

Visualizing Big Data

David J. Kahle, Ph.D.

Assistant Professor

Department of Statistical Science

eCOTS 2014



“

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

A screenshot of a spreadsheet application showing the "diamonds" dataset. The window title is "diamonds". The left sidebar shows the file path "diamonds.csv" and "Source". Below that is a list of columns: carat, cut, color, clarity, depth, table, price, x, y, z. At the bottom of the sidebar, there are sections for "Rows" (All rows: 53940, Selected: 0, Excluded: 0, Hidden: 0, Labelled: 0) and a summary row count of 53940.

	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)

CSV file: diamonds.csv

Source: diamonds

Columns (10/0)

Rows: All rows (53940), Selected (0), Excluded (0), Hidden (0), Labelled (0)

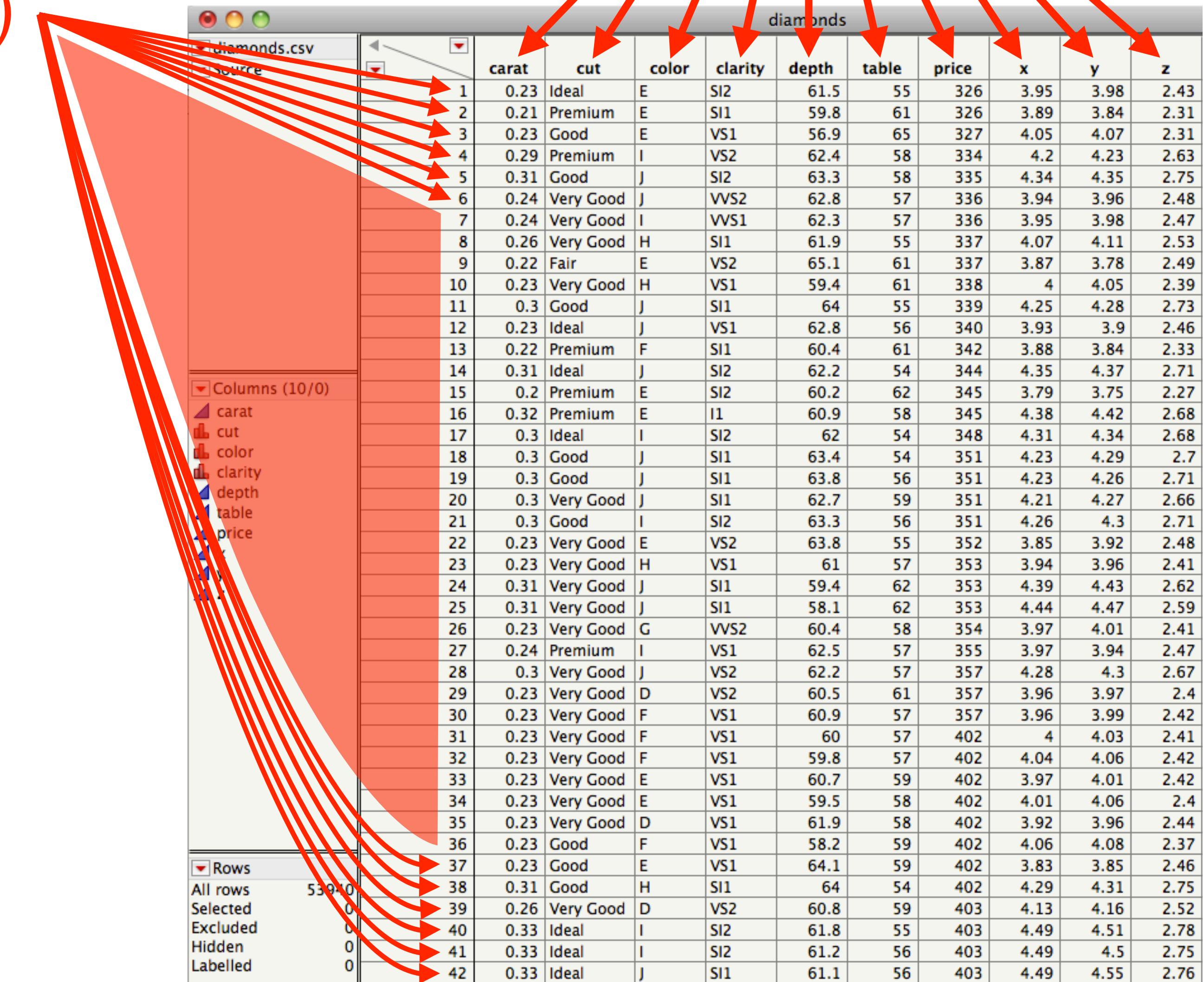
	carat	cut	color	clarity	depth	table	price	x	y	z
1	0.23	Ideal	E	SI2	61.5	55	326	3.95	3.98	2.43
2	0.21	Premium	E	SI1	59.8	61	326	3.89	3.84	2.31
3	0.23	Good	E	VS1	56.9	65	327	4.05	4.07	2.31
4	0.29	Premium	I	VS2	62.4	58	334	4.2	4.23	2.63
5	0.31	Good	J	SI2	63.3	58	335	4.34	4.35	2.75
6	0.24	Very Good	J	VVS2	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	VVS2	63.8	55	352	3.85	3.92	2.48
23	0.23	Very Good	H	VVS1	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	VVS2	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	VVS2	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)



	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	VVS2	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	VVS2	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

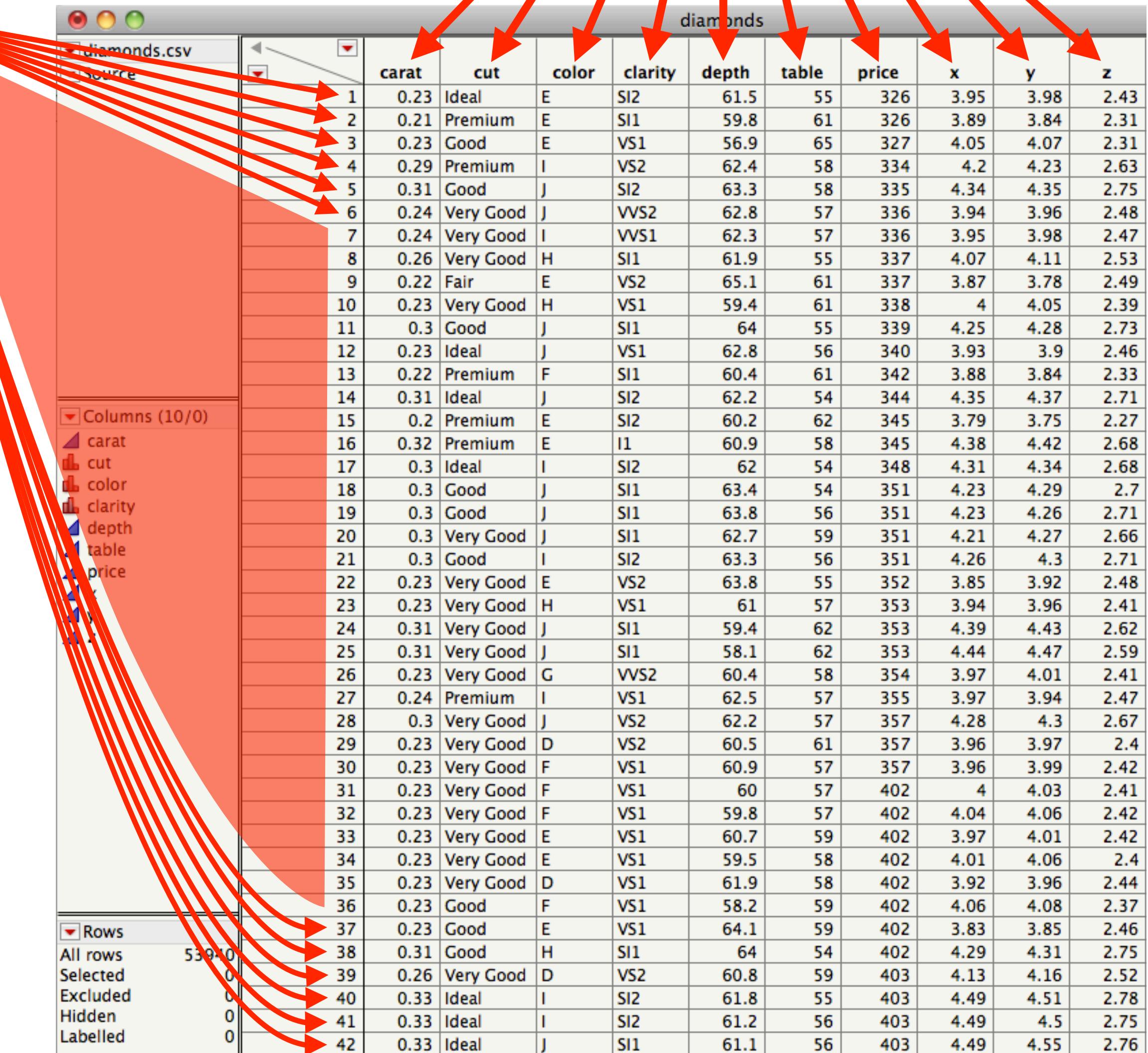
Spreadsheet-type datasets

Rows = subjects (n)

individuals
schools
school districts
counties
census tracts
...

The easiest data has $n \gg p$!

Columns = variables (p)



	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	VVS2	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	VVS2	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

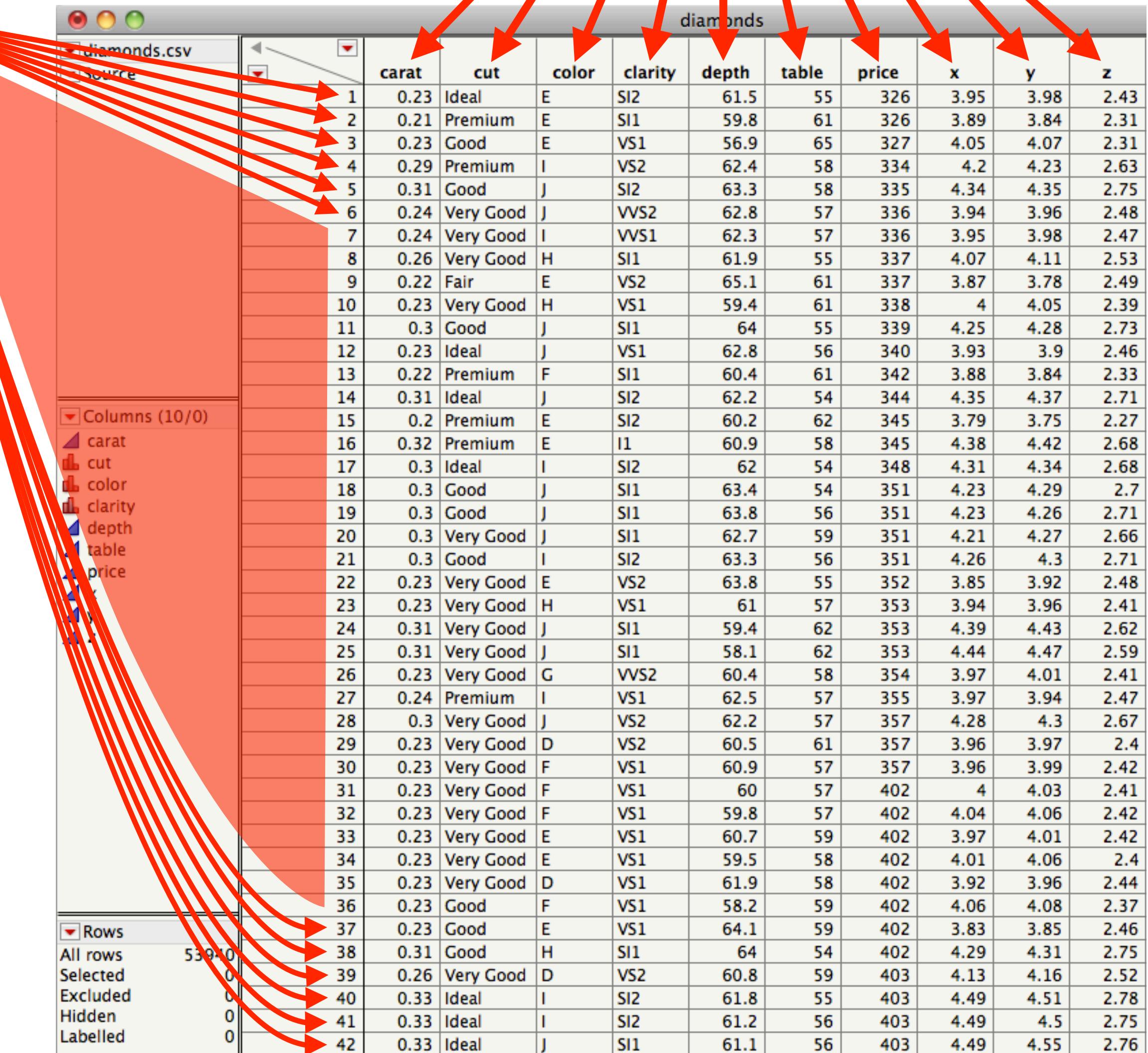
Rows = subjects (n)

individuals
schools
school districts
counties
census tracts
 \dots

The easiest data has $n \gg p$!

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

Columns = variables (p)



	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	VVS2	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	VVS2	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

Rows = subjects (n)

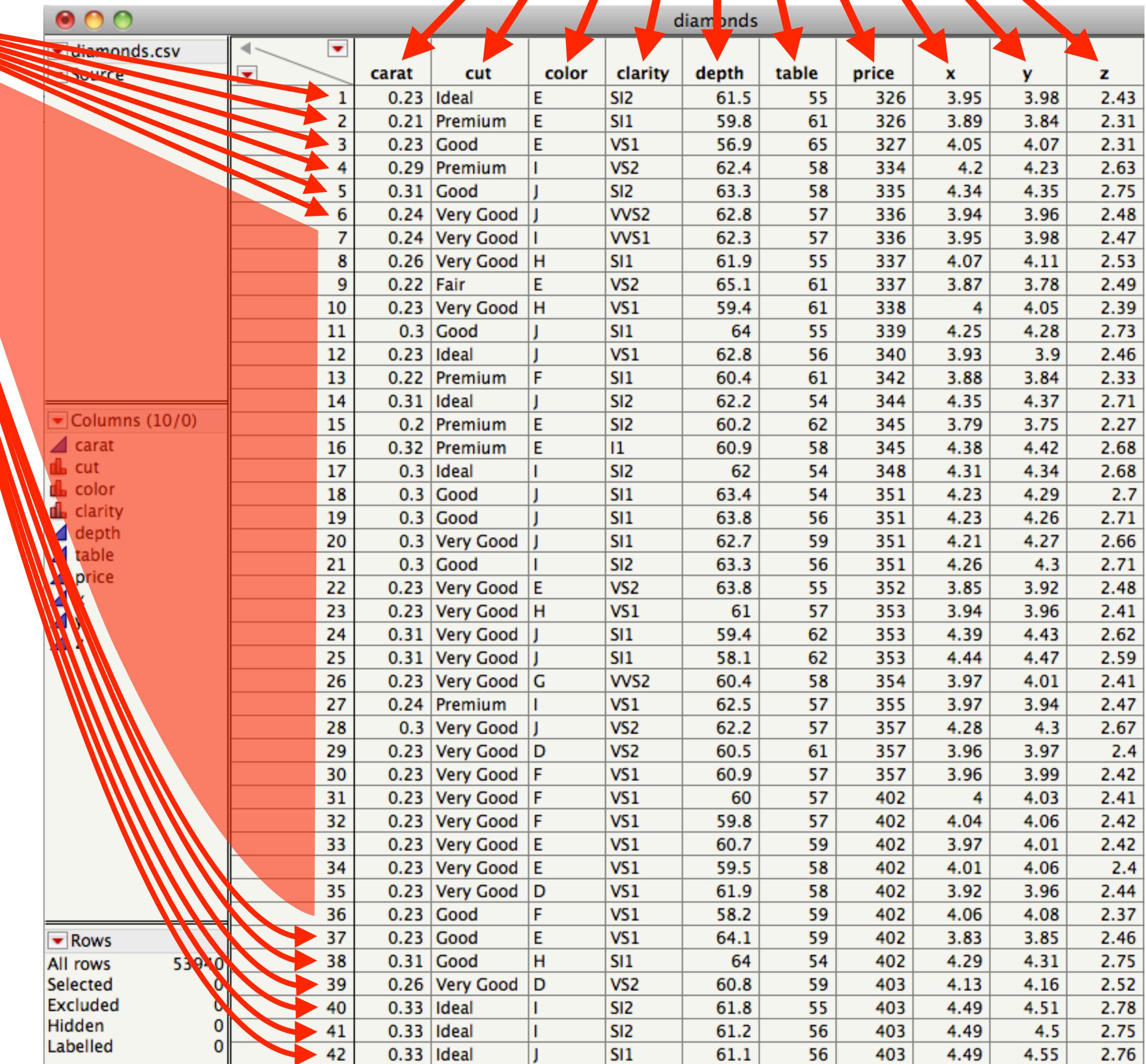
- individuals
- schools
- school districts
- counties
- census tracts
- ...

The easiest data has $n \gg p$!

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

What's big data?

Columns = variables (p)



A screenshot of a data visualization tool showing the 'diamonds' dataset from R. The interface includes a sidebar for columns and rows, and a bottom navigation bar.

	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	VVS2	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	VVS2	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
...

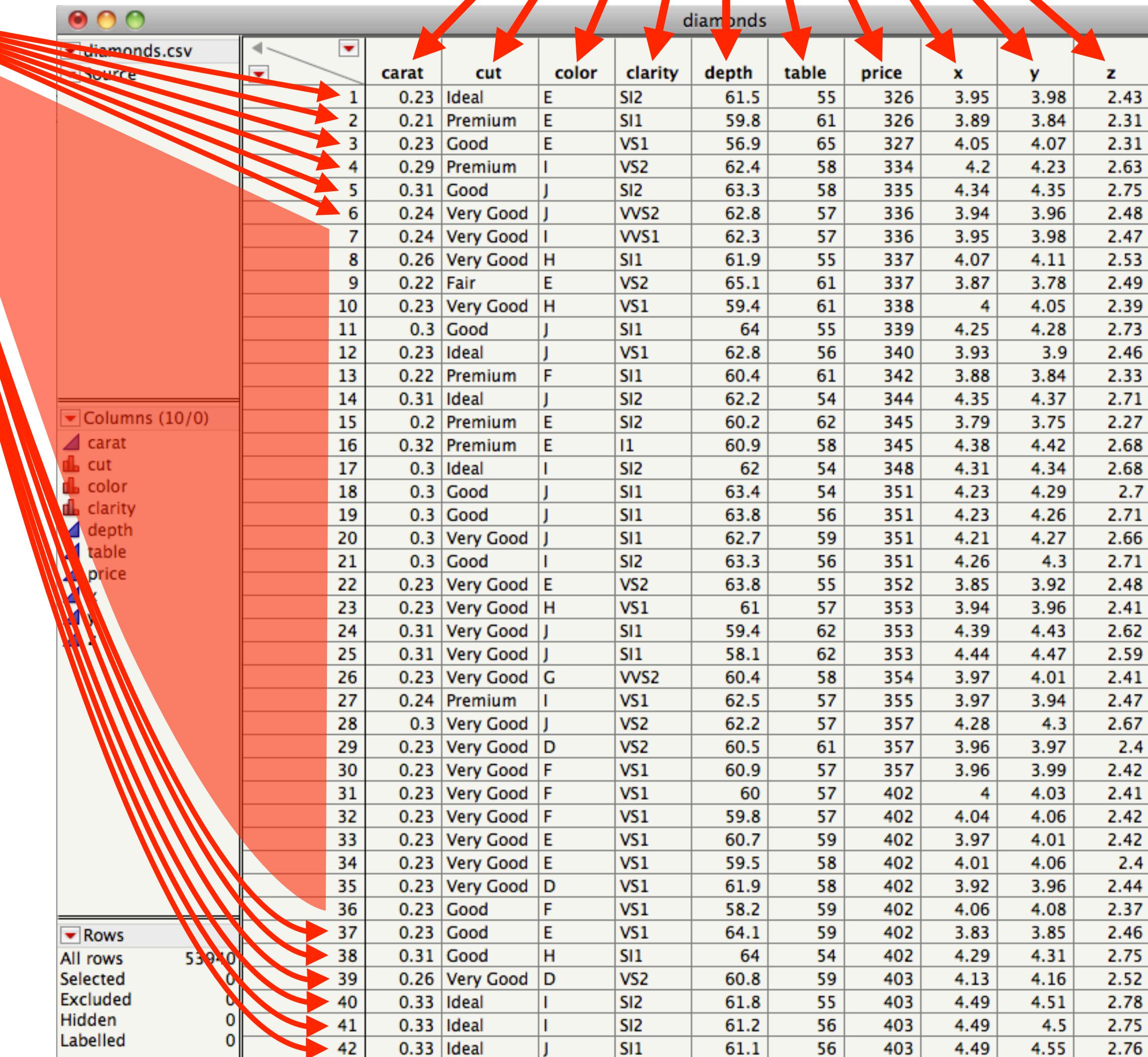
The easiest data has $n \gg p$!

...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)



	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	VVS2	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	VVS2	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?

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

Columns = variables (p)

The screenshot shows a data visualization application window titled "diamonds.csv". The main area displays a table of diamond data with 39 rows and 10 columns. Red arrows point from the column names in the sidebar to their corresponding columns in the table. A callout box in the bottom right corner contains the text: "Efforts are being made to keep these computations under 5 seconds. (On a standard desktop computer.)".

	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	VVS2	63.8	55	352	3.85	3.92	2.48		
23	0.23	Very Good	E	VVS1	60	57	353	3.9	3.96	2.41		
24	0.31	Very Good	J	VS1	59.4	52	354	4.3	4.23	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	3.87	4.01	2.41		
27	0.24	Premium	I	VS1	62.5	57	353	3.67	3.7	2.47		
28	0.3	Very Good	J	VVS2	62.2	57	357	4.28	4.3	2.67		
29	0.23	Very Good	D	VVS2	60.2	51	357	3.96	3.97	2.4		
30	0.23	Very Good	F	VVS1	60.3	57	357	3.96	3.99	2.42		
31	0.23	Very Good	F	VS1	60	57	402	4	4.02	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		
All rows	53940	38	0.31	Good	H	SI1	64	54	402	4.29	4.31	2.75
Selected	0	39	0.26	Very Good	D	VS2	60.8	59	403	4.13	4.16	2.52
Excluded	0	40	0.33	Ideal	I	SI2	61.8	55	403	4.49	4.51	2.78
Hidden	0	41	0.33	Ideal	I	SI2	61.2	56	403	4.49	4.5	2.75
Labelled	0	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

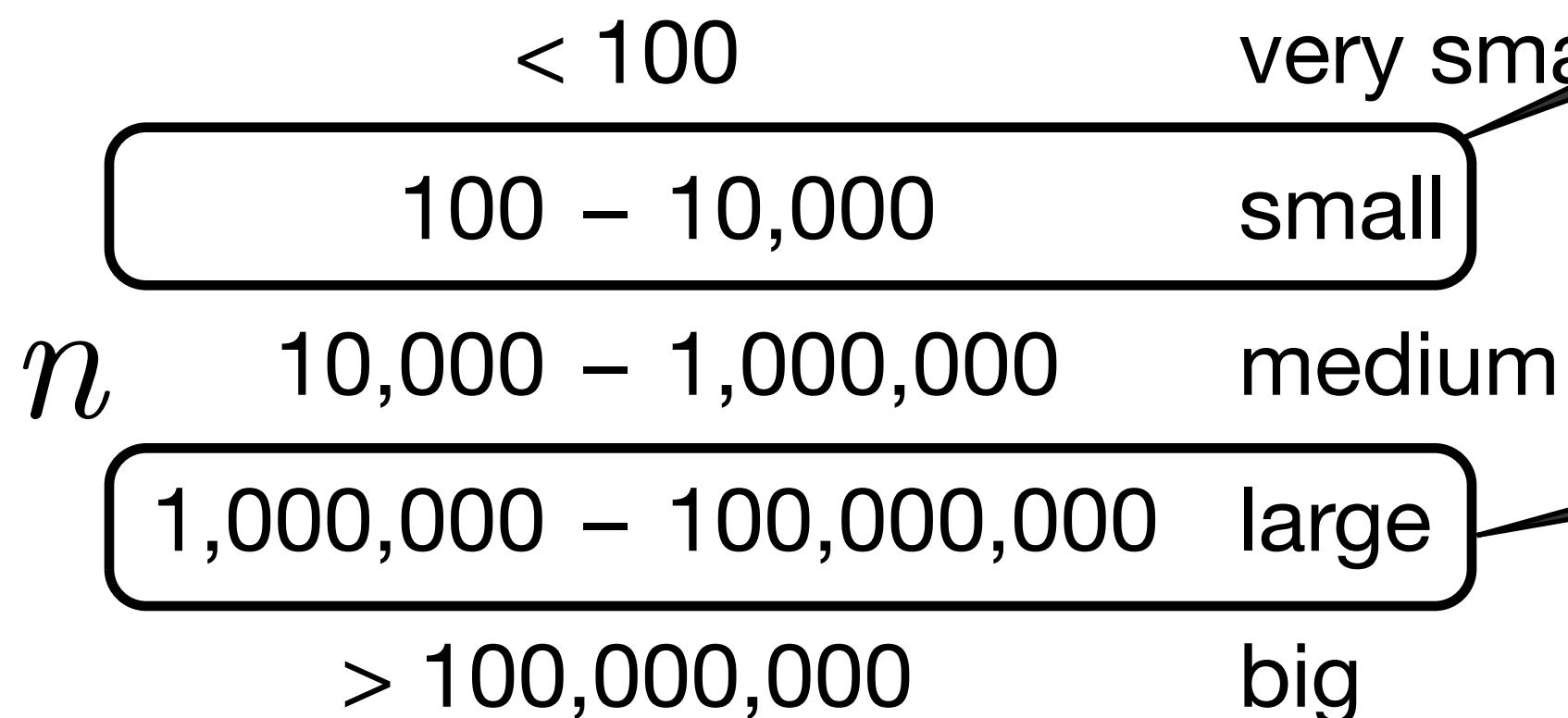
Rows = subjects (n)

individuals
schools
school districts
counties
census tracts
...

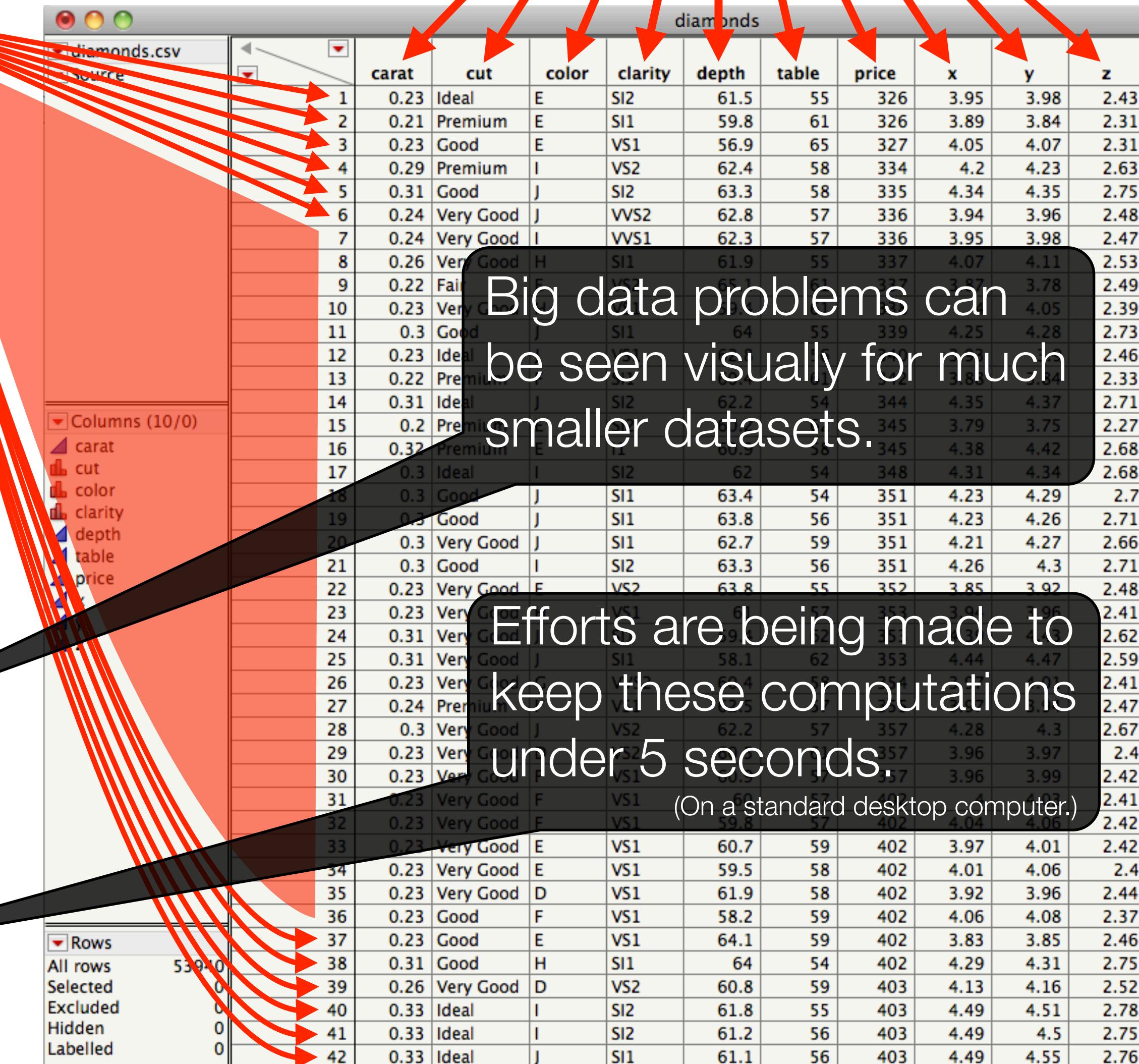
The easiest data has $n \gg p$!

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

What's big data?



Columns = variables (p)



A screenshot of a spreadsheet application showing the diamonds dataset. The table has 39 rows and 10 columns. The columns are labeled: carat, cut, color, clarity, depth, table, price, x, y, and z. Red arrows point from the column labels to their respective columns in the table. A callout box points to the 'x' column with the text: "Big data problems can be seen visually for much smaller datasets." Another callout box points to the 'x' column with the text: "Efforts are being made to keep these computations under 5 seconds." (On a standard desktop computer.)

	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	327	3.87	3.78	2.49
10	0.23	Very Good	I	SI1	69.1	59	339	4.05	4.05	2.39
11	0.3	Good	J	SI1	64	55	339	4.25	4.28	2.73
12	0.23	Ideal	VVS1	VS1	61.1	50	340	4.18	4.24	2.46
13	0.22	Premium	I	VS1	61.1	51	340	4.18	4.23	2.33
14	0.31	Ideal	J	SI2	62.2	54	344	4.35	4.37	2.71
15	0.2	Premium	E	SI1	60.5	56	345	3.79	3.75	2.27
16	0.32	Premium	E	SI1	60.5	56	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	VVS2	63.8	55	352	3.85	3.92	2.48
23	0.23	Very Good	E	VVS1	66	57	352	3.8	3.96	2.41
24	0.31	Very Good	J	SI1	69.1	52	355	4.1	4.2	2.62
25	0.31	Very Good	J	SI1	58.1	62	353	4.44	4.47	2.59
26	0.23	Very Good	C	VVS1	60.4	58	354	3.8	3.91	2.41
27	0.24	Premium	V	S1	60.5	55	355	4.1	4.27	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	SI2	60.1	51	357	3.96	3.97	2.4
30	0.23	Very Good	F	VVS1	60.2	57	357	3.96	3.99	2.42
31	0.23	Very Good	F	VS1	60	57	357	4.04	4.06	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

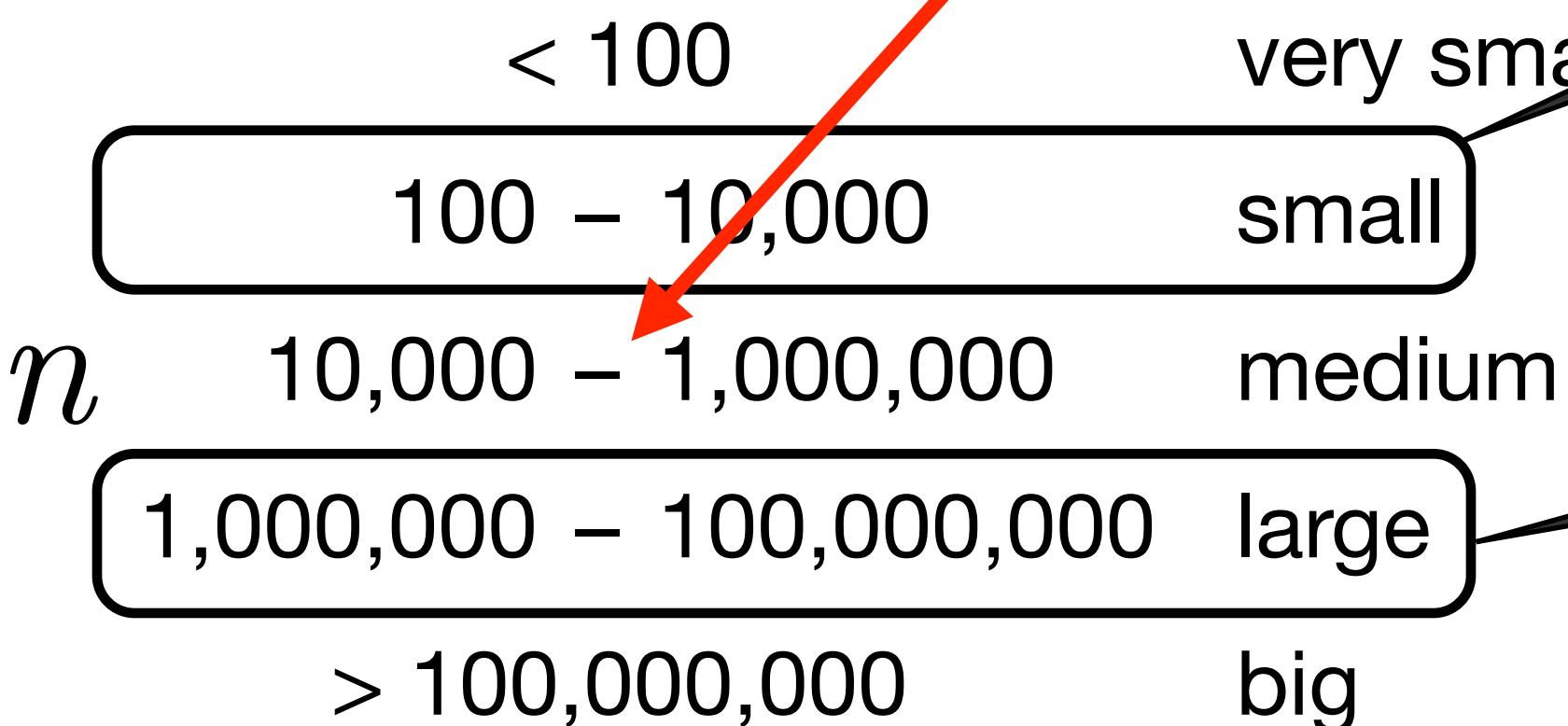
Rows = subjects (n)

individuals
schools
school districts
counties
census tracts
...

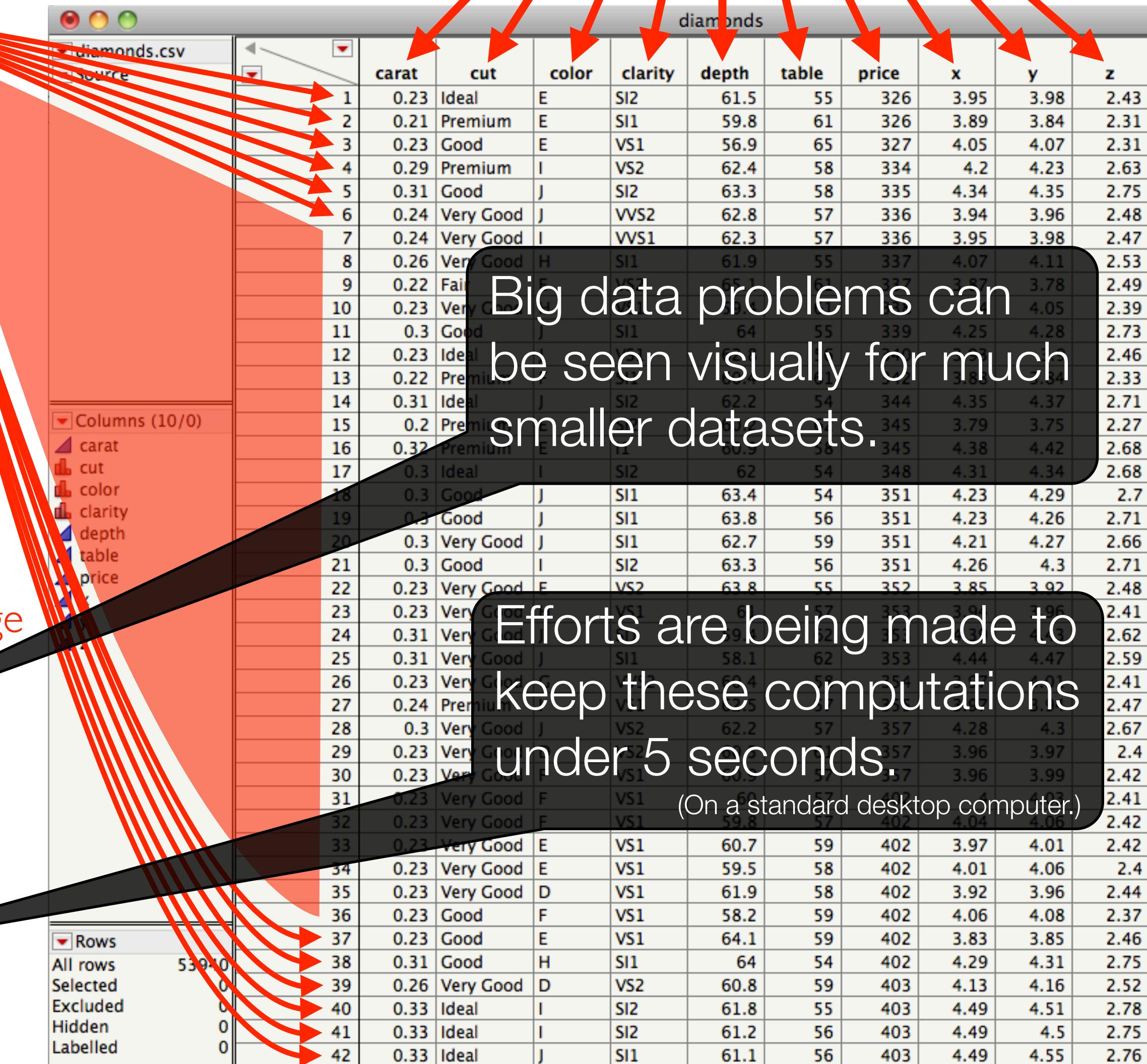
The easiest data has $n \gg p$!

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

What's big data?



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



A screenshot of the diamonds dataset in RStudio. The dataset is a CSV file named 'diamonds.csv'. The interface shows the first few rows of data and a column browser on the left. Red arrows point from the text 'Rows = subjects (n)' to the rows in the table, and from the text 'Columns = variables (p)' to the columns. A callout box on the right says: 'Big data problems can be seen visually for much smaller datasets.' Another callout box at the bottom right says: 'Efforts are being made to keep these computations under 5 seconds. (On a standard desktop computer.)'

	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	327	3.87	3.78	2.49
10	0.23	Very Good	I	SI1	69.1	59	339	4.05	4.05	2.39
11	0.3	Good	J	SI1	64	55	339	4.25	4.28	2.73
12	0.23	Ideal	VVS1	VS1	61.9	55	339	4.18	4.24	2.46
13	0.22	Premium	I	VS1	61.9	55	339	4.23	4.26	2.33
14	0.31	Ideal	J	SI2	62.2	54	344	4.35	4.37	2.71
15	0.2	Premium	E	SI1	61.9	55	345	3.79	3.75	2.27
16	0.32	Premium	E	SI1	60.2	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	VVS2	63.8	55	352	3.85	3.92	2.48
23	0.23	Very Good	E	VVS1	61	57	352	3.8	3.96	2.41
24	0.31	Very Good	I	SI1	69.1	52	355	4.2	4.2	2.62
25	0.31	Very Good	J	SI1	58.1	62	353	4.44	4.47	2.59
26	0.23	Very Good	C	VVS1	60.4	58	354	4.28	4.31	2.41
27	0.24	Premium	V	SI1	55.5	55	355	4.2	4.27	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	SI2	60.1	51	357	3.96	3.97	2.4
30	0.23	Very Good	F	VVS1	59.2	57	357	3.96	3.99	2.42
31	0.23	Very Good	F	VVS1	60	57	357	4.04	4.06	2.41
32	0.23	Very Good	F	VVS1	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

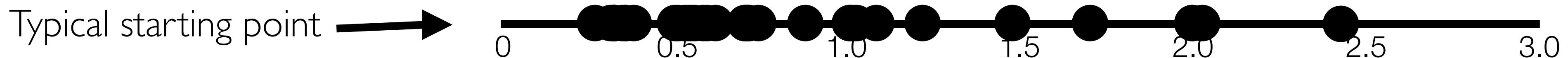
36 Diamond Carat Weights (sorted)

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

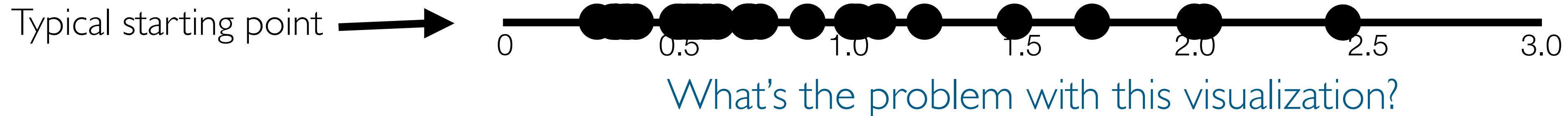
Q : How can we visualize the dataset?



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

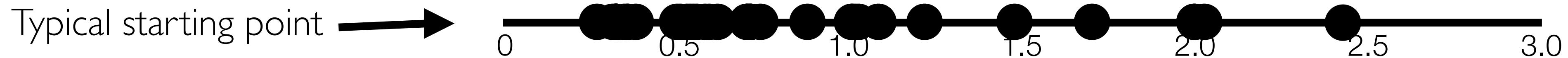
Q : How can we visualize the dataset?



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?



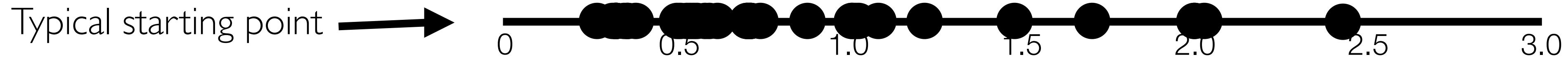
What's the problem with this visualization?

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?



What's the problem with this visualization?

Overplotting : visual confusion caused by plotting too much data.

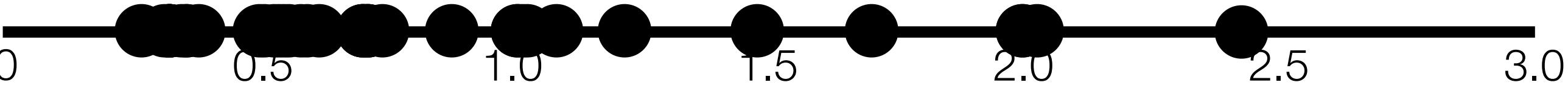
36 Diamond Carat Weights (sorted)

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

Solicit solutions from students!

Q : How can we visualize the dataset?

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

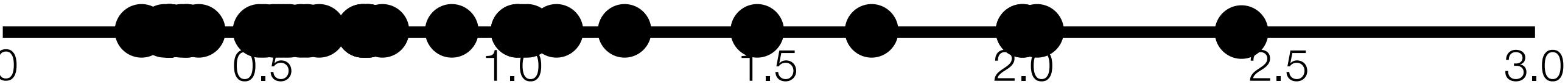
Solicit solutions from students!

Offset from line, resize points



Q : How can we visualize the dataset?

Typical starting point →



What's the problem with this visualization?

Overplotting : visual confusion caused by plotting too much data.

36 Diamond Carat Weights (sorted)

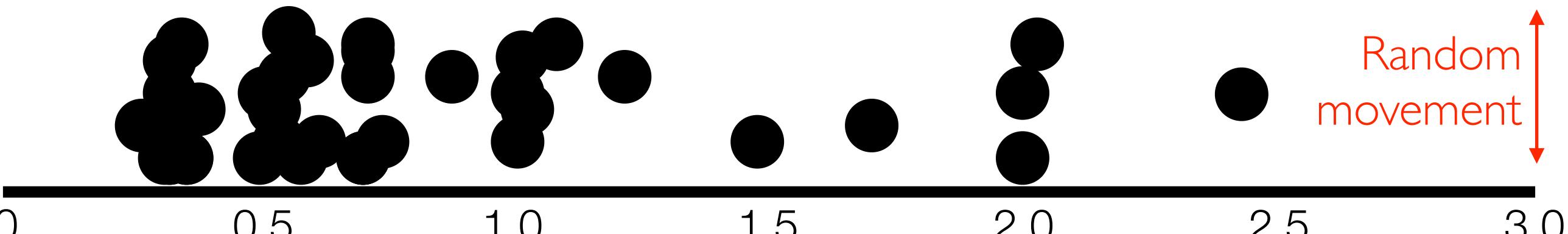
0.26	0.30	0.31	0.31	0.31	0.31
0.35	0.36	0.36	0.50	0.51	0.53
0.54	0.55	0.59	0.61	0.70	0.70
0.70	0.71	0.71	0.71	0.74	0.87
1.00	1.00	1.01	1.02	1.08	1.21
1.47	1.70	2.01	2.01	2.03	2.43

Solicit solutions from students!

Offset from line, resize points

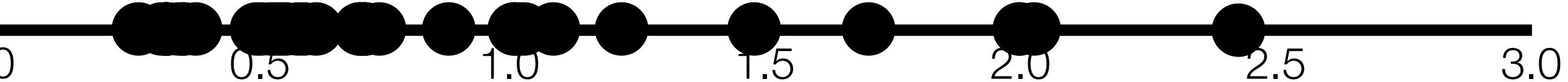


Jitter points



Q : How can we visualize the dataset?

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

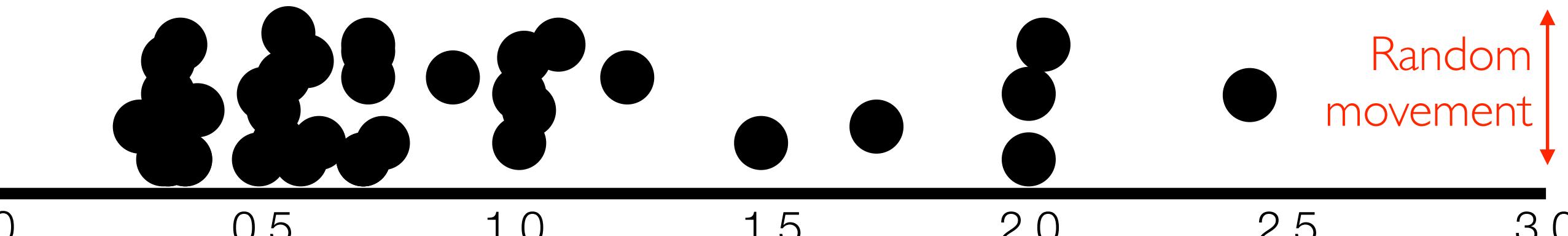
Q : How can we visualize the dataset?

Solicit solutions from students!

Offset from line, resize points



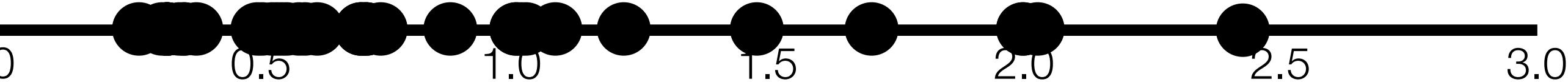
Jitter points



Alpha blend the points (make them semi-transparent)



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

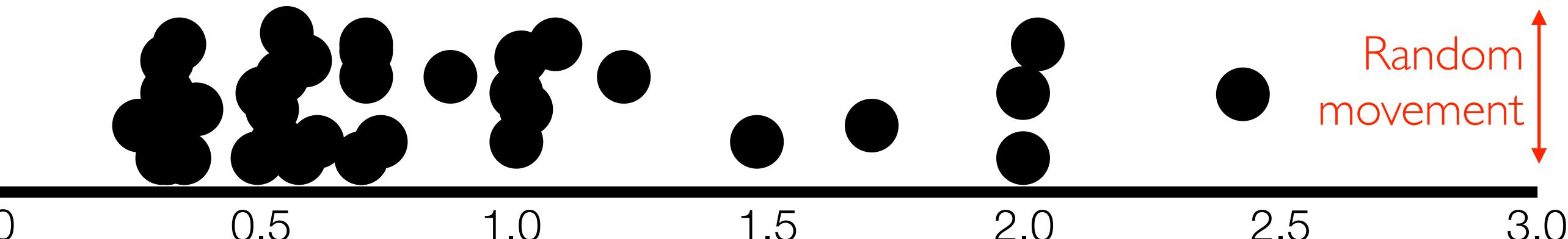
Q : How can we visualize the dataset?

Solicit solutions from students!

Offset from line, resize points



Jitter points

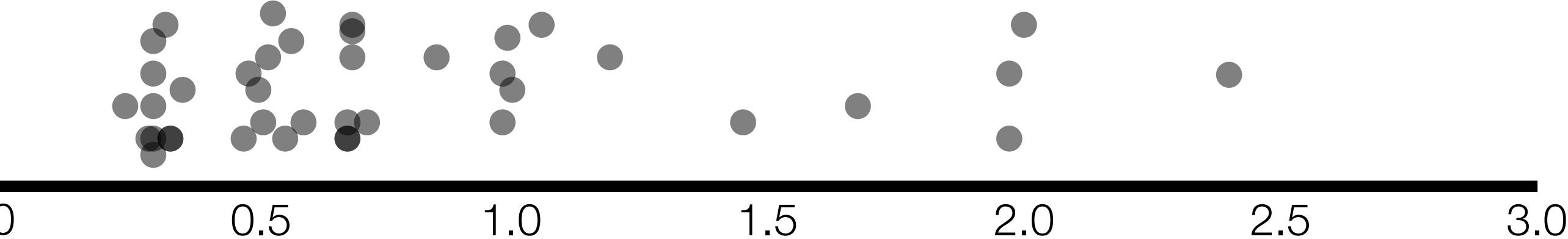


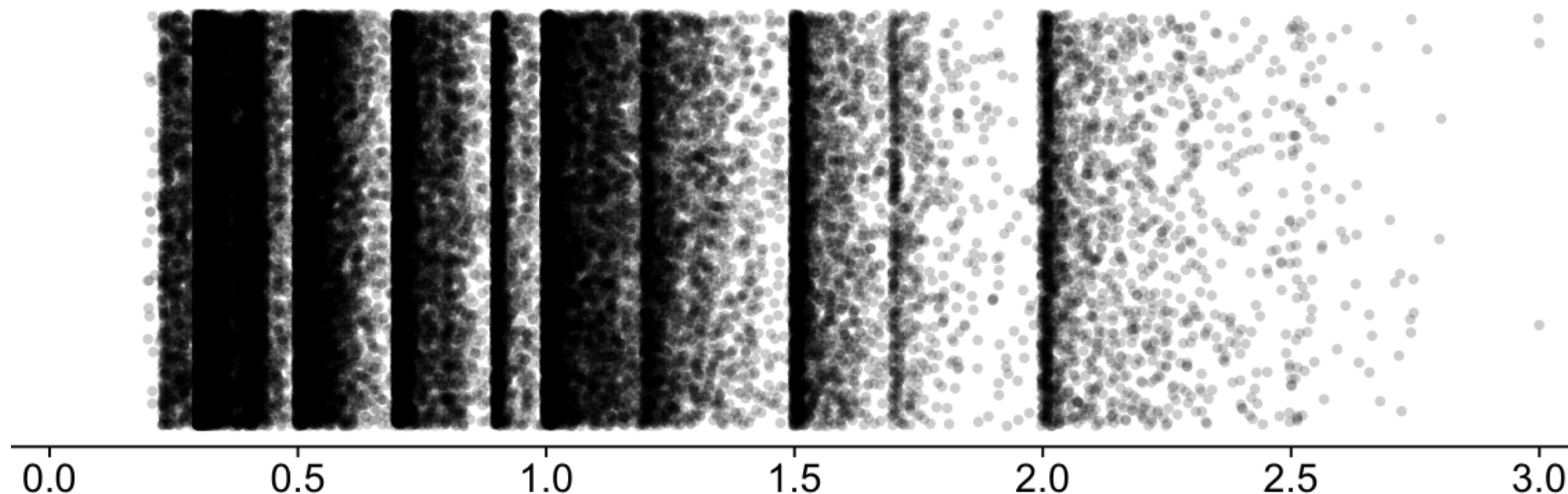
Random movement

Alpha blend the points (make them semi-transparent)

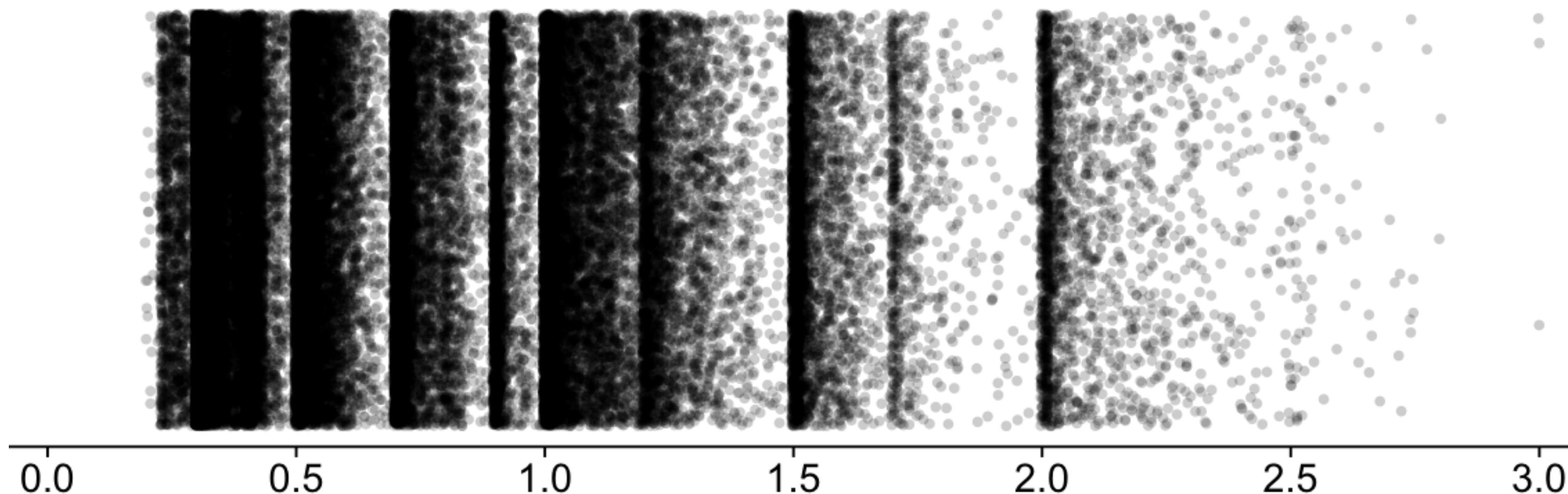


Combo





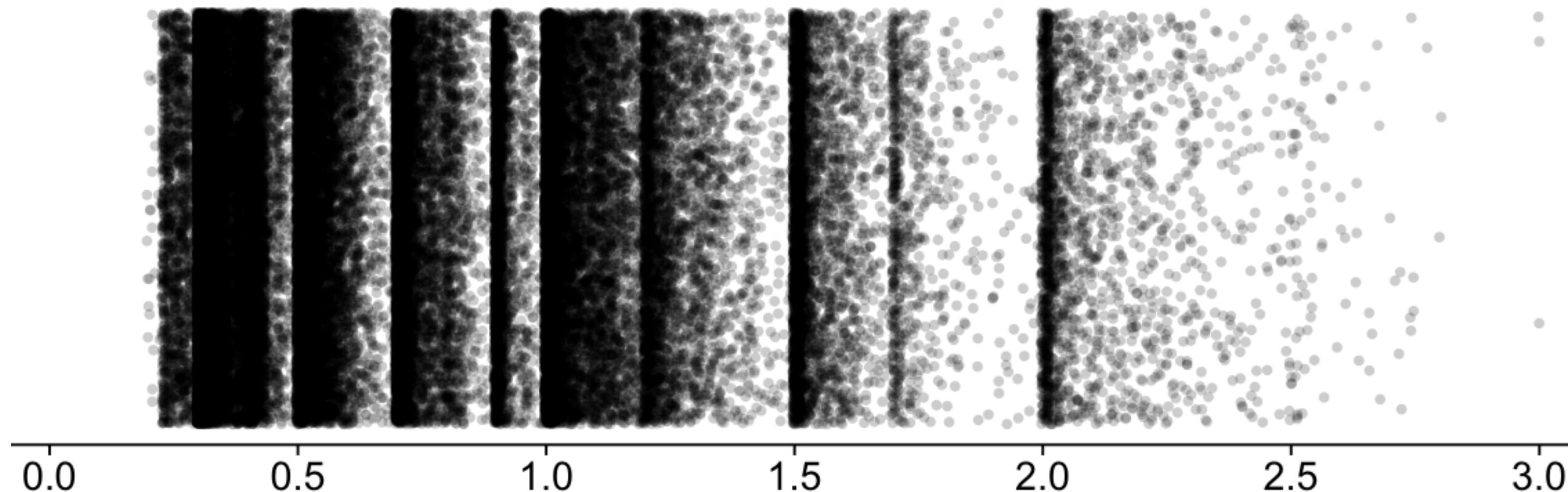
Problem : for the entire dataset of 55,000 diamonds,
we still can't get away from the problem.



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

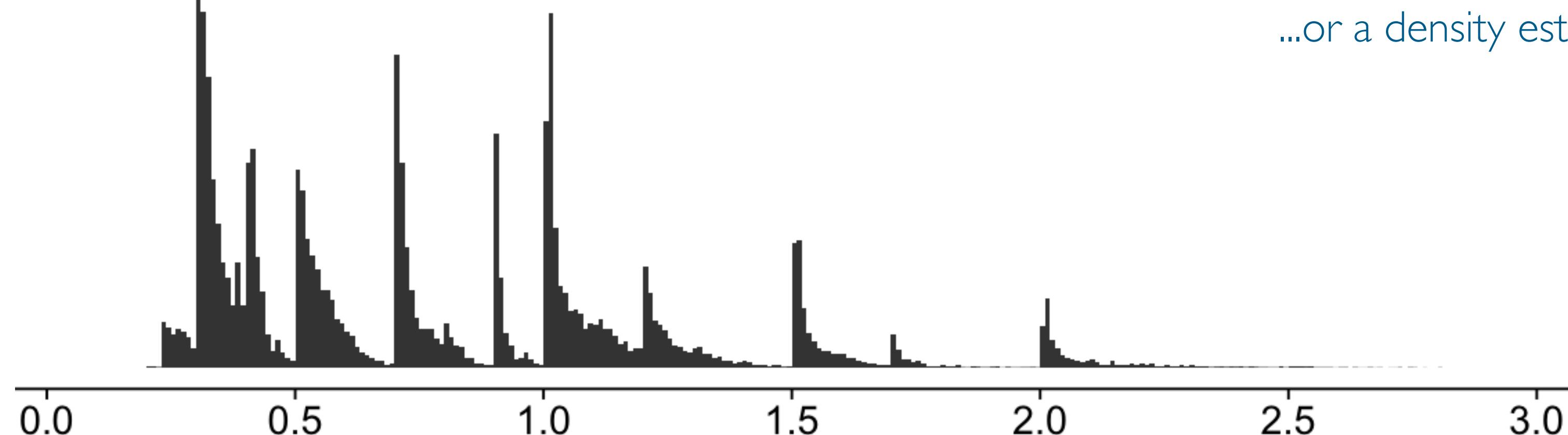
...or a density estimate, boxplot, etc.

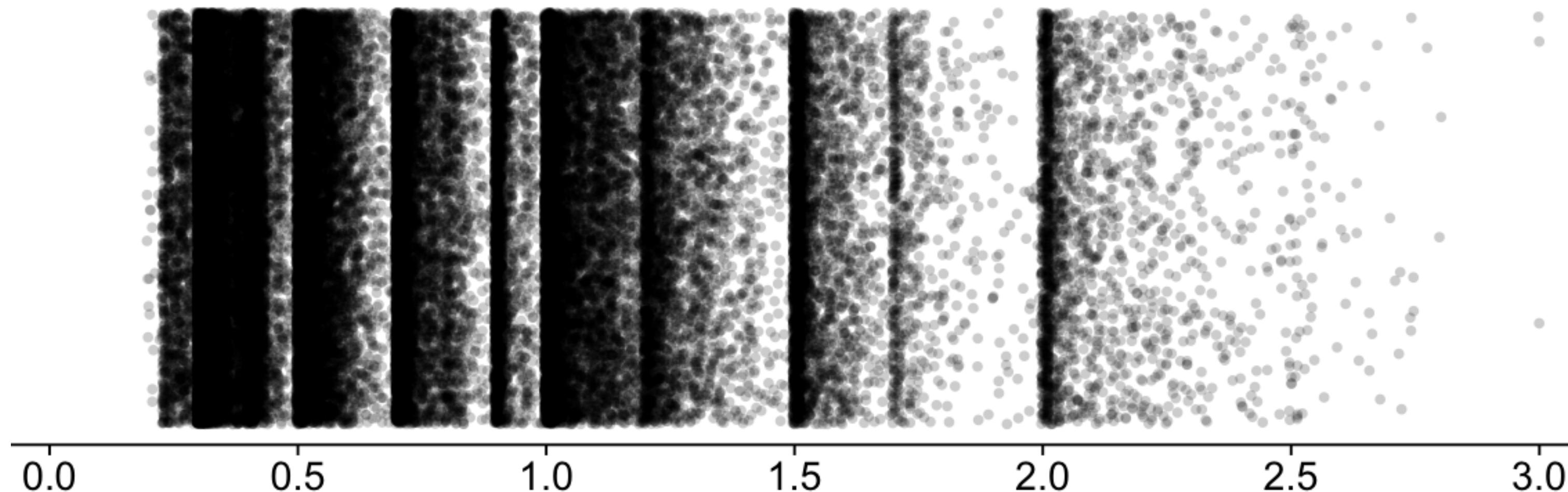


Problem : for the entire dataset of 55,000 diamonds,
we still can't get away from the problem.

“Solution” : Summarize dataset with bins as histogram

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



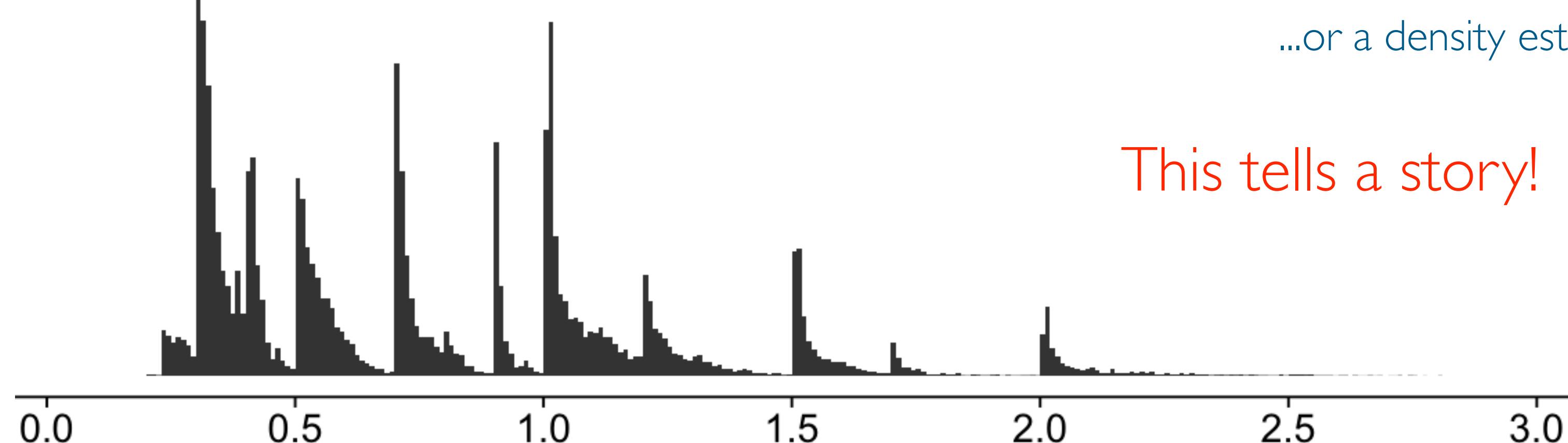


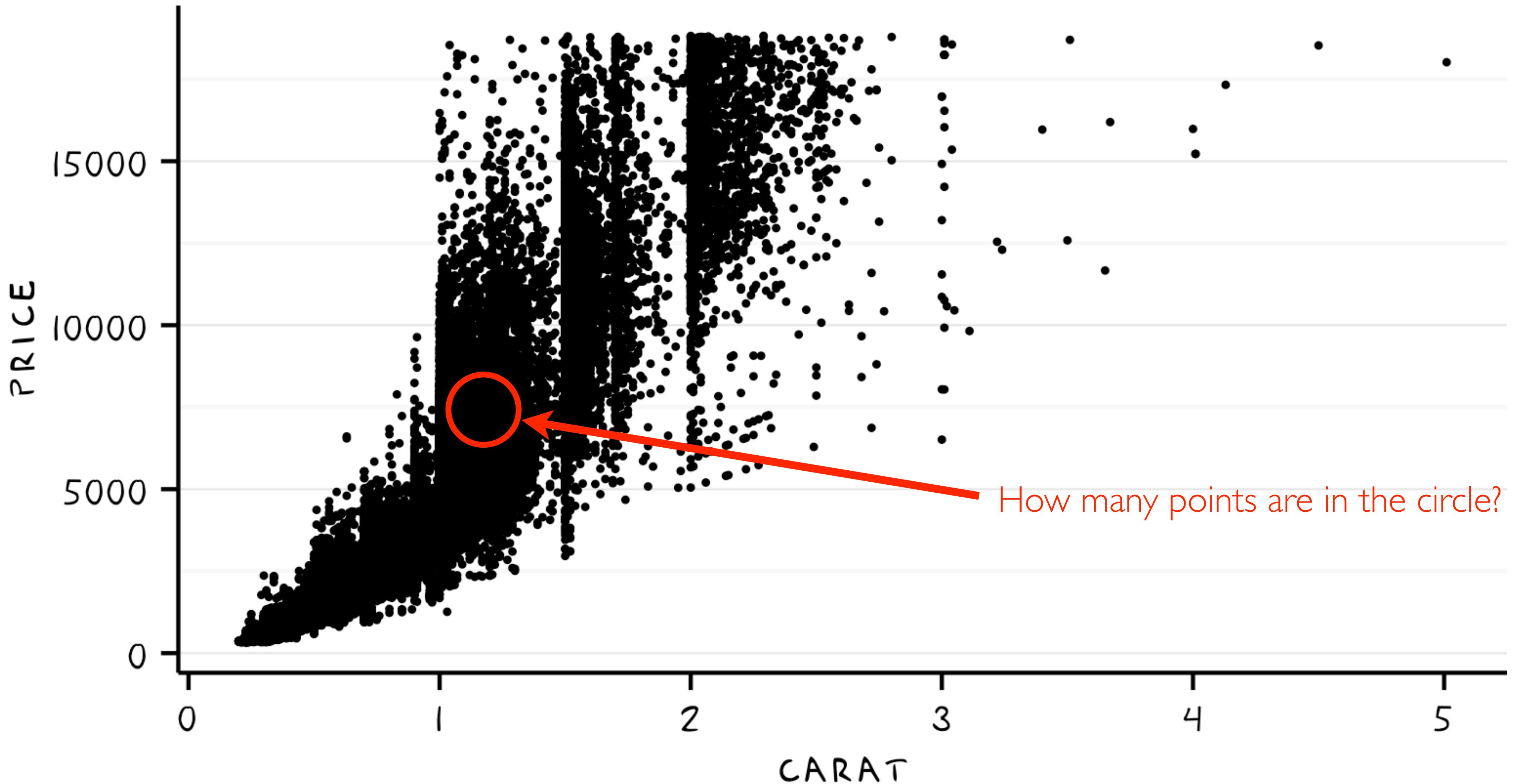
Problem : for the entire dataset of 55,000 diamonds,
we still can't get away from the problem.

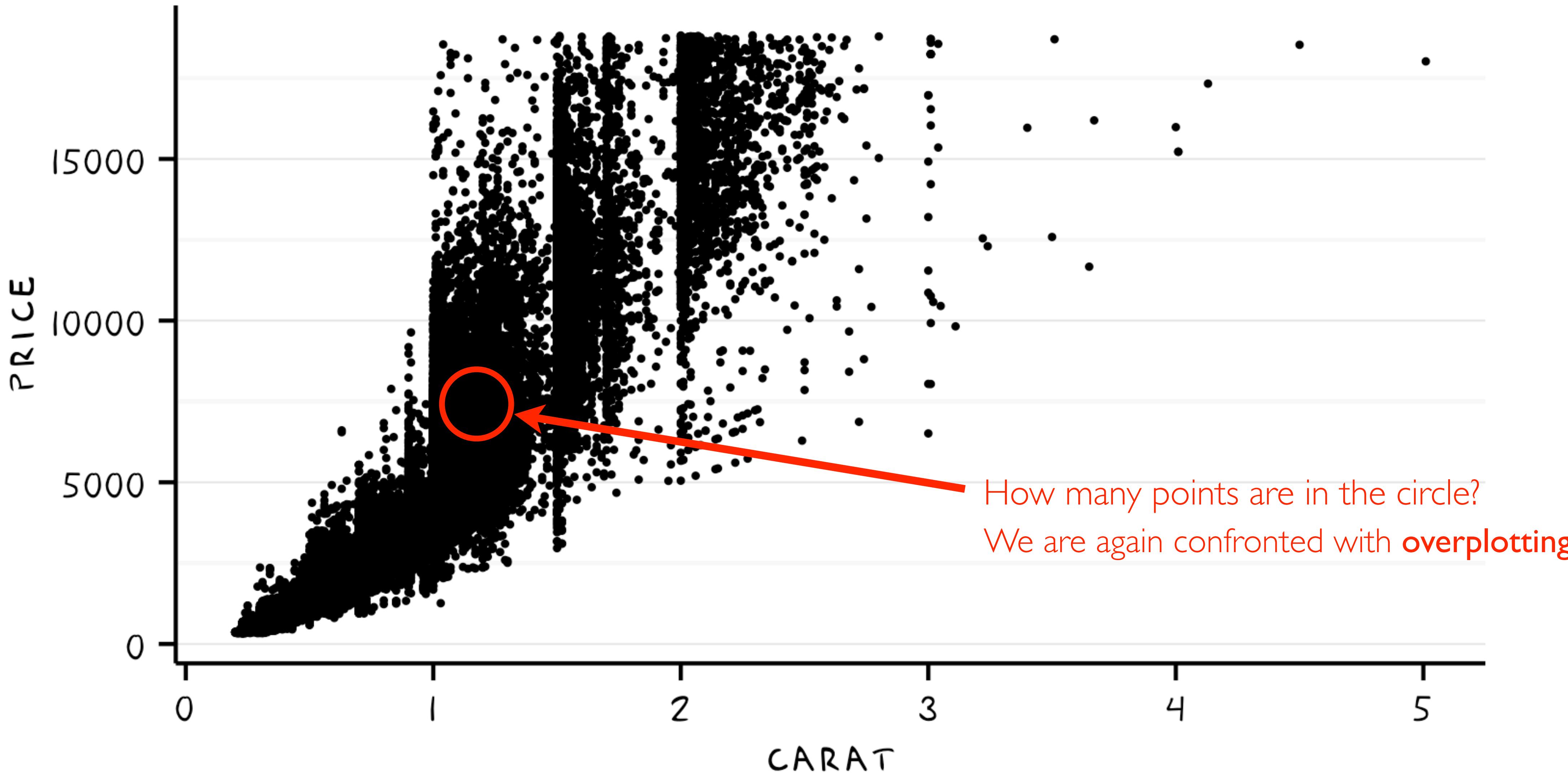
“Solution” : Summarize dataset with bins as histogram

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

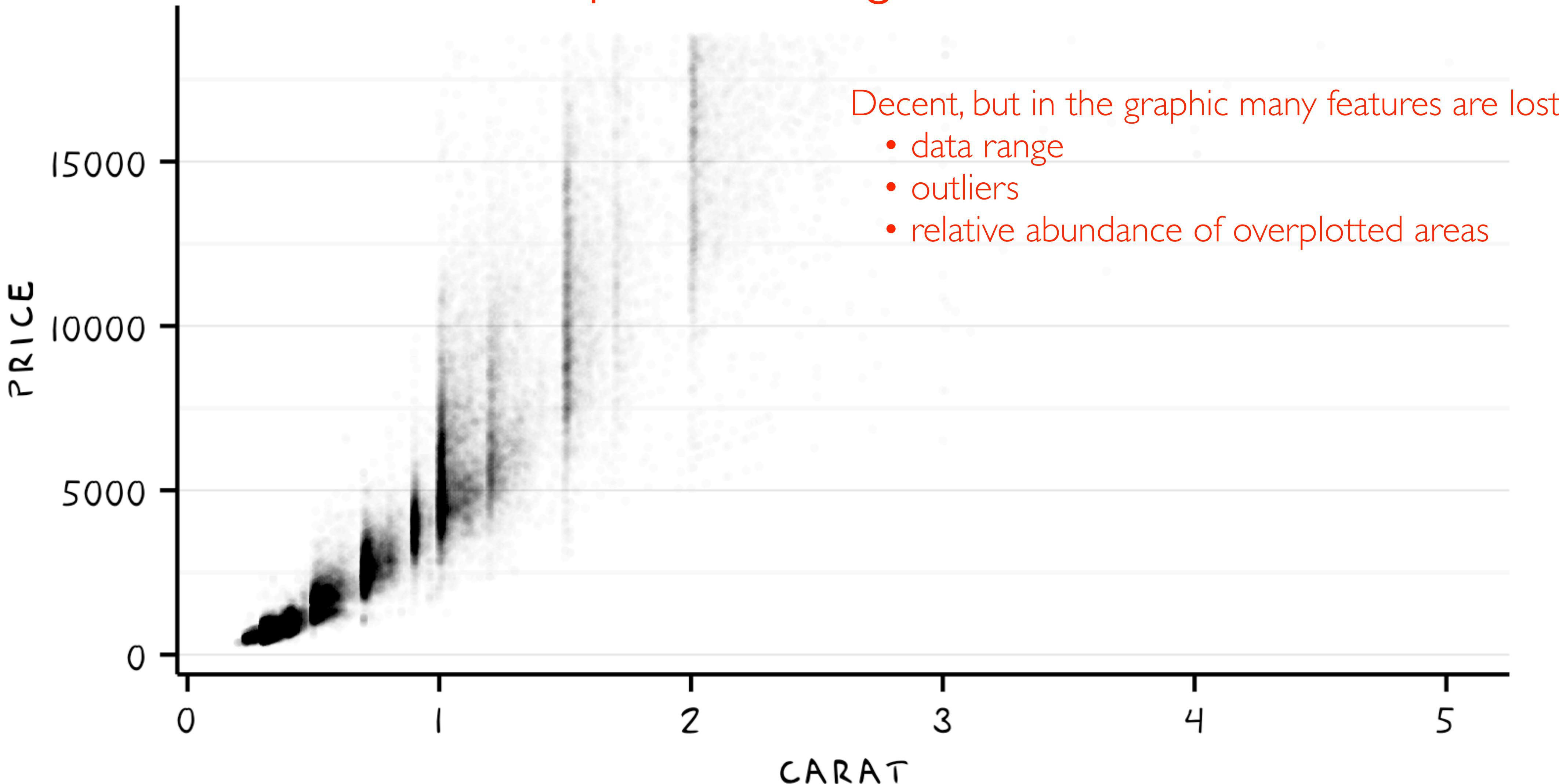
This tells a story!



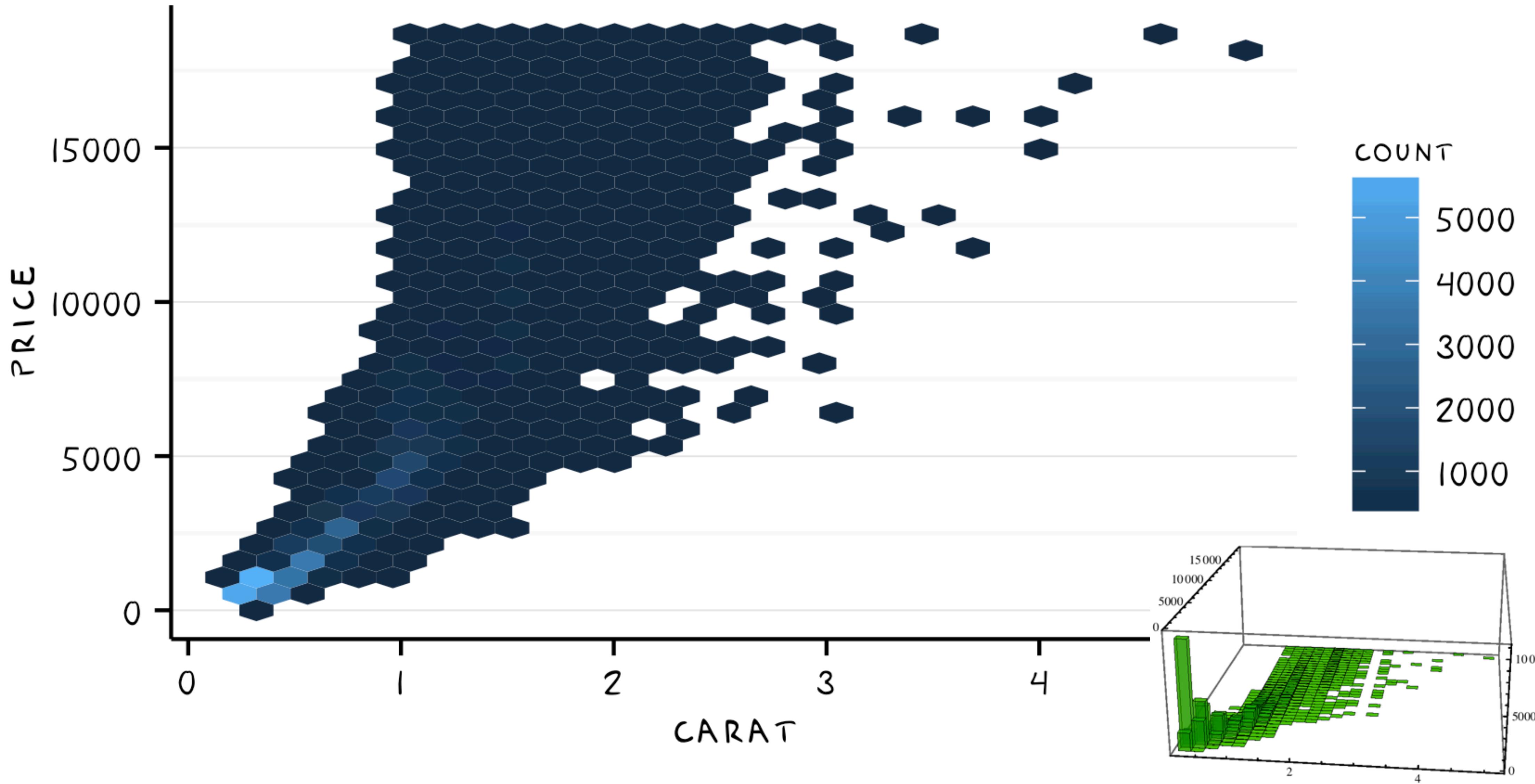




alpha blending = 100

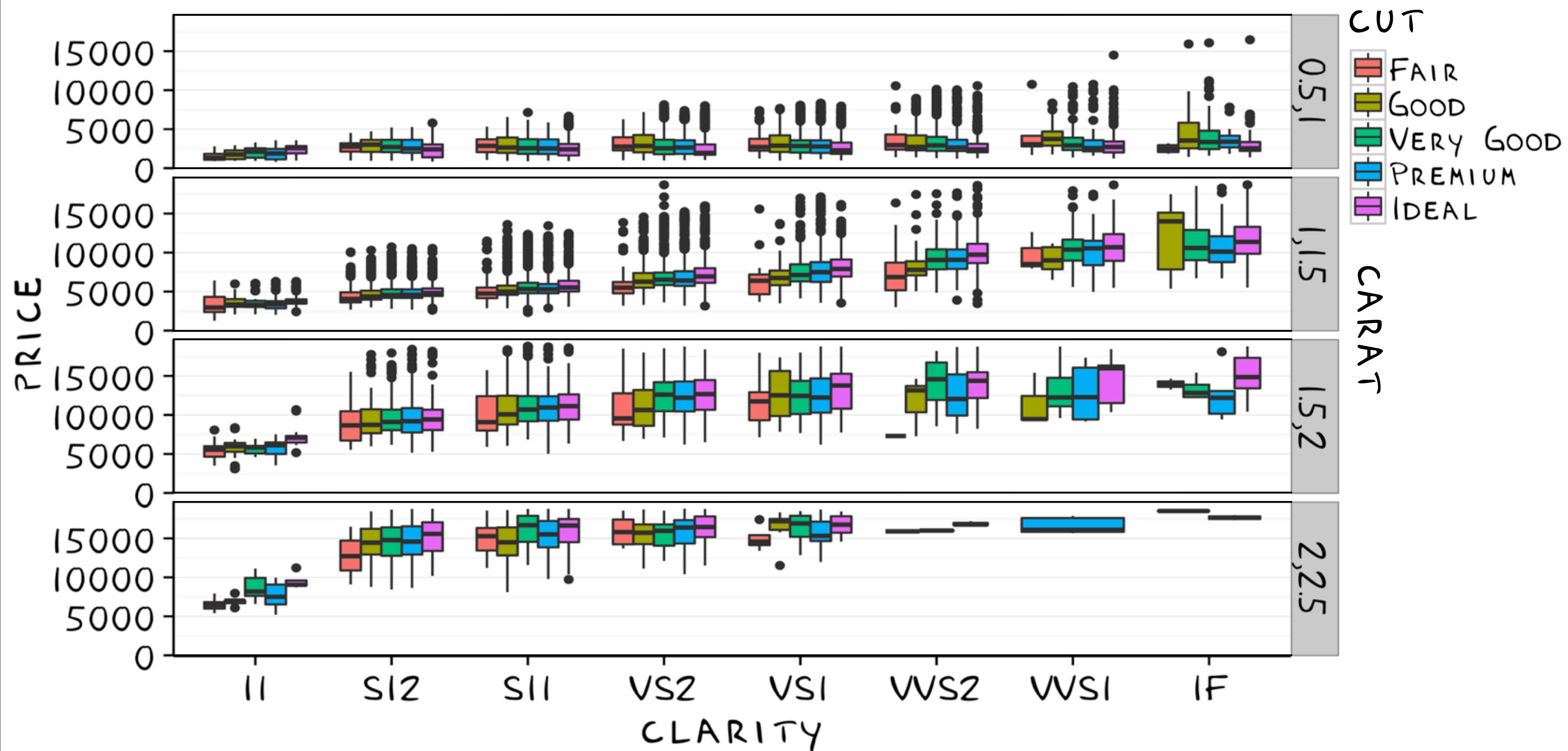


2d graphics

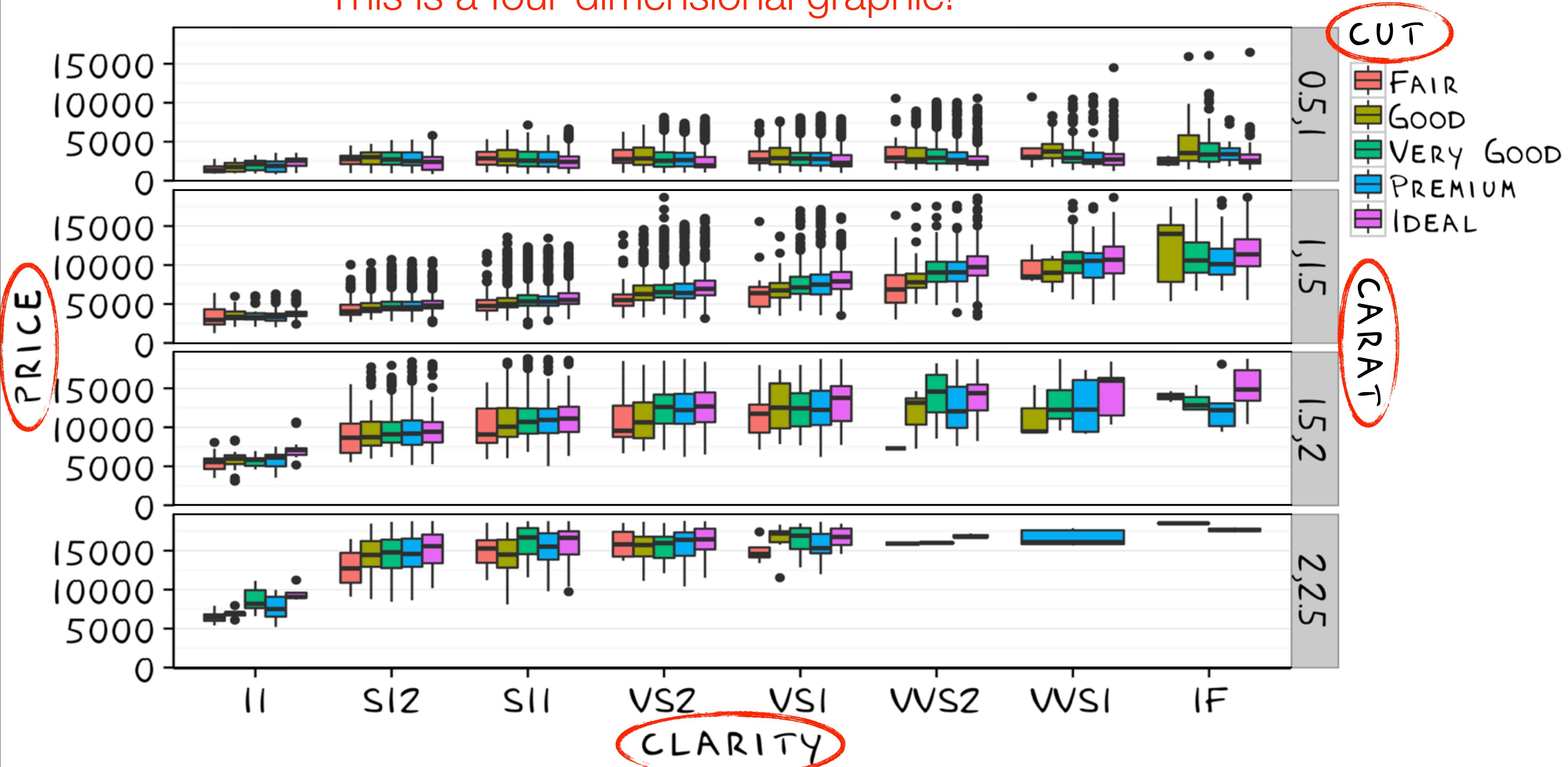


pd graphics – Faceting

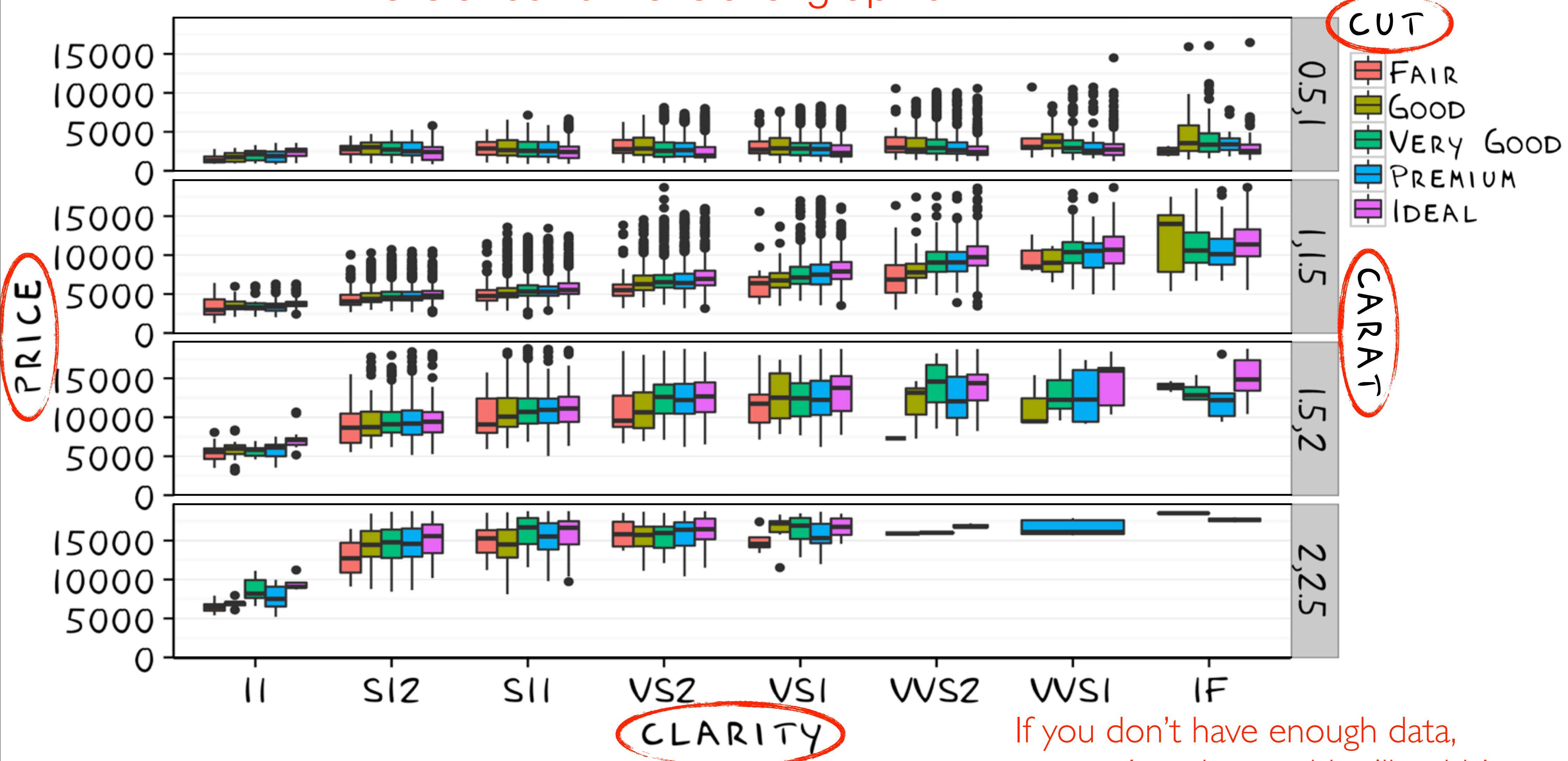
Conditional plots on discrete variables



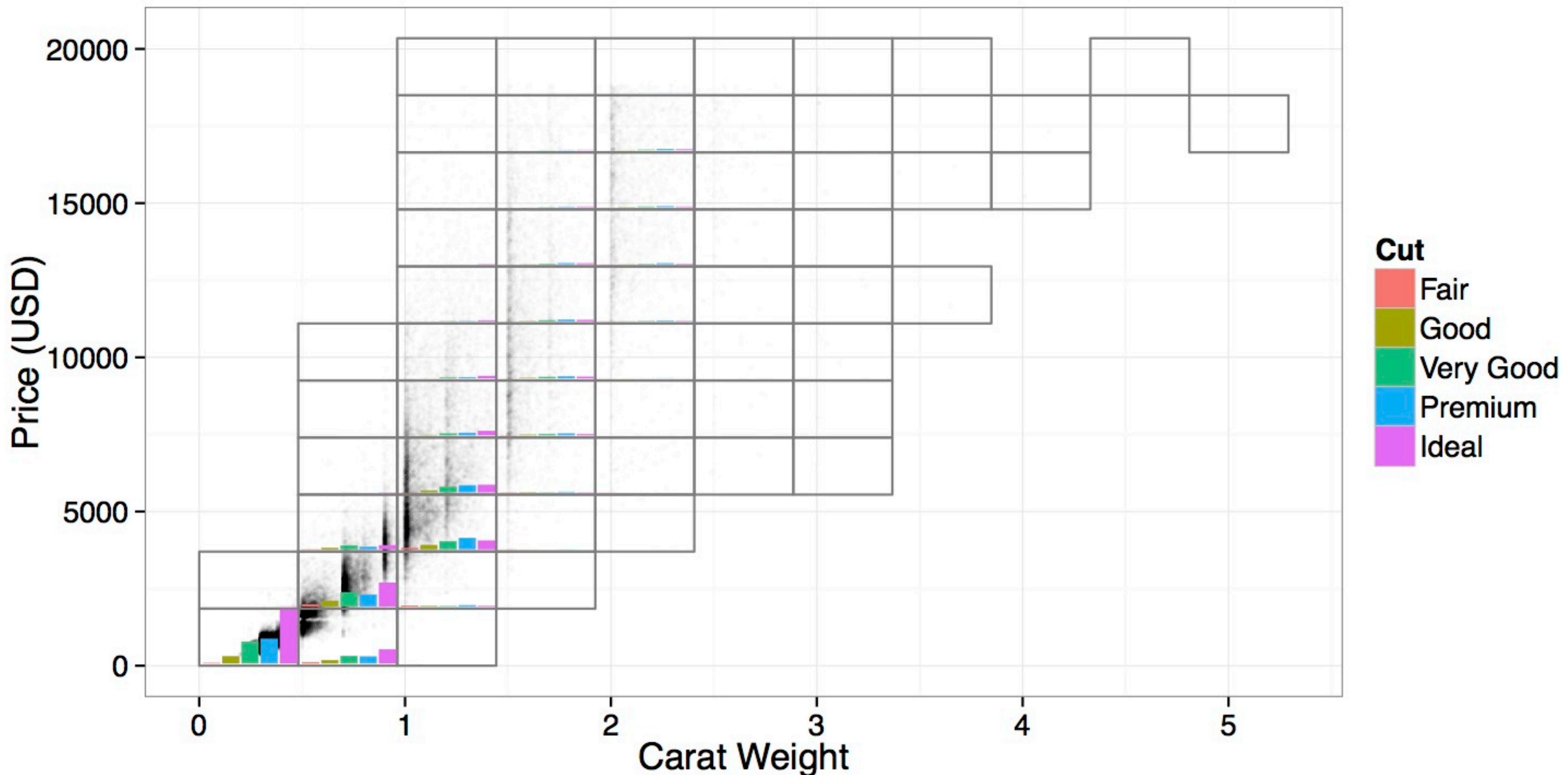
This is a four-dimensional graphic!

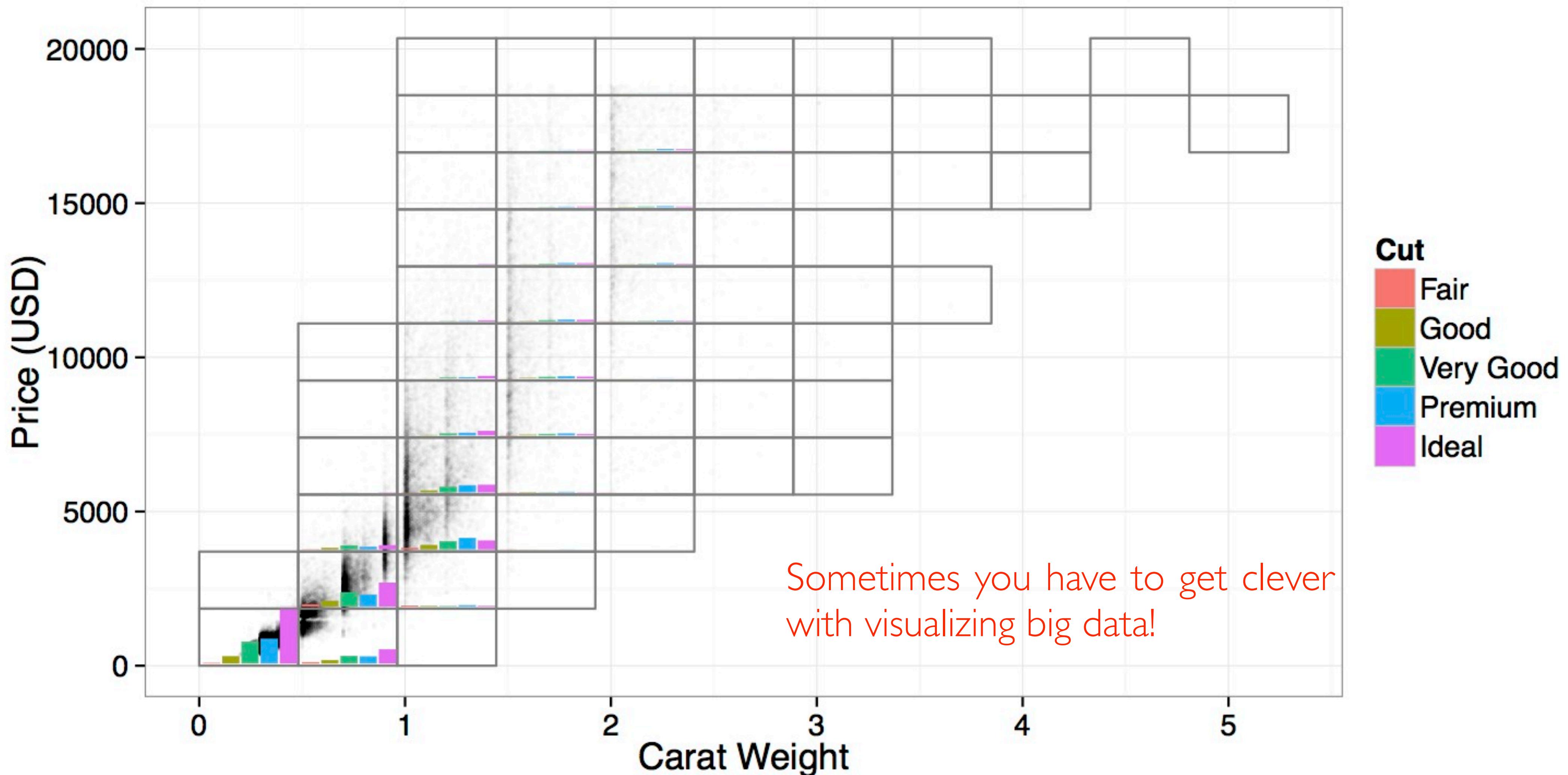


This is a four-dimensional graphic!



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





Thanks!

 creative commons

Attribution-ShareAlike 4.0 International (cc BY-SA 4.0)

This is a human-readable summary of (and not a substitute for) the [license](#).

[Disclaimer](#)



You are free to:

- Share** — copy and redistribute the material in any medium or format
- Adapt** — remix, transform, and build upon the material for any purpose, even commercially.

The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:

-  **Attribution** — You must give [appropriate credit](#), provide a link to the license, and [indicate if changes were made](#). You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.
-  **ShareAlike** — If you remix, transform, or build upon the material, you must distribute your contributions under the [same license](#) as the original.

No additional restrictions — You may not apply legal terms or [technological measures](#) that legally restrict others from doing anything the license permits.

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 / ggs subplot
 - googleVis
 - Tableau
 - JMP's Graph Builder
 - Mondrian