Visualizing Big Data in the Introductory Course

David J. Kahle, Ph.D.
Assistant Professor
Department of Statistical Science

eCOTS 2014

BAYLOR UNIVERSITY
DEPARTMENT OF STATISTICAL SCIENCE

Why graphics?

> Modern data graphics can do much more than simply substitute for small statistical tables.
>
> At their best, graphics are **instruments for reasoning** about quantitative information.
>
> Often the most effective way to describe, explore, and summarize a set of numbers–even a very large set–is to look at pictures of those numbers.

Edward Tufte
The Visual Display of Quantitative Information, 2001.
Emphasis added.

Spreadsheet-type datasets

diamonds

diamonds.csv
Source

Columns (10/0)
carat
cut
color
clarity
depth
table
price
x
y
z

Rows
All rows 53940
Selected 0
Excluded 0
Hidden 0
Labelled 0

	carat	cut	color	clarity	depth	table	price	x	y	z
1	0.23	Ideal	E	SI2	61.5	55	326	3.95	3.98	2.43
2	0.21	Premium	E	SI1	59.8	61	326	3.89	3.84	2.31
3	0.23	Good	E	VS1	56.9	65	327	4.05	4.07	2.31
4	0.29	Premium	I	VS2	62.4	58	334	4.2	4.23	2.63
5	0.31	Good	J	SI2	63.3	58	335	4.34	4.35	2.75
6	0.24	Very Good	J	VVS2	62.8	57	336	3.94	3.96	2.48
7	0.24	Very Good	I	VVS1	62.3	57	336	3.95	3.98	2.47
8	0.26	Very Good	H	SI1	61.9	55	337	4.07	4.11	2.53
9	0.22	Fair	E	VS2	65.1	61	337	3.87	3.78	2.49
10	0.23	Very Good	H	VS1	59.4	61	338	4	4.05	2.39
11	0.3	Good	J	SI1	64	55	339	4.25	4.28	2.73
12	0.23	Ideal	J	VS1	62.8	56	340	3.93	3.9	2.46
13	0.22	Premium	F	SI1	60.4	61	342	3.88	3.84	2.33
14	0.31	Ideal	J	SI2	62.2	54	344	4.35	4.37	2.71
15	0.2	Premium	E	SI2	60.2	62	345	3.79	3.75	2.27
16	0.32	Premium	E	I1	60.9	58	345	4.38	4.42	2.68
17	0.3	Ideal	I	SI2	62	54	348	4.31	4.34	2.68
18	0.3	Good	J	SI1	63.4	54	351	4.23	4.29	2.7
19	0.3	Good	J	SI1	63.8	56	351	4.23	4.26	2.71
20	0.3	Very Good	J	SI1	62.7	59	351	4.21	4.27	2.66
21	0.3	Good	I	SI2	63.3	56	351	4.26	4.3	2.71
22	0.23	Very Good	E	VS2	63.8	55	352	3.85	3.92	2.48
23	0.23	Very Good	H	VS1	61	57	353	3.94	3.96	2.41
24	0.31	Very Good	J	SI1	59.4	62	353	4.39	4.43	2.62
25	0.31	Very Good	J	SI1	58.1	62	353	4.44	4.47	2.59
26	0.23	Very Good	G	VVS2	60.4	58	354	3.97	4.01	2.41
27	0.24	Premium	I	VS1	62.5	57	355	3.97	3.94	2.47
28	0.3	Very Good	J	VS2	62.2	57	357	4.28	4.3	2.67
29	0.23	Very Good	D	VS2	60.5	61	357	3.96	3.97	2.4
30	0.23	Very Good	F	VS1	60.9	57	357	3.96	3.99	2.42
31	0.23	Very Good	F	VS1	60	57	402	4	4.03	2.41
32	0.23	Very Good	F	VS1	59.8	57	402	4.04	4.06	2.42
33	0.23	Very Good	E	VS1	60.7	59	402	3.97	4.01	2.42
34	0.23	Very Good	E	VS1	59.5	58	402	4.01	4.06	2.4
35	0.23	Very Good	D	VS1	61.9	58	402	3.92	3.96	2.44
36	0.23	Good	F	VS1	58.2	59	402	4.06	4.08	2.37
37	0.23	Good	E	VS1	64.1	59	402	3.83	3.85	2.46
38	0.31	Good	H	SI1	64	54	402	4.29	4.31	2.75
39	0.26	Very Good	D	VS2	60.8	59	403	4.13	4.16	2.52
40	0.33	Ideal	I	SI2	61.8	55	403	4.49	4.51	2.78
41	0.33	Ideal	I	SI2	61.2	56	403	4.49	4.5	2.75
42	0.33	Ideal	J	SI1	61.1	56	403	4.49	4.55	2.76

Spreadsheet-type datasets

Columns = variables (p)

diamonds

diamonds.csv
Source

Columns (10/0)
carat
cut
color
clarity
depth
table
price
x
y
z

Rows
All rows 53940
Selected 0
Excluded 0
Hidden 0
Labelled 0

	carat	cut	color	clarity	depth	table	price	x	y	z
1	0.23	Ideal	E	SI2	61.5	55	326	3.95	3.98	2.43
2	0.21	Premium	E	SI1	59.8	61	326	3.89	3.84	2.31
3	0.23	Good	E	VS1	56.9	65	327	4.05	4.07	2.31
4	0.29	Premium	I	VS2	62.4	58	334	4.2	4.23	2.63
5	0.31	Good	J	SI2	63.3	58	335	4.34	4.35	2.75
6	0.24	Very Good	J	VVS2	62.8	57	336	3.94	3.96	2.48
7	0.24	Very Good	I	VVS1	62.3	57	336	3.95	3.98	2.47
8	0.26	Very Good	H	SI1	61.9	55	337	4.07	4.11	2.53
9	0.22	Fair	E	VS2	65.1	61	337	3.87	3.78	2.49
10	0.23	Very Good	H	VS1	59.4	61	338	4	4.05	2.39
11	0.3	Good	J	SI1	64	55	339	4.25	4.28	2.73
12	0.23	Ideal	J	VS1	62.8	56	340	3.93	3.9	2.46
13	0.22	Premium	F	SI1	60.4	61	342	3.88	3.84	2.33
14	0.31	Ideal	J	SI2	62.2	54	344	4.35	4.37	2.71
15	0.2	Premium	E	SI2	60.2	62	345	3.79	3.75	2.27
16	0.32	Premium	E	I1	60.9	58	345	4.38	4.42	2.68
17	0.3	Ideal	I	SI2	62	54	348	4.31	4.34	2.68
18	0.3	Good	J	SI1	63.4	54	351	4.23	4.29	2.7
19	0.3	Good	J	SI1	63.8	56	351	4.23	4.26	2.71
20	0.3	Very Good	J	SI1	62.7	59	351	4.21	4.27	2.66
21	0.3	Good	I	SI2	63.3	56	351	4.26	4.3	2.71
22	0.23	Very Good	E	VS2	63.8	55	352	3.85	3.92	2.48
23	0.23	Very Good	H	VS1	61	57	353	3.94	3.96	2.41
24	0.31	Very Good	J	SI1	59.4	62	353	4.39	4.43	2.62
25	0.31	Very Good	J	SI1	58.1	62	353	4.44	4.47	2.59
26	0.23	Very Good	G	VVS2	60.4	58	354	3.97	4.01	2.41
27	0.24	Premium	I	VS1	62.5	57	355	3.97	3.94	2.47
28	0.3	Very Good	J	VS2	62.2	57	357	4.28	4.3	2.67
29	0.23	Very Good	D	VS2	60.5	61	357	3.96	3.97	2.4
30	0.23	Very Good	F	VS1	60.9	57	357	3.96	3.99	2.42
31	0.23	Very Good	F	VS1	60	57	402	4	4.03	2.41
32	0.23	Very Good	F	VS1	59.8	57	402	4.04	4.06	2.42
33	0.23	Very Good	E	VS1	60.7	59	402	3.97	4.01	2.42
34	0.23	Very Good	E	VS1	59.5	58	402	4.01	4.06	2.4
35	0.23	Very Good	D	VS1	61.9	58	402	3.92	3.96	2.44
36	0.23	Good	F	VS1	58.2	59	402	4.06	4.08	2.37
37	0.23	Good	E	VS1	64.1	59	402	3.83	3.85	2.46
38	0.31	Good	H	SI1	64	54	402	4.29	4.31	2.75
39	0.26	Very Good	D	VS2	60.8	59	403	4.13	4.16	2.52
40	0.33	Ideal	I	SI2	61.8	55	403	4.49	4.51	2.78
41	0.33	Ideal	I	SI2	61.2	56	403	4.49	4.5	2.75
42	0.33	Ideal	J	SI1	61.1	56	403	4.49	4.55	2.76

Spreadsheet-type datasets

Rows = subjects (n)

individuals
schools
school districts
counties
census tracts
...

Columns = variables (p)

diamonds

diamonds.csv
source

Columns (10/0)
carat
cut
color
clarity
depth
table
price
x
y
z

Rows
All rows 53940
Selected 0
Excluded 0
Hidden 0
Labelled 0

	carat	cut	color	clarity	depth	table	price	x	y	z
1	0.23	Ideal	E	SI2	61.5	55	326	3.95	3.98	2.43
2	0.21	Premium	E	SI1	59.8	61	326	3.89	3.84	2.31
3	0.23	Good	E	VS1	56.9	65	327	4.05	4.07	2.31
4	0.29	Premium	I	VS2	62.4	58	334	4.2	4.23	2.63
5	0.31	Good	J	SI2	63.3	58	335	4.34	4.35	2.75
6	0.24	Very Good	J	VVS2	62.8	57	336	3.94	3.96	2.48
7	0.24	Very Good	I	VVS1	62.3	57	336	3.95	3.98	2.47
8	0.26	Very Good	H	SI1	61.9	55	337	4.07	4.11	2.53
9	0.22	Fair	E	VS2	65.1	61	337	3.87	3.78	2.49
10	0.23	Very Good	H	VS1	59.4	61	338	4	4.05	2.39
11	0.3	Good	J	SI1	64	55	339	4.25	4.28	2.73
12	0.23	Ideal	J	VS1	62.8	56	340	3.93	3.9	2.46
13	0.22	Premium	F	SI1	60.4	61	342	3.88	3.84	2.33
14	0.31	Ideal	J	SI2	62.2	54	344	4.35	4.37	2.71
15	0.2	Premium	E	SI2	60.2	62	345	3.79	3.75	2.27
16	0.32	Premium	E	I1	60.9	58	345	4.38	4.42	2.68
17	0.3	Ideal	I	SI2	62	54	348	4.31	4.34	2.68
18	0.3	Good	J	SI1	63.4	54	351	4.23	4.29	2.7
19	0.3	Good	J	SI1	63.8	56	351	4.23	4.26	2.71
20	0.3	Very Good	J	SI1	62.7	59	351	4.21	4.27	2.66
21	0.3	Good	I	SI2	63.3	56	351	4.26	4.3	2.71
22	0.23	Very Good	E	VS2	63.8	55	352	3.85	3.92	2.48
23	0.23	Very Good	H	VS1	61	57	353	3.94	3.96	2.41
24	0.31	Very Good	J	SI1	59.4	62	353	4.39	4.43	2.62
25	0.31	Very Good	J	SI1	58.1	62	353	4.44	4.47	2.59
26	0.23	Very Good	G	VVS2	60.4	58	354	3.97	4.01	2.41
27	0.24	Premium	I	VS1	62.5	57	355	3.97	3.94	2.47
28	0.3	Very Good	J	VS2	62.2	57	357	4.28	4.3	2.67
29	0.23	Very Good	D	VS2	60.5	61	357	3.96	3.97	2.4
30	0.23	Very Good	F	VS1	60.9	57	357	3.96	3.99	2.42
31	0.23	Very Good	F	VS1	60	57	402	4	4.03	2.41
32	0.23	Very Good	F	VS1	59.8	57	402	4.04	4.06	2.42
33	0.23	Very Good	E	VS1	60.7	59	402	3.97	4.01	2.42
34	0.23	Very Good	E	VS1	59.5	58	402	4.01	4.06	2.4
35	0.23	Very Good	D	VS1	61.9	58	402	3.92	3.96	2.44
36	0.23	Good	F	VS1	58.2	59	402	4.06	4.08	2.37
37	0.23	Good	E	VS1	64.1	59	402	3.83	3.85	2.46
38	0.31	Good	H	SI1	64	54	402	4.29	4.31	2.75
39	0.26	Very Good	D	VS2	60.8	59	403	4.13	4.16	2.52
40	0.33	Ideal	I	SI2	61.8	55	403	4.49	4.51	2.78
41	0.33	Ideal	I	SI2	61.2	56	403	4.49	4.5	2.75
42	0.33	Ideal	J	SI1	61.1	56	403	4.49	4.55	2.76

Spreadsheet-type datasets

Rows = subjects (n)

individuals
schools
school districts
counties
census tracts
...

Columns = variables (p)

The easiest data has $n \gg p$!

diamonds.csv
source

Columns (10/0)
carat
cut
color
clarity
depth
table
price
x
y
z

Rows
All rows 53940
Selected 0
Excluded 0
Hidden 0
Labelled 0

diamonds

	carat	cut	color	clarity	depth	table	price	x	y	z
1	0.23	Ideal	E	SI2	61.5	55	326	3.95	3.98	2.43
2	0.21	Premium	E	SI1	59.8	61	326	3.89	3.84	2.31
3	0.23	Good	E	VS1	56.9	65	327	4.05	4.07	2.31
4	0.29	Premium	I	VS2	62.4	58	334	4.2	4.23	2.63
5	0.31	Good	J	SI2	63.3	58	335	4.34	4.35	2.75
6	0.24	Very Good	J	VVS2	62.8	57	336	3.94	3.96	2.48
7	0.24	Very Good	I	VVS1	62.3	57	336	3.95	3.98	2.47
8	0.26	Very Good	H	SI1	61.9	55	337	4.07	4.11	2.53
9	0.22	Fair	E	VS2	65.1	61	337	3.87	3.78	2.49
10	0.23	Very Good	H	VS1	59.4	61	338	4	4.05	2.39
11	0.3	Good	J	SI1	64	55	339	4.25	4.28	2.73
12	0.23	Ideal	J	VS1	62.8	56	340	3.93	3.9	2.46
13	0.22	Premium	F	SI1	60.4	61	342	3.88	3.84	2.33
14	0.31	Ideal	J	SI2	62.2	54	344	4.35	4.37	2.71
15	0.2	Premium	E	SI2	60.2	62	345	3.79	3.75	2.27
16	0.32	Premium	E	I1	60.9	58	345	4.38	4.42	2.68
17	0.3	Ideal	I	SI2	62	54	348	4.31	4.34	2.68
18	0.3	Good	J	SI1	63.4	54	351	4.23	4.29	2.7
19	0.3	Good	J	SI1	63.8	56	351	4.23	4.26	2.71
20	0.3	Very Good	J	SI1	62.7	59	351	4.21	4.27	2.66
21	0.3	Good	I	SI2	63.3	56	351	4.26	4.3	2.71
22	0.23	Very Good	E	VS2	63.8	55	352	3.85	3.92	2.48
23	0.23	Very Good	H	VS1	61	57	353	3.94	3.96	2.41
24	0.31	Very Good	J	SI1	59.4	62	353	4.39	4.43	2.62
25	0.31	Very Good	J	SI1	58.1	62	353	4.44	4.47	2.59
26	0.23	Very Good	G	VVS2	60.4	58	354	3.97	4.01	2.41
27	0.24	Premium	I	VS1	62.5	57	355	3.97	3.94	2.47
28	0.3	Very Good	J	VS2	62.2	57	357	4.28	4.3	2.67
29	0.23	Very Good	D	VS2	60.5	61	357	3.96	3.97	2.4
30	0.23	Very Good	F	VS1	60.9	57	357	3.96	3.99	2.42
31	0.23	Very Good	F	VS1	60	57	402	4	4.03	2.41
32	0.23	Very Good	F	VS1	59.8	57	402	4.04	4.06	2.42
33	0.23	Very Good	E	VS1	60.7	59	402	3.97	4.01	2.42
34	0.23	Very Good	E	VS1	59.5	58	402	4.01	4.06	2.4
35	0.23	Very Good	D	VS1	61.9	58	402	3.92	3.96	2.44
36	0.23	Good	F	VS1	58.2	59	402	4.06	4.08	2.37
37	0.23	Good	E	VS1	64.1	59	402	3.83	3.85	2.46
38	0.31	Good	H	SI1	64	54	402	4.29	4.31	2.75
39	0.26	Very Good	D	VS2	60.8	59	403	4.13	4.16	2.52
40	0.33	Ideal	I	SI2	61.8	55	403	4.49	4.51	2.78
41	0.33	Ideal	I	SI2	61.2	56	403	4.49	4.5	2.75
42	0.33	Ideal	J	SI1	61.1	56	403	4.49	4.55	2.76

Spreadsheet-type datasets

Columns = variables (p)

Rows = subjects (n)

individuals
schools
school districts
counties
census tracts
...

The easiest data has $n \gg p$!

...but we ***can*** work with $p \gg n$ data

diamonds.csv
source

Columns (10/0)
carat
cut
color
clarity
depth
table
price
x
y
z

Rows
All rows 53940
Selected 0
Excluded 0
Hidden 0
Labelled 0

diamonds

	carat	cut	color	clarity	depth	table	price	x	y	z
1	0.23	Ideal	E	SI2	61.5	55	326	3.95	3.98	2.43
2	0.21	Premium	E	SI1	59.8	61	326	3.89	3.84	2.31
3	0.23	Good	E	VS1	56.9	65	327	4.05	4.07	2.31
4	0.29	Premium	I	VS2	62.4	58	334	4.2	4.23	2.63
5	0.31	Good	J	SI2	63.3	58	335	4.34	4.35	2.75
6	0.24	Very Good	J	VVS2	62.8	57	336	3.94	3.96	2.48
7	0.24	Very Good	I	VVS1	62.3	57	336	3.95	3.98	2.47
8	0.26	Very Good	H	SI1	61.9	55	337	4.07	4.11	2.53
9	0.22	Fair	E	VS2	65.1	61	337	3.87	3.78	2.49
10	0.23	Very Good	H	VS1	59.4	61	338	4	4.05	2.39
11	0.3	Good	J	SI1	64	55	339	4.25	4.28	2.73
12	0.23	Ideal	J	VS1	62.8	56	340	3.93	3.9	2.46
13	0.22	Premium	F	SI1	60.4	61	342	3.88	3.84	2.33
14	0.31	Ideal	J	SI2	62.2	54	344	4.35	4.37	2.71
15	0.2	Premium	E	SI2	60.2	62	345	3.79	3.75	2.27
16	0.32	Premium	E	I1	60.9	58	345	4.38	4.42	2.68
17	0.3	Ideal	I	SI2	62	54	348	4.31	4.34	2.68
18	0.3	Good	J	SI1	63.4	54	351	4.23	4.29	2.7
19	0.3	Good	J	SI1	63.8	56	351	4.23	4.26	2.71
20	0.3	Very Good	J	SI1	62.7	59	351	4.21	4.27	2.66
21	0.3	Good	I	SI2	63.3	56	351	4.26	4.3	2.71
22	0.23	Very Good	E	VS2	63.8	55	352	3.85	3.92	2.48
23	0.23	Very Good	H	VS1	61	57	353	3.94	3.96	2.41
24	0.31	Very Good	J	SI1	59.4	62	353	4.39	4.43	2.62
25	0.31	Very Good	J	SI1	58.1	62	353	4.44	4.47	2.59
26	0.23	Very Good	G	VVS2	60.4	58	354	3.97	4.01	2.41
27	0.24	Premium	I	VS1	62.5	57	355	3.97	3.94	2.47
28	0.3	Very Good	J	VS2	62.2	57	357	4.28	4.3	2.67
29	0.23	Very Good	D	VS2	60.5	61	357	3.96	3.97	2.4
30	0.23	Very Good	F	VS1	60.9	57	357	3.96	3.99	2.42
31	0.23	Very Good	F	VS1	60	57	402	4	4.03	2.41
32	0.23	Very Good	F	VS1	59.8	57	402	4.04	4.06	2.42
33	0.23	Very Good	E	VS1	60.7	59	402	3.97	4.01	2.42
34	0.23	Very Good	E	VS1	59.5	58	402	4.01	4.06	2.4
35	0.23	Very Good	D	VS1	61.9	58	402	3.92	3.96	2.44
36	0.23	Good	F	VS1	58.2	59	402	4.06	4.08	2.37
37	0.23	Good	E	VS1	64.1	59	402	3.83	3.85	2.46
38	0.31	Good	H	SI1	64	54	402	4.29	4.31	2.75
39	0.26	Very Good	D	VS2	60.8	59	403	4.13	4.16	2.52
40	0.33	Ideal	I	SI2	61.8	55	403	4.49	4.51	2.78
41	0.33	Ideal	I	SI2	61.2	56	403	4.49	4.5	2.75
42	0.33	Ideal	J	SI1	61.1	56	403	4.49	4.55	2.76

Spreadsheet-type datasets

Rows = subjects (n)

individuals
schools
school districts
counties
census tracts
...

The easiest data has $n \gg p$!

...but we ***can*** work with $p \gg n$ data

What's big data?

Columns = variables (p)

diamonds

diamonds.csv
source

Columns (10/0)
carat
cut
color
clarity
depth
table
price
x
y
z

Rows
All rows 53940
Selected 0
Excluded 0
Hidden 0
Labelled 0

	carat	cut	color	clarity	depth	table	price	x	y	z
1	0.23	Ideal	E	SI2	61.5	55	326	3.95	3.98	2.43
2	0.21	Premium	E	SI1	59.8	61	326	3.89	3.84	2.31
3	0.23	Good	E	VS1	56.9	65	327	4.05	4.07	2.31
4	0.29	Premium	I	VS2	62.4	58	334	4.2	4.23	2.63
5	0.31	Good	J	SI2	63.3	58	335	4.34	4.35	2.75
6	0.24	Very Good	J	VVS2	62.8	57	336	3.94	3.96	2.48
7	0.24	Very Good	I	VVS1	62.3	57	336	3.95	3.98	2.47
8	0.26	Very Good	H	SI1	61.9	55	337	4.07	4.11	2.53
9	0.22	Fair	E	VS2	65.1	61	337	3.87	3.78	2.49
10	0.23	Very Good	H	VS1	59.4	61	338	4	4.05	2.39
11	0.3	Good	J	SI1	64	55	339	4.25	4.28	2.73
12	0.23	Ideal	J	VS1	62.8	56	340	3.93	3.9	2.46
13	0.22	Premium	F	SI1	60.4	61	342	3.88	3.84	2.33
14	0.31	Ideal	J	SI2	62.2	54	344	4.35	4.37	2.71
15	0.2	Premium	E	SI2	60.2	62	345	3.79	3.75	2.27
16	0.32	Premium	E	I1	60.9	58	345	4.38	4.42	2.68
17	0.3	Ideal	I	SI2	62	54	348	4.31	4.34	2.68
18	0.3	Good	J	SI1	63.4	54	351	4.23	4.29	2.7
19	0.3	Good	J	SI1	63.8	56	351	4.23	4.26	2.71
20	0.3	Very Good	J	SI1	62.7	59	351	4.21	4.27	2.66
21	0.3	Good	I	SI2	63.3	56	351	4.26	4.3	2.71
22	0.23	Very Good	E	VS2	63.8	55	352	3.85	3.92	2.48
23	0.23	Very Good	H	VS1	61	57	353	3.94	3.96	2.41
24	0.31	Very Good	J	SI1	59.4	62	353	4.39	4.43	2.62
25	0.31	Very Good	J	SI1	58.1	62	353	4.44	4.47	2.59
26	0.23	Very Good	G	VVS2	60.4	58	354	3.97	4.01	2.41
27	0.24	Premium	I	VS1	62.5	57	355	3.97	3.94	2.47
28	0.3	Very Good	J	VS2	62.2	57	357	4.28	4.3	2.67
29	0.23	Very Good	D	VS2	60.5	61	357	3.96	3.97	2.4
30	0.23	Very Good	F	VS1	60.9	57	357	3.96	3.99	2.42
31	0.23	Very Good	F	VS1	60	57	402	4	4.03	2.41
32	0.23	Very Good	F	VS1	59.8	57	402	4.04	4.06	2.42
33	0.23	Very Good	E	VS1	60.7	59	402	3.97	4.01	2.42
34	0.23	Very Good	E	VS1	59.5	58	402	4.01	4.06	2.4
35	0.23	Very Good	D	VS1	61.9	58	402	3.92	3.96	2.44
36	0.23	Good	F	VS1	58.2	59	402	4.06	4.08	2.37
37	0.23	Good	E	VS1	64.1	59	402	3.83	3.85	2.46
38	0.31	Good	H	SI1	64	54	402	4.29	4.31	2.75
39	0.26	Very Good	D	VS2	60.8	59	403	4.13	4.16	2.52
40	0.33	Ideal	I	SI2	61.8	55	403	4.49	4.51	2.78
41	0.33	Ideal	I	SI2	61.2	56	403	4.49	4.5	2.75
42	0.33	Ideal	J	SI1	61.1	56	403	4.49	4.55	2.76

Spreadsheet-type datasets

Columns = variables (p)

Rows = subjects (n)

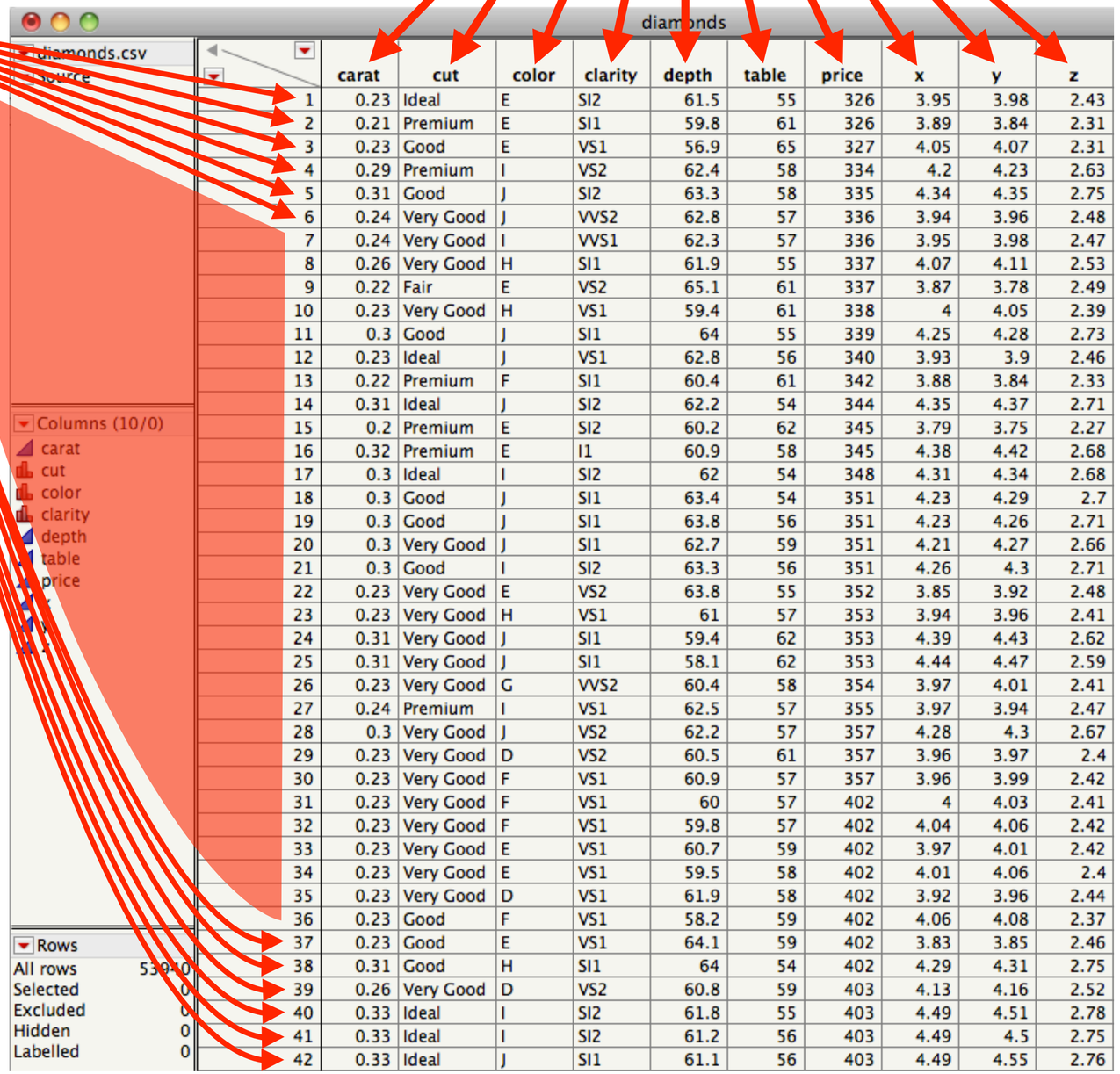
individuals
schools
school districts
counties
census tracts
...

The easiest data has $n \gg p$!

...but we ***can*** work with $p \gg n$ data

What's big data?

n	< 100	very small
	100 – 10,000	small
	10,000 – 1,000,000	medium
	1,000,000 – 100,000,000	large
	> 100,000,000	big



	carat	cut	color	clarity	depth	table	price	x	y	z
1	0.23	Ideal	E	SI2	61.5	55	326	3.95	3.98	2.43
2	0.21	Premium	E	SI1	59.8	61	326	3.89	3.84	2.31
3	0.23	Good	E	VS1	56.9	65	327	4.05	4.07	2.31
4	0.29	Premium	I	VS2	62.4	58	334	4.2	4.23	2.63
5	0.31	Good	J	SI2	63.3	58	335	4.34	4.35	2.75
6	0.24	Very Good	J	VVS2	62.8	57	336	3.94	3.96	2.48
7	0.24	Very Good	I	VVS1	62.3	57	336	3.95	3.98	2.47
8	0.26	Very Good	H	SI1	61.9	55	337	4.07	4.11	2.53
9	0.22	Fair	E	VS2	65.1	61	337	3.87	3.78	2.49
10	0.23	Very Good	H	VS1	59.4	61	338	4	4.05	2.39
11	0.3	Good	J	SI1	64	55	339	4.25	4.28	2.73
12	0.23	Ideal	J	VS1	62.8	56	340	3.93	3.9	2.46
13	0.22	Premium	F	SI1	60.4	61	342	3.88	3.84	2.33
14	0.31	Ideal	J	SI2	62.2	54	344	4.35	4.37	2.71
15	0.2	Premium	E	SI2	60.2	62	345	3.79	3.75	2.27
16	0.32	Premium	E	I1	60.9	58	345	4.38	4.42	2.68
17	0.3	Ideal	I	SI2	62	54	348	4.31	4.34	2.68
18	0.3	Good	J	SI1	63.4	54	351	4.23	4.29	2.7
19	0.3	Good	J	SI1	63.8	56	351	4.23	4.26	2.71
20	0.3	Very Good	J	SI1	62.7	59	351	4.21	4.27	2.66
21	0.3	Good	I	SI2	63.3	56	351	4.26	4.3	2.71
22	0.23	Very Good	E	VS2	63.8	55	352	3.85	3.92	2.48
23	0.23	Very Good	H	VS1	61	57	353	3.94	3.96	2.41
24	0.31	Very Good	J	SI1	59.4	62	353	4.39	4.43	2.62
25	0.31	Very Good	J	SI1	58.1	62	353	4.44	4.47	2.59
26	0.23	Very Good	G	VVS2	60.4	58	354	3.97	4.01	2.41
27	0.24	Premium	I	VS1	62.5	57	355	3.97	3.94	2.47
28	0.3	Very Good	J	VS2	62.2	57	357	4.28	4.3	2.67
29	0.23	Very Good	D	VS2	60.5	61	357	3.96	3.97	2.4
30	0.23	Very Good	F	VS1	60.9	57	357	3.96	3.99	2.42
31	0.23	Very Good	F	VS1	60	57	402	4	4.03	2.41
32	0.23	Very Good	F	VS1	59.8	57	402	4.04	4.06	2.42
33	0.23	Very Good	E	VS1	60.7	59	402	3.97	4.01	2.42
34	0.23	Very Good	E	VS1	59.5	58	402	4.01	4.06	2.4
35	0.23	Very Good	D	VS1	61.9	58	402	3.92	3.96	2.44
36	0.23	Good	F	VS1	58.2	59	402	4.06	4.08	2.37
37	0.23	Good	E	VS1	64.1	59	402	3.83	3.85	2.46
38	0.31	Good	H	SI1	64	54	402	4.29	4.31	2.75
39	0.26	Very Good	D	VS2	60.8	59	403	4.13	4.16	2.52
40	0.33	Ideal	I	SI2	61.8	55	403	4.49	4.51	2.78
41	0.33	Ideal	I	SI2	61.2	56	403	4.49	4.5	2.75
42	0.33	Ideal	J	SI1	61.1	56	403	4.49	4.55	2.76

Spreadsheet-type datasets

Rows = subjects (n)

individuals
schools
school districts
counties
census tracts
...

The easiest data has $n \gg p$!

...but we ***can*** work with $p \gg n$ data

What's big data?

	< 100	very small
	100 – 10,000	small
n	10,000 – 1,000,000	medium
	1,000,000 – 100,000,000	large
	> 100,000,000	big

Columns = variables (p)

diamonds

diamonds.csv
source

Columns (10/0)
carat
cut
color
clarity
depth
table
price

Rows
All rows 53940
Selected 0
Excluded 0
Hidden 0
Labelled 0

	carat	cut	color	clarity	depth	table	price	x	y	z
1	0.23	Ideal	E	SI2	61.5	55	326	3.95	3.98	2.43
2	0.21	Premium	E	SI1	59.8	61	326	3.89	3.84	2.31
3	0.23	Good	E	VS1	56.9	65	327	4.05	4.07	2.31
4	0.29	Premium	I	VS2	62.4	58	334	4.2	4.23	2.63
5	0.31	Good	J	SI2	63.3	58	335	4.34	4.35	2.75
6	0.24	Very Good	J	VVS2	62.8	57	336	3.94	3.96	2.48
7	0.24	Very Good	I	VVS1	62.3	57	336	3.95	3.98	2.47
8	0.26	Very Good	H	SI1	61.9	55	337	4.07	4.11	2.53
9	0.22	Fair	E	VS2	65.1	61	337	3.87	3.78	2.49
10	0.23	Very Good	H	VS1	59.4	61	338	4	4.05	2.39
11	0.3	Good	J	SI1	64	55	339	4.25	4.28	2.73
12	0.23	Ideal	J	VS1	62.8	56	340	3.93	3.9	2.46
13	0.22	Premium	F	SI1	60.4	61	342	3.88	3.84	2.33
14	0.31	Ideal	J	SI2	62.2	54	344	4.35	4.37	2.71
15	0.2	Premium	E	SI2	60.2	62	345	3.79	3.75	2.27
16	0.32	Premium	E	I1	60.9	58	345	4.38	4.42	2.68
17	0.3	Ideal	I	SI2	62	54	348	4.31	4.34	2.68
18	0.3	Good	J	SI1	63.4	54	351	4.23	4.29	2.7
19	0.3	Good	J	SI1	63.8	56	351	4.23	4.26	2.71
20	0.3	Very Good	J	SI1	62.7	59	351	4.21	4.27	2.66
21	0.3	Good	I	SI2	63.3	56	351	4.26	4.3	2.71
22	0.23	Very Good	E	VS2	63.8	55	352	3.85	3.92	2.48
23	0.23	Very Good	H	VS1	61	57	353	3.94	3.96	2.41
24	0.31	Very Good	J	SI2	59.4	62	353	4.39	4.43	2.62
25	0.31	Very Good	J	SI1	58.1	62	353	4.44	4.47	2.59
26	0.23	Very Good	G	VS1	60.4	58	354	3.97	4.01	2.41
27	0.24	Premium	I	VS1	62.5	57	355	3.97	4	2.47
28	0.3	Very Good	J	VS2	62.2	57	357	4.28	4.3	2.67
29	0.23	Very Good	D	VS2	60.5	58	357	3.96	3.97	2.4
30	0.23	Very Good	H	VS1	60.9	57	357	3.96	3.99	2.42
31	0.23	Very Good	F	VS1	60.9	57	402	4	4.03	2.41
32	0.23	Very Good	F	VS1	59.8	57	402	4.04	4.06	2.42
33	0.23	Very Good	E	VS1	60.7	59	402	3.97	4.01	2.42
34	0.23	Very Good	E	VS1	59.5	58	402	4.01	4.06	2.4
35	0.23	Very Good	D	VS1	61.9	58	402	3.92	3.96	2.44
36	0.23	Good	F	VS1	58.2	59	402	4.06	4.08	2.37
37	0.23	Good	E	VS1	64.1	59	402	3.83	3.85	2.46
38	0.31	Good	H	SI1	64	54	402	4.29	4.31	2.75
39	0.26	Very Good	D	VS2	60.8	59	403	4.13	4.16	2.52
40	0.33	Ideal	I	SI2	61.8	55	403	4.49	4.51	2.78
41	0.33	Ideal	I	SI2	61.2	56	403	4.49	4.5	2.75
42	0.33	Ideal	J	SI1	61.1	56	403	4.49	4.55	2.76

Efforts are being made to keep these computations under 5 seconds.

(On a standard desktop computer.)

Spreadsheet-type datasets

Rows = subjects (n)

individuals
schools
school districts
counties
census tracts
...

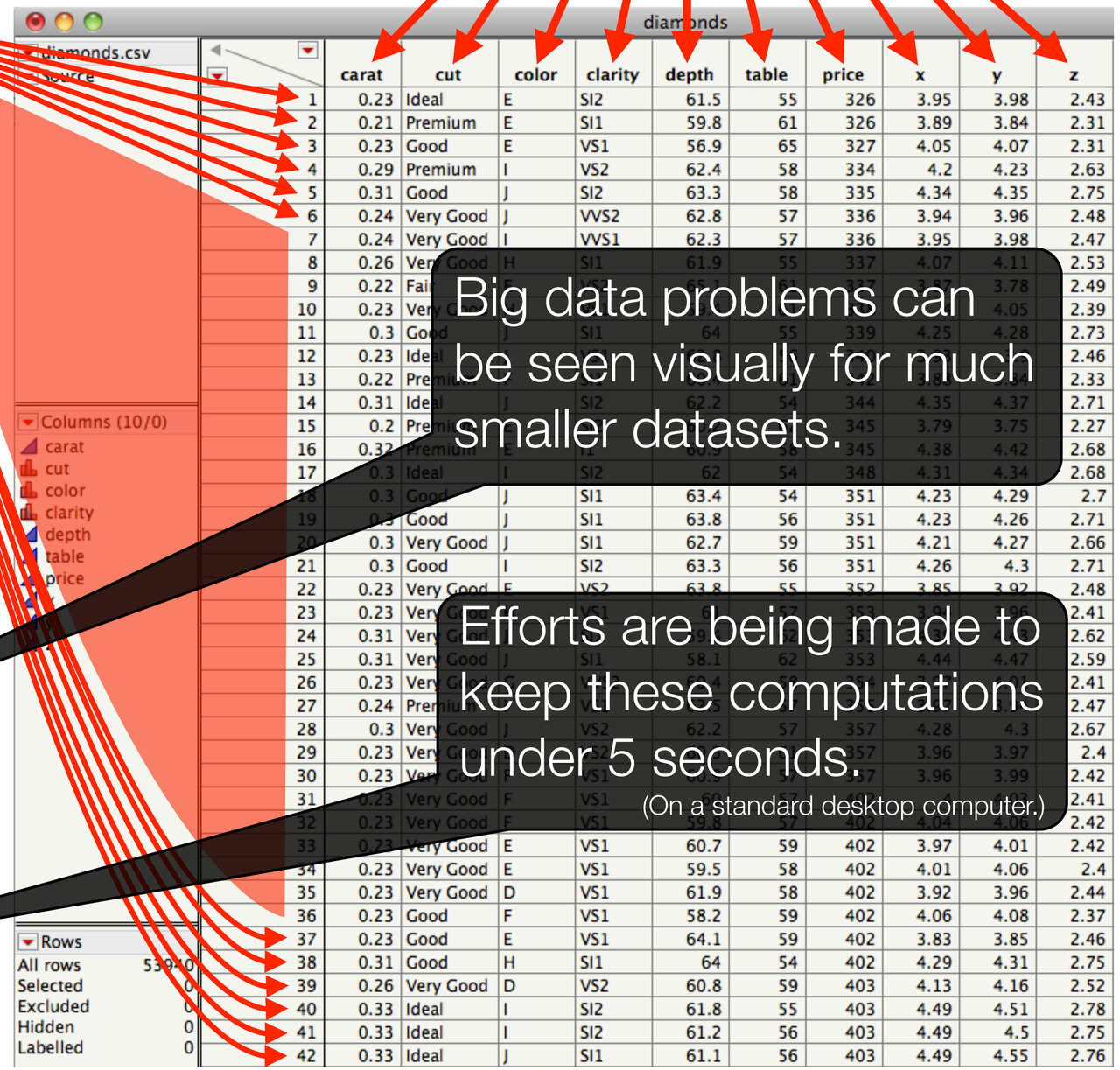
Columns = variables (p)

The easiest data has $n \gg p$!

...but we ***can*** work with $p \gg n$ data

What's big data?

n	< 100	very small
	100 – 10,000	small
	10,000 – 1,000,000	medium
	1,000,000 – 100,000,000	large
	> 100,000,000	big



Big data problems can be seen visually for much smaller datasets.

Efforts are being made to keep these computations under 5 seconds.
(On a standard desktop computer.)

Spreadsheet-type datasets

Rows = subjects (n)

individuals
schools
school districts
counties
census tracts
...

Columns = variables (p)

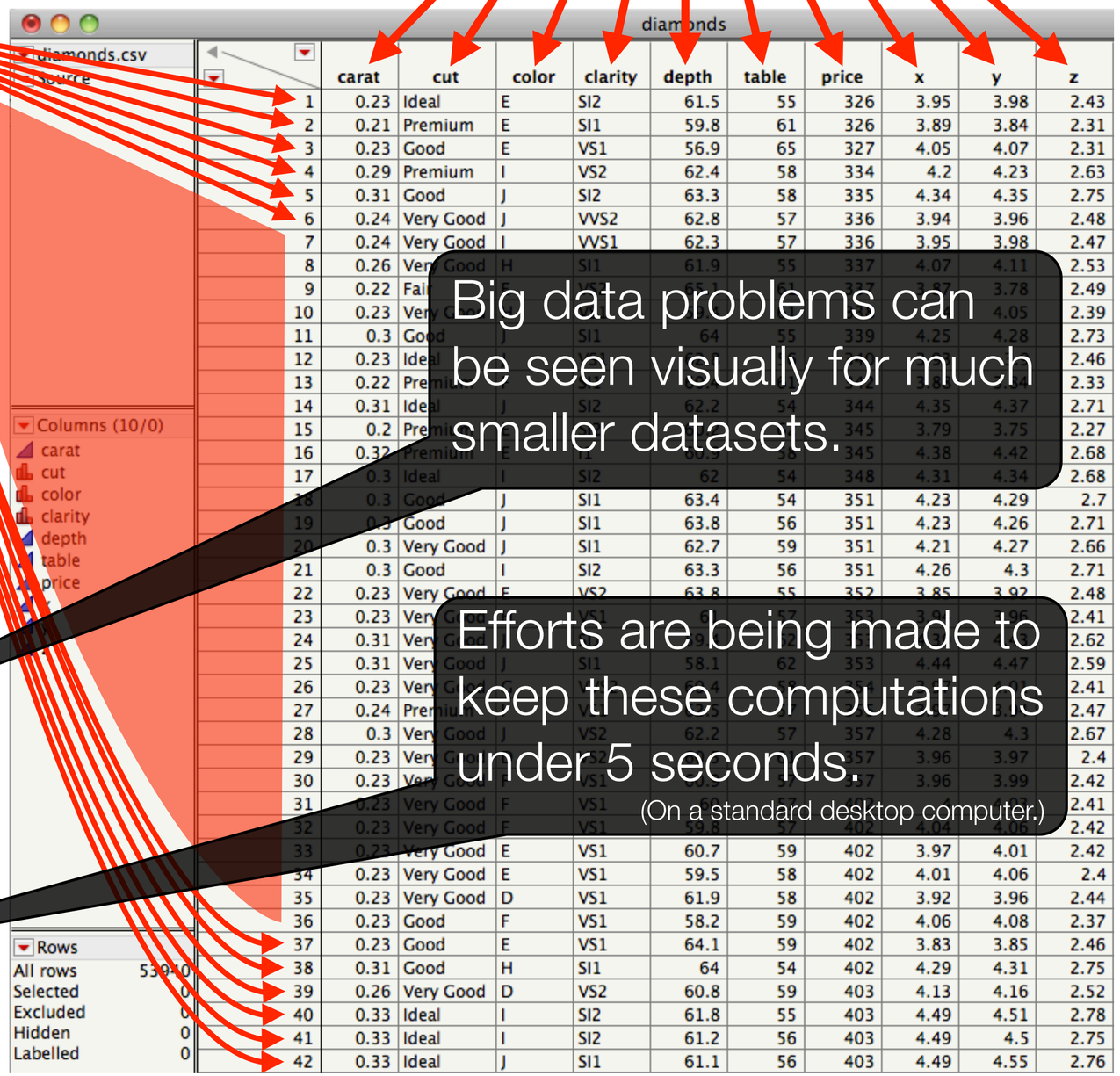
The easiest data has $n \gg p$!

...but we ***can*** work with $p \gg n$ data

What's big data?

This dataset has $n = 55$k
Free in R's ggplot2 package

n	< 100	very small
	100 – 10,000	small
	10,000 – 1,000,000	medium
	1,000,000 – 100,000,000	large
	> 100,000,000	big



Big data problems can be seen visually for much smaller datasets.

Efforts are being made to keep these computations under 5 seconds.
(On a standard desktop computer.)

36 Diamond Carat Weights (sorted)

0.26	0.30	0.31	0.31	0.31	0.31
0.35	0.36	0.36	0.50	0.51	0.53
0.54	0.55	0.59	0.61	0.70	0.70
0.70	0.71	0.71	0.71	0.74	0.87
1.00	1.00	1.01	1.02	1.08	1.21
1.47	1.70	2.01	2.01	2.03	2.43

36 Diamond Carat Weights (sorted)

0.26	0.30	0.31	0.31	0.31	0.31
0.35	0.36	0.36	0.50	0.51	0.53
0.54	0.55	0.59	0.61	0.70	0.70
0.70	0.71	0.71	0.71	0.74	0.87
1.00	1.00	1.01	1.02	1.08	1.21
1.47	1.70	2.01	2.01	2.03	2.43

Q : How can we visualize the dataset?

1d graphics

36 Diamond Carat Weights (sorted)

0.26	0.30	0.31	0.31	0.31	0.31
0.35	0.36	0.36	0.50	0.51	0.53
0.54	0.55	0.59	0.61	0.70	0.70
0.70	0.71	0.71	0.71	0.74	0.87
1.00	1.00	1.01	1.02	1.08	1.21
1.47	1.70	2.01	2.01	2.03	2.43

Q : How can we visualize the dataset?

1d graphics

Typical starting point →

What's the problem with this visualization?

36 Diamond Carat Weights (sorted)

0.26	0.30	0.31	0.31	0.31	0.31
0.35	0.36	0.36	0.50	0.51	0.53
0.54	0.55	0.59	0.61	0.70	0.70
0.70	0.71	0.71	0.71	0.74	0.87
1.00	1.00	1.01	1.02	1.08	1.21
1.47	1.70	2.01	2.01	2.03	2.43

Q : How can we visualize the dataset?

1d graphics

Typical starting point →

What's the problem with this visualization?

Overplotting : visual confusion caused by plotting too much data.

36 Diamond Carat Weights (sorted)

0.26	0.30	0.31	0.31	0.31	0.31
0.35	0.36	0.36	0.50	0.51	0.53
0.54	0.55	0.59	0.61	0.70	0.70
0.70	0.71	0.71	0.71	0.74	0.87
1.00	1.00	1.01	1.02	1.08	1.21
1.47	1.70	2.01	2.01	2.03	2.43

Q : How can we visualize the dataset?

1d graphics

Typical starting point →

What's the problem with this visualization?

Overplotting : visual confusion caused by plotting too much data.

Solicit solutions from students!

36 Diamond Carat Weights (sorted)

0.26	0.30	0.31	0.31	0.31	0.31
0.35	0.36	0.36	0.50	0.51	0.53
0.54	0.55	0.59	0.61	0.70	0.70
0.70	0.71	0.71	0.71	0.74	0.87
1.00	1.00	1.01	1.02	1.08	1.21
1.47	1.70	2.01	2.01	2.03	2.43

Q : How can we visualize the dataset?

1d graphics

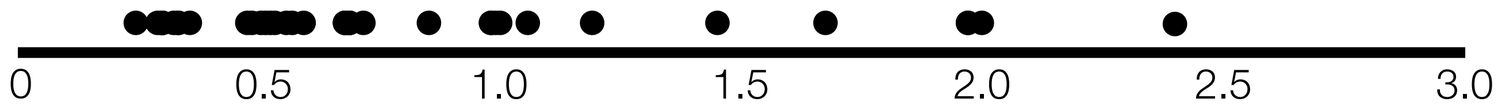
Typical starting point →

What's the problem with this visualization?

Overplotting : visual confusion caused by plotting too much data.

Solicit solutions from students!

Offset from line, resize points



36 Diamond Carat Weights (sorted)

0.26	0.30	0.31	0.31	0.31	0.31
0.35	0.36	0.36	0.50	0.51	0.53
0.54	0.55	0.59	0.61	0.70	0.70
0.70	0.71	0.71	0.71	0.74	0.87
1.00	1.00	1.01	1.02	1.08	1.21
1.47	1.70	2.01	2.01	2.03	2.43

Q : How can we visualize the dataset?

1d graphics

Typical starting point →

What's the problem with this visualization?

Overplotting : visual confusion caused by plotting too much data.

36 Diamond Carat Weights (sorted)

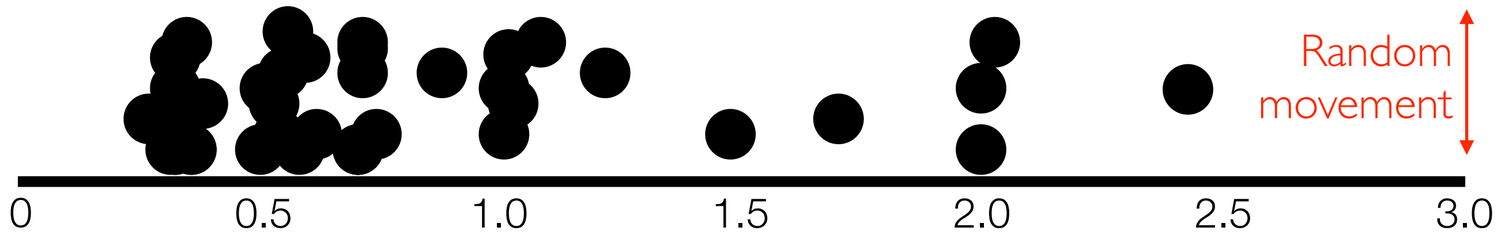
0.26	0.30	0.31	0.31	0.31	0.31
0.35	0.36	0.36	0.50	0.51	0.53
0.54	0.55	0.59	0.61	0.70	0.70
0.70	0.71	0.71	0.71	0.74	0.87
1.00	1.00	1.01	1.02	1.08	1.21
1.47	1.70	2.01	2.01	2.03	2.43

Q : How can we visualize the dataset?

Solicit solutions from students!

Offset from line, resize points

Jitter points



1d graphics

Typical starting point →

What's the problem with this visualization?

Overplotting : visual confusion caused by plotting too much data.

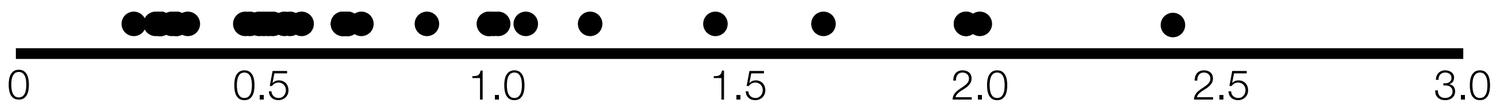
36 Diamond Carat Weights (sorted)

0.26	0.30	0.31	0.31	0.31	0.31
0.35	0.36	0.36	0.50	0.51	0.53
0.54	0.55	0.59	0.61	0.70	0.70
0.70	0.71	0.71	0.71	0.74	0.87
1.00	1.00	1.01	1.02	1.08	1.21
1.47	1.70	2.01	2.01	2.03	2.43

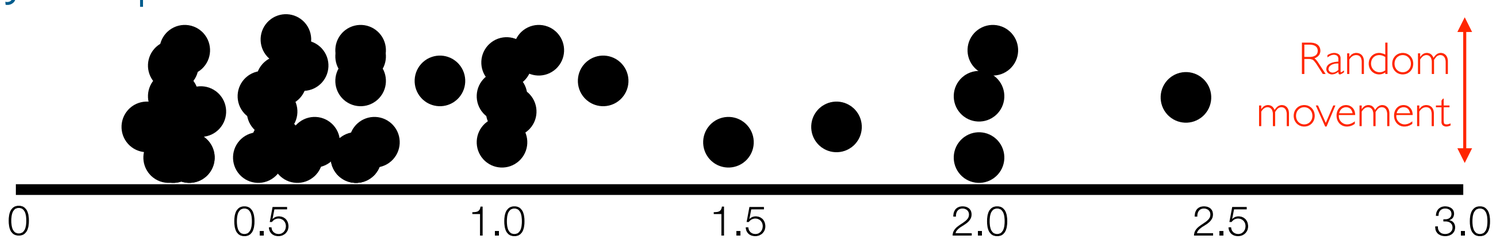
Q : How can we visualize the dataset?

Solicit solutions from students!

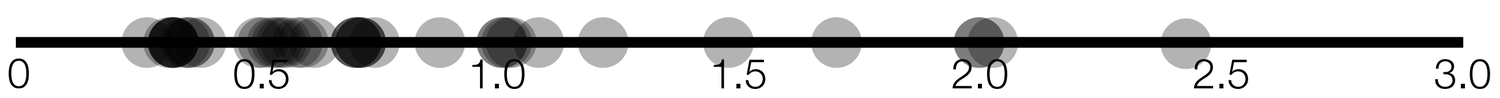
Offset from line, resize points



Jitter points



Alpha blend the points (make them semi-transparent)



1d graphics

Typical starting point →

What's the problem with this visualization?

Overplotting : visual confusion caused by plotting too much data.

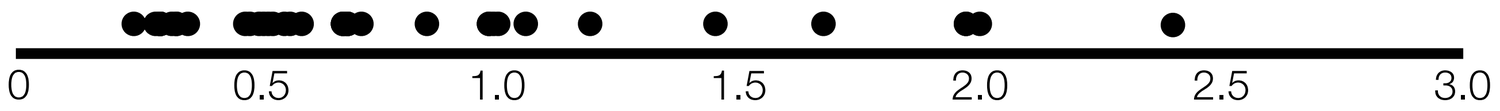
36 Diamond Carat Weights (sorted)

0.26	0.30	0.31	0.31	0.31	0.31
0.35	0.36	0.36	0.50	0.51	0.53
0.54	0.55	0.59	0.61	0.70	0.70
0.70	0.71	0.71	0.71	0.74	0.87
1.00	1.00	1.01	1.02	1.08	1.21
1.47	1.70	2.01	2.01	2.03	2.43

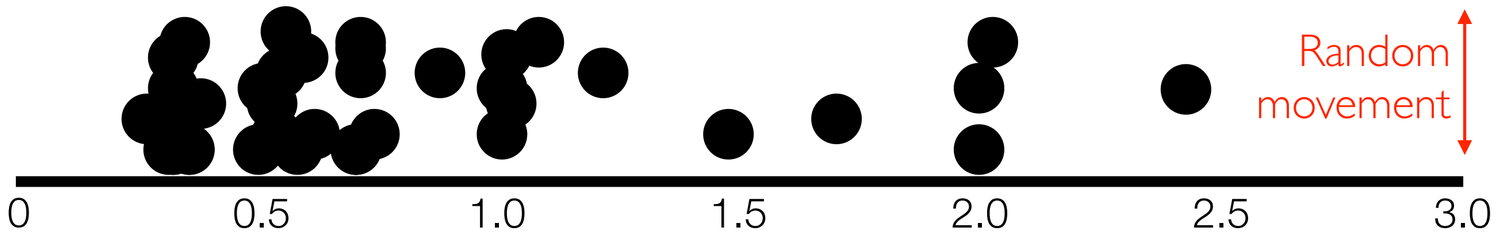
Q : How can we visualize the dataset?

Solicit solutions from students!

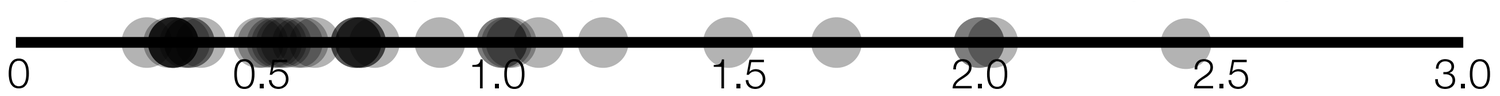
Offset from line, resize points



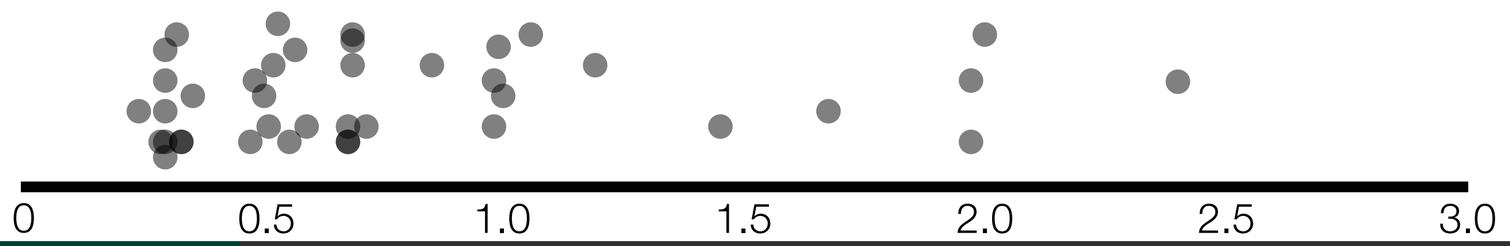
Jitter points



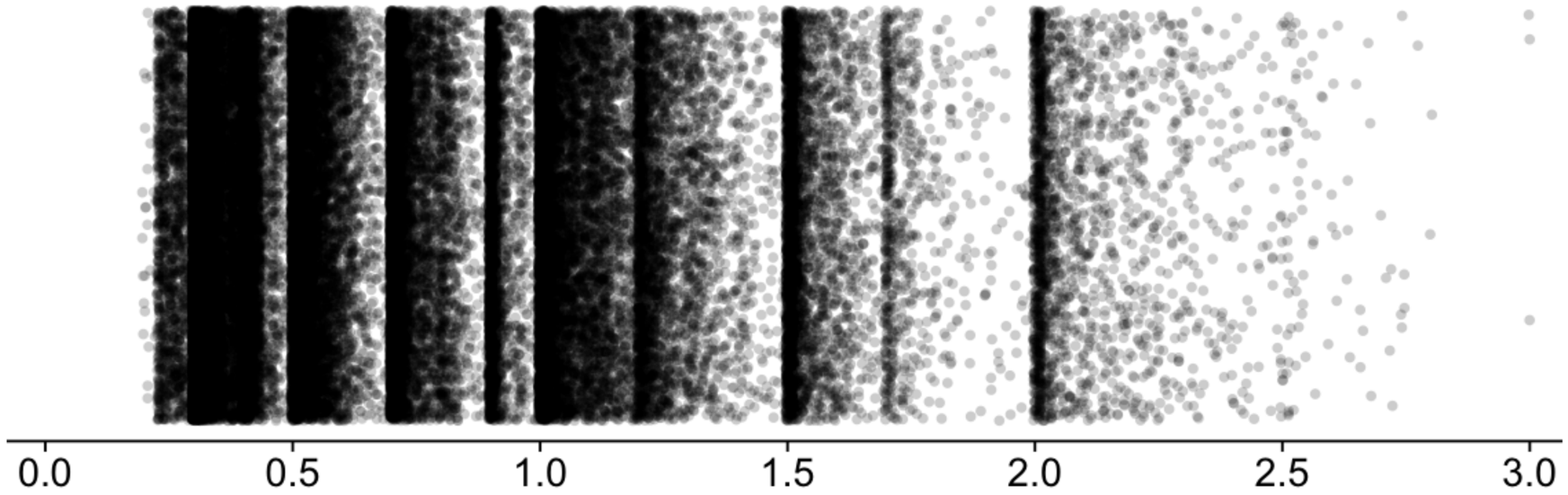
Alpha blend the points (make them semi-transparent)



Combo

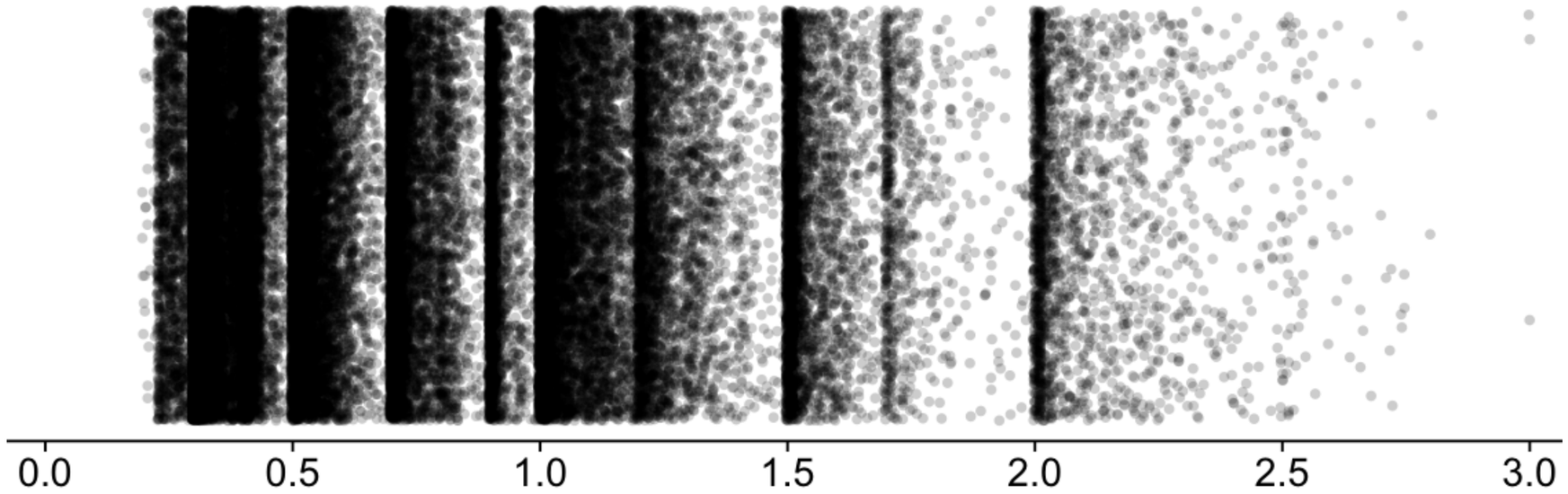


1d graphics



Problem : for the entire dataset of 55,000 diamonds, we still can't get away from the problem.

1d graphics

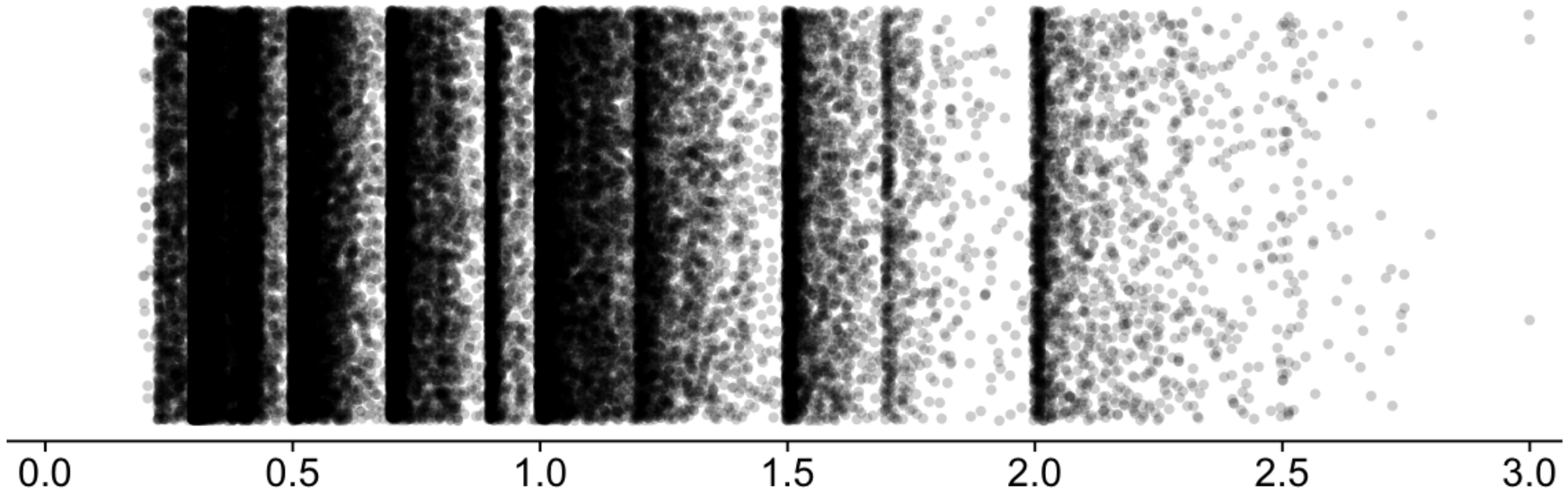


Problem : for the entire dataset of 55,000 diamonds, we still can't get away from the problem.

"Solution" : Summarize dataset with bins as histogram

...or a density estimate, boxplot, etc.

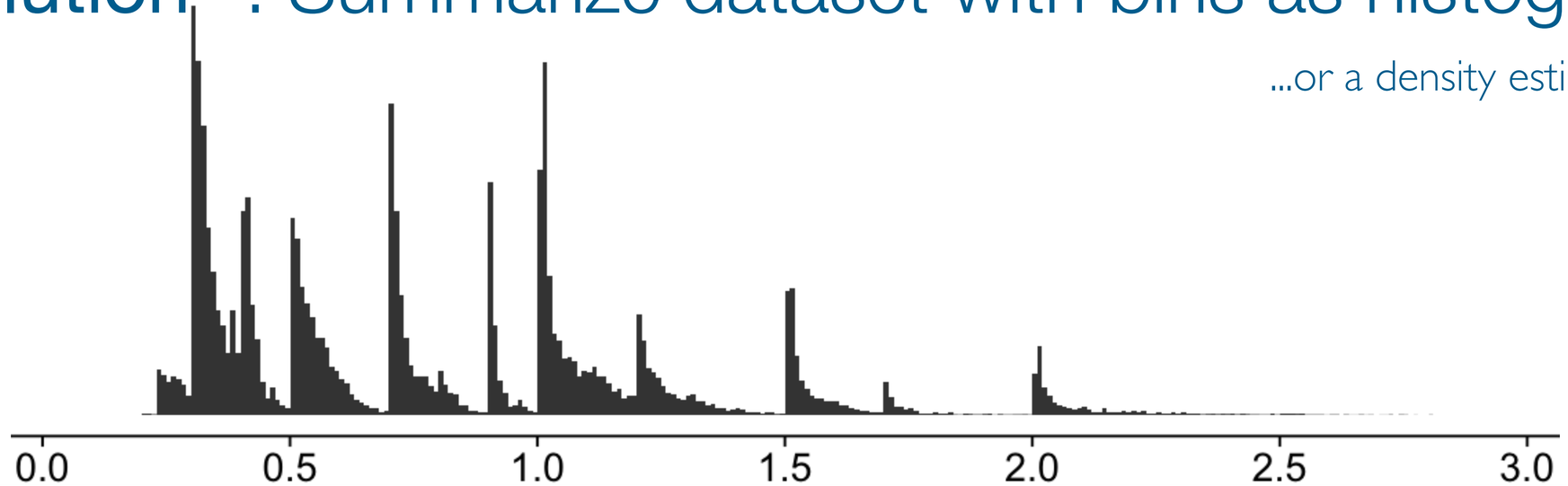
1d graphics



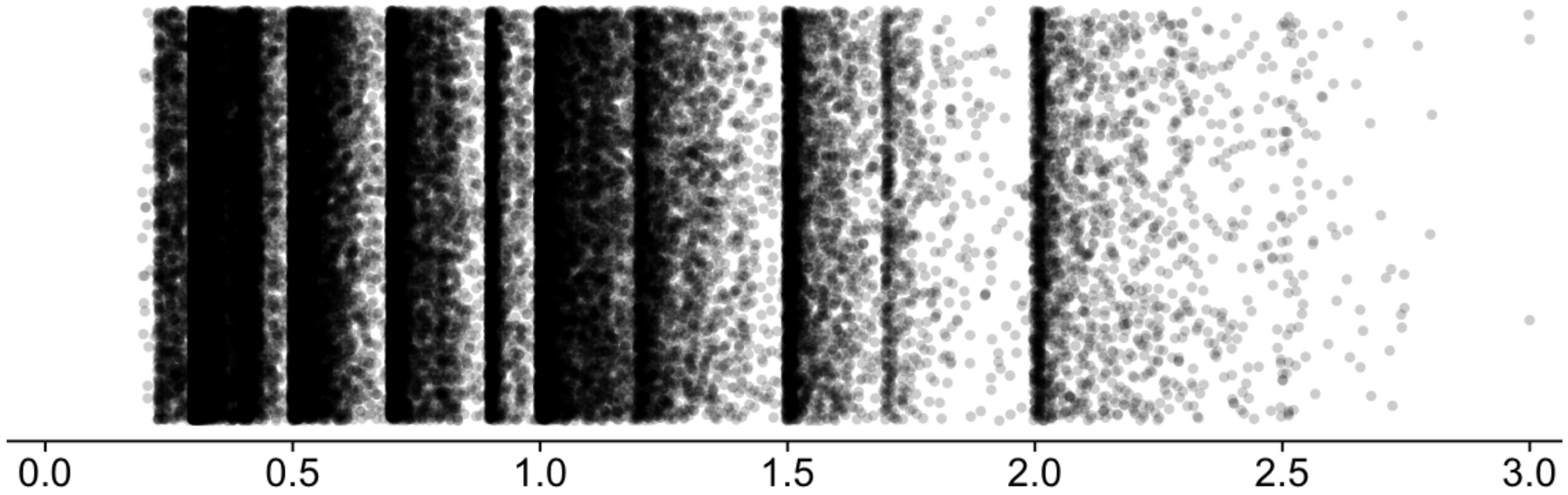
Problem : for the entire dataset of 55,000 diamonds, we still can't get away from the problem.

"Solution" : Summarize dataset with bins as histogram

...or a density estimate, boxplot, etc.



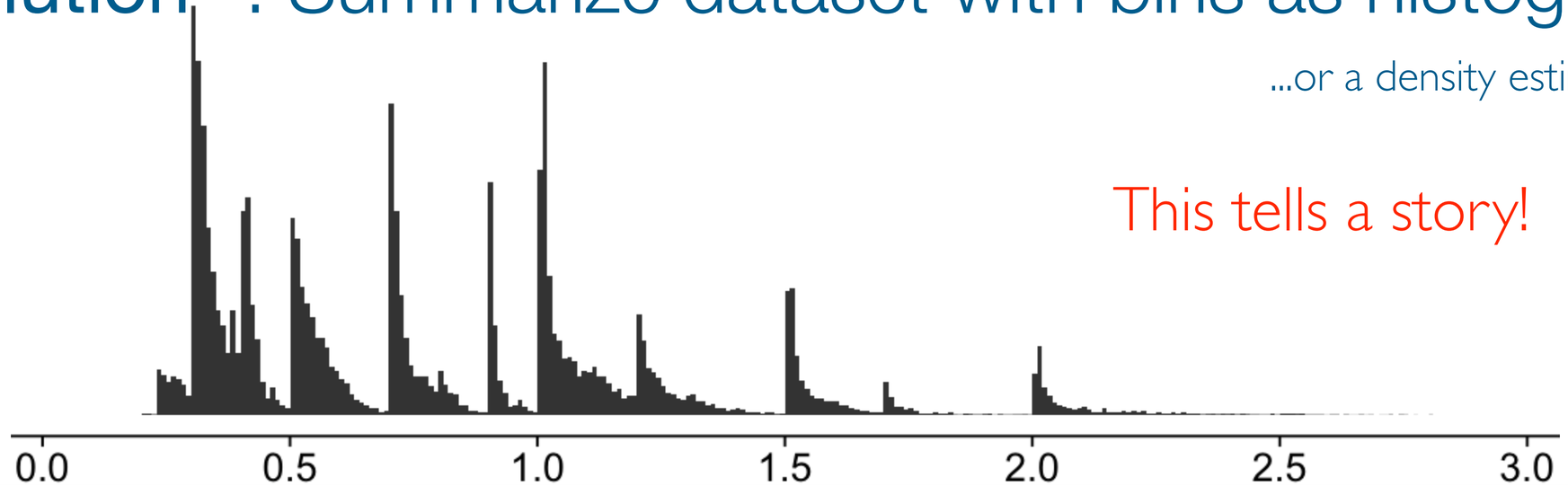
1d graphics



Problem : for the entire dataset of 55,000 diamonds, we still can't get away from the problem.

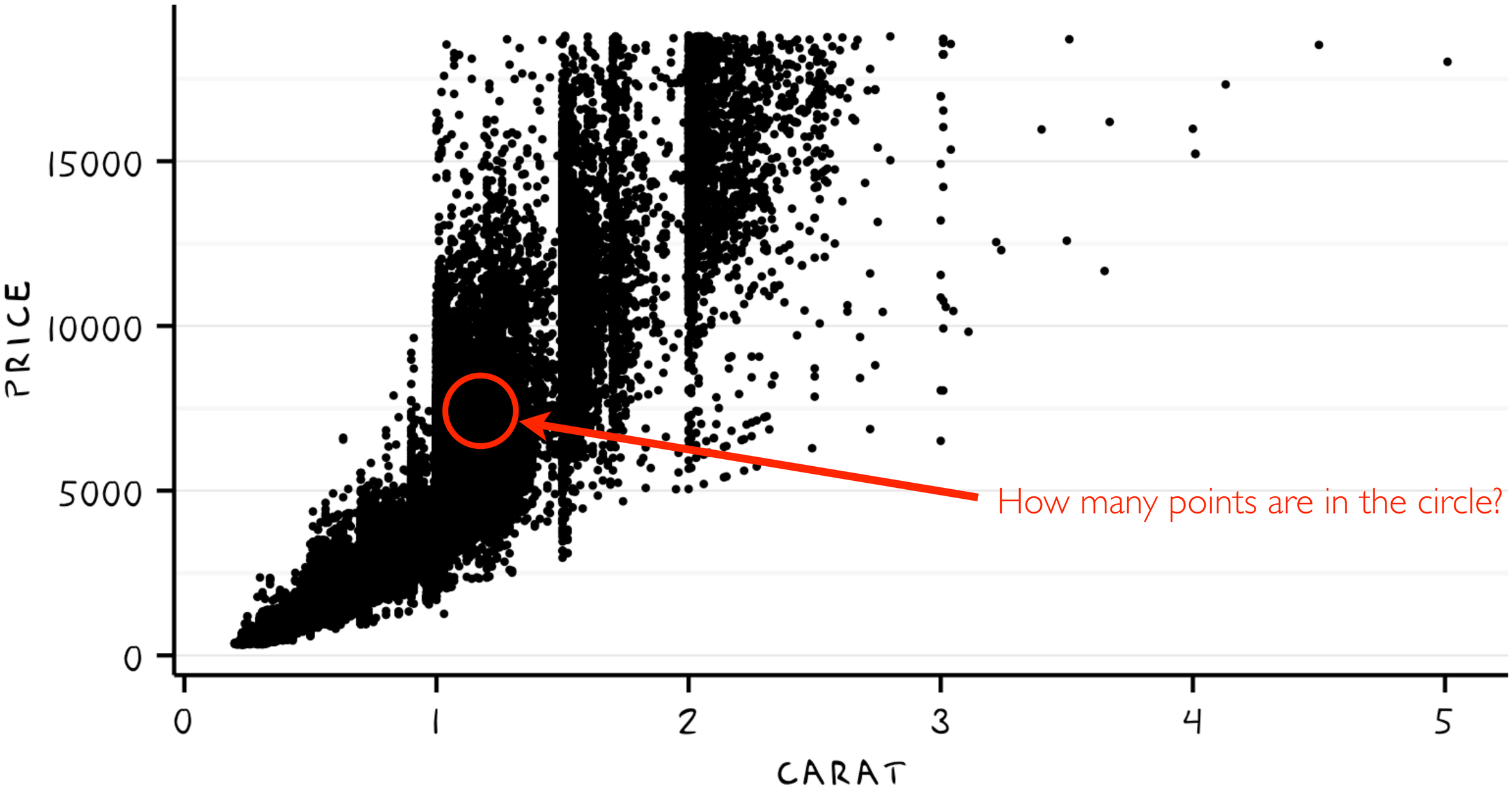
"Solution" : Summarize dataset with bins as histogram

...or a density estimate, boxplot, etc.

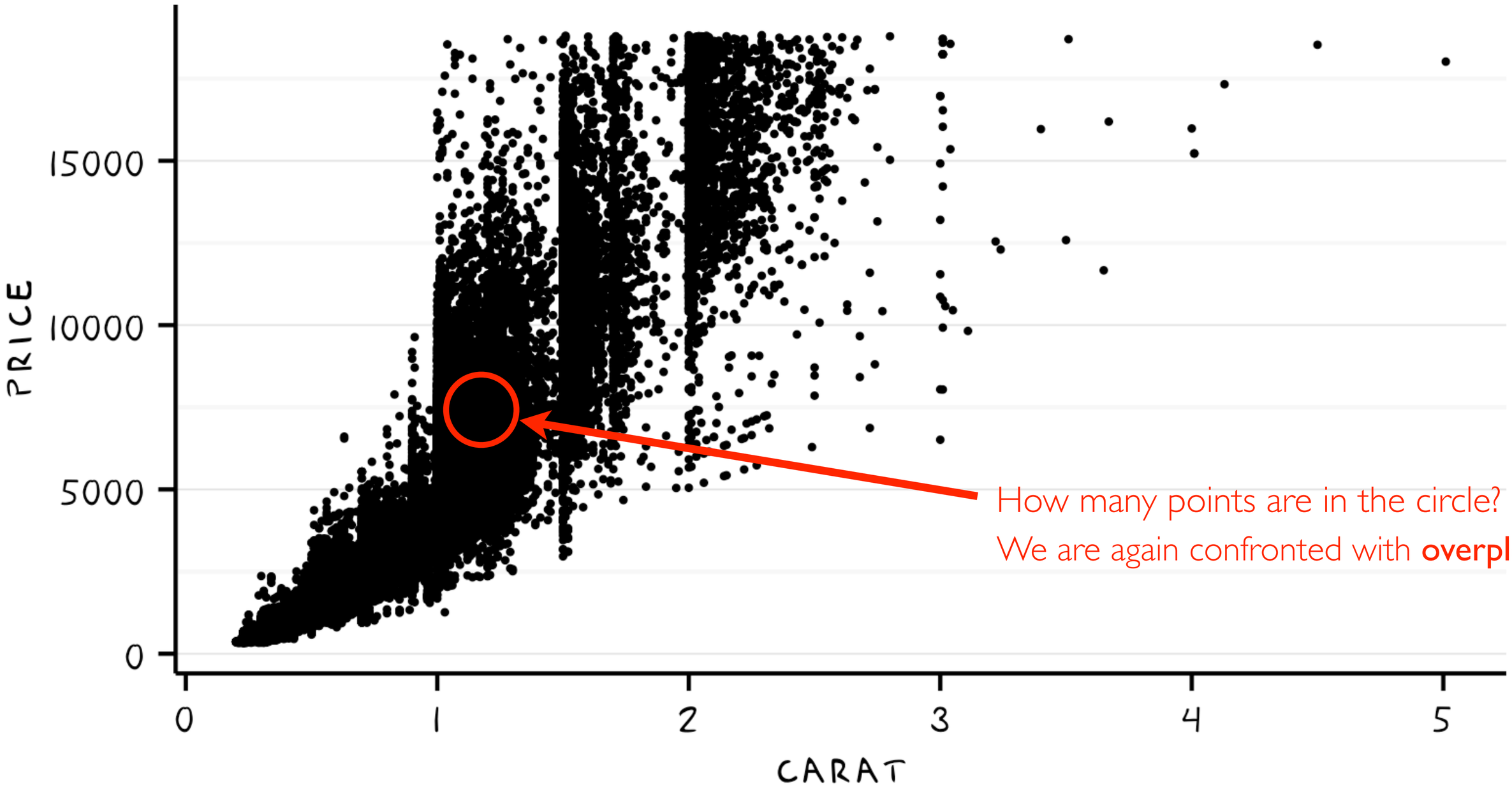


This tells a story!

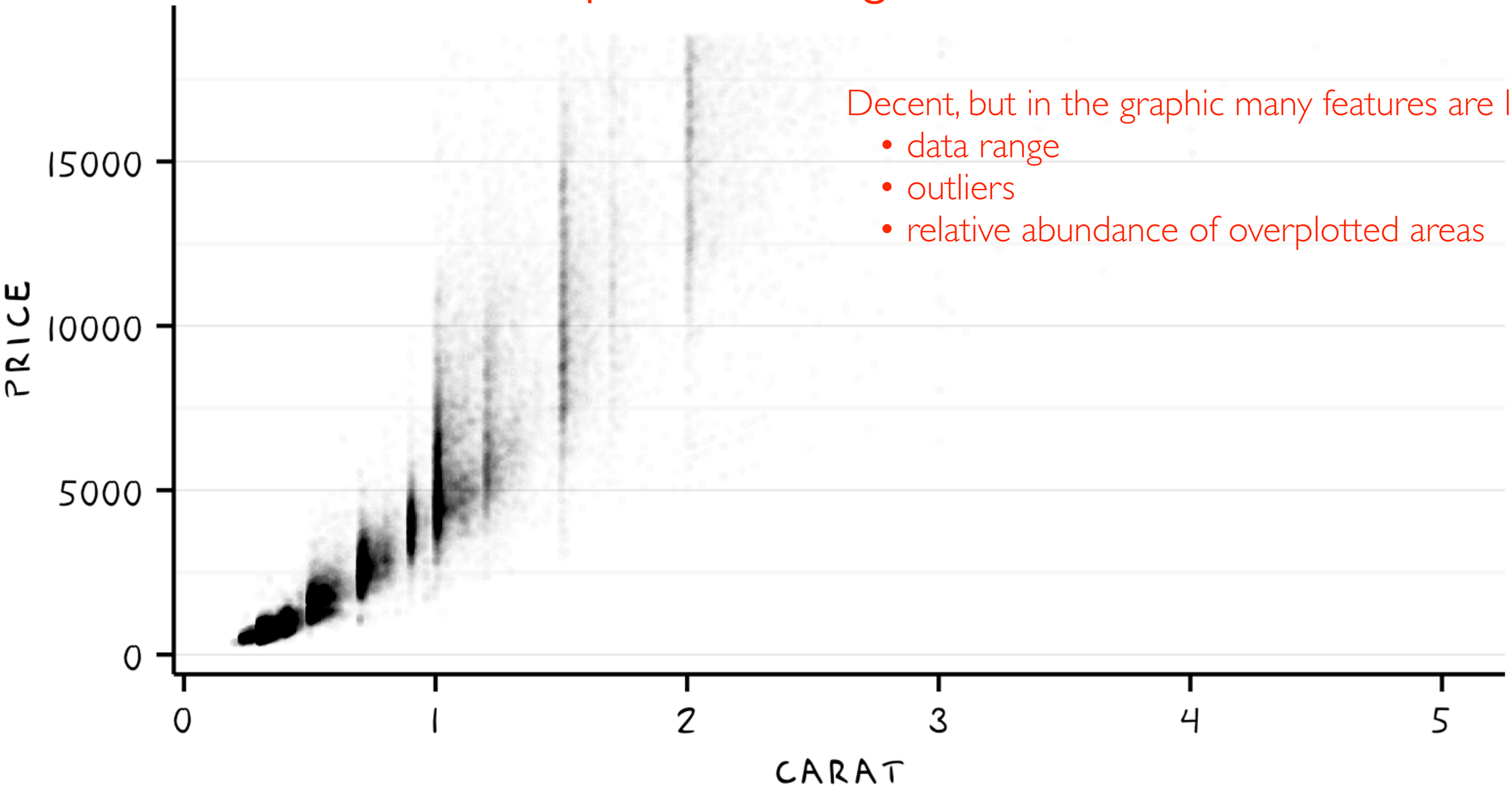
2d graphics


15000
10000
5000
0
PRICE
0
1
2
3
4
5
CARAT
How many points are in the circle?

2d graphics


15000
10000
5000
0
PRICE
0
1
2
3
4
5
CARAT
How many points are in the circle?
We are again confronted with **overplotting**

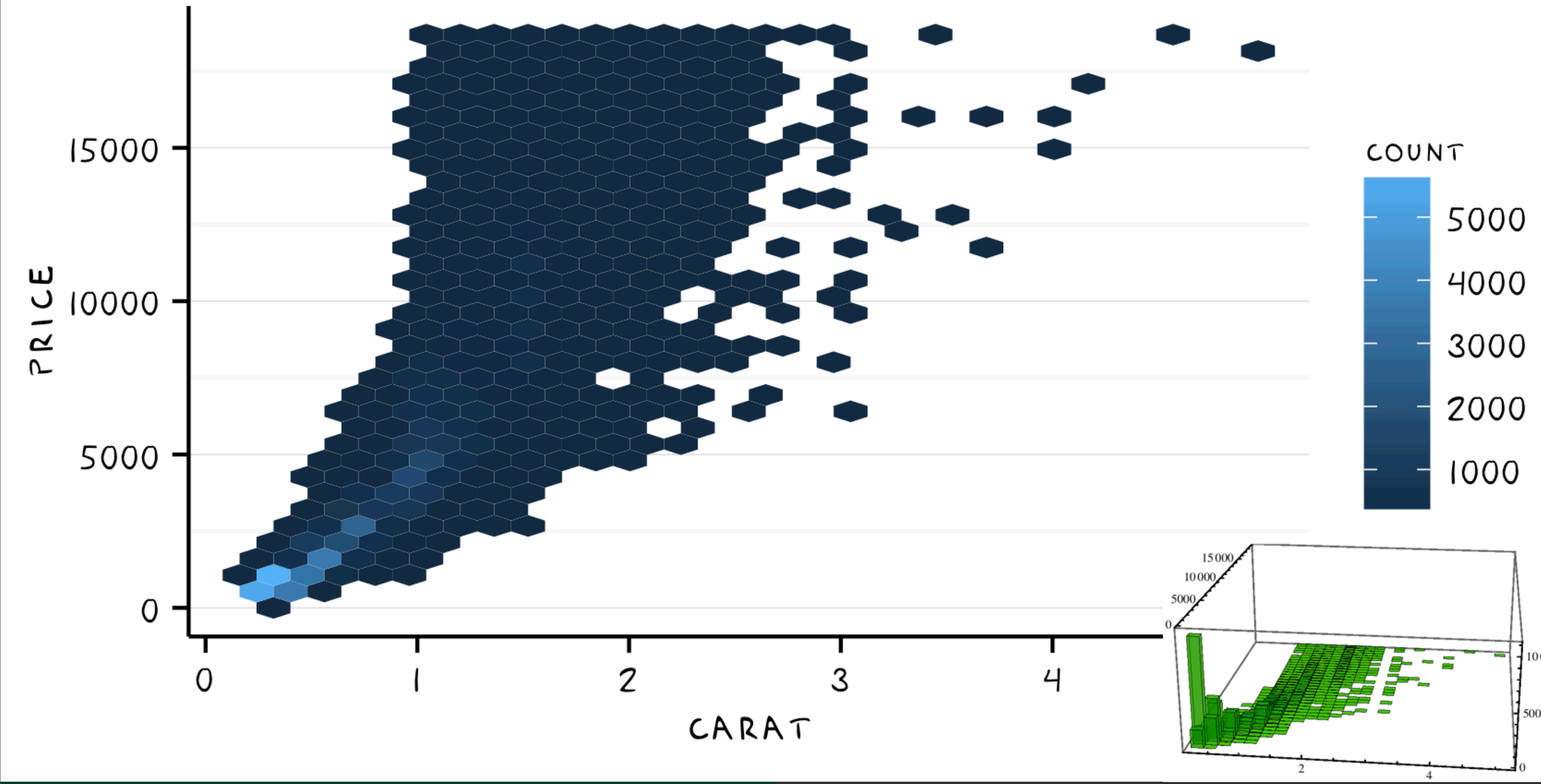
alpha blending = 100



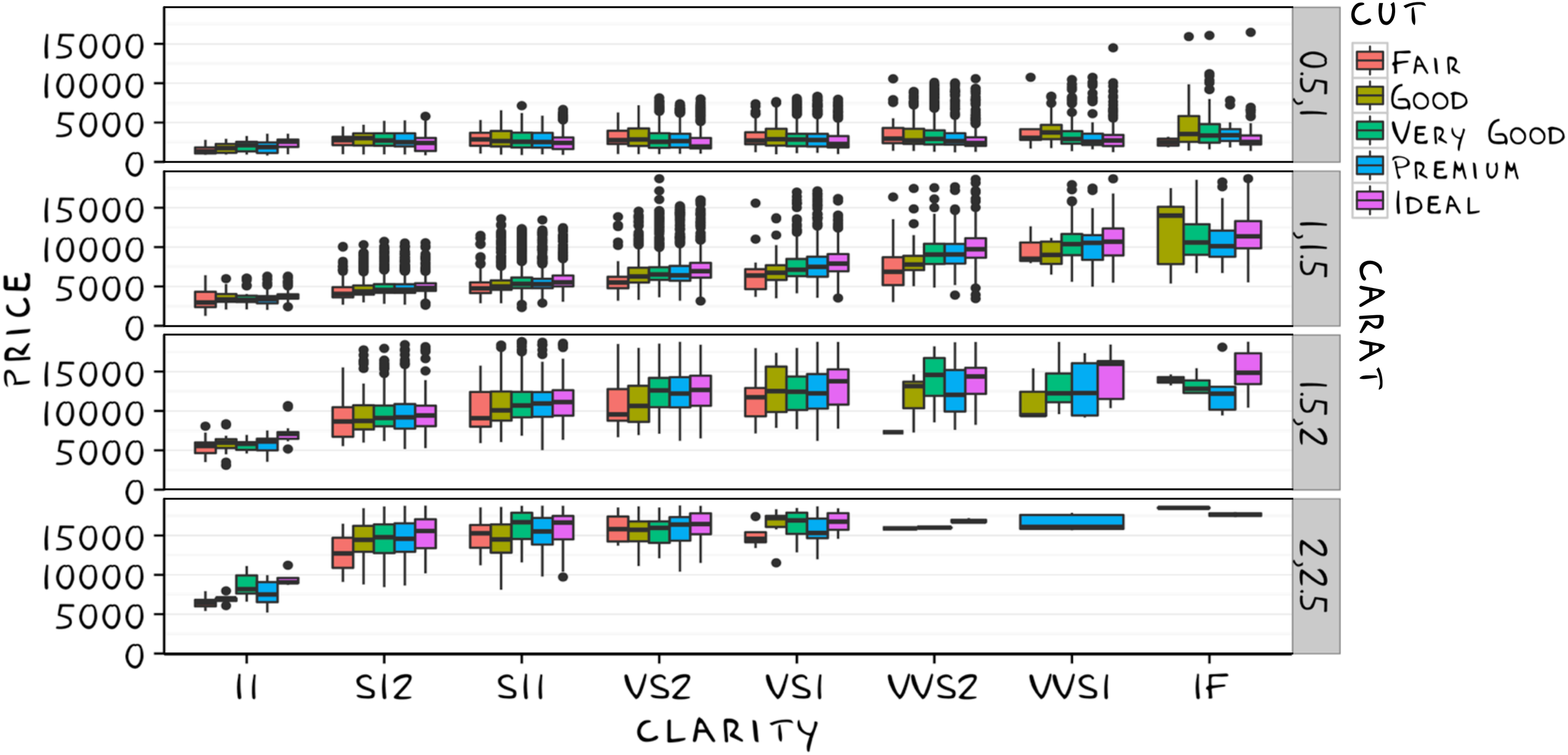
Decent, but in the graphic many features are lost

- data range
- outliers
- relative abundance of overplotted areas

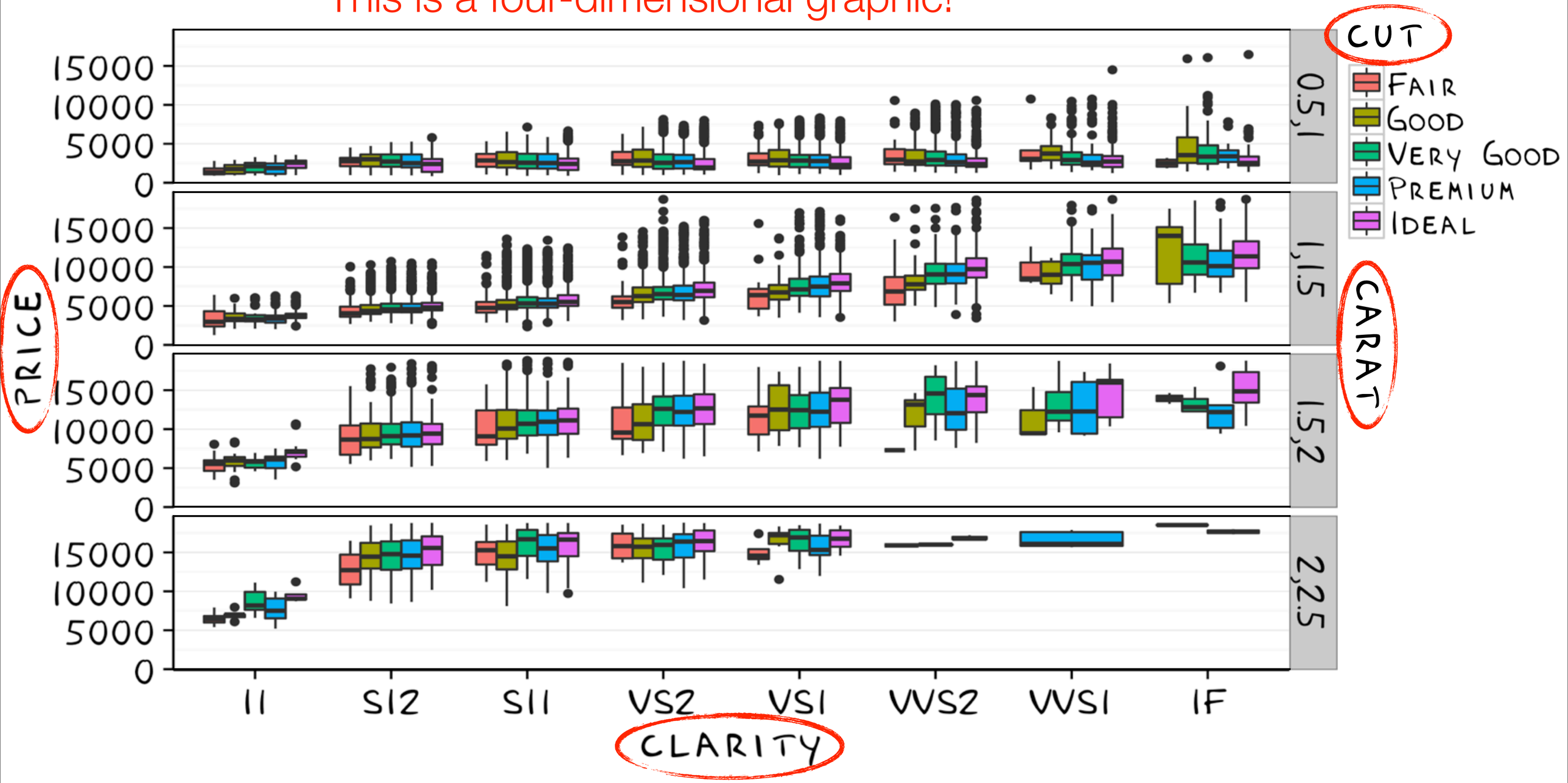
2d graphics


PRICE
15000
10000
5000
0
0
1
2
3
4
CARAT
COUNT
5000
4000
3000
2000
1000
15000
10000
5000
0
2
4
10000
5000
0

pd graphics – Faceting Conditional plots on discrete variables

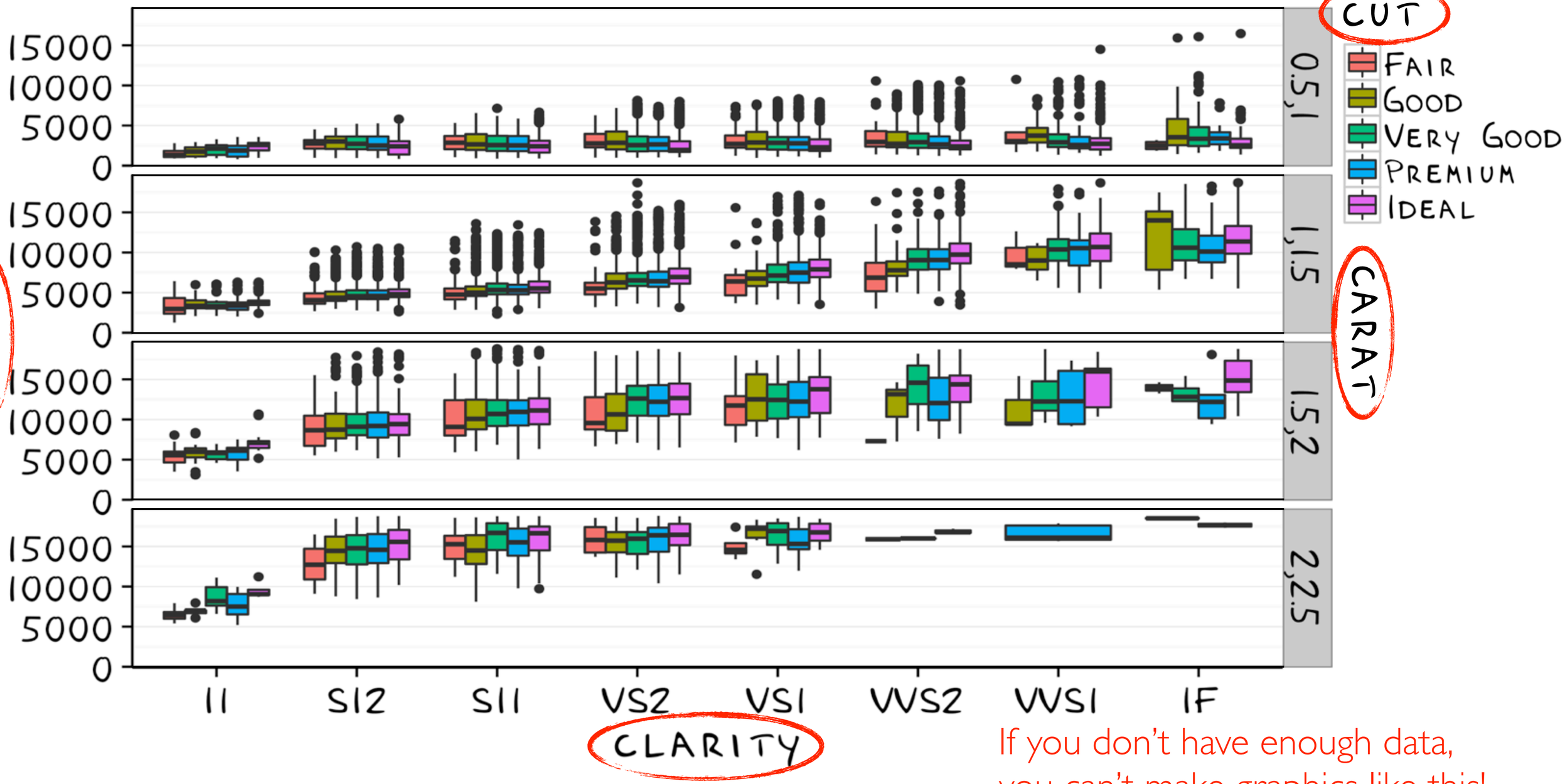


This is a four-dimensional graphic!

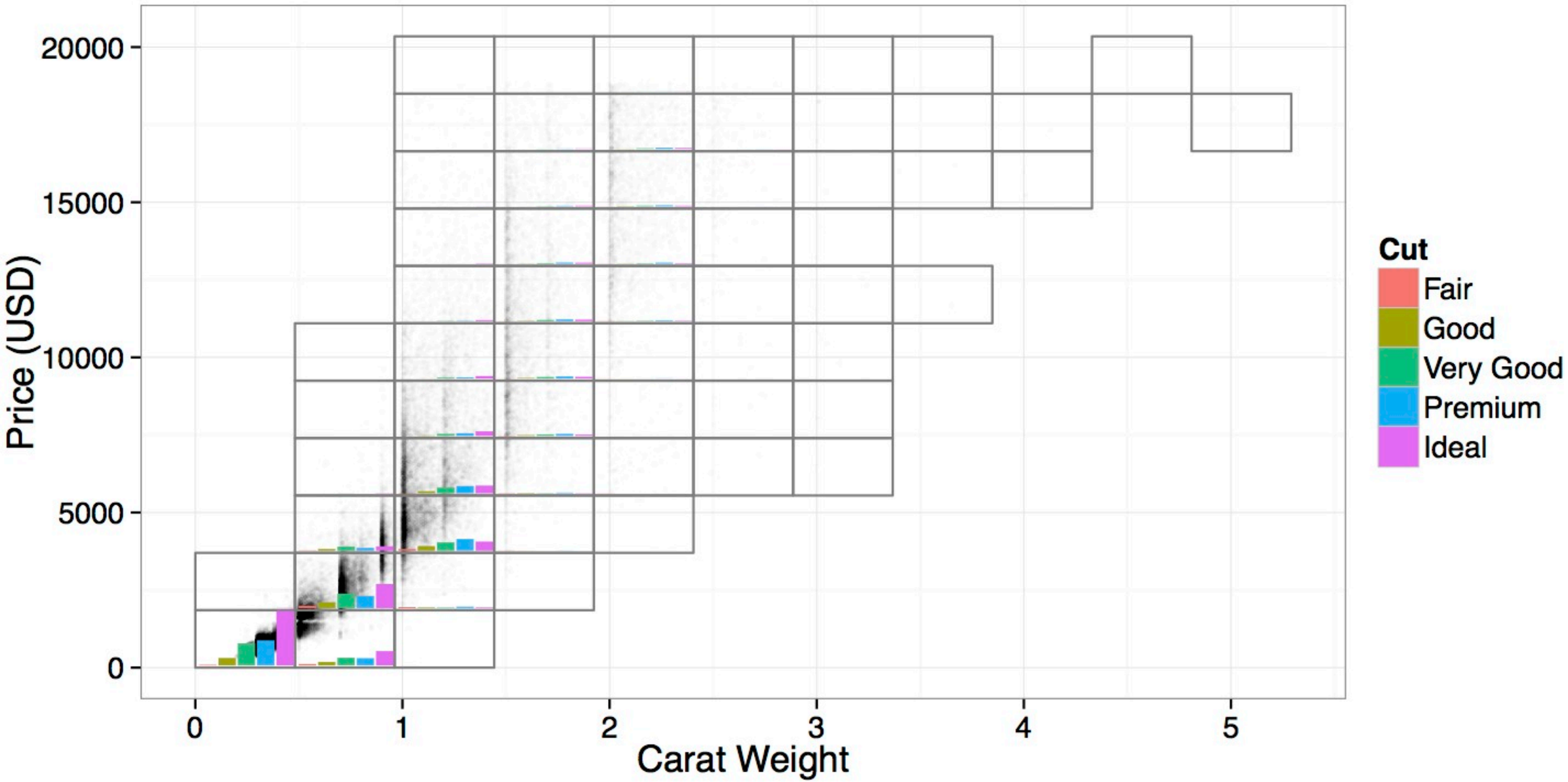


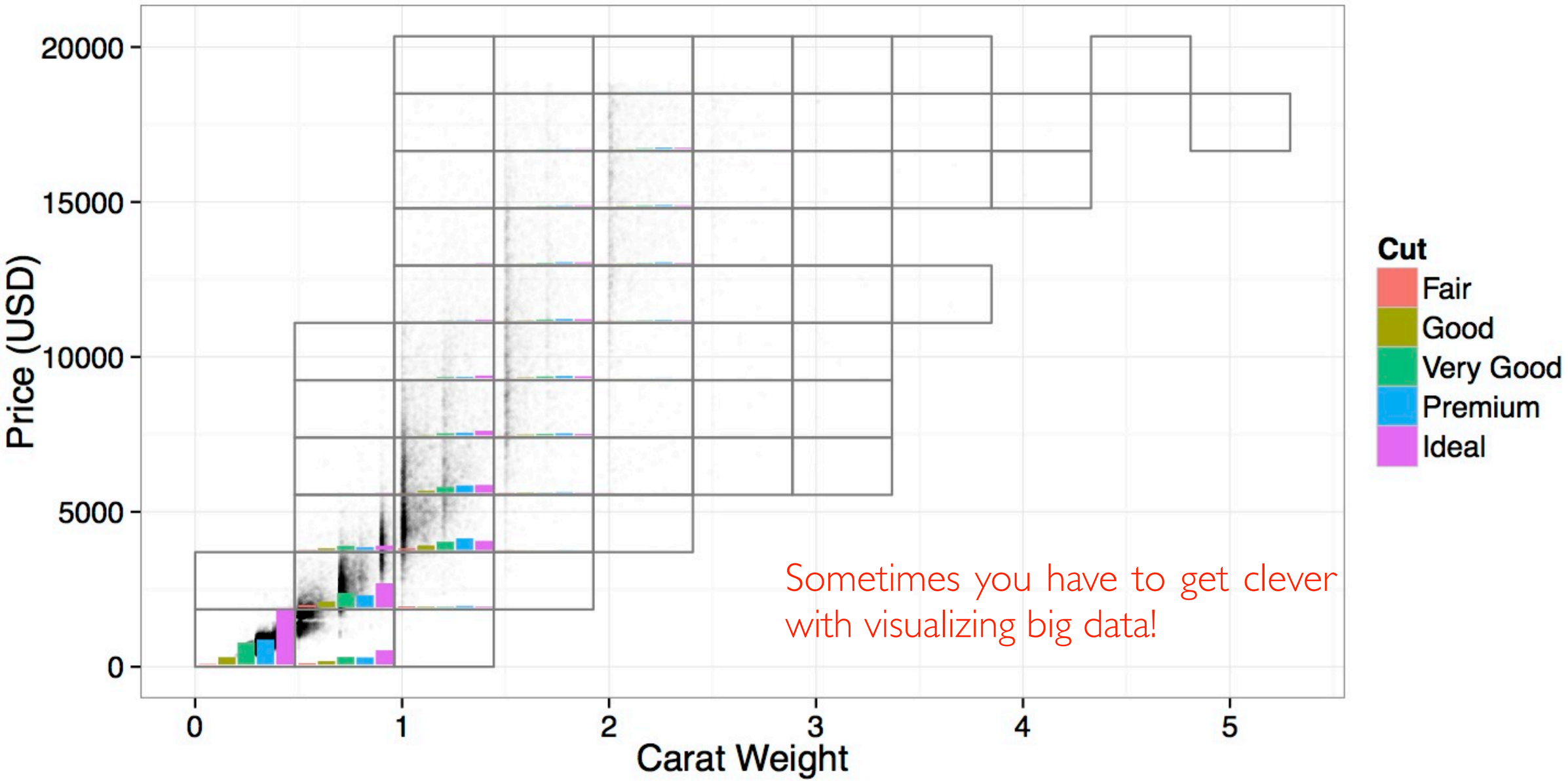
BAYLOR UNIVERSITY 1845

This is a four-dimensional graphic!



If you don't have enough data, you can't make graphics like this!


20000
15000
10000
5000
0
Price (USD)
0
1
2
3
4
5
Carat Weight
Cut
Fair
Good
Very Good
Premium
Ideal


20000
15000
10000
5000
0
Price (USD)
0
1
2
3
4
5
Carat Weight
Cut
Fair
Good
Very Good
Premium
Ideal
Sometimes you have to get clever
with visualizing big data!

Thanks!

Thanks for your time!

If you have any thoughts/comments, feel free to contact me at david_kahle@baylor.edu

Useful tools :

- Free! R and packages
  - ggplot2 / ggsubplot
  - googleVis
- Tableau
- JMP's Graph Builder
- Free! Mondrian