Overview

Dataset statistics

Number of variables	8
Number of observations	1032
Missing cells	31
Missing cells (%)	0.4%
Duplicate rows	256
Duplicate rows (%)	24.8%
Total size in memory	104.9 KiB
Average record size in memory	104.0 B

Variable types

Categorical	4
Numeric	4

Alerts

Dataset has 256 (24.8%) duplicate rows	Duplicates
`sex` has 23 (2.2%) missing values	Missing
`species` is uniformly distributed	Uniform

Reproduction

Analysis started	2023-08-28 18:34:35.503975
Analysis finished	2023-08-28 18:34:44.889878
Duration	9.39 seconds
Software version	pandas-profiling v3.6.6 (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%

Variables

Select Columns

species
Categorical

Distinct	3
Distinct (%)	0.3%
Missing	0
Missing (%)	0.0%
Memory size	16.1 KiB

Length

Max length	9
Median length	6
Mean length	7
Min length	6

Characters and Unicode

Total characters	7224	
Distinct characters	15	
Distinct categories	2 (https://en.wikipedia.org/wiki/Unicode_character_property#General_Category)	?
Distinct scripts	1 (https://en.wikipedia.org/wiki/Script_(Unicode)#List_of_scripts_in_Unicode)	?
Distinct blocks	1 (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	0	?
Unique (%)	0.0%	

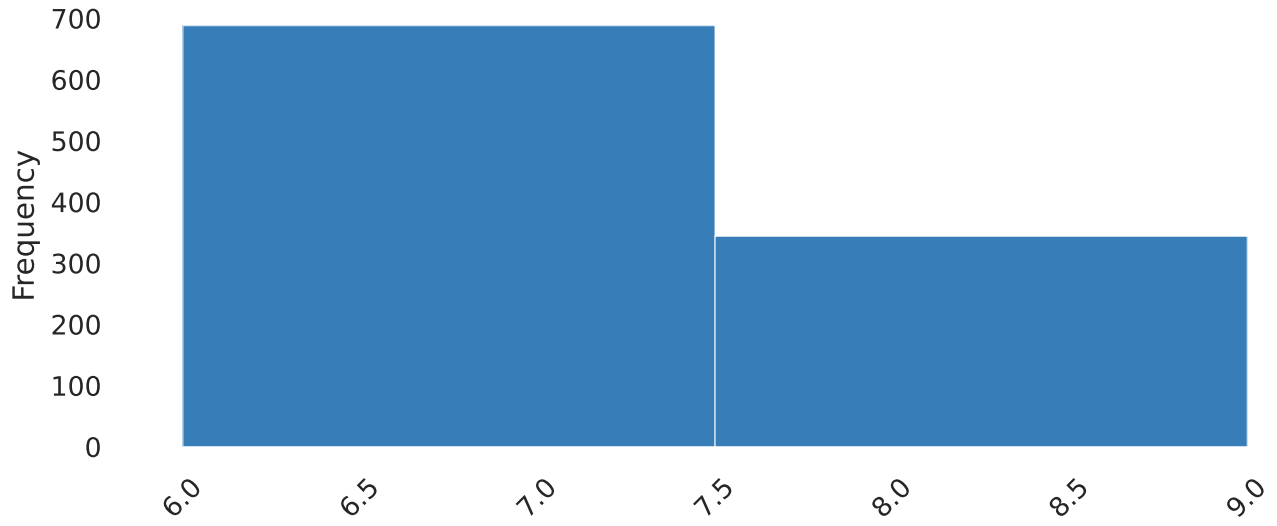
Sample

1st row	Adelie
2nd row	Adelie
3rd row	Adelie
4th row	Adelie
5th row	Adelie

Common Values

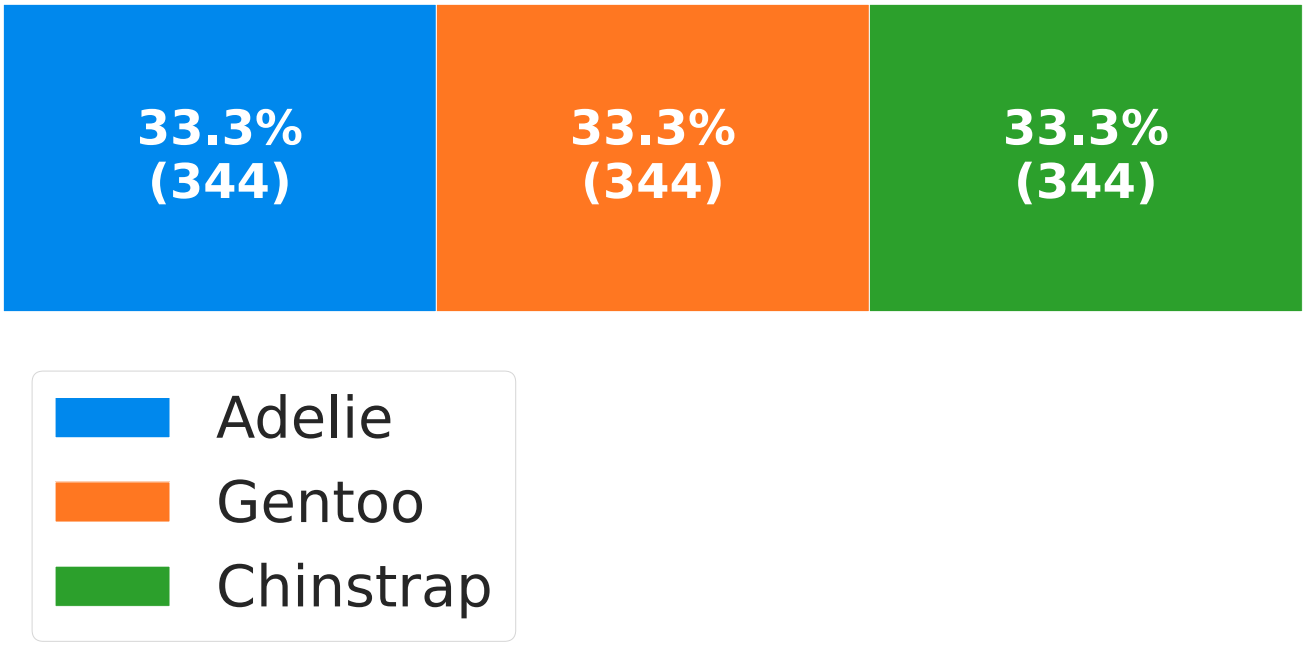
Value	Count	Frequency (%)
Adelie	344	33.3%
Gentoo	344	33.3%
Chinstrap	344	33.3%

Length



Histogram of lengths of the category

Common Values (Plot)



Value	Count	Frequency (%)
adelie	344	33.3%
gentoo	344	33.3%
chinstrap	344	33.3%

Most occurring characters

Value	Count	Frequency (%)
e	1032	14.3%
i	688	9.5%
n	688	9.5%
t	688	9.5%
o	688	9.5%
A	344	4.8%
d	344	4.8%
l	344	4.8%
G	344	4.8%
C	344	4.8%
Other values (5)	1720	23.8%

Most occurring categories

Value	Count	Frequency (%)
Lowercase Letter	6192	85.7%
Uppercase Letter	1032	14.3%

Most frequent character per category

Lowercase Letter

Value	Count	Frequency (%)
e	1032	16.7%
i	688	11.1%
n	688	11.1%
t	688	11.1%
o	688	11.1%
d	344	5.6%
l	344	5.6%
h	344	5.6%
s	344	5.6%
r	344	5.6%
Other values (2)	688	11.1%

Uppercase Letter

Value	Count	Frequency (%)
A	344	33.3%
G	344	33.3%
C	344	33.3%

Most occurring scripts

Value	Count	Frequency (%)
Latin	7224	100.0%

Most frequent character per script

Latin

Value	Count	Frequency (%)
e	1032	14.3%
i	688	9.5%
n	688	9.5%
t	688	9.5%
o	688	9.5%
A	344	4.8%
d	344	4.8%
l	344	4.8%
G	344	4.8%
C	344	4.8%
Other values (5)	1720	23.8%

Most occurring blocks

Value	Count	Frequency (%)
ASCII	7224	100.0%

Most frequent character per block

ASCII

Value	Count	Frequency (%)
e	1032	14.3%
i	688	9.5%
n	688	9.5%
t	688	9.5%
o	688	9.5%
A	344	4.8%
d	344	4.8%
l	344	4.8%
G	344	4.8%
C	344	4.8%
Other values (5)	1720	23.8%

island
Categorical

Distinct	3
Distinct (%)	0.3%
Missing	0
Missing (%)	0.0%
Memory size	16.1 KiB

Length

Max length	9
Median length	6
Mean length	5.8343023
Min length	5

Characters and Unicode

Total characters	6021	
Distinct characters	13	
Distinct categories	2 (https://en.wikipedia.org/wiki/Unicode_character_property#General_Category)	?
Distinct scripts	1 (https://en.wikipedia.org/wiki/Script_(Unicode)#List_of_scripts_in_Unicode)	?
Distinct blocks	1 (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	0	?
Unique (%)	0.0%	

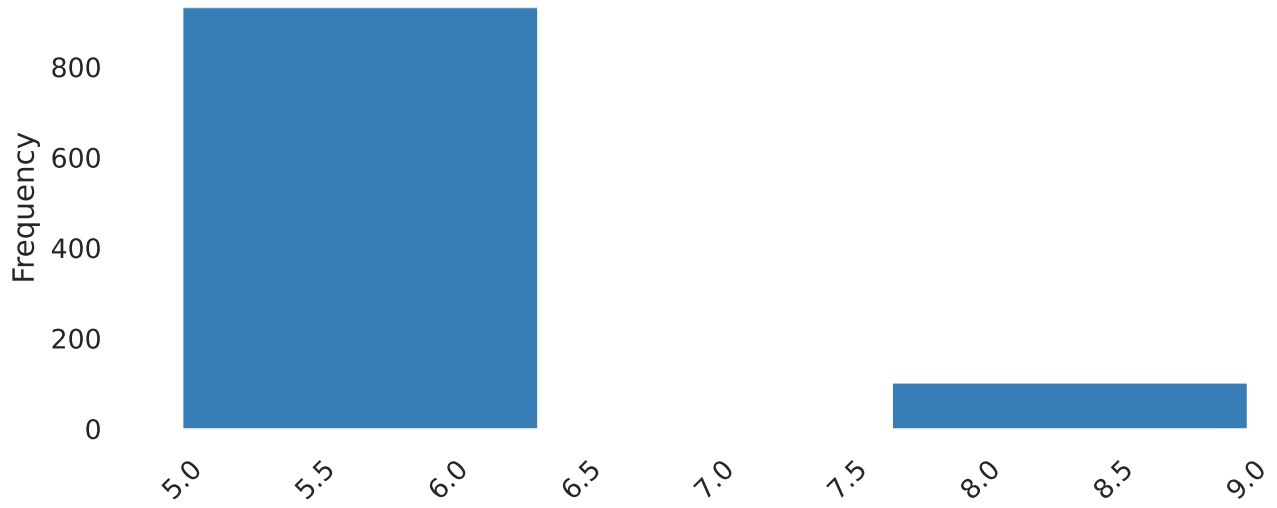
Sample

1st row	Biscoe
2nd row	Dream
3rd row	Torgersen
4th row	Biscoe
5th row	Torgersen

Common Values

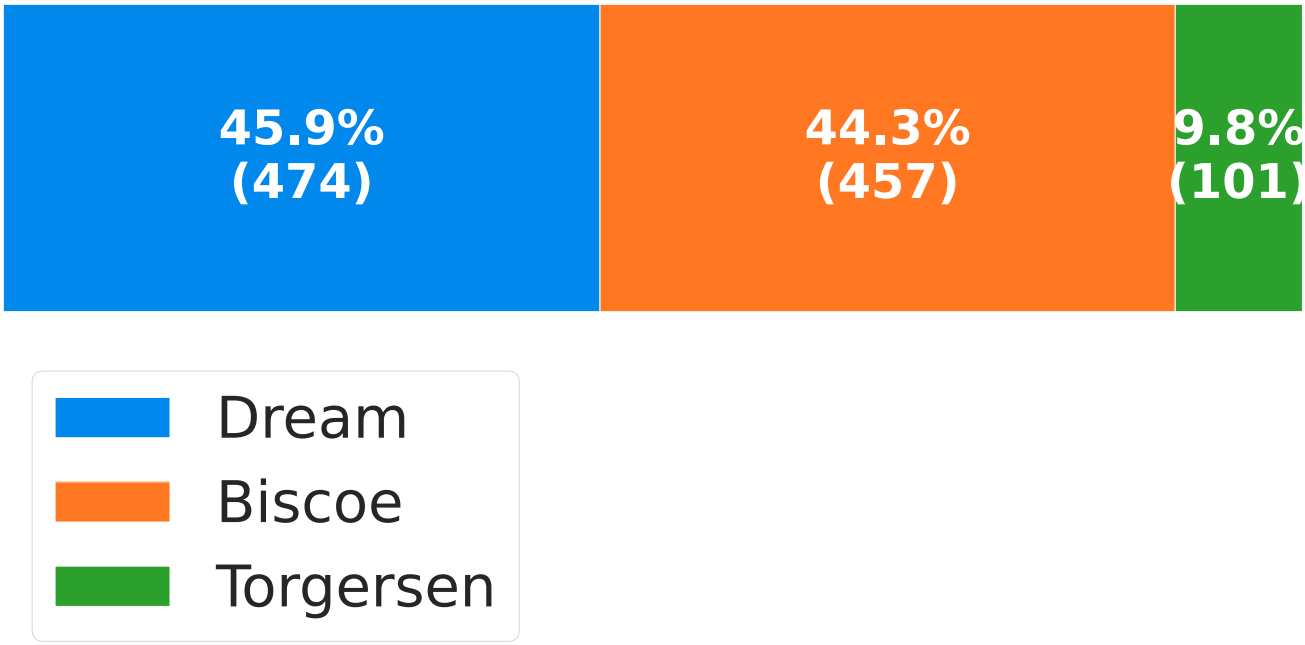
Value	Count	Frequency (%)
Dream	474	45.9%
Biscoe	457	44.3%
Torgersen	101	9.8%

Length



Histogram of lengths of the category

Common Values (Plot)



Value	Count	Frequency (%)
dream	474	45.9%
biscoe	457	44.3%
torgersen	101	9.8%

Most occurring characters

Value	Count	Frequency (%)
e	1133	18.8%
r	676	11.2%
s	558	9.3%
o	558	9.3%
D	474	7.9%
a	474	7.9%
m	474	7.9%
B	457	7.6%
i	457	7.6%
c	457	7.6%
Other values (3)	303	5.0%

Most occurring categories

Value	Count	Frequency (%)
Lowercase Letter	4989	82.9%
Uppercase Letter	1032	17.1%

Most frequent character per category

Lowercase Letter

Value	Count	Frequency (%)
e	1133	22.7%
r	676	13.5%
s	558	11.2%
o	558	11.2%
a	474	9.5%
m	474	9.5%
i	457	9.2%
c	457	9.2%
g	101	2.0%
n	101	2.0%

Uppercase Letter

Value	Count	Frequency (%)
D	474	45.9%
B	457	44.3%
T	101	9.8%

Most occurring scripts

Value	Count	Frequency (%)
Latin	6021	100.0%

Most frequent character per script

Latin

Value	Count	Frequency (%)
e	1133	18.8%
r	676	11.2%
s	558	9.3%
o	558	9.3%
D	474	7.9%
a	474	7.9%
m	474	7.9%
B	457	7.6%
i	457	7.6%
c	457	7.6%
Other values (3)	303	5.0%

Most occurring blocks

Value	Count	Frequency (%)
ASCII	6021	100.0%

Most frequent character per block

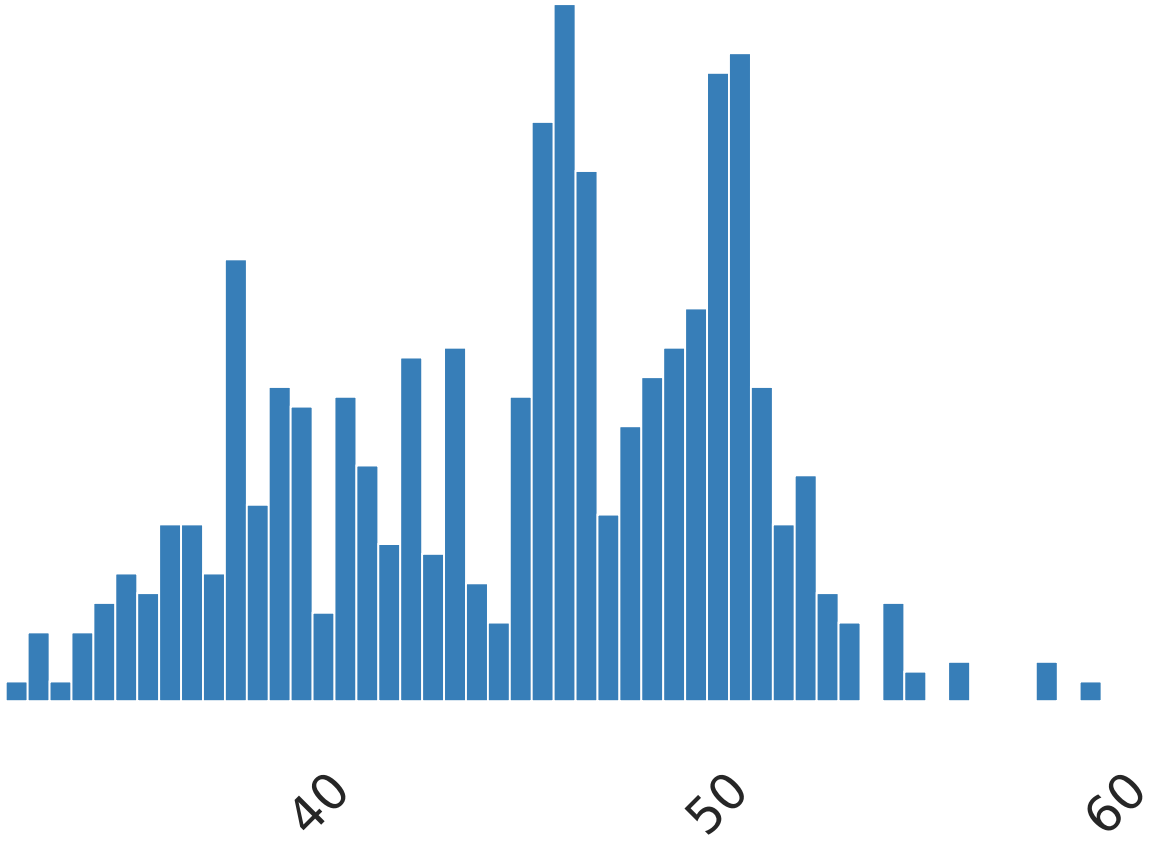
ASCII

Value	Count	Frequency (%)
e	1133	18.8%
r	676	11.2%
s	558	9.3%
o	558	9.3%
D	474	7.9%
a	474	7.9%
m	474	7.9%
B	457	7.6%
i	457	7.6%
c	457	7.6%
Other values (3)	303	5.0%

bill_length_mm
Real number (ℝ)

Distinct	158
Distinct (%)	15.3%
Missing	2
Missing (%)	0.2%
Infinite	0
Infinite (%)	0.0%
Mean	45.036505

Minimum	32.1
Maximum	59.6
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	16.1 KiB



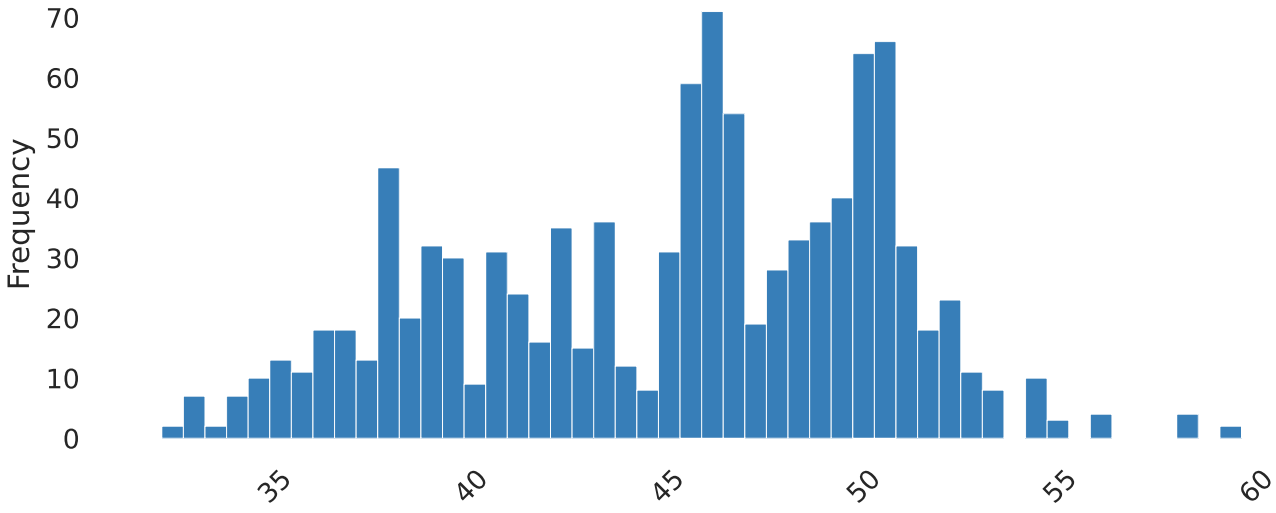
Quantile statistics

Minimum	32.1
5-th percentile	35.945
Q1	40.7
median	46

Q3	49.575
95-th percentile	52.1
Maximum	59.6
Range	27.5
Interquartile range (IQR)	8.875

Descriptive statistics

Standard deviation	5.3264862
Coefficient of variation (CV)	0.11827042
Kurtosis	-0.73306213
Mean	45.036505
Median Absolute Deviation (MAD)	4
Skewness	-0.28596892
Sum	46387.6
Variance	28.371455
Monotonicity	Not monotonic



Histogram with fixed size bins (bins=50)

Value	Count	Frequency (%)
50	26	2.5%
46.2	24	2.3%
46.8	21	2.0%
51.3	20	1.9%
37.8	17	1.6%
46.4	17	1.6%
42.5	16	1.6%
45.5	16	1.6%
45.7	16	1.6%
50.5	16	1.6%
Other values (148)	841	81.5%

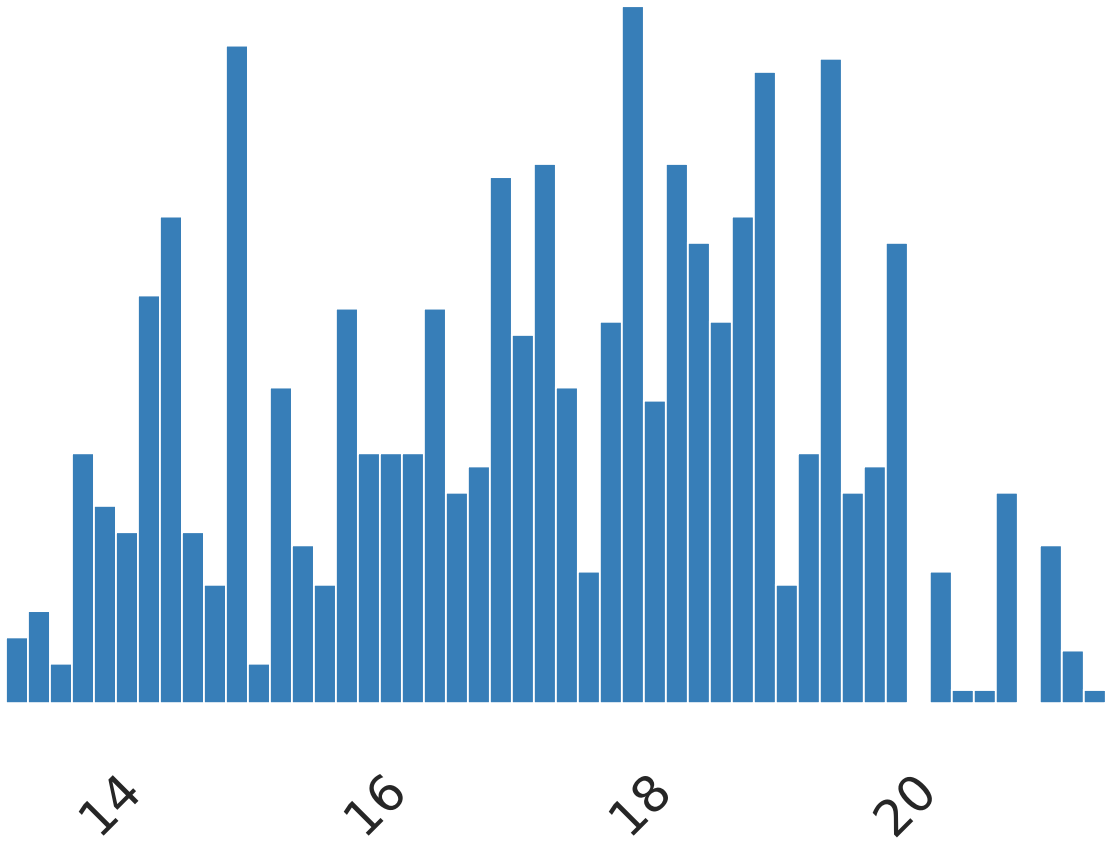
Value	Count	Frequency (%)
32.1	2	0.2%
33.1	7	0.7%
33.5	2	0.2%
34	4	0.4%
34.1	3	0.3%
34.5	1	0.1%
34.6	9	0.9%
35	8	0.8%
35.1	3	0.3%
35.2	1	0.1%

Value	Count	Frequency (%)
59.6	2	0.2%
58	4	0.4%
55.8	4	0.4%
55.1	3	0.3%
54.3	2	0.2%
54.2	8	0.8%
53.5	3	0.3%
53.4	5	0.5%
52.8	6	0.6%
52.7	3	0.3%

bill_depth_mm
Real number (ℝ)

Distinct	77
Distinct (%)	7.5%
Missing	2
Missing (%)	0.2%
Infinite	0
Infinite (%)	0.0%
Mean	17.24932

Minimum	13.2
Maximum	21.5
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	16.1 KiB



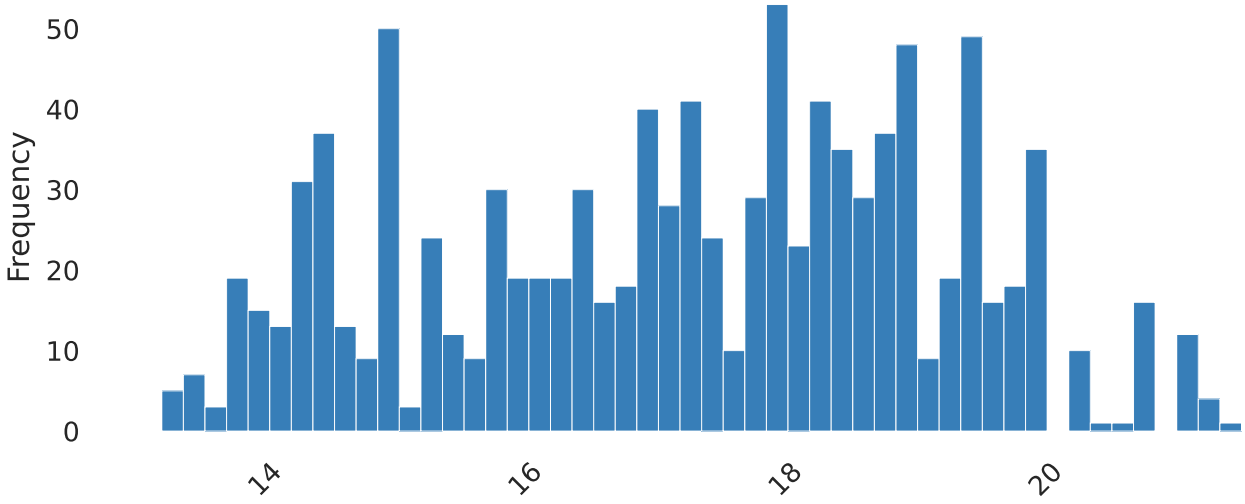
Quantile statistics

Minimum	13.2
5-th percentile	14.1
Q1	15.7
median	17.5

Q3	18.8
95-th percentile	20
Maximum	21.5
Range	8.3
Interquartile range (IQR)	3.1

Descriptive statistics

Standard deviation	1.9369595
Coefficient of variation (CV)	0.11229193
Kurtosis	-0.91090306
Mean	17.24932
Median Absolute Deviation (MAD)	1.4
Skewness	-0.16506733
Sum	17766.8
Variance	3.751812
Monotonicity	Not monotonic



Histogram with fixed size bins (bins=50)

Value	Count	Frequency (%)
17.9	41	4.0%
15	38	3.7%
17	37	3.6%
17.3	30	2.9%
18.6	29	2.8%
17.1	28	2.7%
14.5	25	2.4%
19.5	25	2.4%
19.4	24	2.3%
20	24	2.3%
Other values (67)	729	70.6%

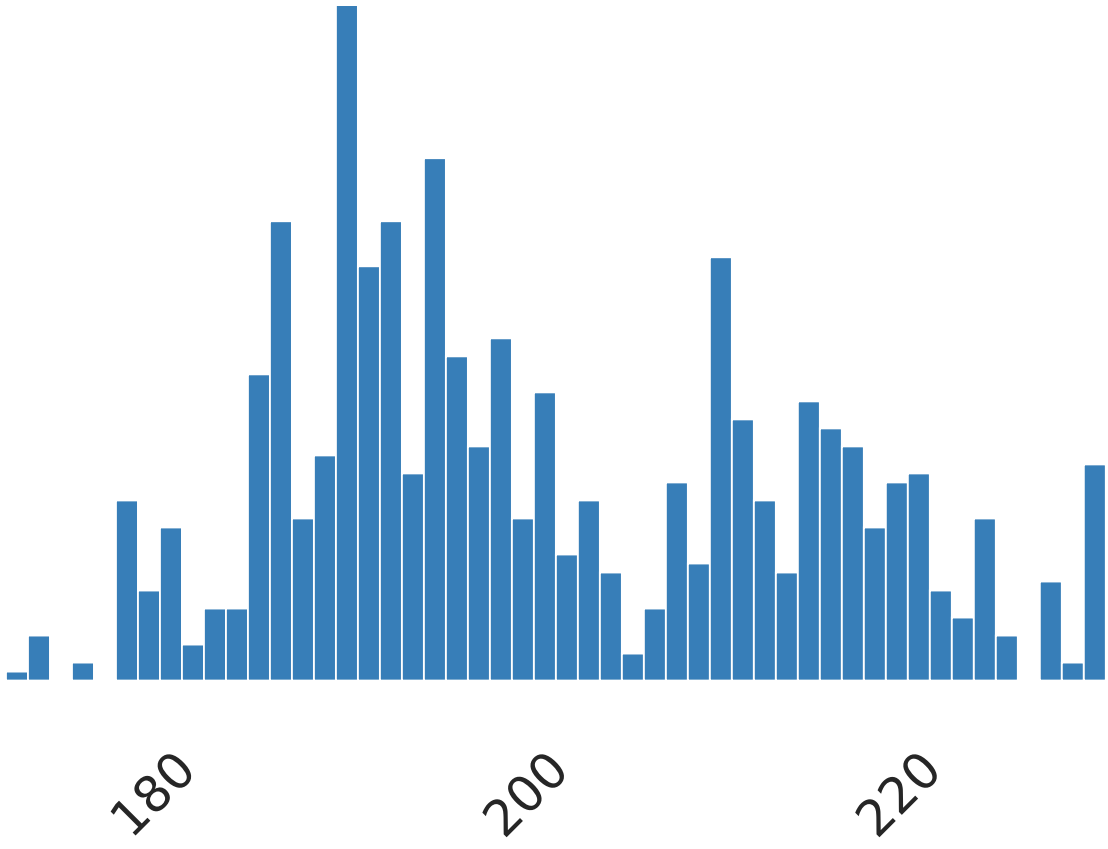
Value	Count	Frequency (%)
13.2	4	0.4%
13.3	1	0.1%
13.4	4	0.4%
13.5	3	0.3%
13.6	3	0.3%
13.7	12	1.2%
13.8	7	0.7%
13.9	7	0.7%
14	8	0.8%
14.1	13	1.3%

Value	Count	Frequency (%)
21.5	1	0.1%
21.2	4	0.4%
21.1	12	1.2%
20.8	8	0.8%
20.7	8	0.8%
20.6	1	0.1%
20.5	1	0.1%
20.3	10	1.0%
20	24	2.3%
19.9	11	1.1%

flipper_length_mm
Real number (ℝ)

Distinct	55
Distinct (%)	5.3%
Missing	2
Missing (%)	0.2%
Infinite	0
Infinite (%)	0.0%
Mean	201.05437

Minimum	172
Maximum	231
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	16.1 KiB



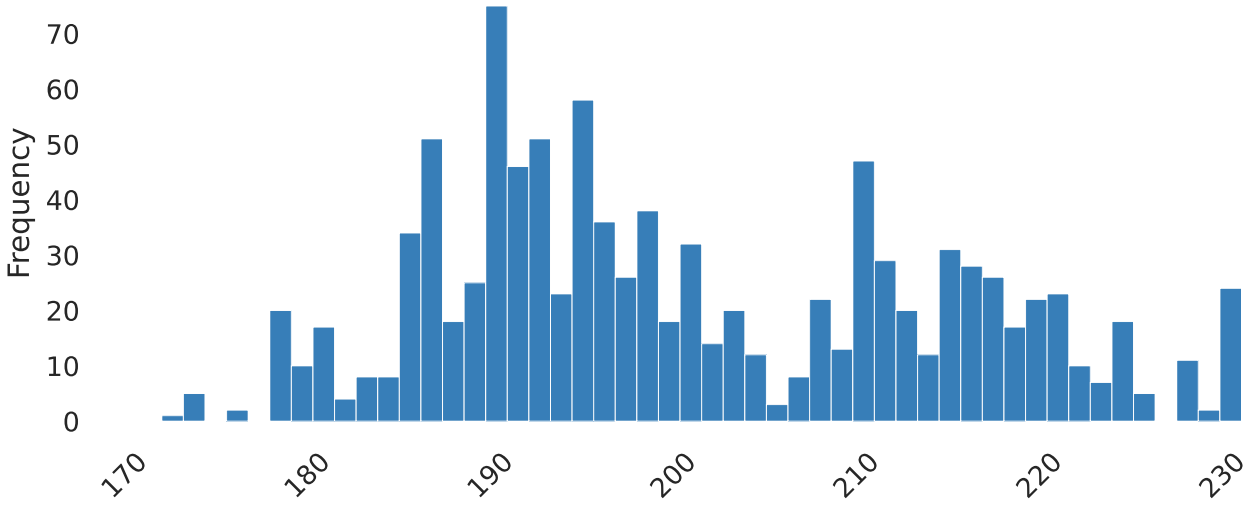
Quantile statistics

Minimum	172
5-th percentile	181
Q1	190
median	197

Q3	212
95-th percentile	224.55
Maximum	231
Range	59
Interquartile range (IQR)	22

Descriptive statistics

Standard deviation	13.50402
Coefficient of variation (CV)	0.067166011
Kurtosis	-0.87033254
Mean	201.05437
Median Absolute Deviation (MAD)	10
Skewness	0.35917001
Sum	207086
Variance	182.35856
Monotonicity	Not monotonic



Histogram with fixed size bins (bins=50)

Value	Count	Frequency (%)
190	75	7.3%
195	58	5.6%
187	51	4.9%
193	51	4.9%
210	47	4.6%
196	36	3.5%
201	32	3.1%
215	31	3.0%
216	28	2.7%
197	26	2.5%
Other values (45)	595	57.7%

Value	Count	Frequency (%)
172	1	0.1%
174	5	0.5%
176	2	0.2%
178	14	1.4%
179	6	0.6%
180	10	1.0%
181	17	1.6%
182	4	0.4%
183	8	0.8%
184	8	0.8%

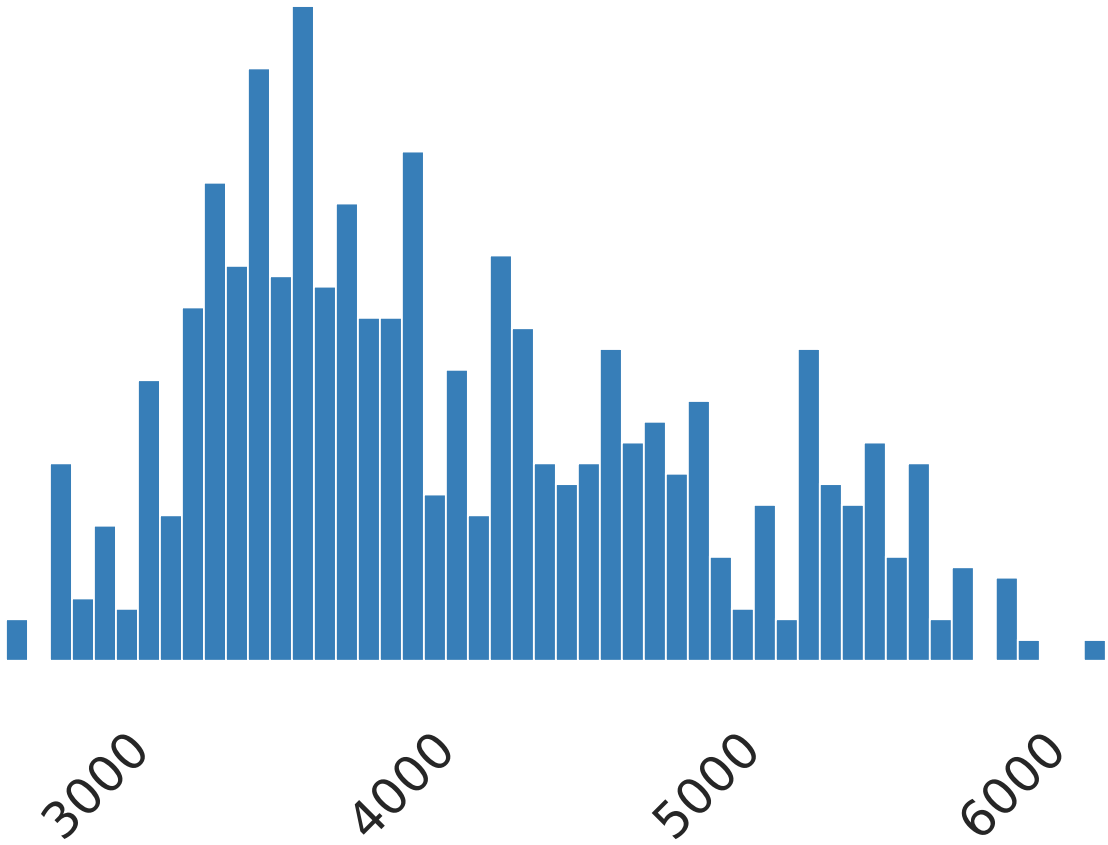
Value	Count	Frequency (%)
231	2	0.2%
230	22	2.1%
229	2	0.2%
228	11	1.1%
226	5	0.5%
225	10	1.0%
224	8	0.8%
223	7	0.7%
222	10	1.0%
221	23	2.2%

body_mass_g

Real number (ℝ)

Distinct	92
Distinct (%)	8.9%
Missing	2
Missing (%)	0.2%
Infinite	0
Infinite (%)	0.0%
Mean	4169.2233

Minimum	2700
Maximum	6300
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	16.1 KiB



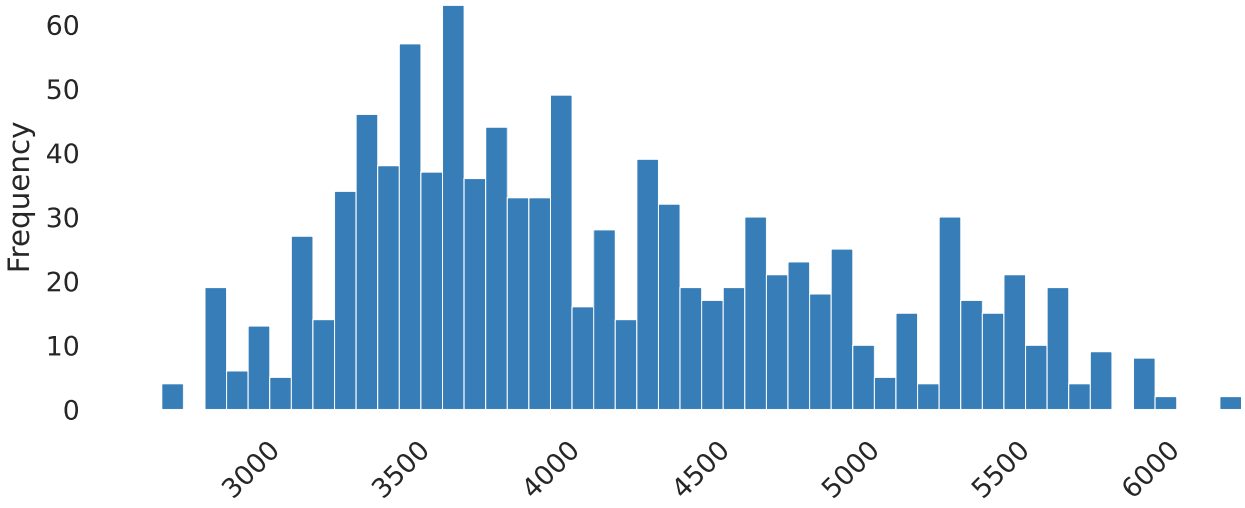
Quantile statistics

Minimum	2700
5-th percentile	3150
Q1	3550
median	4000

Q3	4718.75
95-th percentile	5600
Maximum	6300
Range	3600
Interquartile range (IQR)	1168.75

Descriptive statistics

Standard deviation	775.50221
Coefficient of variation (CV)	0.1860064
Kurtosis	-0.63226696
Mean	4169.2233
Median Absolute Deviation (MAD)	525
Skewness	0.53152124
Sum	4294300
Variance	601403.67
Monotonicity	Not monotonic



Histogram with fixed size bins (bins=50)

Value	Count	Frequency (%)
3650	37	3.6%
3800	37	3.6%
4300	34	3.3%
3950	33	3.2%
4050	33	3.2%
4400	30	2.9%
3900	29	2.8%
3400	29	2.8%
3600	28	2.7%
3500	26	2.5%
Other values (82)	714	69.2%

Value	Count	Frequency (%)
2700	4	0.4%
2850	4	0.4%
2900	15	1.5%
2975	6	0.6%
3000	4	0.4%
3050	9	0.9%
3075	4	0.4%
3100	1	0.1%
3150	6	0.6%
3175	4	0.4%

Value	Count	Frequency (%)
6300	2	0.2%
6050	2	0.2%
6000	4	0.4%
5950	4	0.4%
5850	7	0.7%
5800	2	0.2%
5750	4	0.4%
5700	19	1.8%
5650	7	0.7%
5600	3	0.3%

sex

Categorical

Distinct	2
Distinct (%)	0.2%
Missing	23
Missing (%)	2.2%
Memory size	16.1 KiB

Length

Max length	6
Median length	4
Mean length	4.9851338
Min length	4

Characters and Unicode

Total characters	5030	
Distinct characters	5	
Distinct categories	1 (https://en.wikipedia.org/wiki/Unicode_character_property#General_Category)	?
Distinct scripts	1 (https://en.wikipedia.org/wiki/Script_(Unicode)#List_of_scripts_in_Unicode)	?
Distinct blocks	1 (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	0	?
Unique (%)	0.0%	

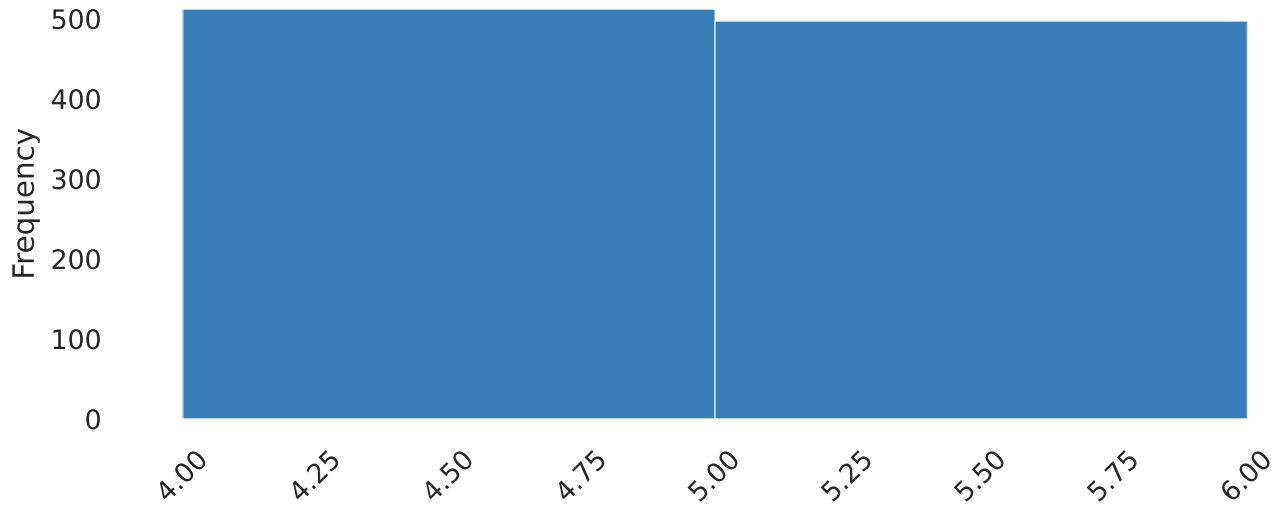
Sample

1st row	female
2nd row	female
3rd row	male
4th row	female
5th row	male

Common Values

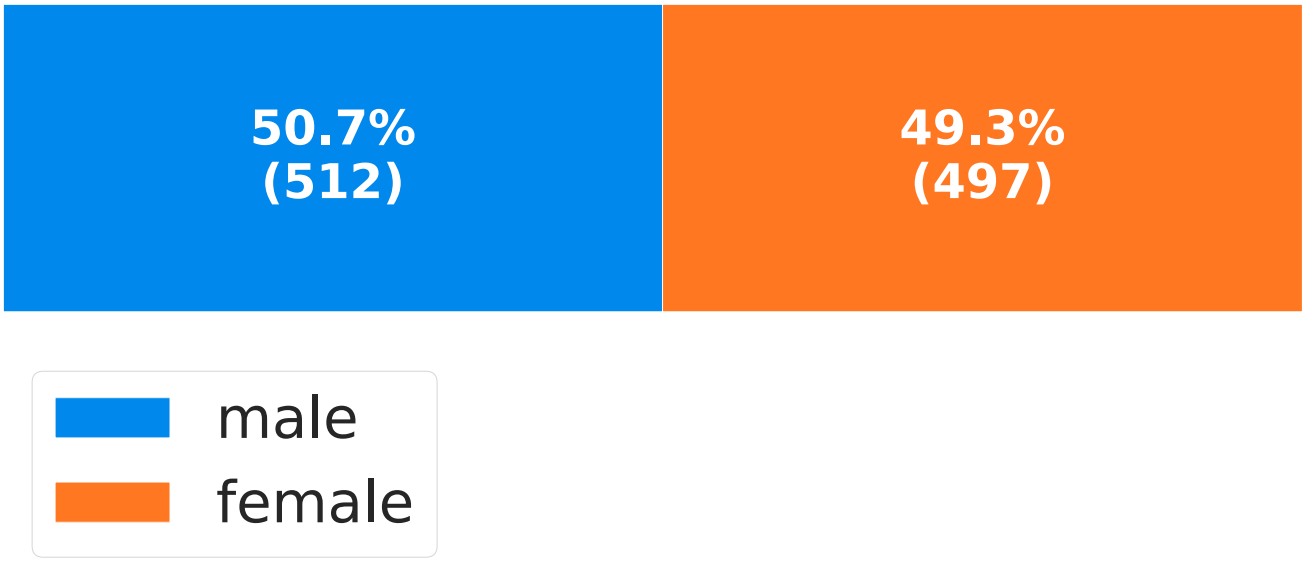
Value	Count	Frequency (%)
male	512	49.6%
female	497	48.2%
(Missing)	23	2.2%

Length



Histogram of lengths of the category

Common Values (Plot)



Value	Count	Frequency (%)
male	512	50.7%
female	497	49.3%

Most occurring characters

Value	Count	Frequency (%)
e	1506	29.9%
m	1009	20.1%
a	1009	20.1%
l	1009	20.1%
f	497	9.9%

Most occurring categories

Value	Count	Frequency (%)
Lowercase Letter	5030	100.0%

Most frequent character per category

Lowercase Letter

Value	Count	Frequency (%)
e	1506	29.9%
m	1009	20.1%
a	1009	20.1%
l	1009	20.1%
f	497	9.9%

Most occurring scripts

Value	Count	Frequency (%)
Latin	5030	100.0%

Most frequent character per script

Latin

Value	Count	Frequency (%)
e	1506	29.9%
m	1009	20.1%
a	1009	20.1%
l	1009	20.1%
f	497	9.9%

Most occurring blocks

Value	Count	Frequency (%)
ASCII	5030	100.0%

Most frequent character per block

ASCII

Value	Count	Frequency (%)
e	1506	29.9%
m	1009	20.1%
a	1009	20.1%
l	1009	20.1%
f	497	9.9%

year

Categorical

Distinct	3
Distinct (%)	0.3%
Missing	0
Missing (%)	0.0%
Memory size	16.1 KiB

Length

Max length	4
Median length	4
Mean length	4
Min length	4

Characters and Unicode

Total characters	4128	
Distinct characters	5	
Distinct categories	1 (https://en.wikipedia.org/wiki/Unicode_character_property#General_Category)	?
Distinct scripts	1 (https://en.wikipedia.org/wiki/Script_(Unicode)#List_of_scripts_in_Unicode)	?
Distinct blocks	1 (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	0	?
Unique (%)	0.0%	

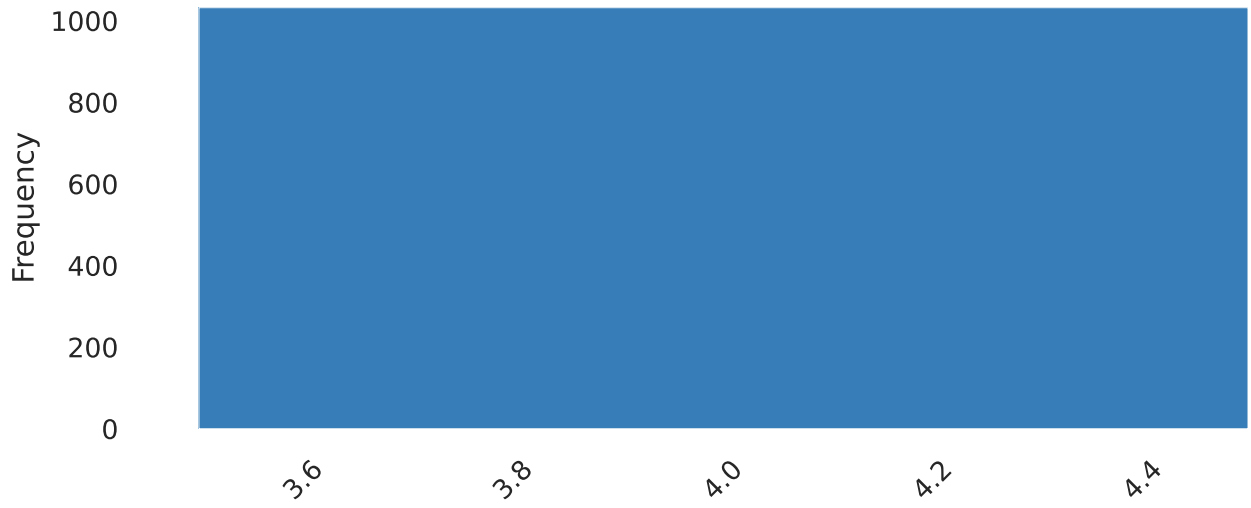
Sample

1st row	2009
2nd row	2008
3rd row	2007
4th row	2009
5th row	2008

Common Values

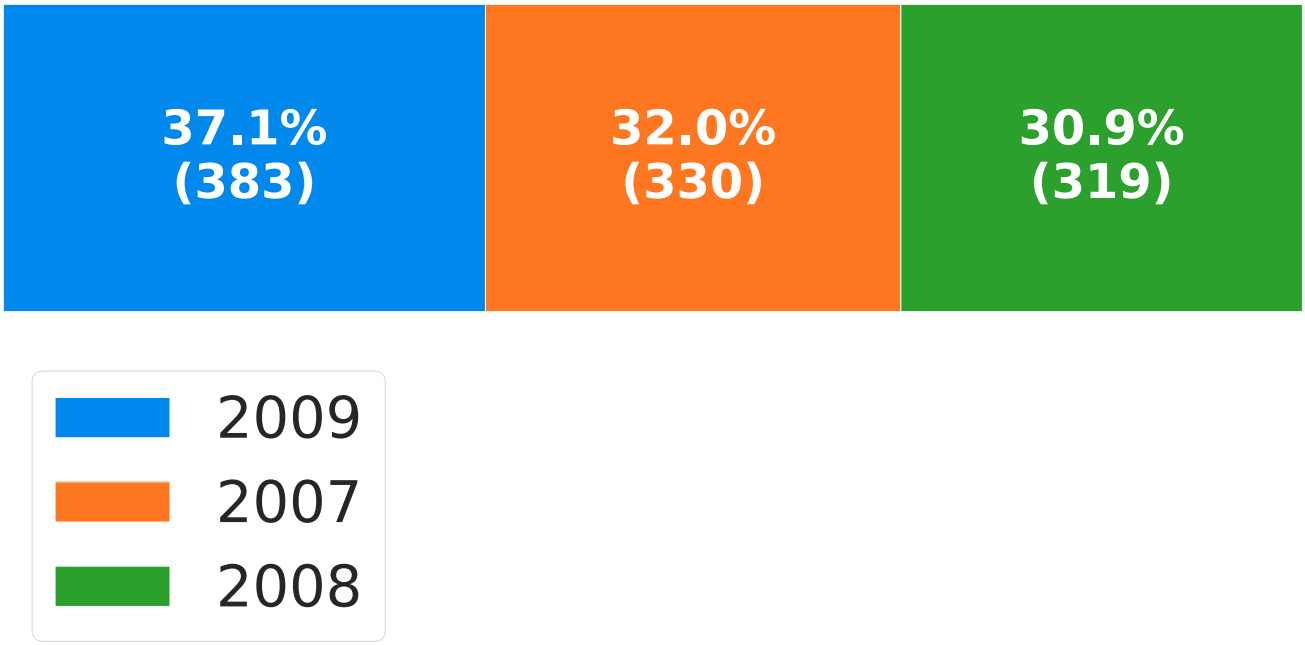
Value	Count	Frequency (%)
2009	383	37.1%
2007	330	32.0%
2008	319	30.9%

Length



Histogram of lengths of the category

Common Values (Plot)



Value	Count	Frequency (%)
2009	383	37.1%
2007	330	32.0%
2008	319	30.9%

Most occurring characters

Value	Count	Frequency (%)
0	2064	50.0%
2	1032	25.0%
9	383	9.3%
7	330	8.0%
8	319	7.7%

Most occurring categories

Value	Count	Frequency (%)
Decimal Number	4128	100.0%

Most frequent character per category

Decimal Number

Value	Count	Frequency (%)
0	2064	50.0%
2	1032	25.0%
9	383	9.3%
7	330	8.0%
8	319	7.7%

Most occurring scripts

Value	Count	Frequency (%)
Common	4128	100.0%

Most frequent character per script

Common

Value	Count	Frequency (%)
0	2064	50.0%
2	1032	25.0%
9	383	9.3%
7	330	8.0%
8	319	7.7%

Most occurring blocks

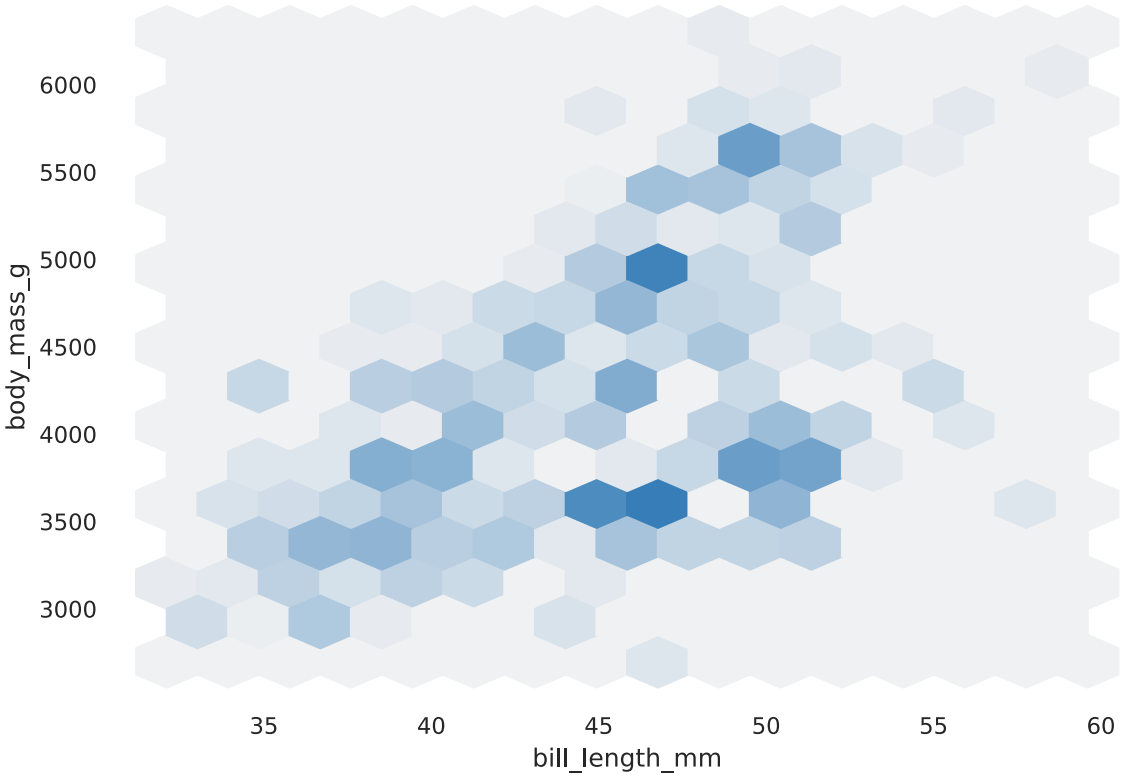
Value	Count	Frequency (%)
ASCII	4128	100.0%

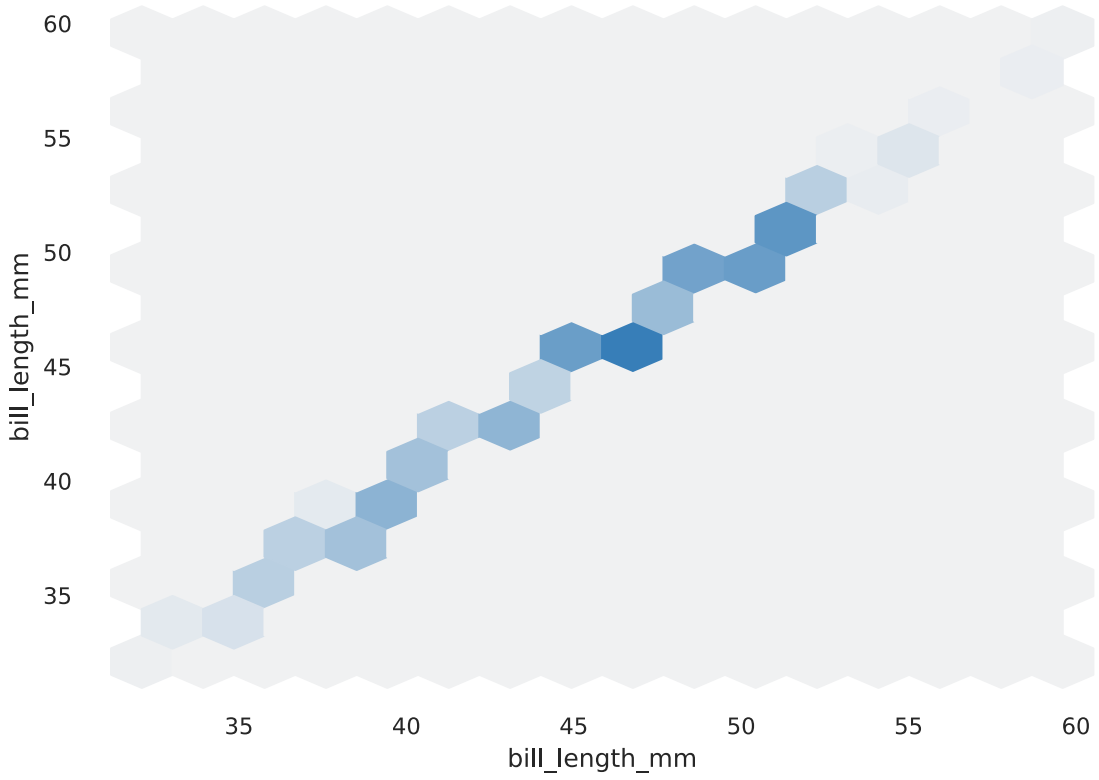
Most frequent character per block

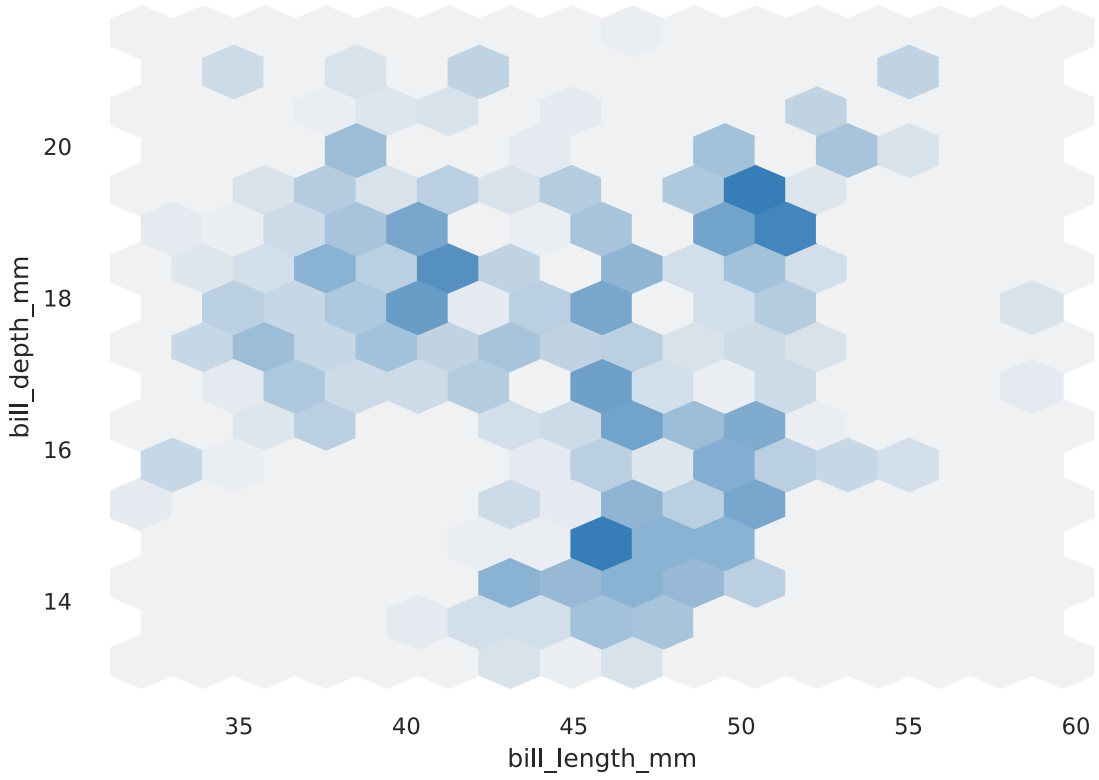
ASCII

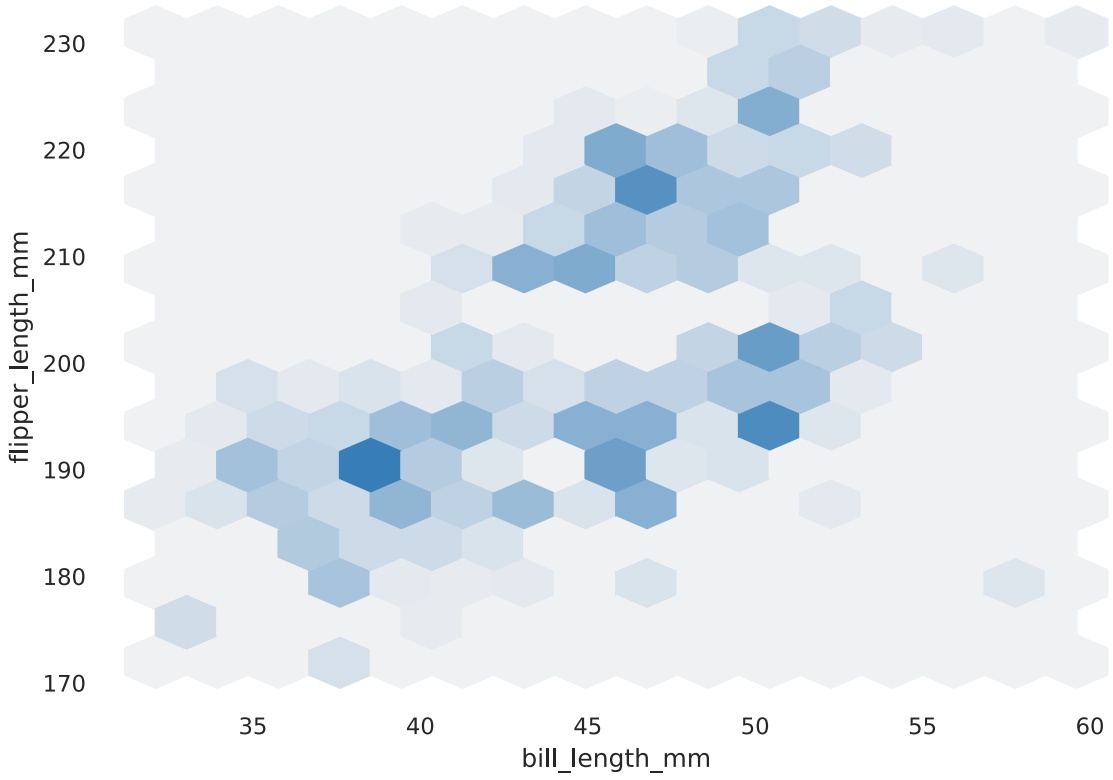
Value	Count	Frequency (%)
0	2064	50.0%
2	1032	25.0%
9	383	9.3%
7	330	8.0%
8	319	7.7%

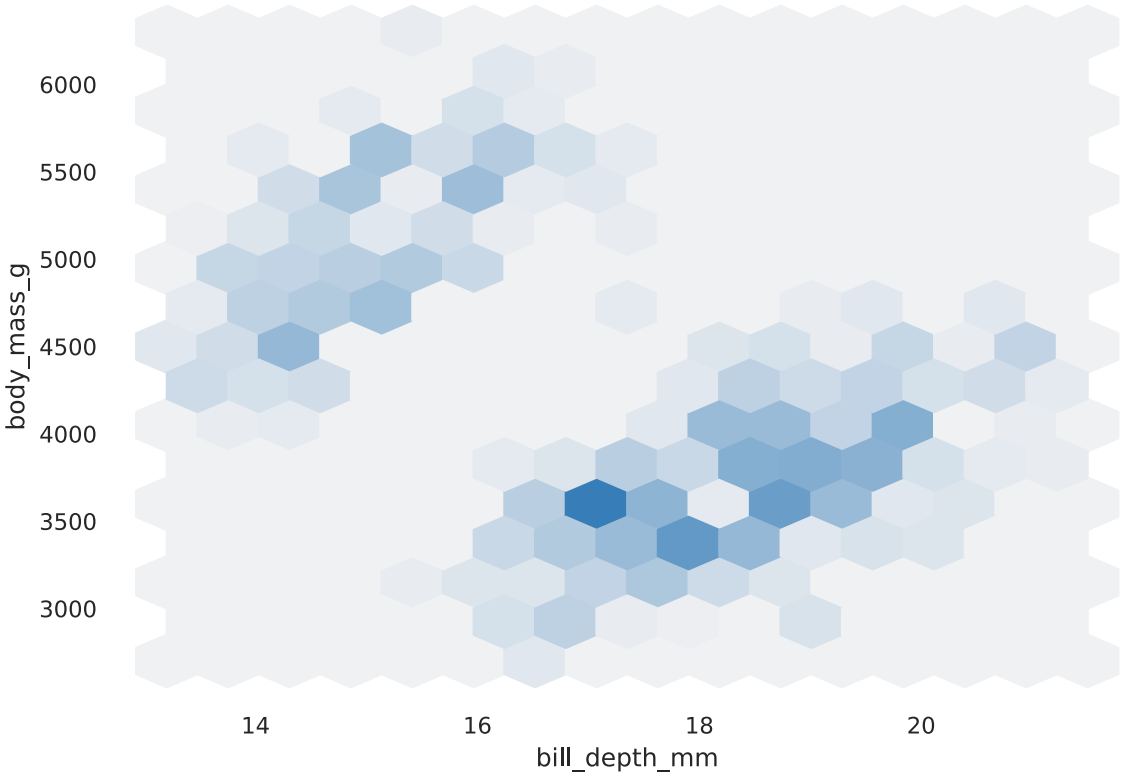
Interactions

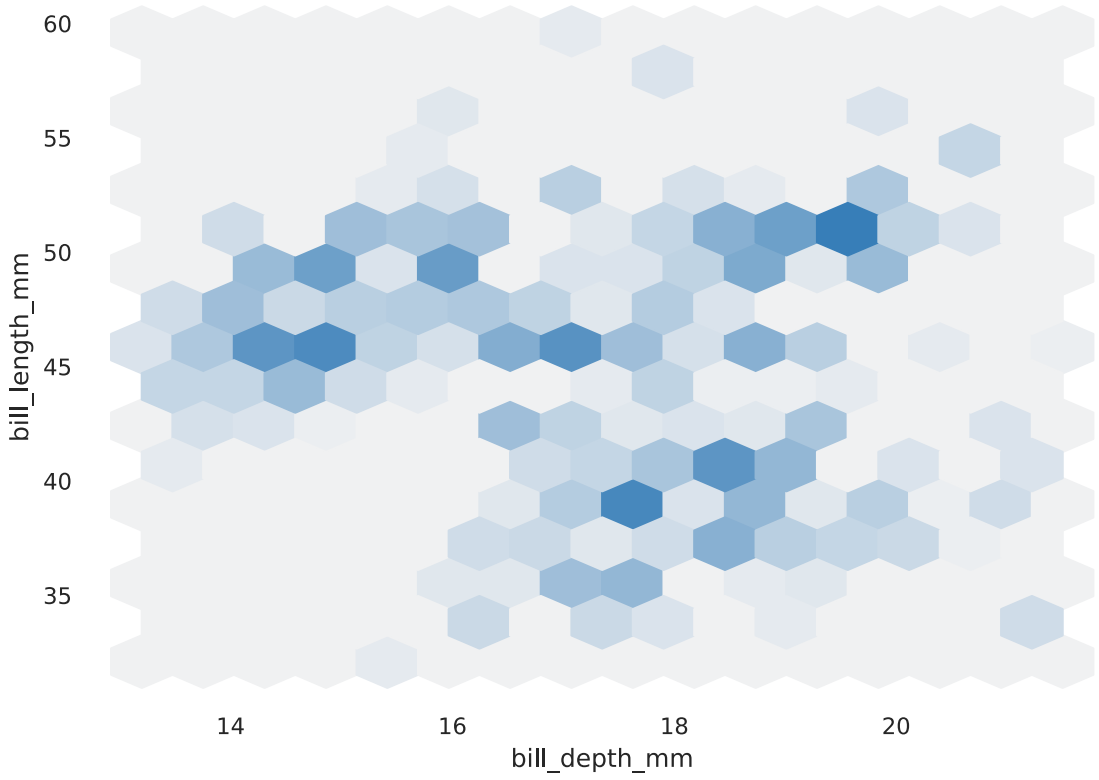


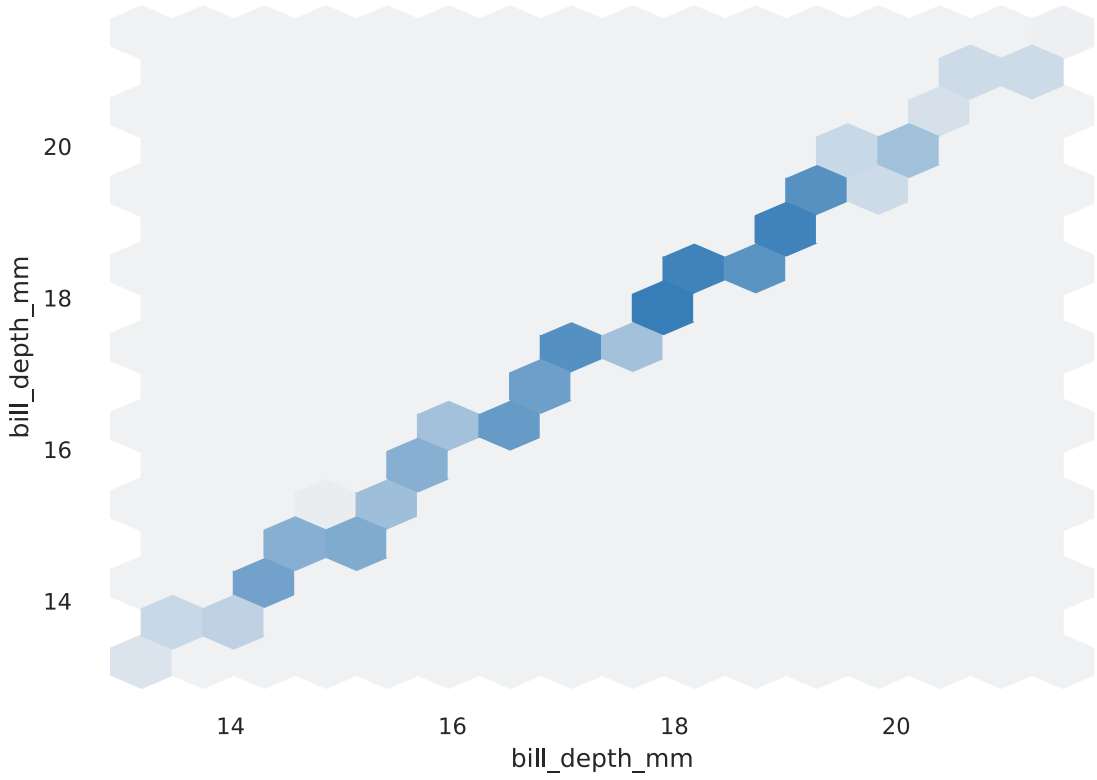

60
55
50
45
40
35
bill_length_mm
35
40
45
50
55
60
bill_length_mm

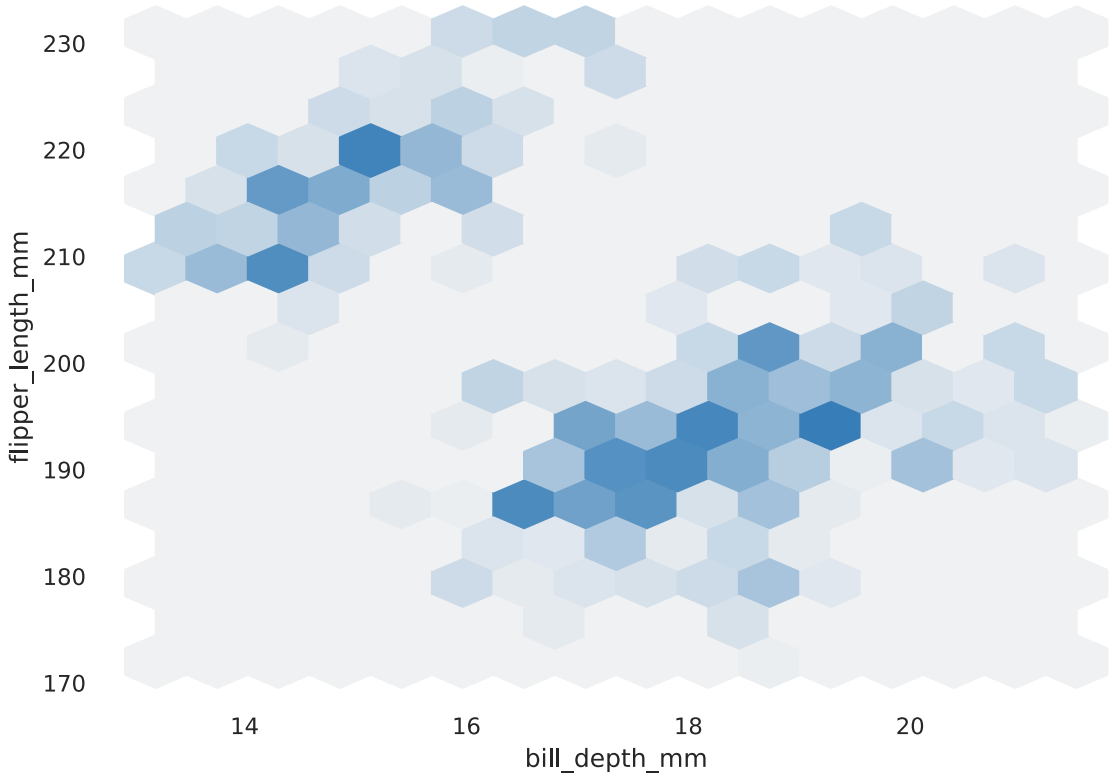

bill_depth_mm
20
18
16
14
35
40
45
50
55
60
bill_length_mm

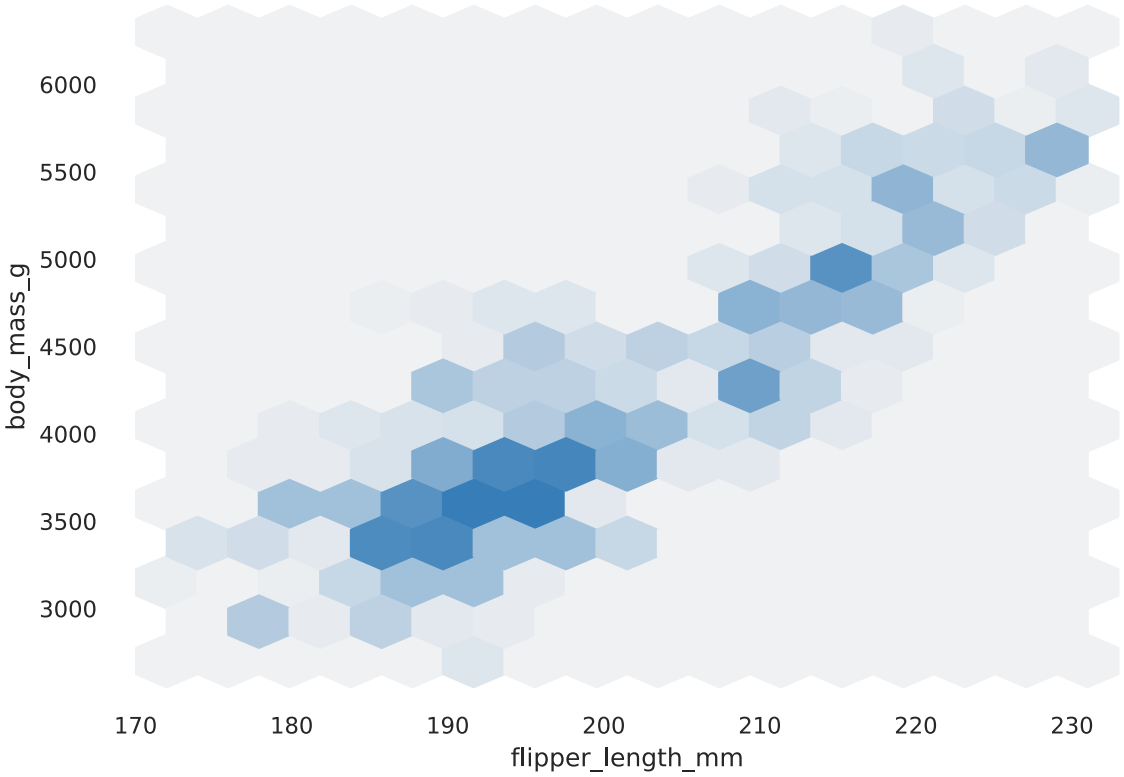

230
220
210
200
190
180
170
flipper_length_mm
35
40
45
50
55
60
bill_length_mm

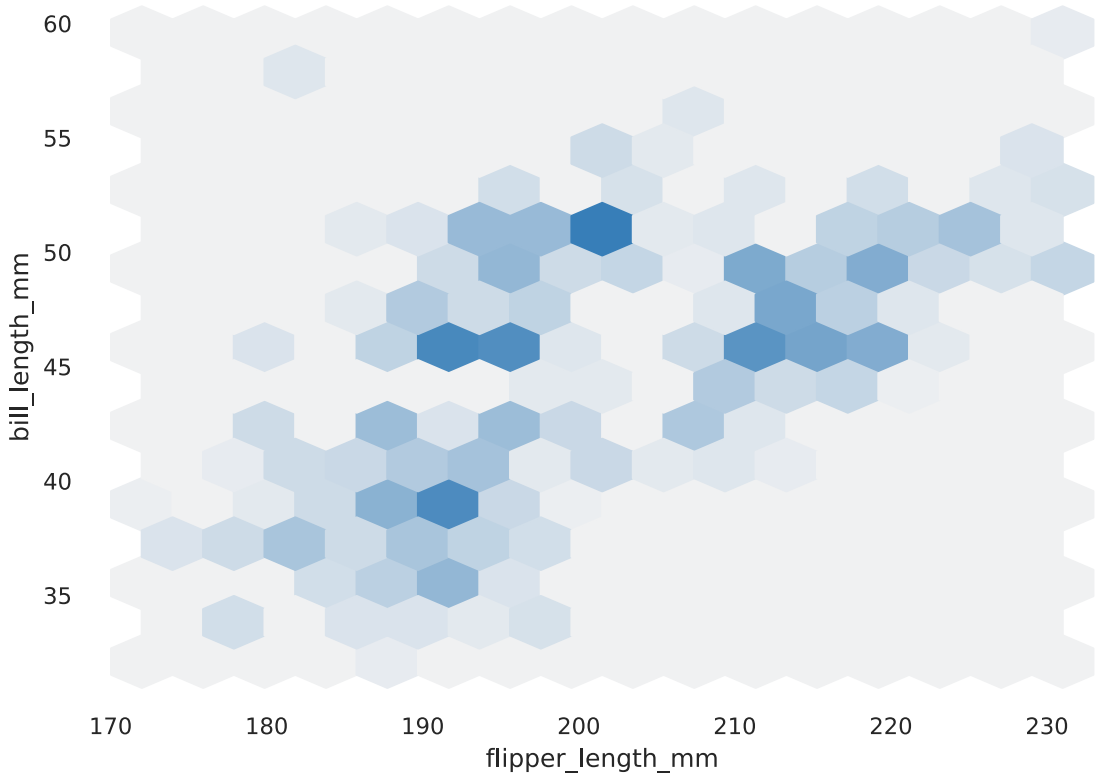

6000
5500
5000
4500
4000
3500
3000
body_mass_g
14
16
18
20
bill_depth_mm

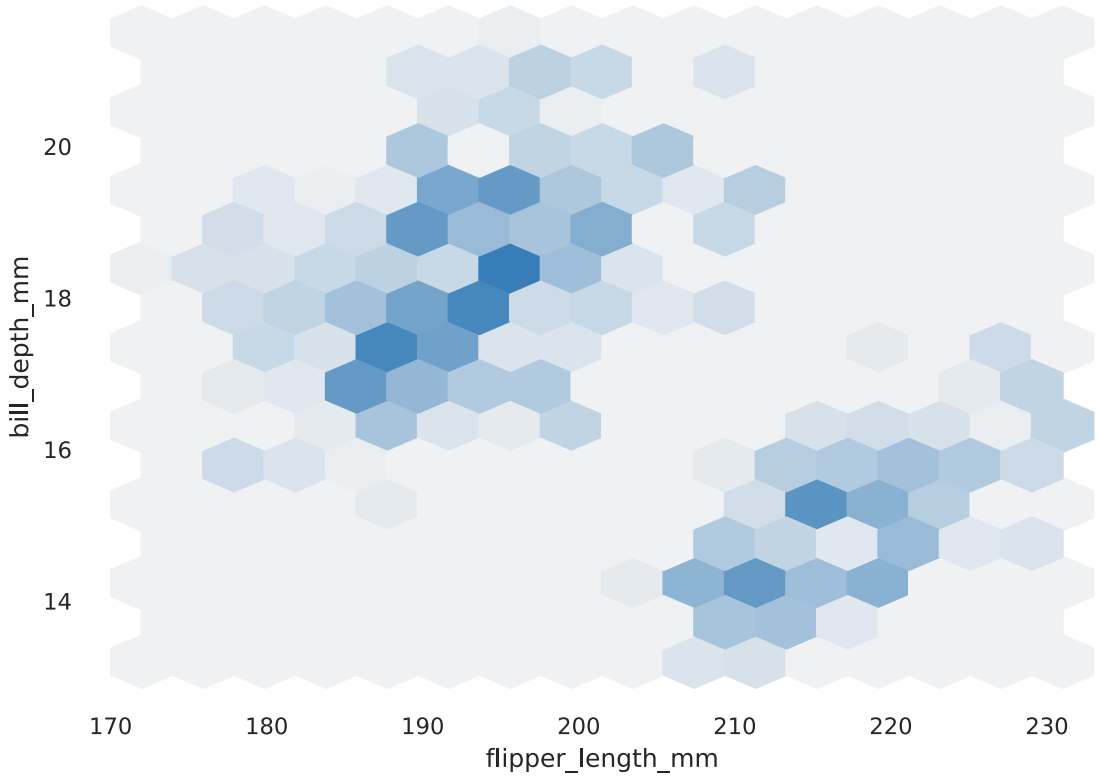

60
55
50
45
40
35
bill_length_mm
14
16
18
20
bill_depth_mm

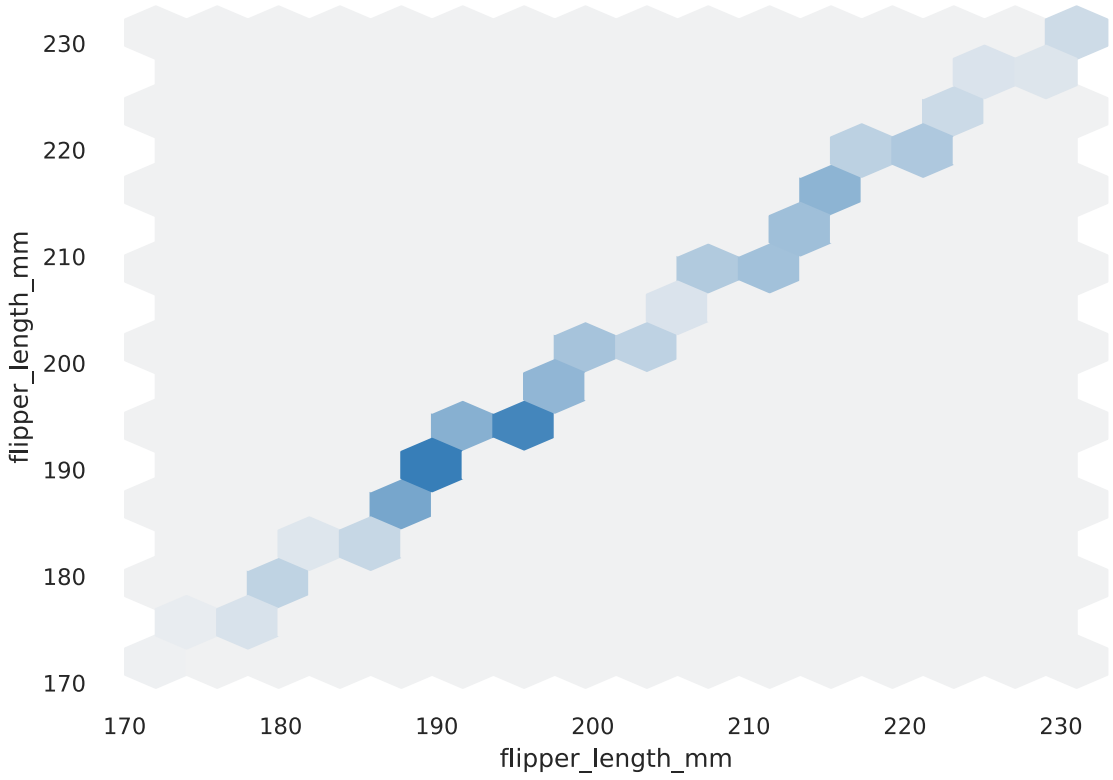

bill_depth_mm
20
18
16
14
14
16
18
20
bill_depth_mm

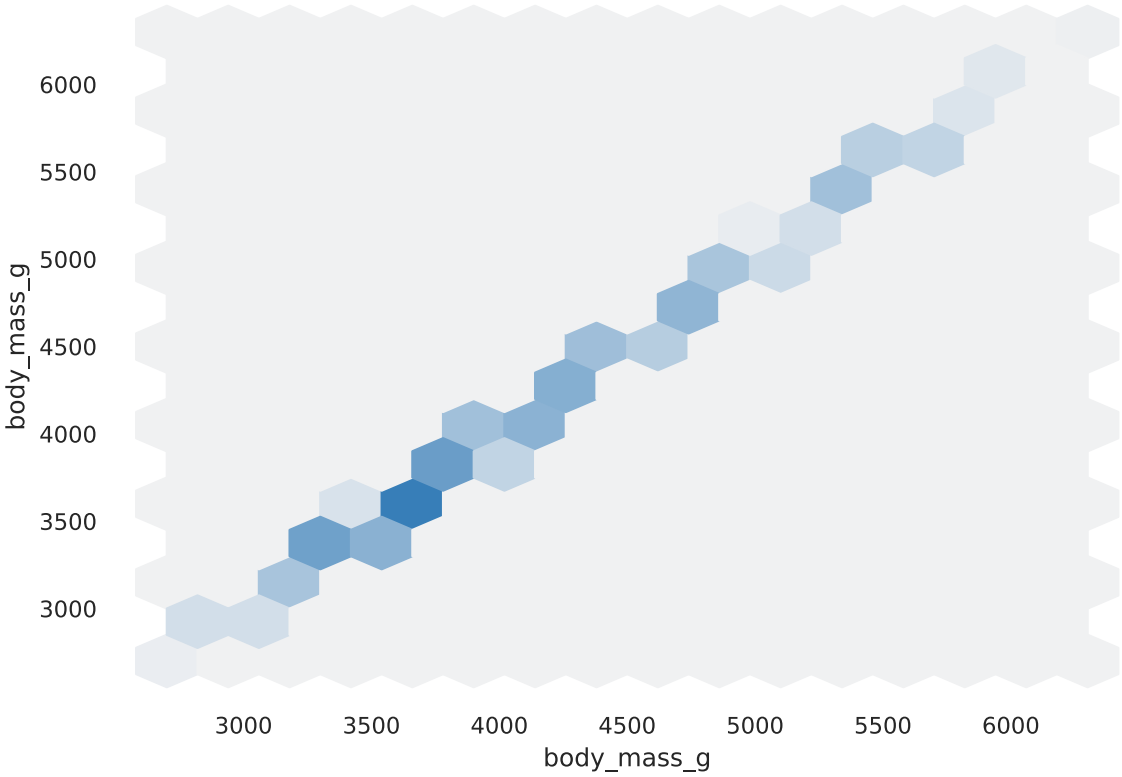

230
220
210
200
190
180
170
flipper_length_mm
14
16
18
20
bill_depth_mm

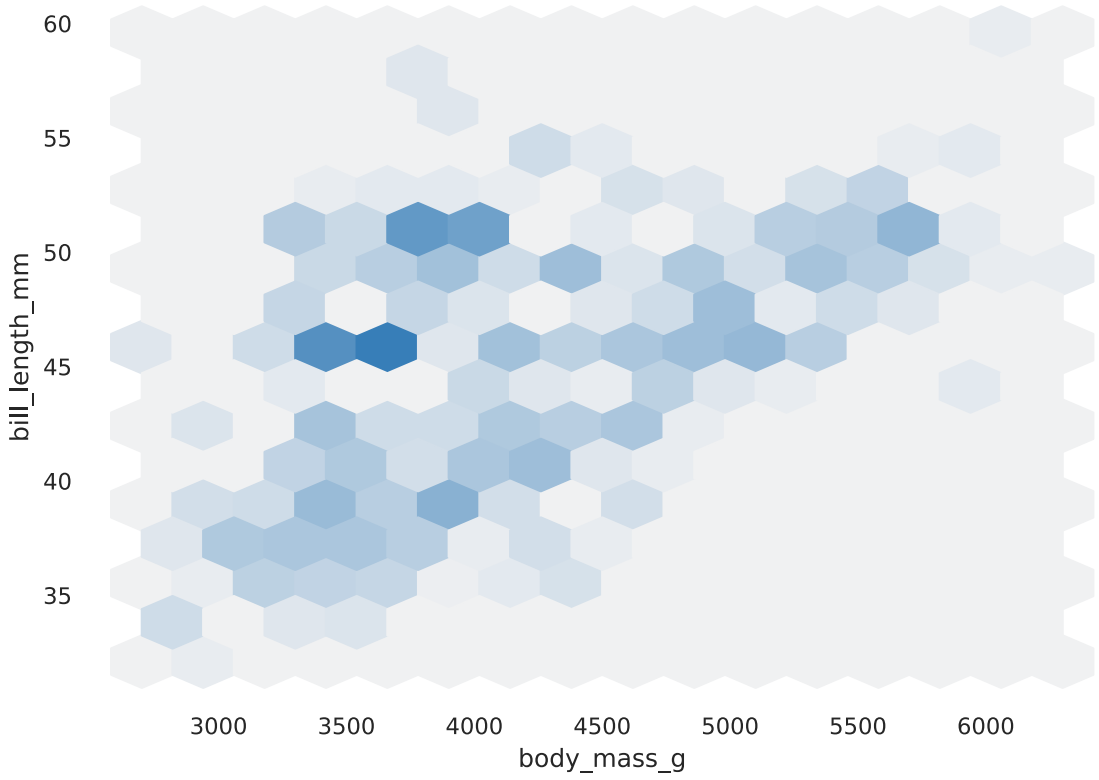

6000
5500
5000
4500
4000
3500
3000
body_mass_g
170
180
190
200
210
220
230
flipper_length_mm

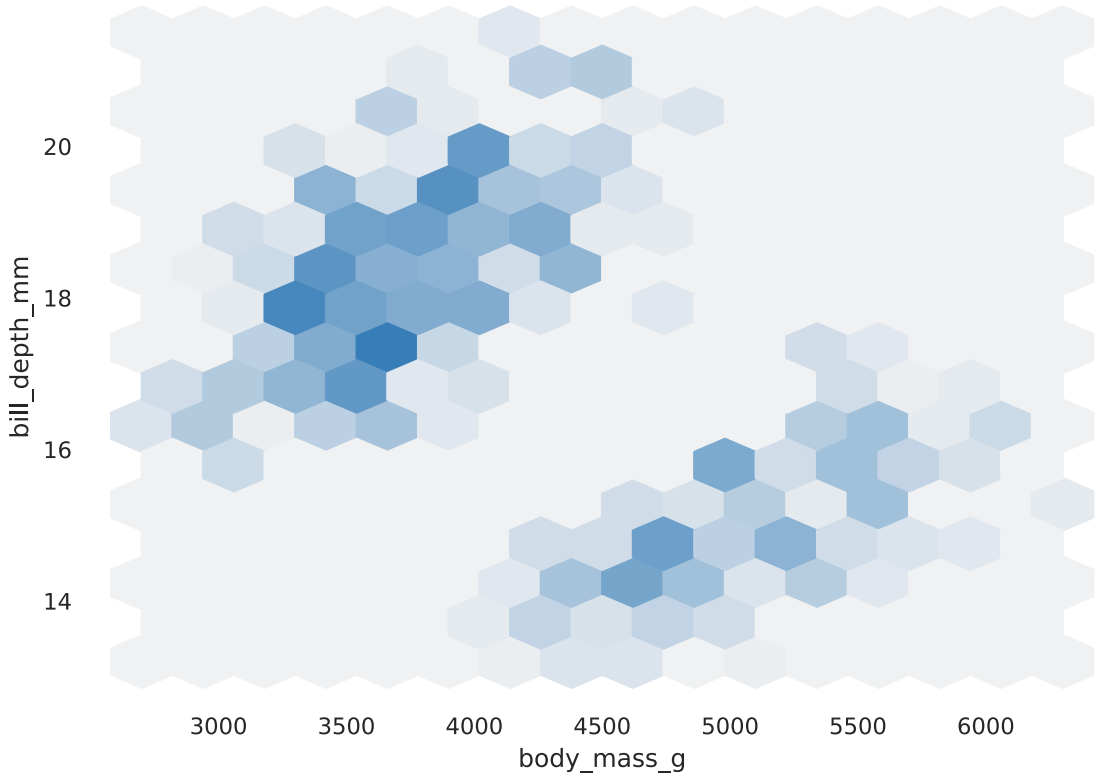

60
55
50
45
40
35
bill_length_mm
170
180
190
200
210
220
230
flipper_length_mm

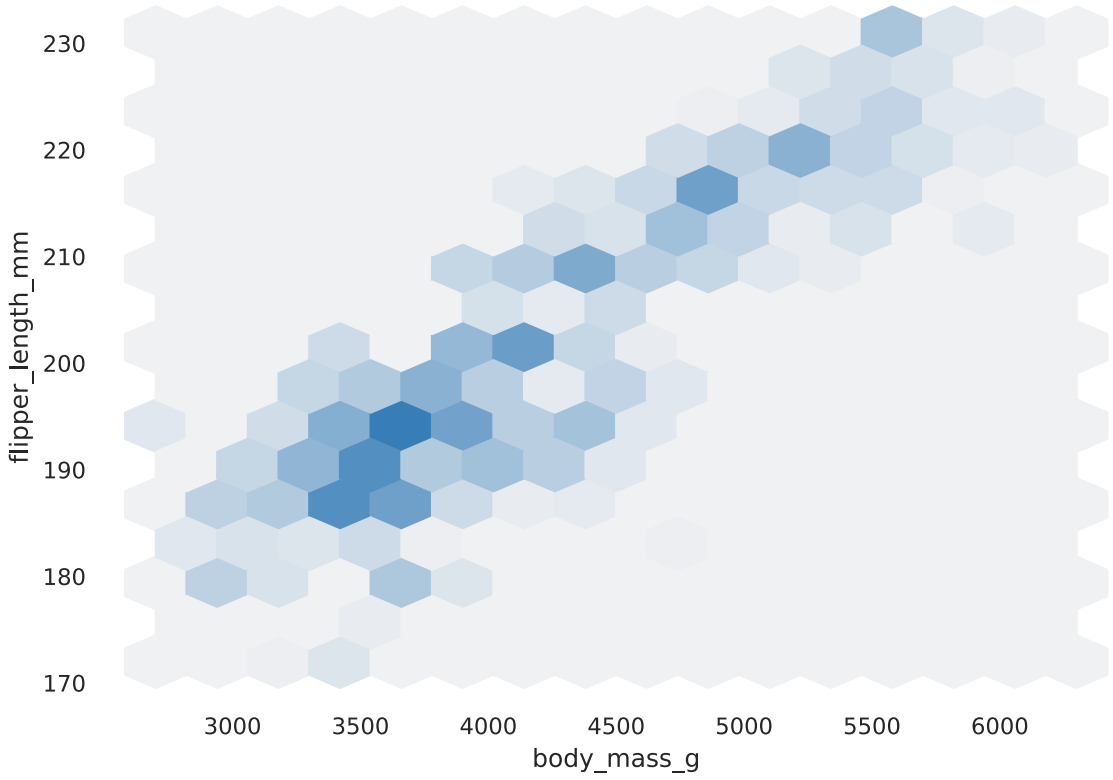

bill_depth_mm
20
18
16
14
170
180
190
200
210
220
230
flipper_length_mm


230
220
210
200
190
180
170
flipper_length_mm
170
180
190
200
210
220
230
flipper_length_mm

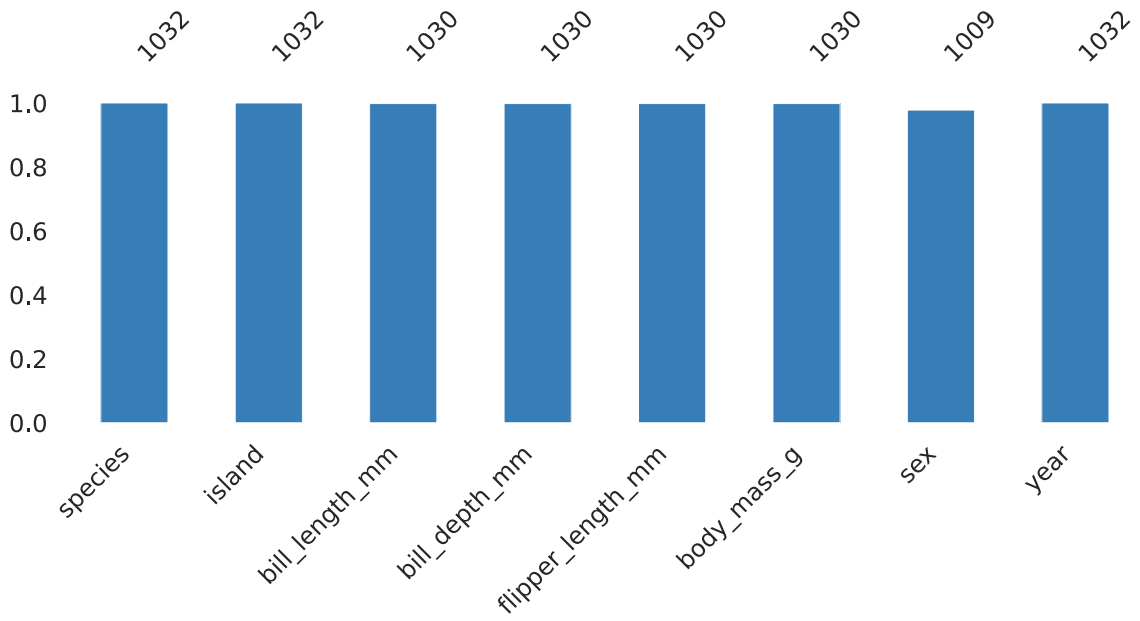

6000
5500
5000
4500
4000
3500
3000
body_mass_g
3000
3500
4000
4500
5000
5500
6000
body_mass_g


60
55
50
45
40
35
bill_length_mm
3000
3500
4000
4500
5000
5500
6000
body_mass_g

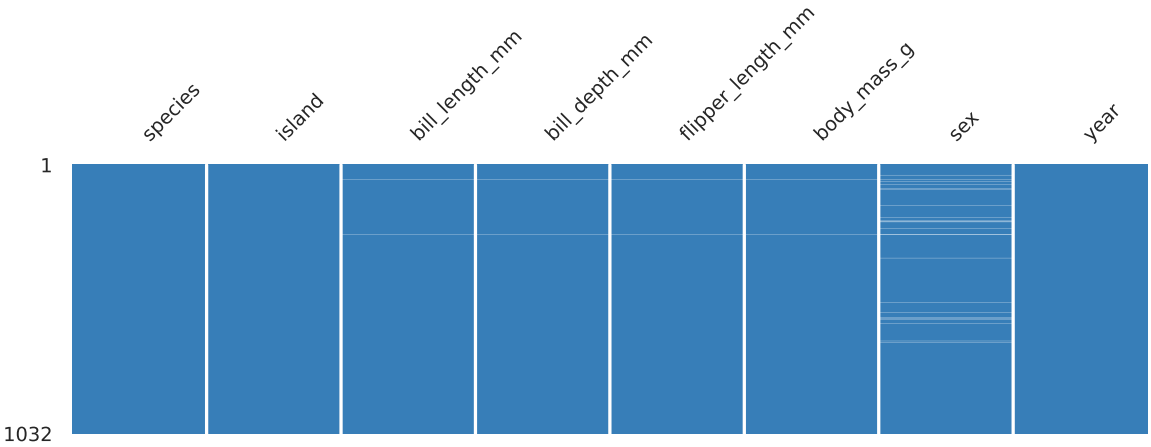

bill_depth_mm
20
18
16
14
3000
3500
4000
4500
5000
5500
6000
body_mass_g


230
220
210
200
190
180
170
flipper_length_mm
3000
3500
4000
4500
5000
5500
6000
body_mass_g

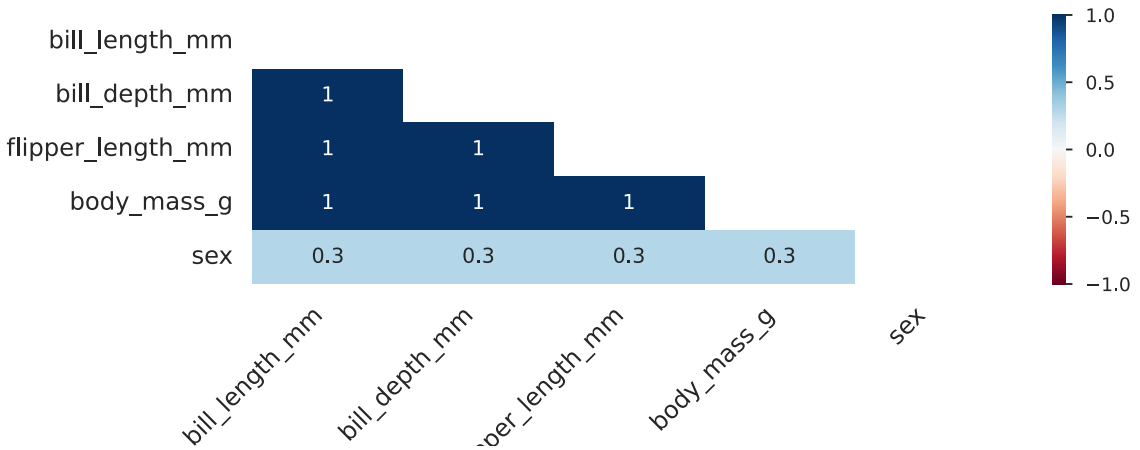
Missing values



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.

Sample

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
102	Adelie	Biscoe	37.7	16.0	183.0	3075.0	female	2009
92	Adelie	Dream	34.0	17.1	185.0	3400.0	female	2008
14	Adelie	Torgersen	34.6	21.1	198.0	4400.0	male	2007
106	Adelie	Biscoe	38.6	17.2	199.0	3750.0	female	2009
71	Adelie	Torgersen	39.7	18.4	190.0	3900.0	male	2008
20	Adelie	Biscoe	37.8	18.3	174.0	3400.0	female	2007
102	Adelie	Biscoe	37.7	16.0	183.0	3075.0	female	2009
121	Adelie	Torgersen	37.7	19.8	198.0	3500.0	male	2009
74	Adelie	Torgersen	35.5	17.5	190.0	3700.0	female	2008
87	Adelie	Dream	36.9	18.6	189.0	3500.0	female	2008

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
338	Chinstrap	Dream	45.7	17.0	195.0	3650.0	female	2009
294	Chinstrap	Dream	46.4	18.6	190.0	3450.0	female	2007
333	Chinstrap	Dream	49.3	19.9	203.0	4050.0	male	2009
330	Chinstrap	Dream	42.5	17.3	187.0	3350.0	female	2009
337	Chinstrap	Dream	46.8	16.5	189.0	3650.0	female	2009
298	Chinstrap	Dream	43.2	16.6	187.0	2900.0	female	2007
284	Chinstrap	Dream	46.0	18.9	195.0	4150.0	female	2007
287	Chinstrap	Dream	51.7	20.3	194.0	3775.0	male	2007
276	Chinstrap	Dream	46.5	17.9	192.0	3500.0	female	2007
333	Chinstrap	Dream	49.3	19.9	203.0	4050.0	male	2009

Duplicate rows

Most frequently occurring

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year	# duplicates
118	Chinstrap	Dream	46.8	16.5	189.0	3650.0	female	2009	14
193	Gentoo	Biscoe	46.2	14.9	221.0	5300.0	male	2008	12
98	Chinstrap	Dream	42.5	16.7	187.0	3350.0	female	2008	10
106	Chinstrap	Dream	45.6	19.4	194.0	3525.0	female	2009	10
133	Chinstrap	Dream	50.0	19.5	196.0	3900.0	male	2007	10
140	Chinstrap	Dream	50.6	19.4	193.0	3800.0	male	2007	9
109	Chinstrap	Dream	45.9	17.1	190.0	3575.0	female	2007	8
126	Chinstrap	Dream	49.0	19.6	212.0	4300.0	male	2009	8
128	Chinstrap	Dream	49.3	19.9	203.0	4050.0	male	2009	8
141	Chinstrap	Dream	50.7	19.7	203.0	4050.0	male	2009	8