QQ-Plot
Sample Quantiles for atbat
Normal Quantiles
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5

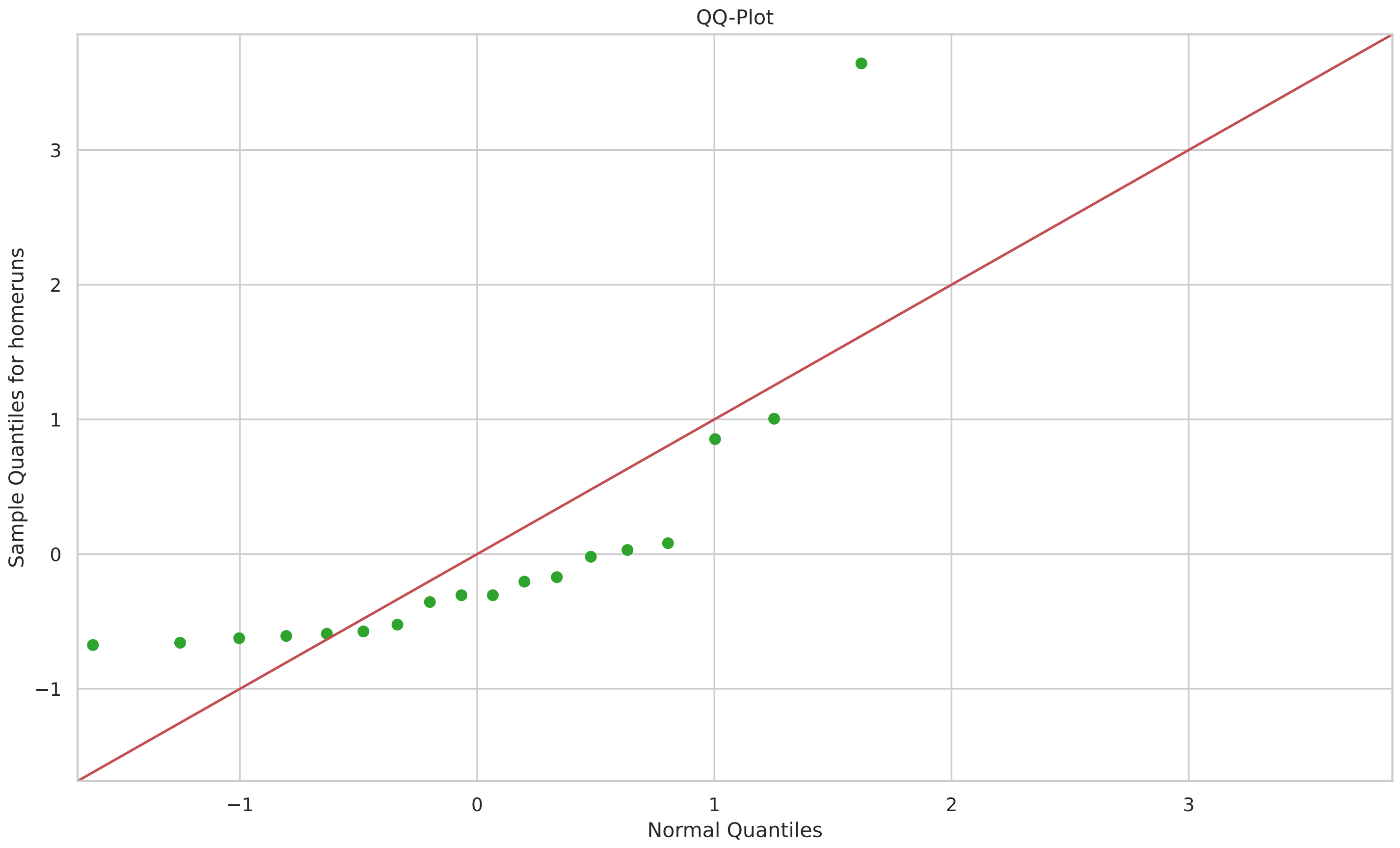

QQ-Plot
Sample Quantiles for atbat86
Normal Quantiles
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

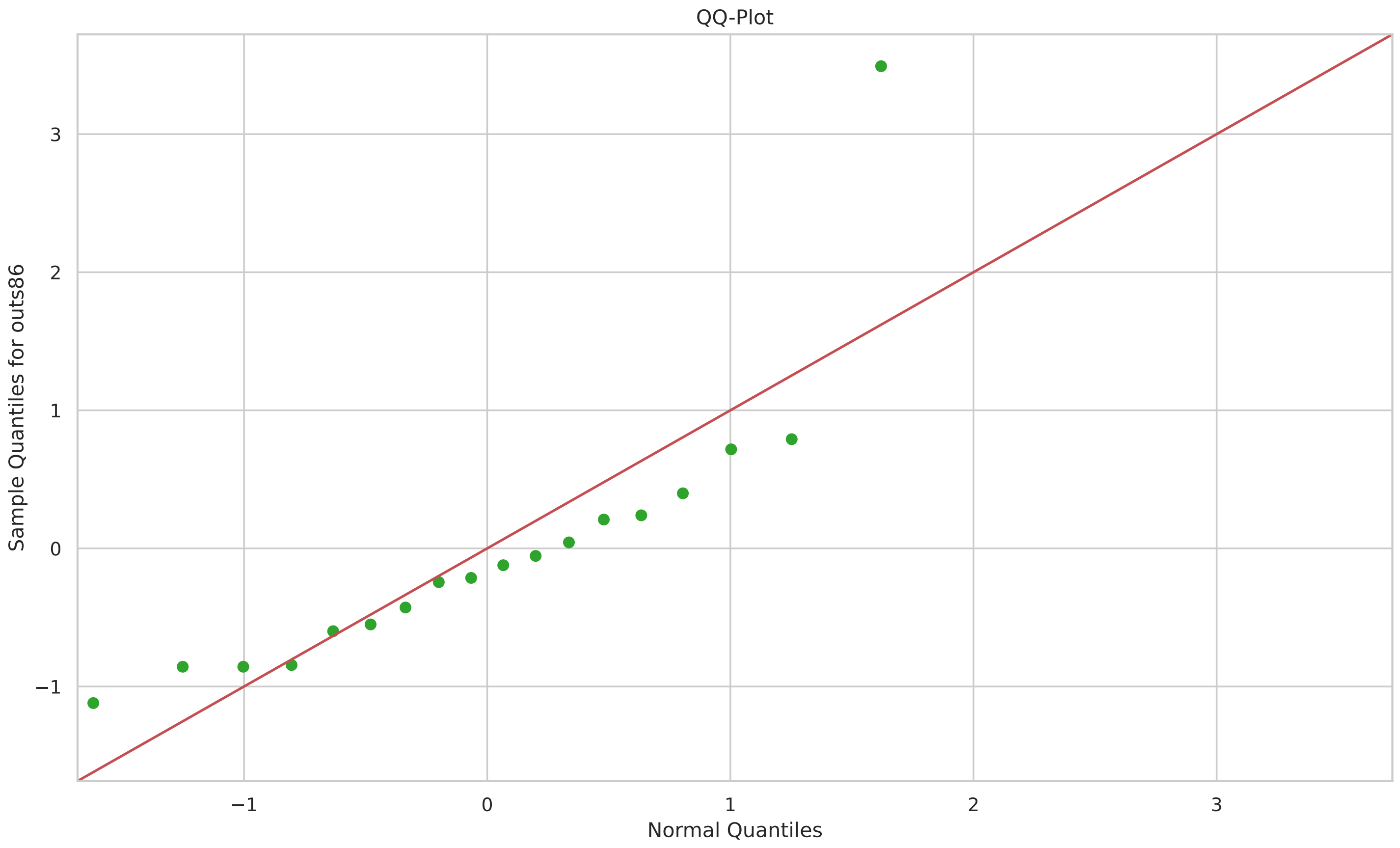

QQ-Plot
Sample Quantiles for error86
Normal Quantiles
2
1
0
−1
−1
0
1
2

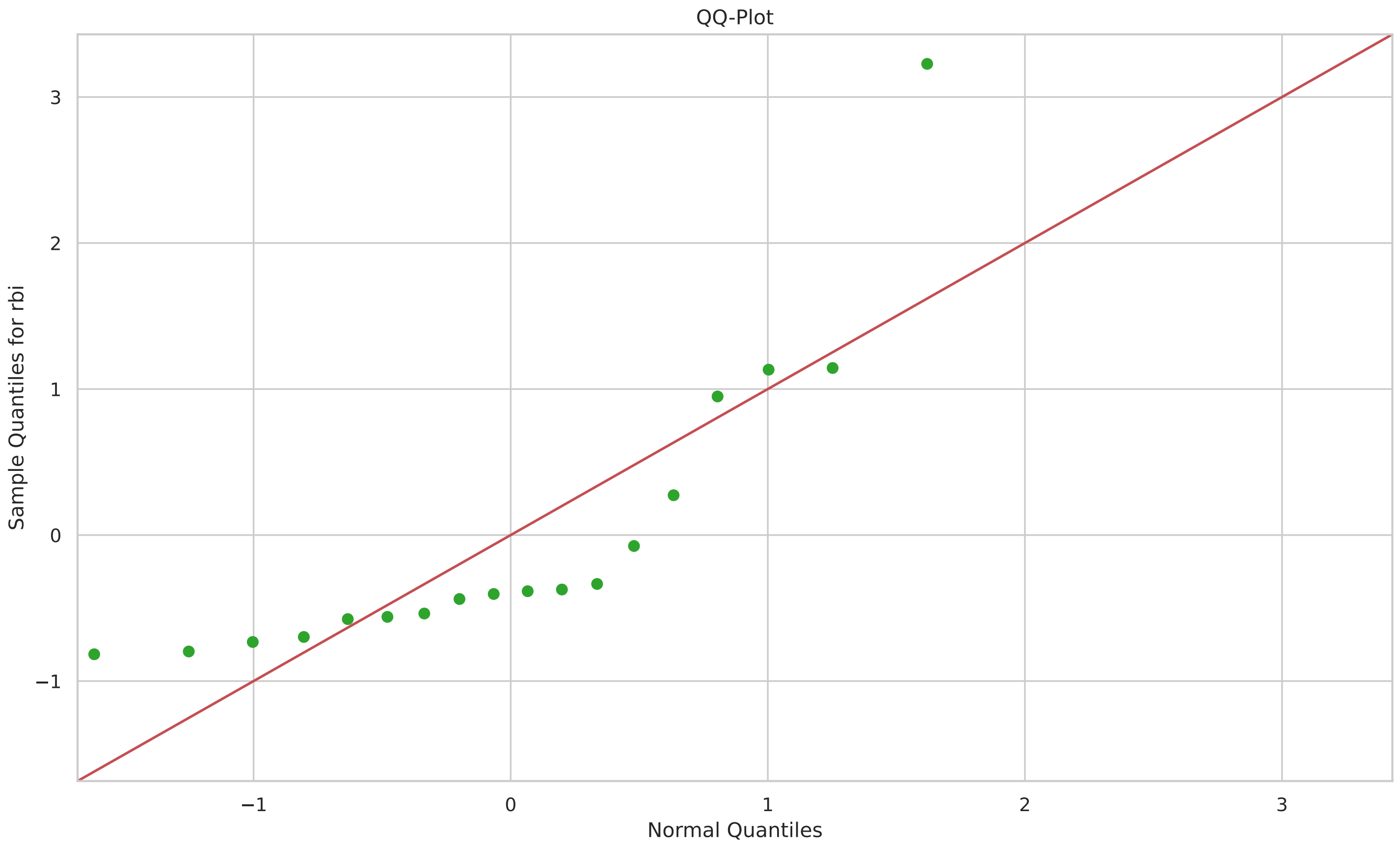

QQ-Plot
Sample Quantiles for hits
Normal Quantiles
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5

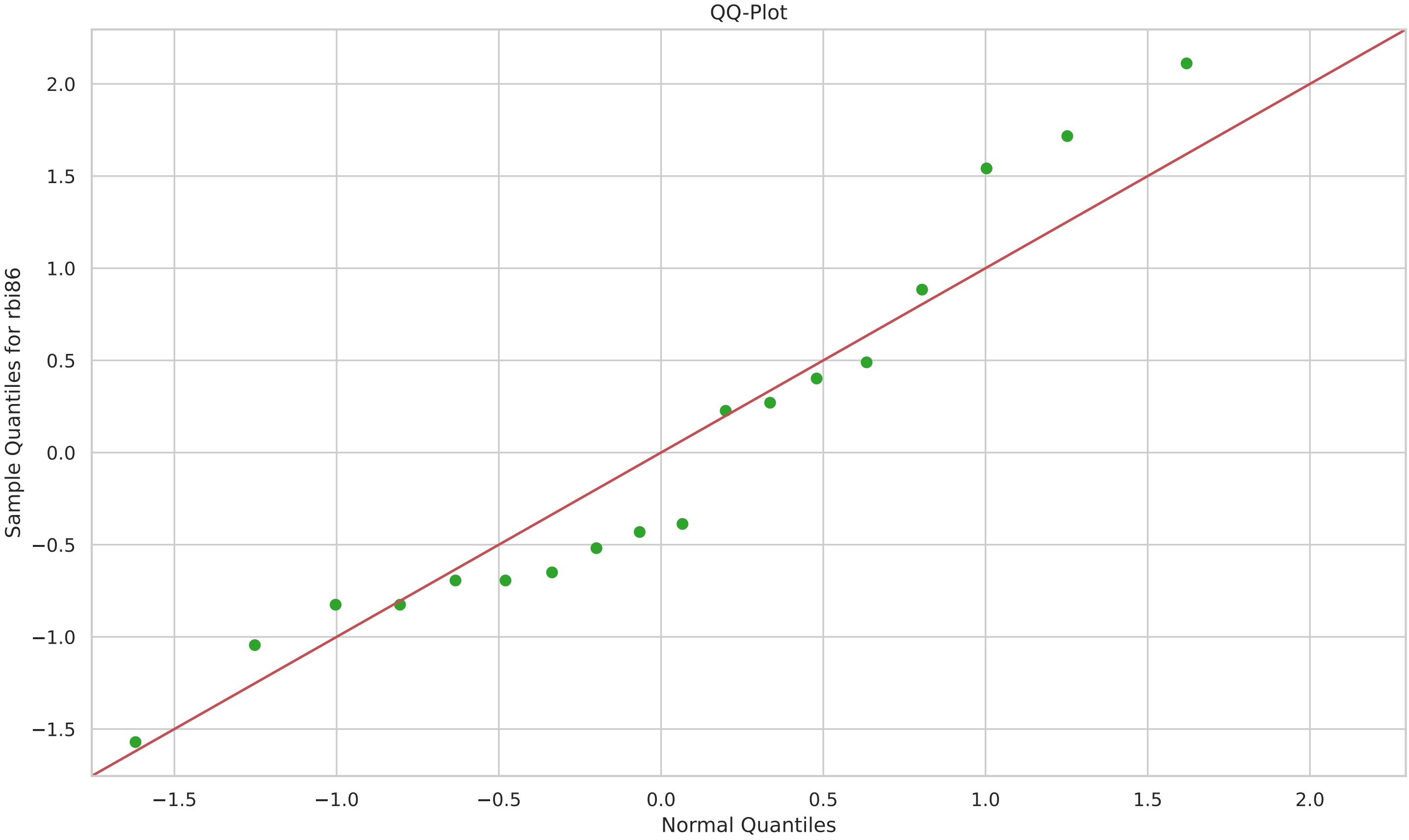

QQ-Plot
Sample Quantiles for hits86
Normal Quantiles
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0

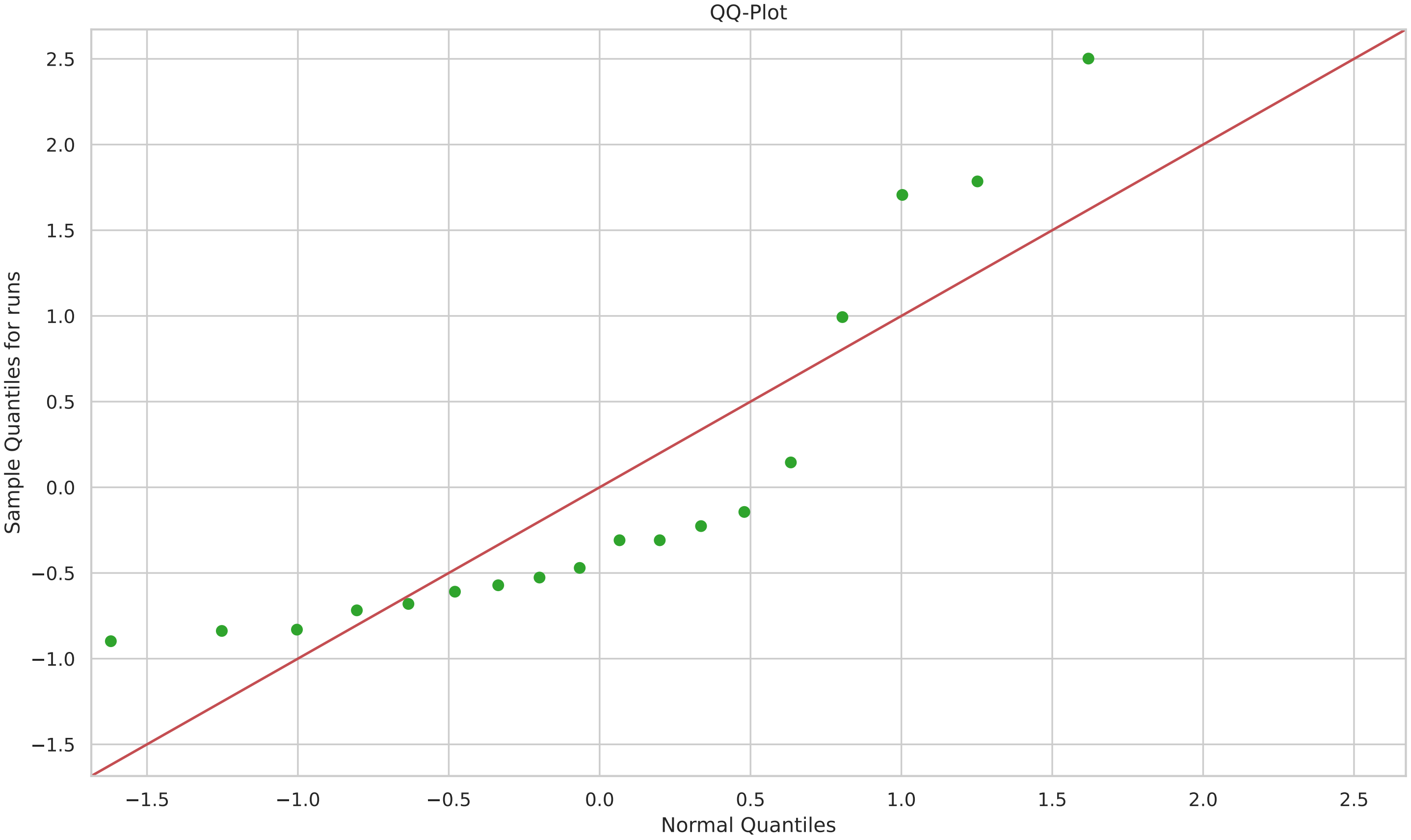

QQ-Plot
Sample Quantiles for homer86
Normal Quantiles
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0

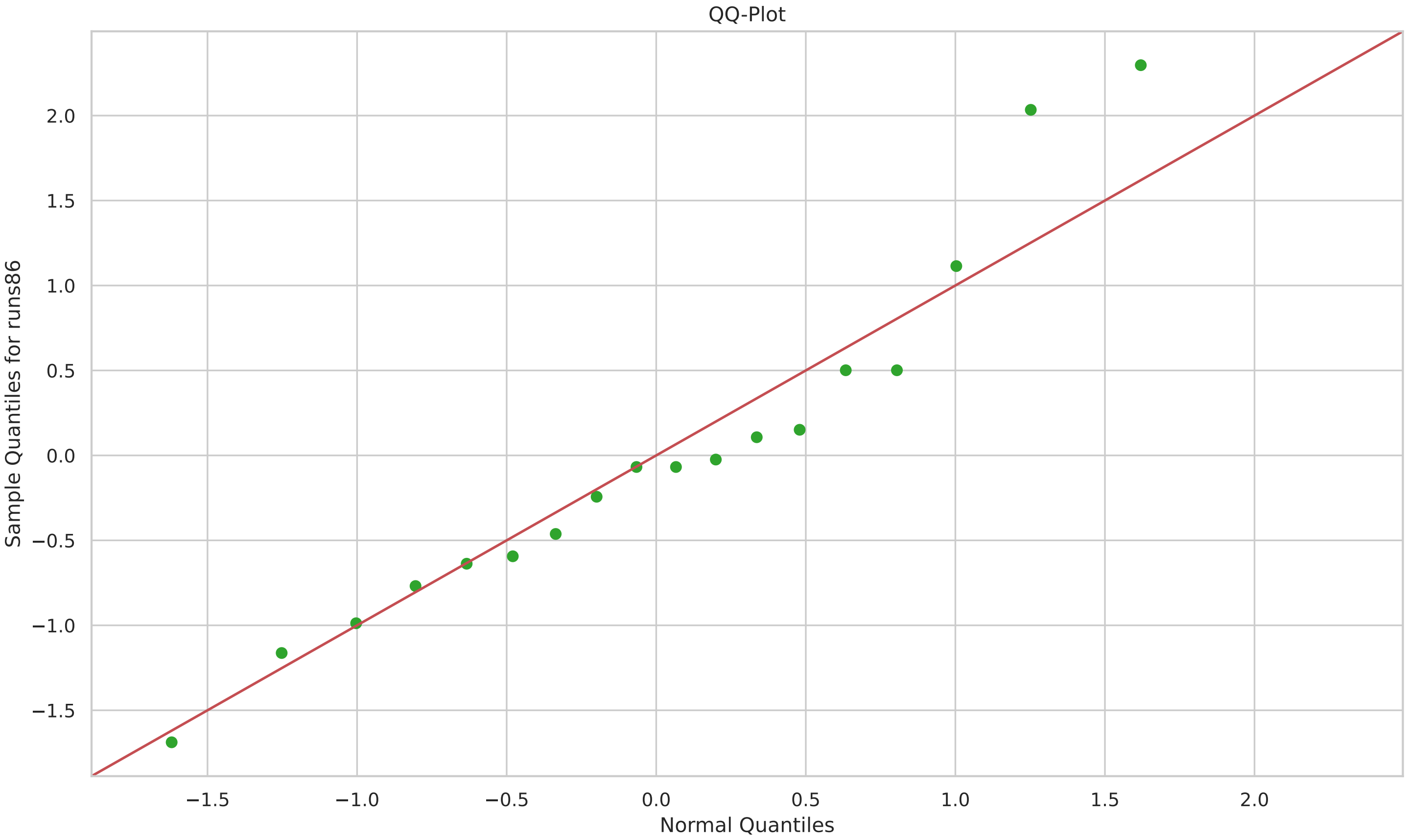

QQ-Plot
Sample Quantiles for homeruns
Normal Quantiles
3
2
1
0
−1
−1
0
1
2
3

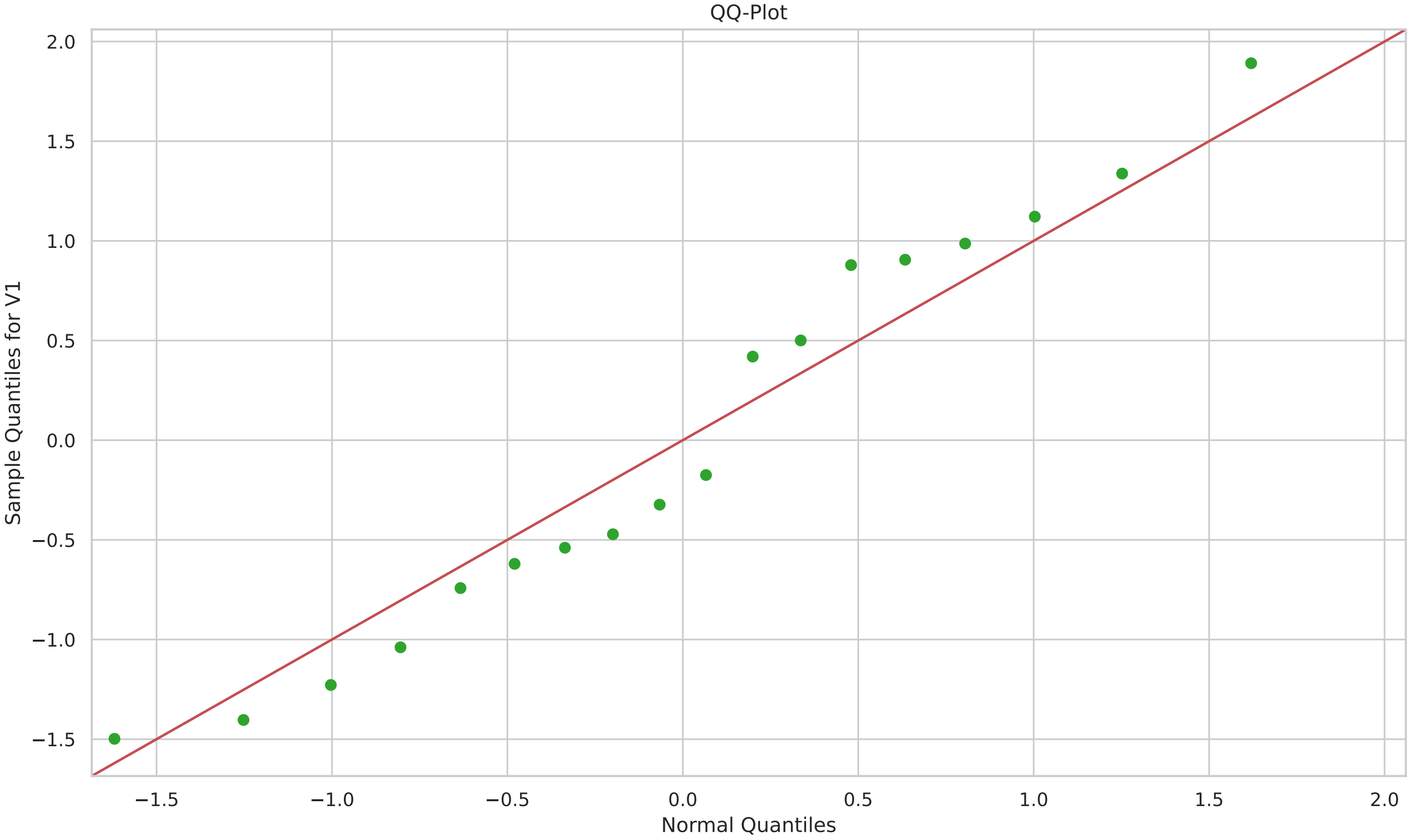

QQ-Plot
Sample Quantiles for outs86
Normal Quantiles
3
2
1
0
−1
−1
0
1
2
3

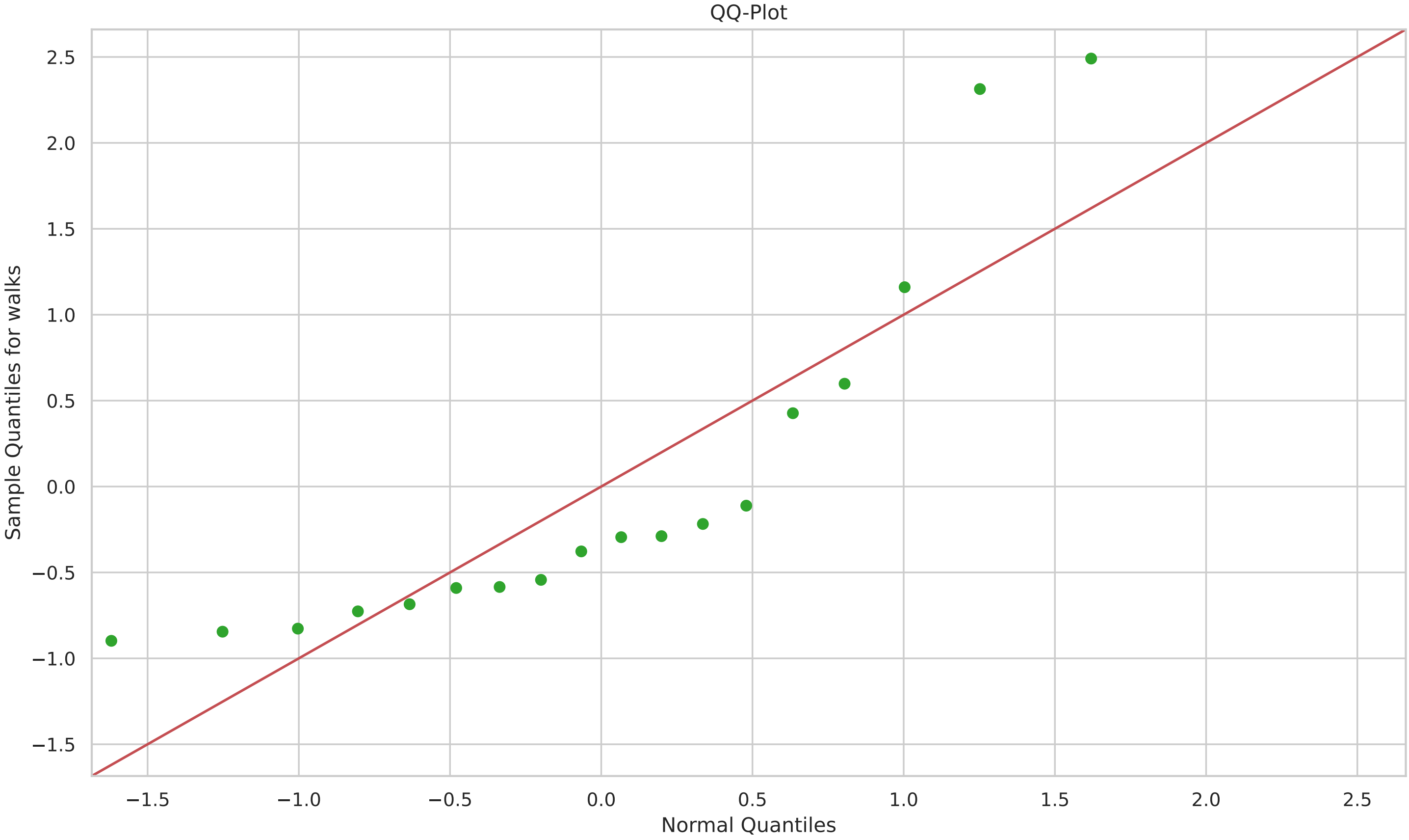

QQ-Plot
Sample Quantiles for rbi
Normal Quantiles
3
2
1
0
−1
−1
0
1
2
3

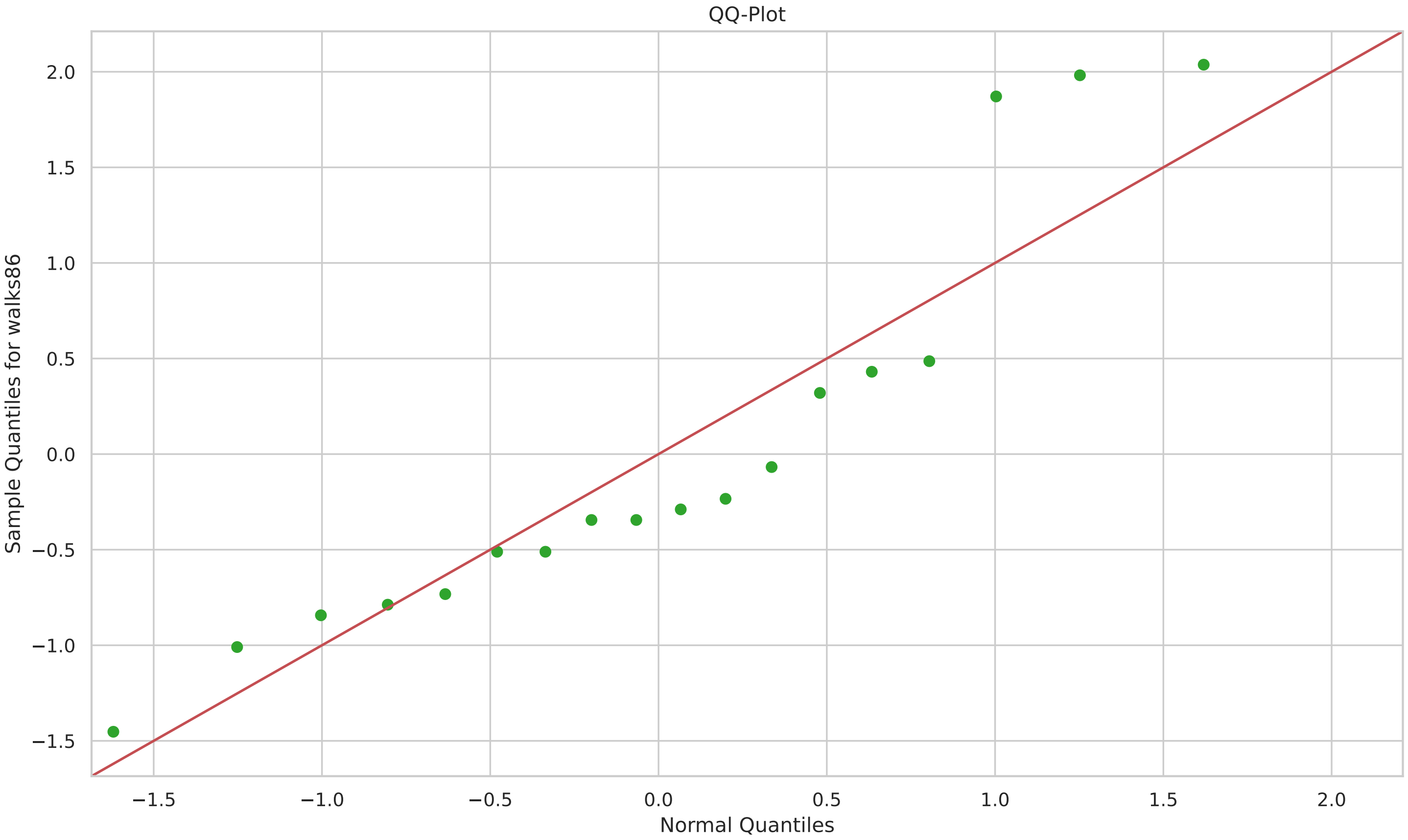

QQ-Plot
Sample Quantiles for rbi86
Normal Quantiles
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0


QQ-Plot
Sample Quantiles for runs
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
Normal Quantiles


QQ-Plot
Sample Quantiles for runs86
Normal Quantiles
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0

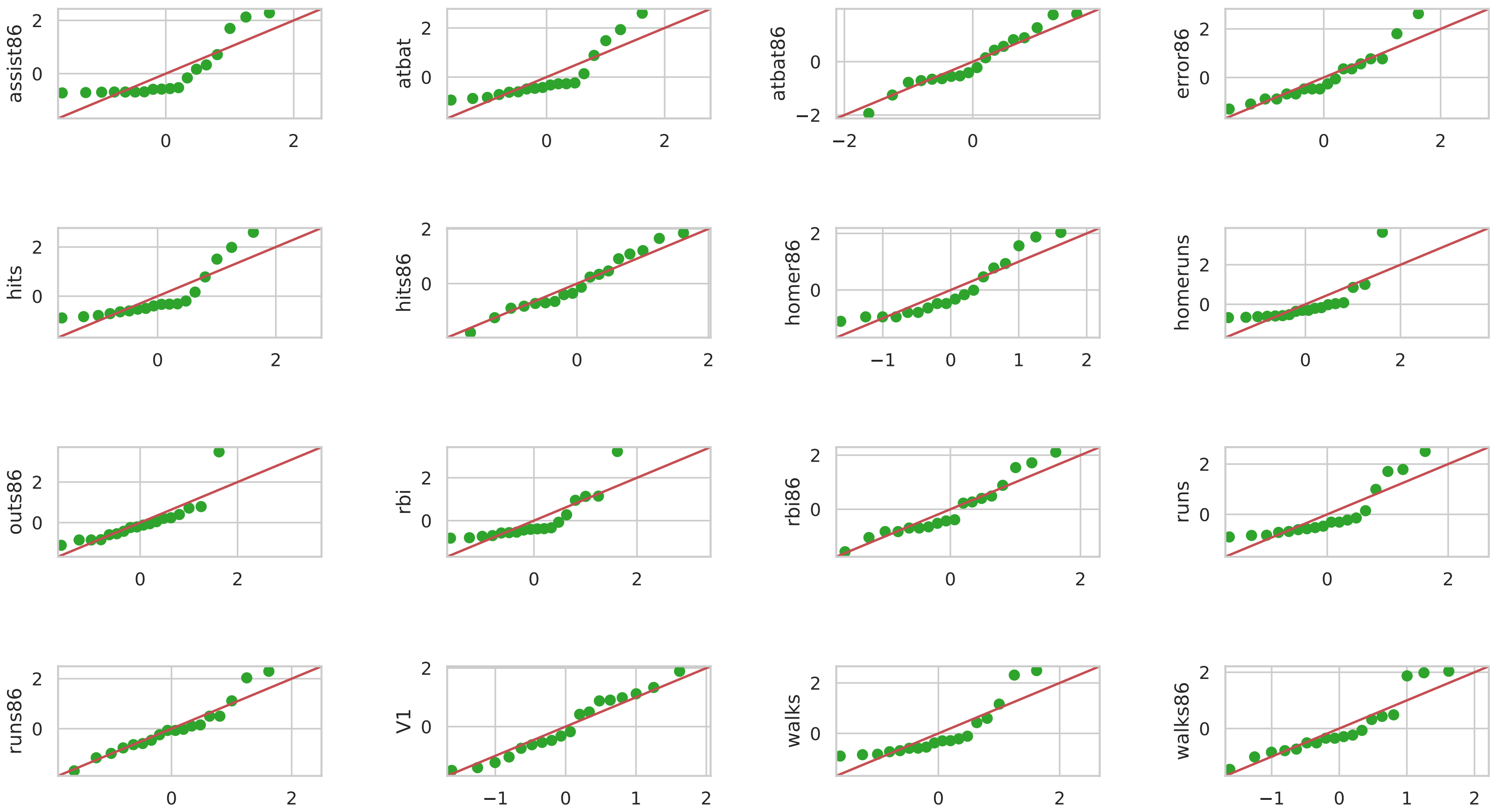

QQ-Plot
Sample Quantiles for V1
Normal Quantiles
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0


QQ-Plot
Sample Quantiles for walks
Normal Quantiles
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5


QQ-Plot
Sample Quantiles for walks86
Normal Quantiles
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0

QQ-Plots Summary

Multiple QQ-Plots of variables in one figure. Variables are sorted alphabetically.



Results for Discrete Variables

Descriptive Statistics

Totals

The table is sorted by the variable name. If any, N Unique contains the missing category.

	N Obs	N Missing	N Valid	% Complete	N Unique
div86	18	0	18	100	2
league86	18	0	18	100	2
name1	18	0	18	100	18
name2	18	0	18	100	18
posit86	18	0	18	100	9
team86	18	0	18	100	14
years	18	0	18	100	11

Frequencies

The table is sorted by the variable name. For each variable, a maximum of 20 unique values are considered, sorted in decreasing order of their frequency. If any, missings are counted as a category.

Variable	Category	Frequency	Percent
div86	W	10	0.555556
div86	E	8	0.444444
league86	A	14	0.777778
league86	N	4	0.222222
name1	Chris	1	0.0555556
name1	Jose	1	0.0555556
name1	Garry	1	0.0555556
name1	Scott	1	0.0555556
name1	John	1	0.0555556
name1	Lloyd	1	0.0555556
name1	Denny	1	0.0555556
name1	Mel	1	0.0555556
name1	Lou	1	0.0555556
name1	Donnie	1	0.0555556
name1	Mike	1	0.0555556
name1	Bruce	1	0.0555556
name1	Bob	1	0.0555556
name1	George	1	0.0555556
name1	Ed	1	0.0555556
name1	Pat	1	0.0555556
name1	Daryl	1	0.0555556
name1	Bill	1	0.0555556
name2	Moseby	1	0.0555556
name2	Hill	1	0.0555556
name2	Bochy	1	0.0555556
name2	Shelby	1	0.0555556
name2	Templeton	1	0.0555556
name2	Walling	1	0.0555556
name2	Boone	1	0.0555556
name2	Kingery	1	0.0555556
name2	Boston	1	0.0555556
name2	Hall	1	0.0555556
name2	Romero	1	0.0555556
name2	Bando	1	0.0555556
name2	Whitaker	1	0.0555556
name2	Bradley	1	0.0555556
name2	Uribe	1	0.0555556

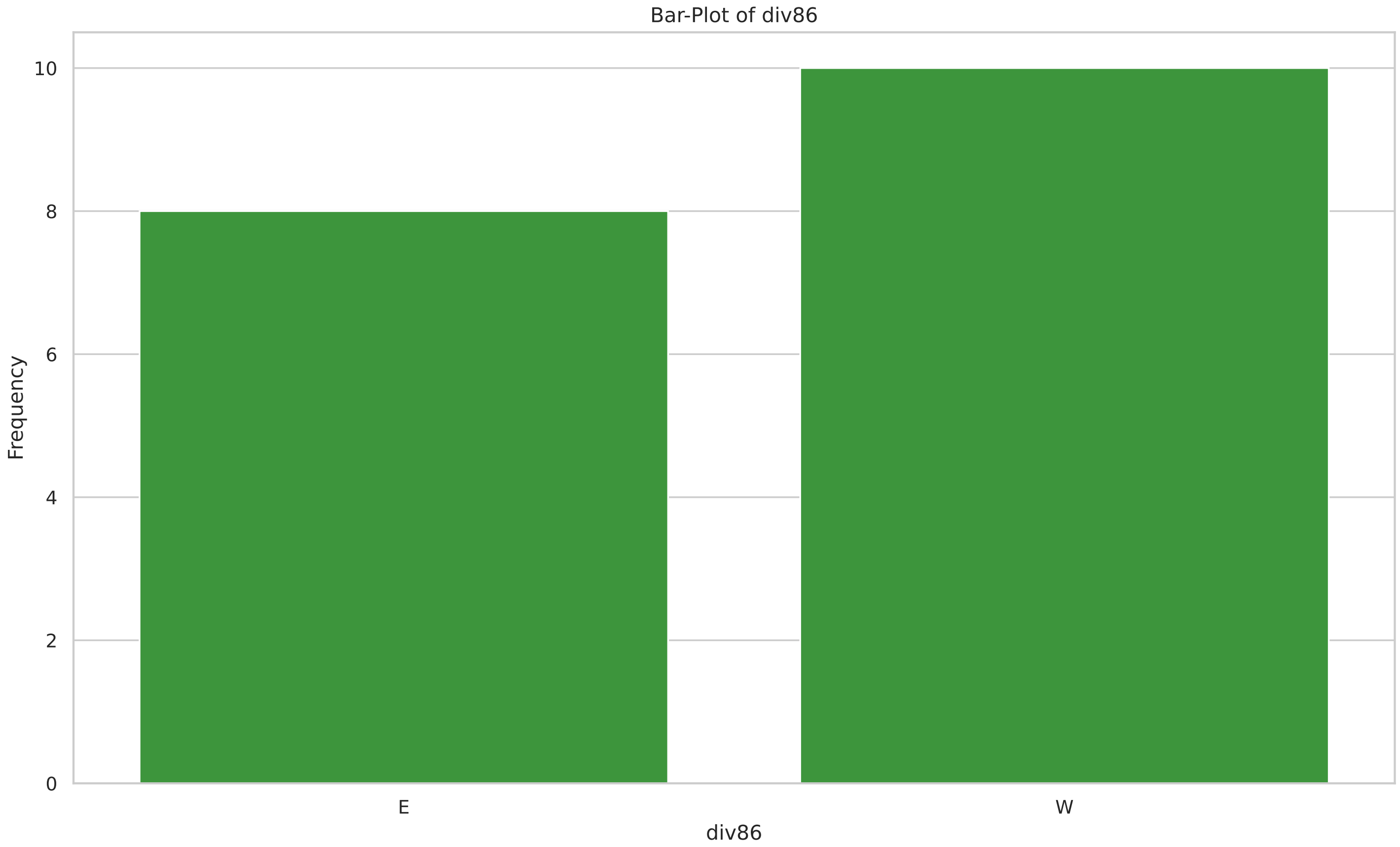
Variable	Category	Frequency	Percent
name2	Schroeder	1	0.0555556
name2	Hendrick	1	0.0555556
name2	Sheridan	1	0.0555556
posit86	C	4	0.222222
posit86	OF	4	0.222222
posit86	SS	3	0.166667
posit86	CF	2	0.111111
posit86	2B	1	0.0555556
posit86	3B	1	0.0555556
posit86	LF	1	0.0555556
posit86	UT	1	0.0555556
posit86	23	1	0.0555556
team86	Cal	2	0.111111
team86	SD	2	0.111111
team86	Det	2	0.111111
team86	Cle	2	0.111111
team86	Chi	1	0.0555556
team86	Bal	1	0.0555556
team86	Bos	1	0.0555556
team86	SF	1	0.0555556
team86	Sea	1	0.0555556
team86	Tor	1	0.0555556
team86	Hou	1	0.0555556
team86	Oak	1	0.0555556
team86	Mil	1	0.0555556
team86	KC	1	0.0555556
years	6	4	0.222222
years	3	3	0.166667
years	8	2	0.111111
years	4	2	0.111111
years	10	1	0.0555556
years	7	1	0.0555556
years	1	1	0.0555556
years	16	1	0.0555556
years	11	1	0.0555556
years	5	1	0.0555556
years	12	1	0.0555556

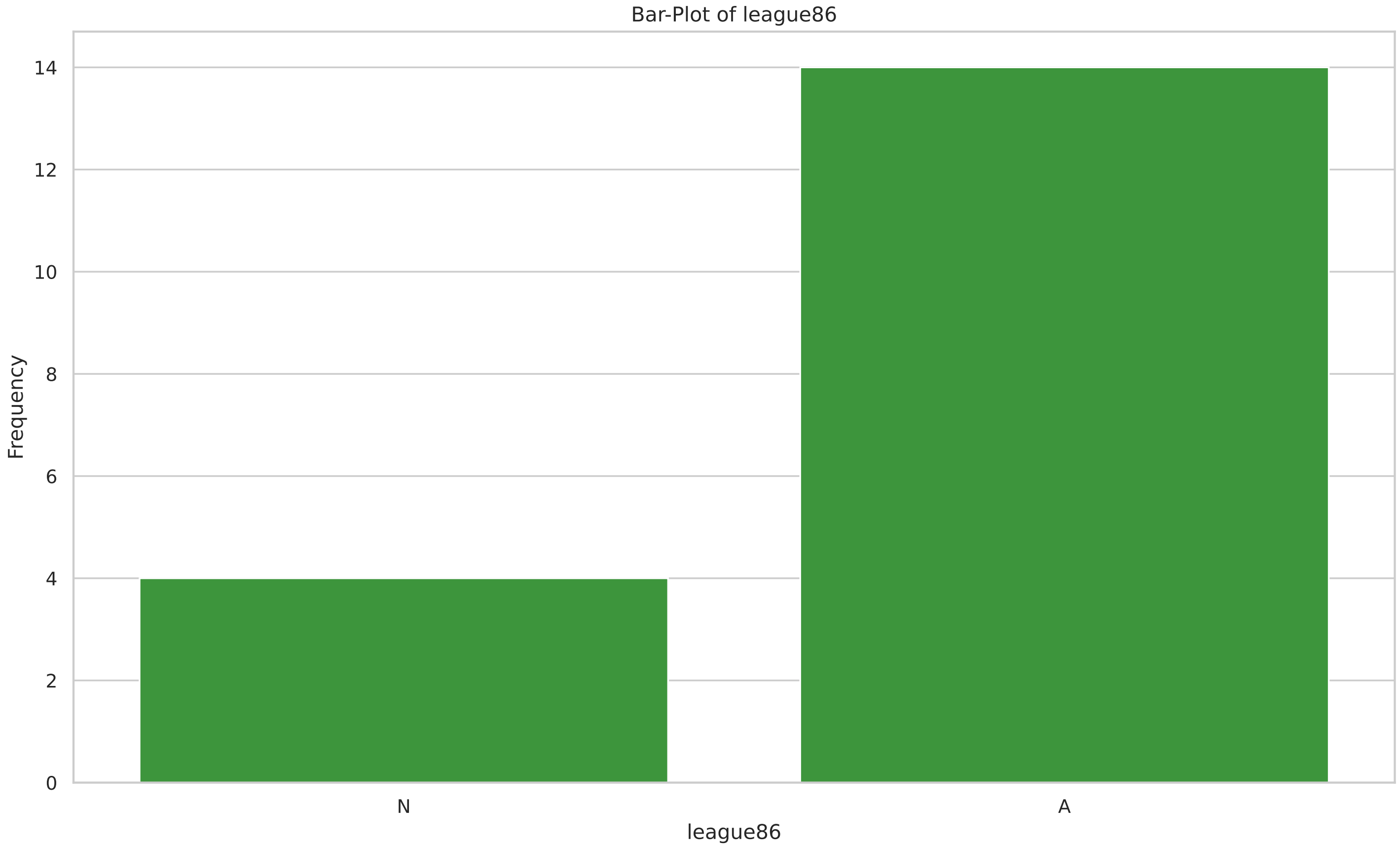
Graphics

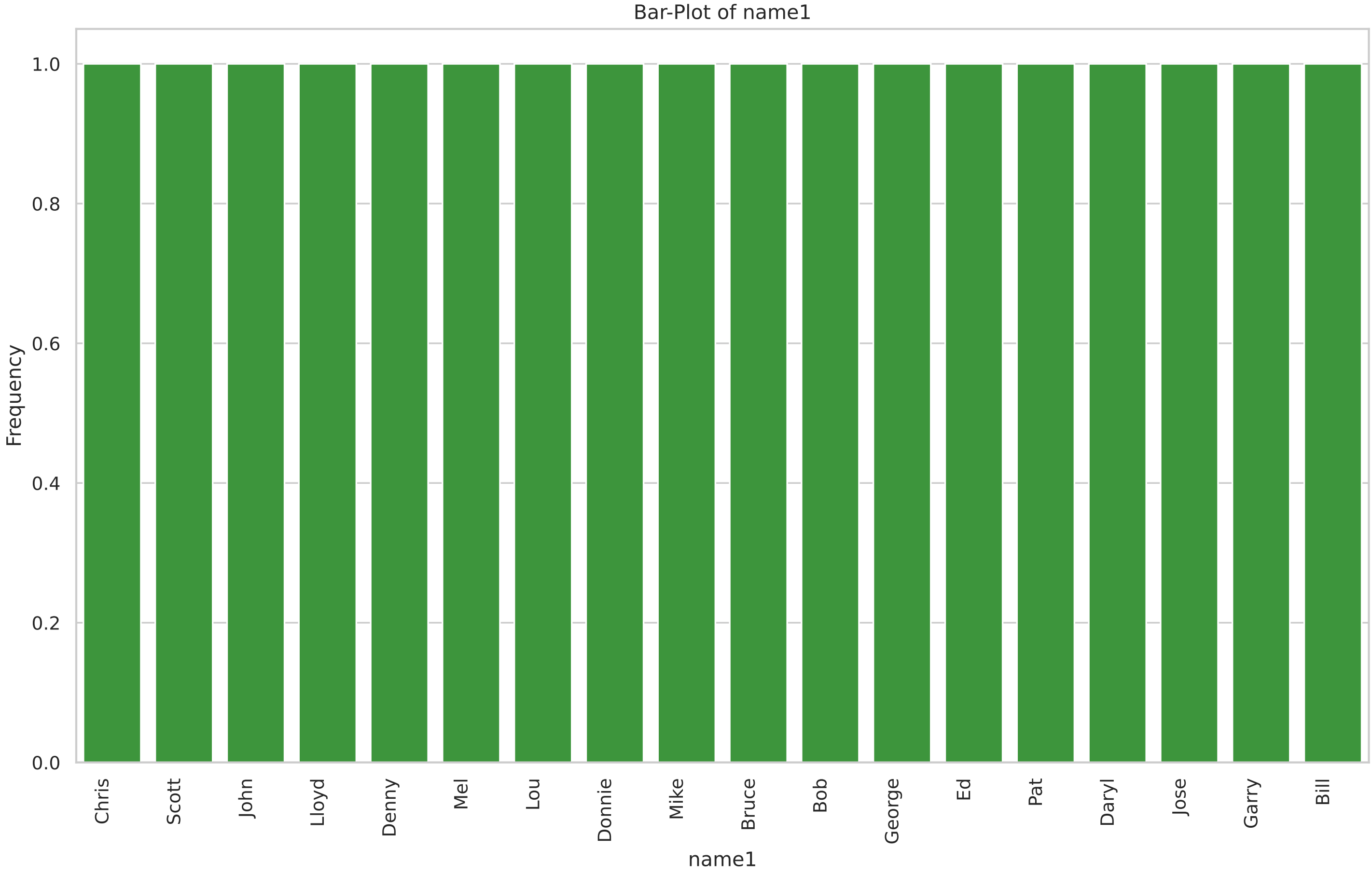
Bar-Plots

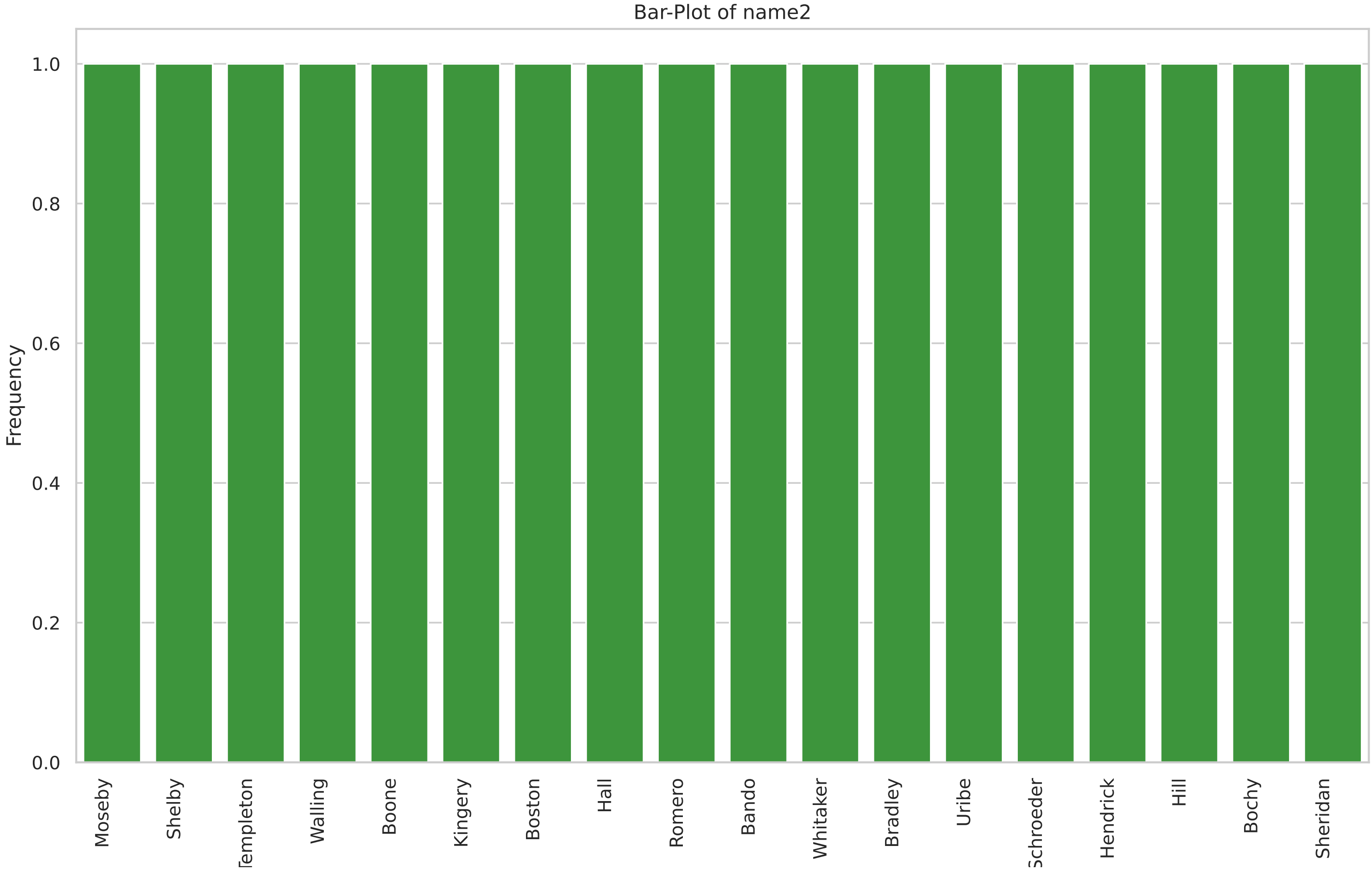
One Bar-Plot per page for each variable. Variables are sorted alphabetically. No labels for variables with more than 40 categories.

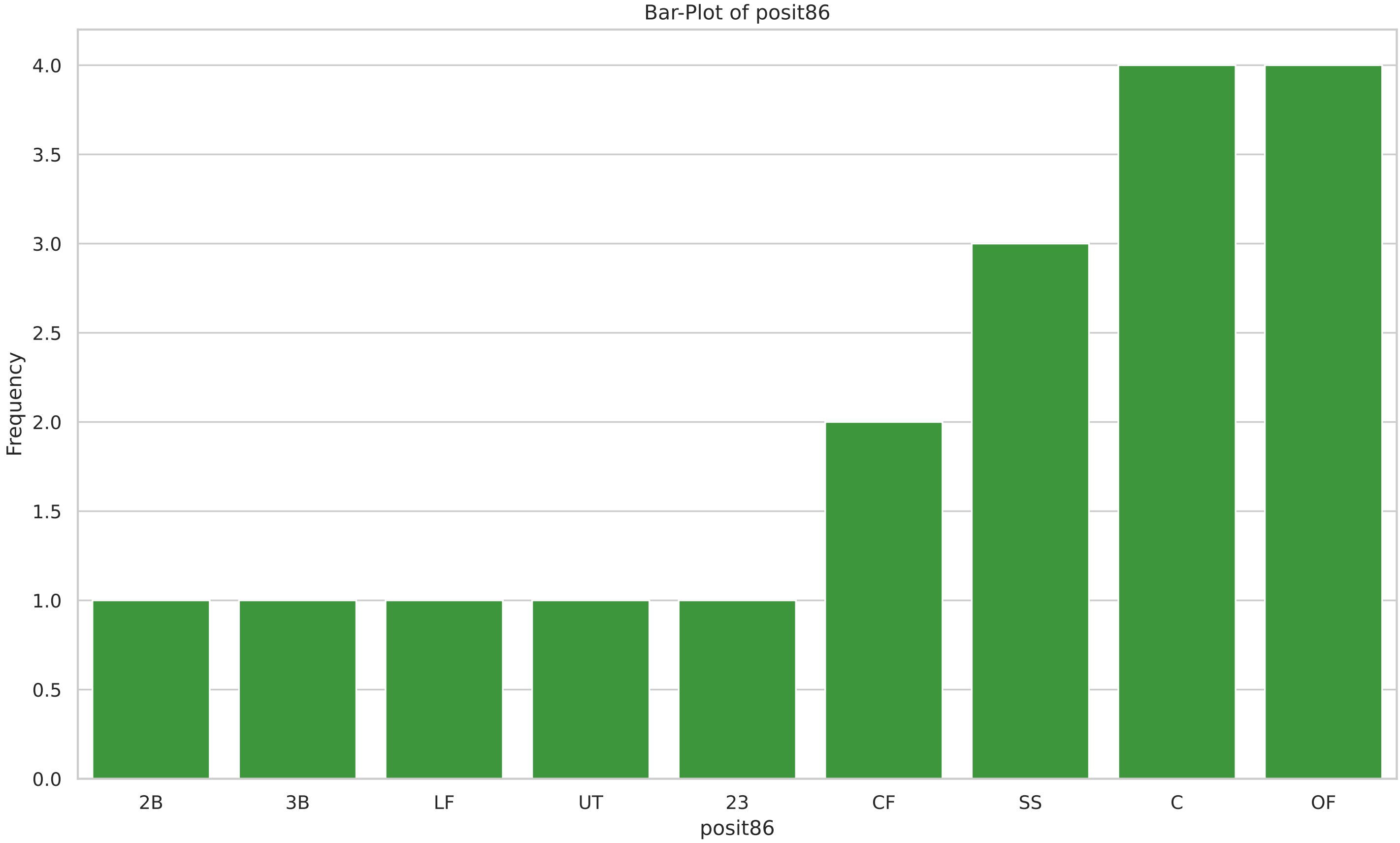
```
[]
```

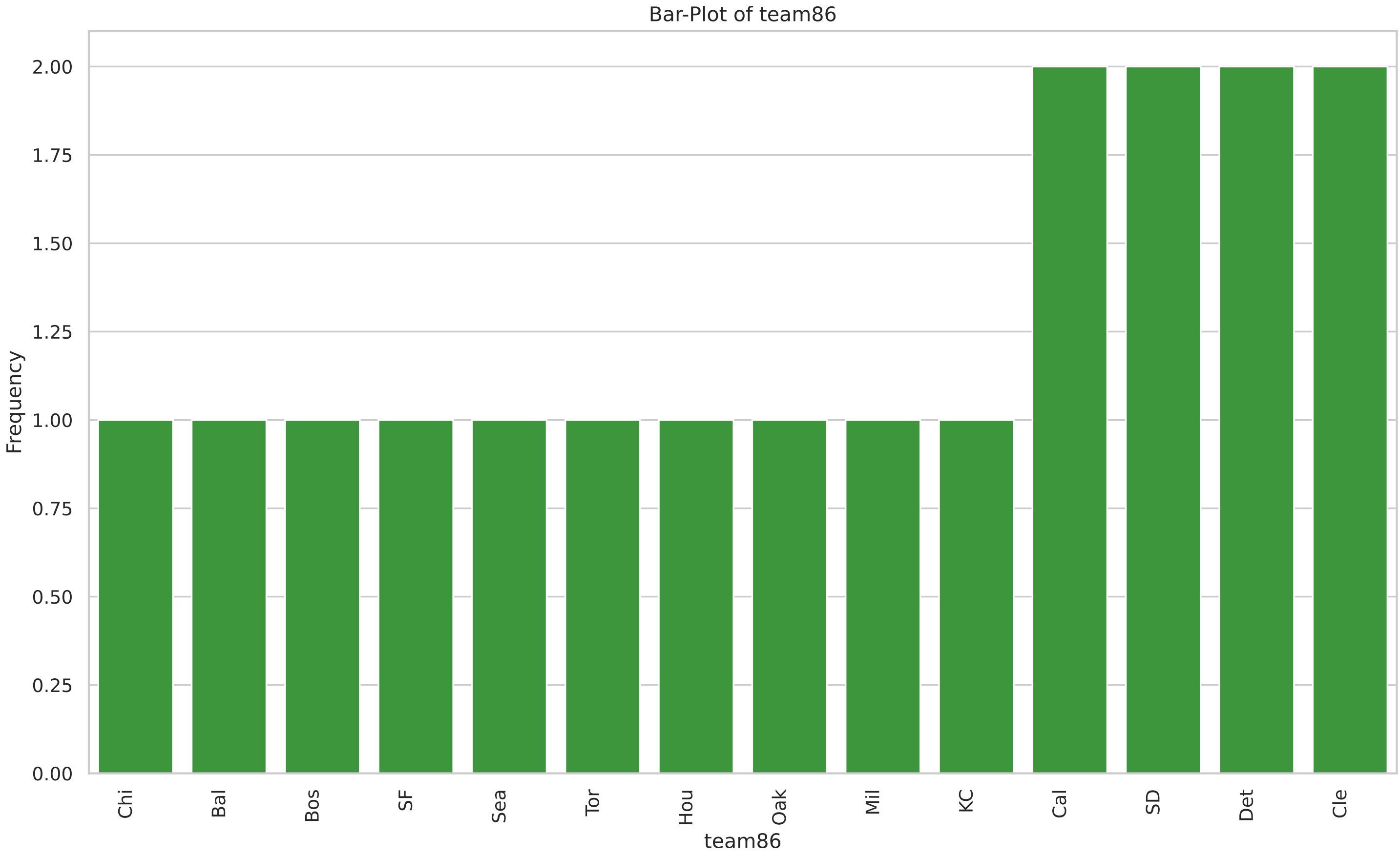


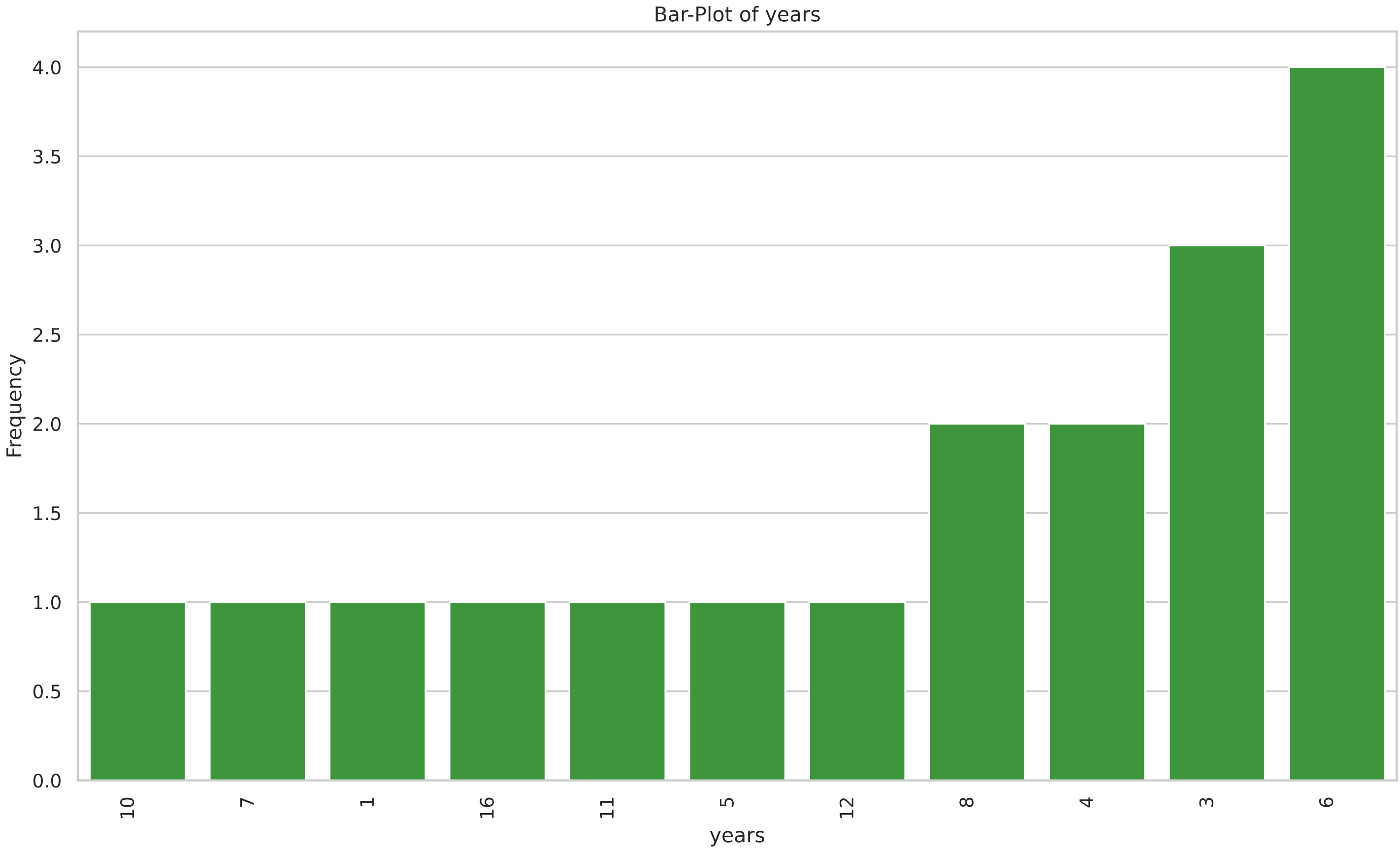

Bar-Plot of league86
Frequency
0
2
4
6
8
10
12
14
N
A
league86


Bar-Plot of name1
Frequency
1.0
0.8
0.6
0.4
0.2
0.0
Chris
Scott
John
Lloyd
Denny
Mel
Lou
Donnie
Mike
Bruce
Bob
George
Ed
Pat
Daryl
Jose
Garry
Bill
name1


Bar-Plot of name2
Frequency
1.0
0.8
0.6
0.4
0.2
0.0
Moseby
Shelby
Templeton
Walling
Boone
Kingery
Boston
Hall
Romero
Bando
Whitaker
Bradley
Uribe
Schroeder
Hendrick
Hill
Bochy
Sheridan

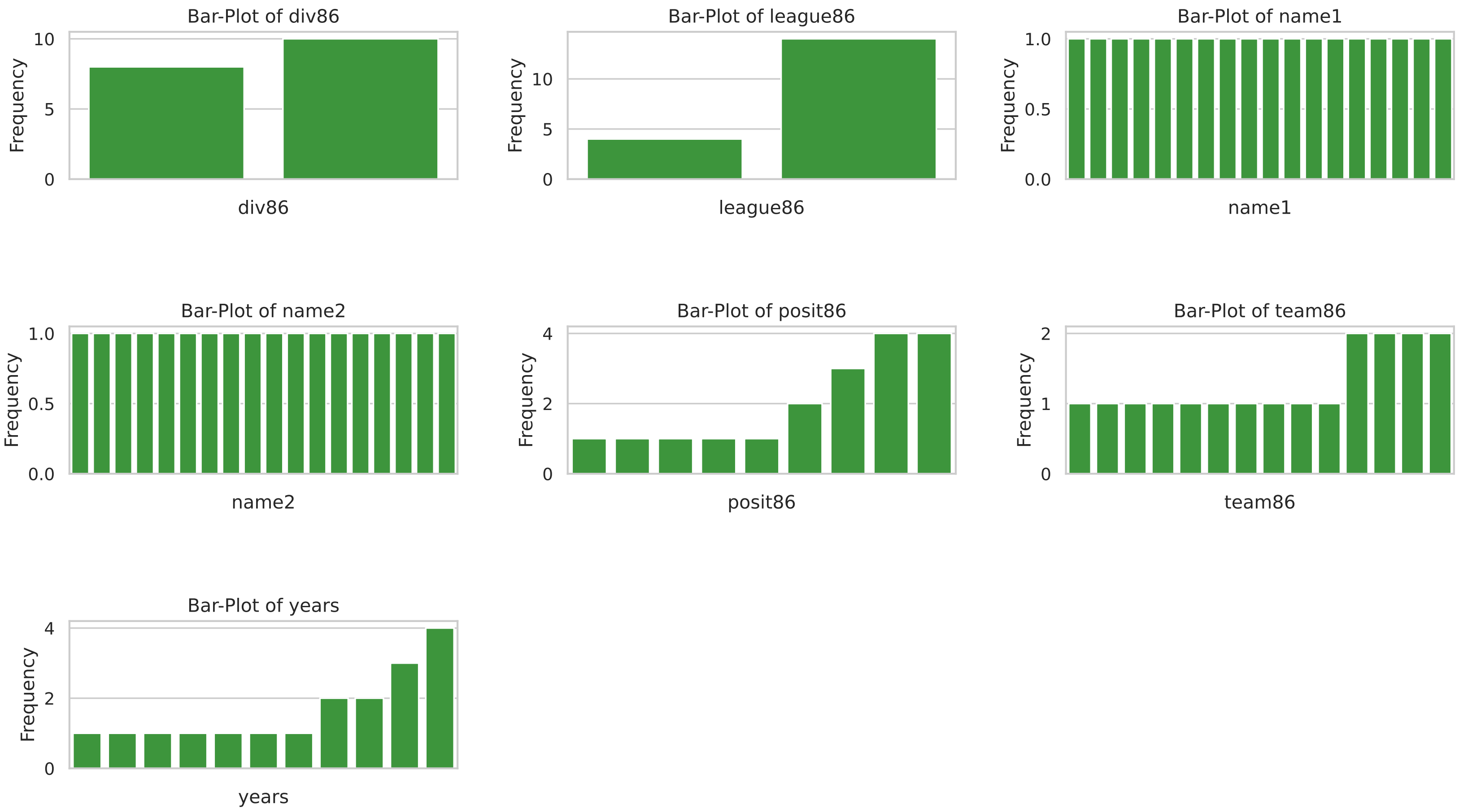

Bar-Plot of posit86
Frequency
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
2B
3B
LF
UT
23
CF
SS
C
OF
posit86


Bar-Plot of team86
Frequency
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Chi
Bal
Bos
SF
Sea
Tor
Hou
Oak
Mil
KC
Cal
SD
Det
Cle
team86


Bar-Plot of years
Frequency
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
10
7
1
16
11
5
12
8
4
3
6
years

Bar-Plots Summary

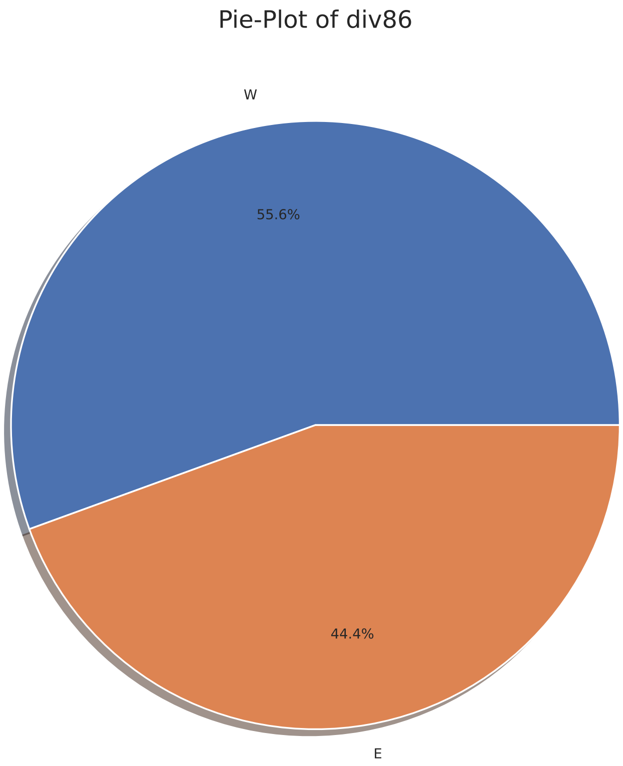
Multiple Bar-Plots of variables in one figure. Variables are sorted alphabetically. No labels displayed.

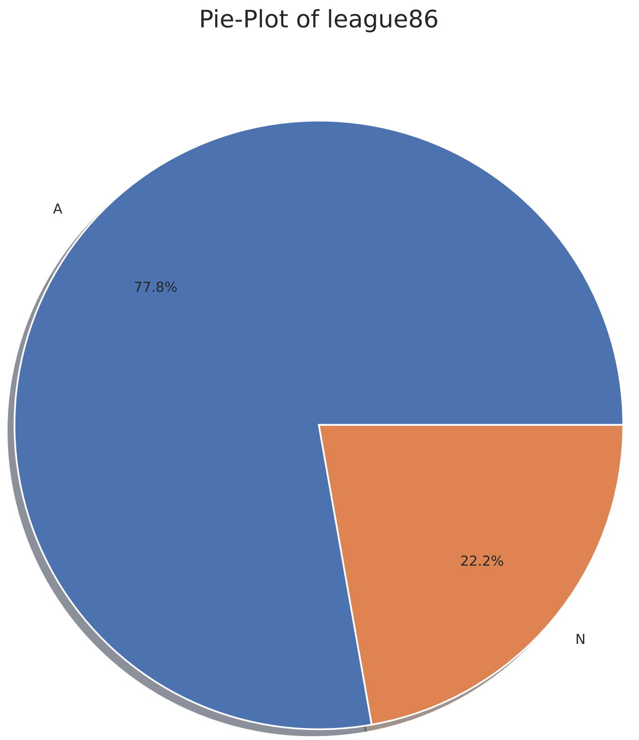


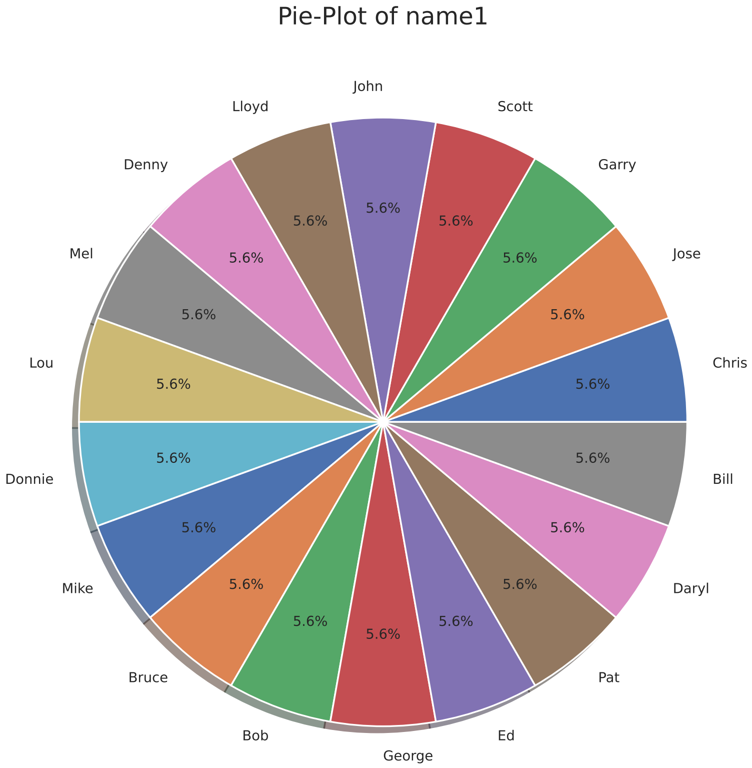
Pie Plots

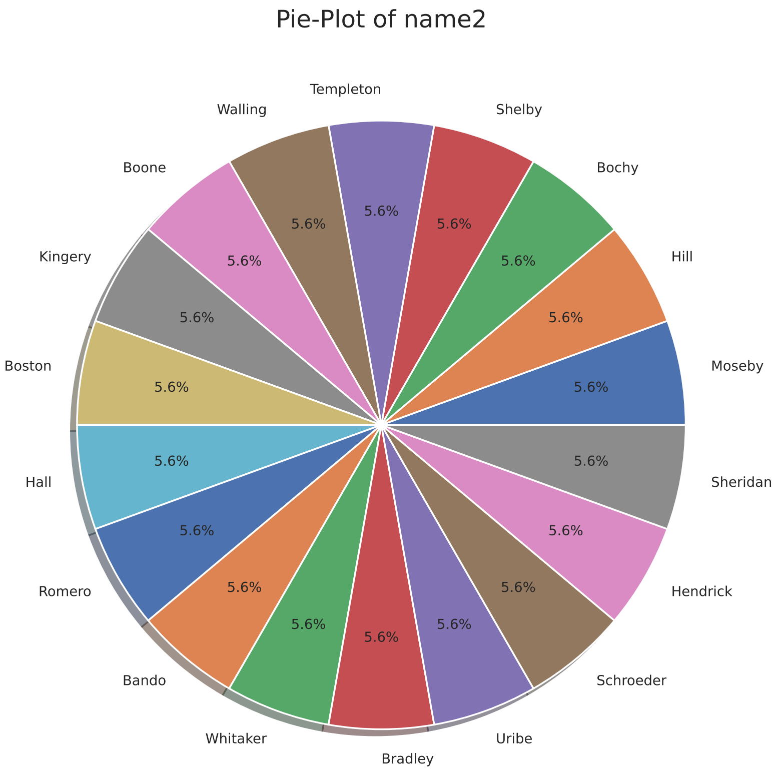
One Pie Plot per page for each variable. Variables are sorted alphabetically.

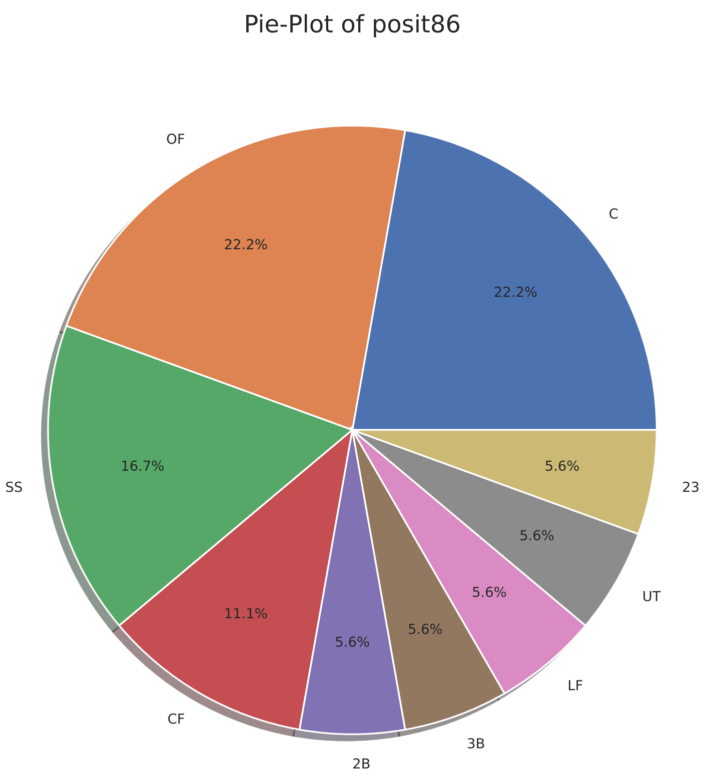
```
[]
```

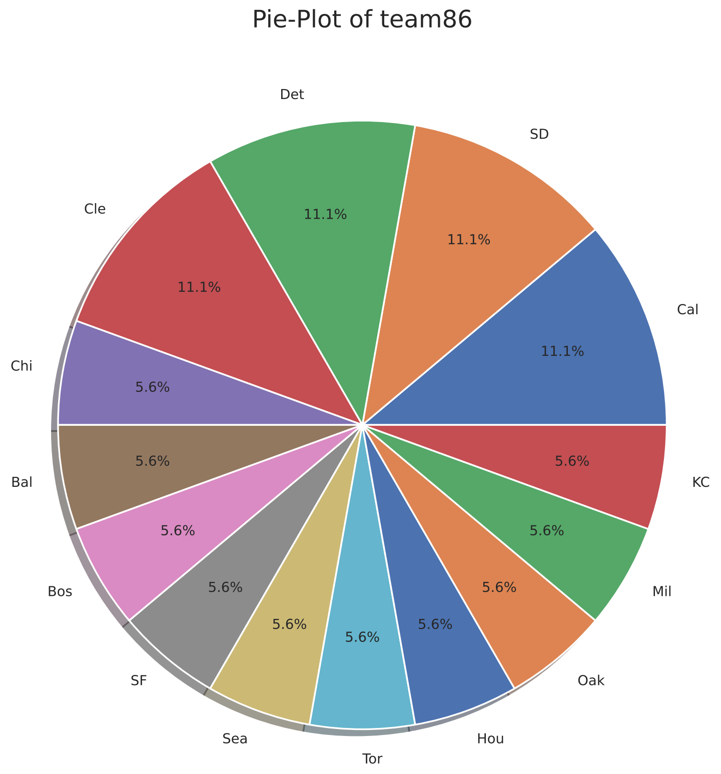


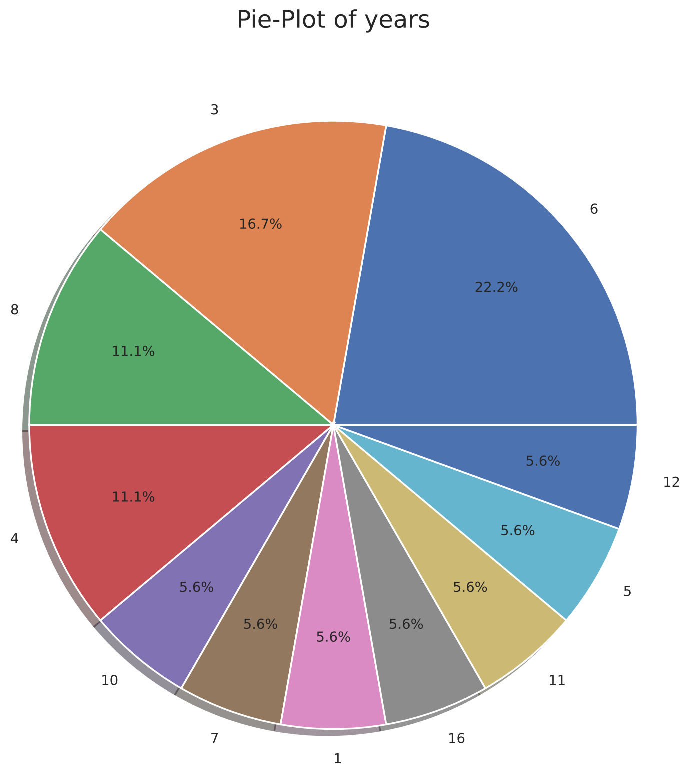

Pie-Plot of league86
A
77.8%
22.2%
N


Pie-Plot of name1
John
Scott
Garry
Jose
Chris
Bill
Daryl
Pat
Ed
George
Bob
Bruce
Mike
Donnie
Lou
Mel
Denny
Lloyd
5.6%


Pie-Plot of name2
Templeton
Walling
Shelby
Boone
Bochy
Kingery
Hill
Boston
Moseby
Hall
Sheridan
Romero
Hendrick
Bando
Schroeder
Whitaker
Uribe
Bradley
5.6%
5.6%
5.6%
5.6%
5.6%
5.6%
5.6%
5.6%
5.6%
5.6%
5.6%
5.6%
5.6%
5.6%
5.6%
5.6%
5.6%
5.6%


Pie-Plot of posit86
OF
22.2%
C
22.2%
SS
16.7%
5.6%
23
5.6%
UT
11.1%
5.6%
5.6%
5.6%
LF
CF
2B
3B

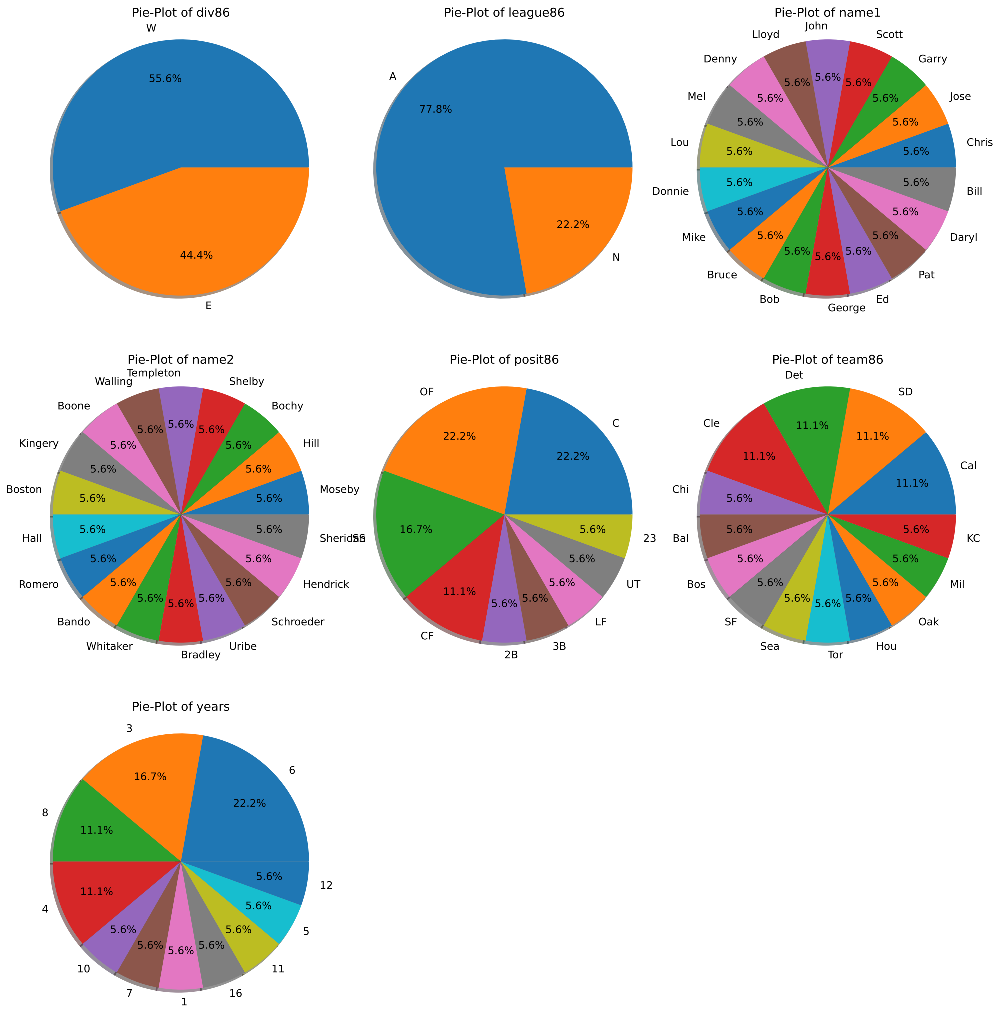

Pie-Plot of team86
Det
11.1%
SD
11.1%
Cle
11.1%
Cal
11.1%
Chi
5.6%
Bal
5.6%
KC
5.6%
Bos
5.6%
Mil
5.6%
SF
5.6%
Oak
5.6%
Sea
5.6%
Tor
5.6%
Hou
5.6%


Pie-Plot of years
3
16.7%
6
22.2%
8
11.1%
11.1%
4
5.6%
12
5.6%
5
5.6%
11
5.6%
16
5.6%
1
5.6%
7
5.6%
10

Pie Plots Summary

Multiple Pie Plots of variables in one figure. Variables are sorted alphabetically.

See figures on next page.


Pie-Plot of div86

Pie-Plot of league86

Pie-Plot of name1

Pie-Plot of name2

Pie-Plot of posit86

Pie-Plot of team86

Pie-Plot of years