INM433: Session 02
Data, Tasks, and Partition-Based Clustering in Visual Analytics

INM433 Visual Analytics

Last week

- Visual Analytics
- Role of interactive visualisation in Visual Analytics
  - Visual variables and when to use
  - Types of visualisation display and when to use
  - Types of interaction
  - Coordinated linked views
- (Practical) how to use Mondrian and Tableau

This week

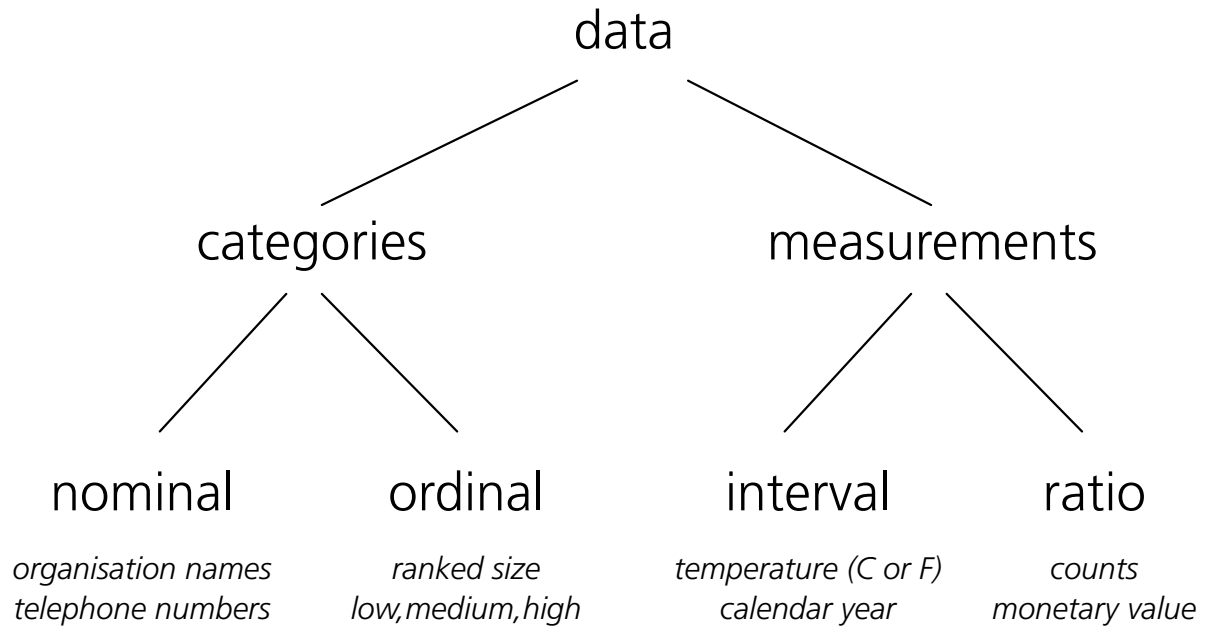
- Data types and structure
  - And how these affect analysis and interpretation
- How **partition-based clustering** combined with interactive visualisation can help dealing with large complex datasets
  - Density-based is the other type of clustering that will be covered later
- Practical: Tableau

Part 1: Data and tasks

Data

Level of Measurement

- "On the theory of scales of measurement" (1946)
  - Stanley Stevens, physiologist
  - Typography of measurements in science



Categories and measurements

- Categories:
  - Ordinal or nominal
  - Tableau: “dimensions”
  - Can be used to identify objects or group data
- Measures:
  - Interval or ratio
  - Tableau: “Measures”
  - Used to describe characteristics

Semantic role of data components

- We can allocate semantic role, depending how we want to consider the data. Based on:
  - the dataset
  - the types of research question we're interested in
- **References**: Discrete things that are described
  - Objects, (discrete/aggregated) times/places/attributes.
- **Characteristics**: Things that characterise them
  - Attributes

Types of reference

- Object
  - No ordering, no distances, discrete
- Time
  - 1D ordering, has distance, continuous
- Space (2D, 3D)
  - 2D ordering, has distances, continuous
- Can be more than one!

potential references (object/place)

attributes

OA	Local Author	oa_lon	oa_lat	Total Popula	Total Numbe	Total Dwelli	Total Housel	Total Popula	Total Popula	Total Popula	Total Person	Total Popula	Total Po
E00000001	City of Londc	-0.1	51.52	194	99	115	115	173	148	148	102	173	
E00000003	City of Londc	-0.1	51.52	250	112	125	125	218	199	199	147	218	
E00000005	City of Londc	-0.1	51.52	367	217	241	241	337	304	304	241	337	
E00000007	City of Londc	-0.1	51.52	123	83	103	103	113	111	111	86	113	
E00000010	City of Londc	-0.1	51.52	102	78	79	79	97	86	86	59	97	
E00000012	City of Londc	-0.1	51.52	213	137	139	139	197	166	166	115	197	
E00000013	City of Londc	-0.1	51.52	216	96	120	120	212	200	200	108	128	
E00000014	City of Londc	-0.1	51.52	154	132	141	141	150	127	127	96	150	
E00000016	City of Londc	-0.09	51.52	281	182	200	200	260	250	250	198	260	
E00000017	City of Londc	-0.09	51.52	290	142	176	176	242	224	224	160	242	
E00000018	City of Londc	-0.09	51.52	218	122	139	139	207	186	186	131	207	
E00000019	City of Londc	-0.09	51.52	139	85	103	103	125	113	113	96	125	
E00000020	City of Londc	-0.09	51.52	226	126	148	148	200	181	181	143	200	
E00000021	City of Londc	-0.09	51.52	282	173	209	210	259	242	242	182	259	
E00000022	City of Londc	-0.1	51.52	223	110	110	110	188	172	172	129	188	
E00000023	City of Londc	-0.1	51.52	183	90	92	93	158	136	136	98	146	
E00000024	City of Londc	-0.1	51.51	276	147	230	237	228	227	227	186	215	
E00000025	City of Londc	-0.11	51.51	320	219	334	334	311	295	295	242	311	
E00000026	City of Londc	-0.1	51.52	270	163	233	233	265	264	264	232	235	
E00000027	City of Londc	-0.1	51.52	358	227	326	329	350	348	348	281	330	
E00000028	City of Londc	-0.11	51.52	151	110	162	164	144	137	137	97	144	
E00000029	City of Londc	-0.08	51.51	225	127	178	178	208	200	200	132	208	
E00000030	City of Londc	-0.08	51.52	254	124	128	128	212	195	195	140	212	
E00000031	City of Londc	-0.08	51.52	137	77	80	80	116	105	105	59	116	
E00000032	City of Londc	-0.07	51.51	369	139	142	142	285	273	273	145	285	
E00000035	City of Londc	-0.08	51.51	227	150	224	226	219	219	219	190	219	
E00000037	Barking and I	0.08	51.54	334	110	112	112	273	262	262	153	273	
E00000038	Barking and I	0.08	51.54	401	116	117	117	309	298	298	183	309	
E00000039	Barking and I	0.07	51.54	358	141	152	152	242	241	241	137	242	
E00000040	Barking and I	0.07	51.54	160	67	70	70	122	121	121	88	122	
E00000041	Barking and I	0.08	51.54	293	89	96	96	229	223	223	113	229	
E00000042	Barking and I	0.08	51.54	223	144	145	145	202	168	168	81	202	
E00000043	Barking and I	0.07	51.54	356	123	126	126	249	246	246	145	249	
E00000045	Barking and I	0.07	51.54	280	102	106	106	190	190	190	100	190	
E00000046	Barking and I	0.08	51.54	357	147	153	153	252	244	244	160	252	

potential references (object/place)

attributes

OA	OACSupergro	OACGroup	OACSubgroup	SuperGroup	SuperGroup	SuperGroup	SuperGroup	SuperGroup	SuperGroup	SuperGroup	SuperGroup	SuperGroup
E00000001	2: Cosmopoli	2d: Aspiring	2d2: Highly-C	1.41271015	2.13593408	1.41271015	1.7174029	1.99952517	1.77662579	2.03003338	1.77353178	2.05459946
E00000003	2: Cosmopoli	2d: Aspiring	2d3: EU Whit	1.2231848	2.14952606	1.2231848	1.53704747	1.70432439	1.65143422	2.00645035	1.79716143	2.01587248
E00000005	2: Cosmopoli	2d: Aspiring	2d2: Highly-C	1.12887041	2.22265946	1.12887041	1.53567011	1.88441316	1.78831511	2.12120877	1.78348253	2.10580848
E00000007	2: Cosmopoli	2d: Aspiring	2d2: Highly-C	1.42540727	2.56007158	1.42540727	1.89859794	2.14532309	2.14316618	2.45321354	2.23235845	2.49471272
E00000010	2: Cosmopoli	2d: Aspiring	2d2: Highly-C	1.30308006	2.50237339	1.30308006	1.32126081	1.89022907	2.04768844	2.44764628	1.74513632	2.20474155
E00000012	3: Ethnicity C	3b: Endeavou	3b3: Multi-Et	1.2050983	2.35383543	1.23974463	1.2050983	1.76718655	1.89993682	2.29798563	1.62374817	2.05625017
E00000013	2: Cosmopoli	2b: Inner-City	2b2: Multicu	1.34388169	2.64009226	1.34388169	1.75929767	2.16656266	2.21367844	2.58459627	2.15422594	2.5031449
E00000014	2: Cosmopoli	2b: Inner-City	2b2: Multicu	1.39201054	2.59907957	1.39201054	1.50421283	2.05215597	2.16817834	2.55901253	1.87201552	2.33116229
E00000016	2: Cosmopoli	2d: Aspiring	2d2: Highly-C	1.24497272	2.36307104	1.24497272	1.69062805	2.03877952	1.92498574	2.26968333	1.90658223	2.24672937
E00000017	2: Cosmopoli	2d: Aspiring	2d2: Highly-C	1.24933253	2.20234425	1.24933253	1.6109549	1.90454797	1.78355787	2.10420848	1.78018497	2.08906133
E00000018	2: Cosmopoli	2d: Aspiring	2d3: EU Whit	1.21653785	2.00275294	1.21653785	1.67302238	1.78097531	1.54938743	1.95084737	1.58116039	1.82053606
E00000019	2: Cosmopoli	2d: Aspiring	2d2: Highly-C	1.30540659	2.45680419	1.30540659	1.70371126	2.07433667	2.02848916	2.38877293	1.99951294	2.33615325
E00000020	2: Cosmopoli	2d: Aspiring	2d2: Highly-C	1.1824554	2.2838429	1.1824554	1.54196108	1.90328351	1.8224928	2.17982787	1.8155879	2.15713373
E00000021	2: Cosmopoli	2d: Aspiring	2d2: Highly-C	1.11584154	2.22507599	1.11584154	1.51024423	1.8333734	1.78292817	2.18110986	1.78968914	2.11055415
E00000022	3: Ethnicity C	3b: Endeavou	3b3: Multi-Et	1.17874207	2.36309233	1.22151279	1.17874207	1.71667761	1.8971902	2.2629575	1.71026181	2.09105052
E00000023	3: Ethnicity C	3b: Endeavou	3b3: Multi-Et	1.3357727	2.38595808	1.43931567	1.3357727	1.79559227	1.95140015	2.33550366	1.71774091	2.109531
E00000024	2: Cosmopoli	2d: Aspiring	2d2: Highly-C	1.22531878	2.33568864	1.22531878	1.68742402	1.95099702	1.91622326	2.30094259	2.05931863	2.3115788
E00000025	2: Cosmopoli	2d: Aspiring	2d2: Highly-C	1.08499676	2.22542315	1.08499676	1.59585575	1.92544511	1.82052526	2.19199659	1.83235978	2.14925877
E00000026	2: Cosmopoli	2b: Inner-City	2b2: Multicu	1.24644789	2.51957252	1.24644789	1.74942774	2.10245116	2.11994199	2.48180554	2.18781728	2.49852177
E00000027	2: Cosmopoli	2d: Aspiring	2d2: Highly-C	1.22463075	2.54299448	1.22463075	1.64244336	2.01952515	2.08577157	2.51346619	2.08322963	2.40949645
E00000028	2: Cosmopoli	2b: Inner-City	2b2: Multicu	1.32156471	2.61665282	1.32156471	1.82943065	2.20224572	2.2131194	2.53453065	2.2816913	2.58493634
E00000029	3: Ethnicity C	3b: Endeavou	3b3: Multi-Et	1.33230927	2.70256497	1.4450588	1.33230927	1.9713525	2.28116154	2.66884978	2.07321373	2.4702233
E00000030	3: Ethnicity C	3b: Endeavou	3b2: Banglad	1.05497651	2.40234823	1.33579164	1.05497651	1.73513984	1.97475378	2.36597909	1.6771211	2.10571285
E00000031	3: Ethnicity C	3b: Endeavou	3b3: Multi-Et	1.32555411	2.72360063	1.54563542	1.32555411	1.96785614	2.27678484	2.65173725	2.01553201	2.4130146
E00000032	3: Ethnicity C	3b: Endeavou	3b2: Banglad	1.28116858	2.69272621	1.79521045	1.28116858	1.8869849	2.28288454	2.61795121	1.91109543	2.31206368
E00000035	2: Cosmopoli	2d: Aspiring	2d2: Highly-C	1.30250728	2.60696232	1.30250728	1.76425003	2.11324062	2.16377408	2.53675489	2.22430377	2.5145841
E00000037	4: Multicultu	4b: Challenge	4b1: Asian Te	1.10707206	2.37398203	1.5965274	1.18781512	1.10707206	1.86054634	2.30327039	1.83864299	2.00155215
E00000038	4: Multicultu	4b: Challenge	4b1: Asian Te	0.97248963	2.19195902	1.62667603	1.3064173	0.97248963	1.69149203	2.11281686	1.71507427	1.81856231
E00000039	3: Ethnicity C	3b: Endeavou	3b2: Banglad	0.97907255	2.6385546	1.50003393	0.97907255	1.44678478	2.1003464	2.56464193	1.96218308	2.27410251
E00000040	3: Ethnicity C	3a: Ethnic Fa	3a2: Young F	1.02365985	2.41677102	1.29891988	1.02365985	1.28142016	1.8918883	2.34980086	1.81230224	2.10000449
E00000041	4: Multicultu	4b: Challenge	4b1: Asian Te	1.15978056	2.27651291	1.62659091	1.39035086	1.15978056	1.78154413	2.1712951	1.96828013	2.03855137
E00000042	3: Ethnicity C	3c: Ethnic Dy	3c1: Constrai	1.16161254	2.55202271	1.56663855	1.16161254	1.53934485	2.06674424	2.50640686	1.71093882	2.14246691
E00000043	3: Ethnicity C	3b: Endeavou	3b1: Striving	0.8423038	2.49272418	1.51496923	0.8423038	1.3687959	1.98819066	2.43210759	1.67151882	2.03779418
E00000045	3: Ethnicity C	3a: Ethnic Fa	3a2: Young F	0.97374161	2.39956385	1.43898773	0.97374161	1.14945498	1.85673679	2.34354815	1.77661249	2.01081879
E00000046	3: Ethnicity C	3a: Ethnic Fa	3a2: Young F	1.03175752	2.39467196	1.438224	1.03175752	1.19772489	1.857876	2.33842206	1.7836068	2.05086767
E00000048	3: Ethnicity C	3a: Ethnic Fa	3a1: Establis	0.81608162	2.47915605	1.54419427	0.81608162	1.29466411	1.94562727	2.41195757	1.63537301	1.97301041
E00000049	4: Multicultu	4b: Challenge	4b1: Asian Te	1.00236013	2.07378323	1.69448444	1.43473271	1.00236013	1.63788895	1.96088293	1.89294674	1.81901439

potential references (time/place)

attributes

year	id	State	Population	Index offenses	Violent crime	Murder	Forcible rape	Robbery	Aggravated assault	Property crime	Burglary	Larceny-theft	Motor vehicle theft
1960	1	Alabama	3266740	39920	6097	406	281	898	4512	33823	11626	19344	2853
1960	2	Alaska	226167	3730	236	23	47	64	102	3494	751	2195	548
1960	4	Arizona	1302161	39243	2704	78	209	706	1711	36539	8926	23207	4406
1960	5	Arkansas	1786272	18472	1924	152	159	443	1170	16548	5399	10250	899
1960	6	California	15717204	546069	37558	616	2859	15287	18796	508511	143102	311956	53453
1960	8	Colorado	1753947	38103	2408	73	229	1362	744	35695	9996	21949	3750
1960	9	Connecticut	2535234	29321	928	41	103	236	548	28393	8452	16653	3288
1960	10	Delaware	446292	9642	375	33	41	157	144	9267	2661	5867	739
1960	11	District of Co	763956	20725	4230	81	111	1072	2966	16495	4587	9905	2003
1960	12	Florida	4951560	133919	11061	527	403	4005	6126	122858	39966	73603	9289
1972	54	West Virgini	1781000	25584	2299	109	146	562	1482	23285	7356	13976	1953
1972	55	Wisconsin	4520000	133382	4358	126	376	1661	2195	129024	28862	89642	10520
1972	56	Wyoming	345000	10461	511	14	48	117	332	9950	2057	7190	703
1973	1	Alabama	3539000	91389	12390	468	751	2809	8362	78999	31754	39206	8039
1973	2	Alaska	330000	16313	1269	33	147	221	868	15044	3852	9456	1736
1973	4	Arizona	2058000	137966	9877	167	637	3031	6042	128089	40301	76560	11228
1973	5	Arkansas	2037000	56149	5905	180	398	1456	3871	50244	18088	29204	2952
1973	6	California	20601000	1298872	116563	1862	8357	49531	56813	1182309	407824	643488	130997
1973	8	Colorado	2437000	133933	10088	193	944	3970	4981	123845	38963	70931	13951
1973	9	Connecticut	3076000	112717	6421	102	342	2589	3388	106296	31661	58742	15893
2000	44	Rhode Island	1048319	36444	3121	45	412	922	1742	33323	6620	22038	4665
2000	45	South Caroli	4012012	209482	32293	233	1511	5883	24666	177189	38888	123094	15207
2000	46	South Dakot	754844	17511	1259	7	305	131	816	16252	2896	12558	798
2000	47	Tennessee	5689283	278218	40233	410	2186	9465	28172	237985	56344	154111	27530
2000	48	Texas	20851820	1033311	113653	1238	7856	30257	74302	919658	188975	637522	93161
2000	49	Utah	2233169	99958	5711	43	863	1242	3563	94247	14348	73438	6461
2000	50	Vermont	608827	18185	691	9	140	117	425	17494	3501	13184	809
2000	51	Virginia	7078515	214348	19943	401	1616	6295	11631	194405	30434	146158	17813
2000	53	Washington	5894121	300932	21788	196	2737	5812	13043	279144	53476	190650	35018
2000	54	West Virgini	1808344	47067	5723	46	331	749	4597	41344	9890	28139	3315
2000	55	Wisconsin	5363675	172124	12700	169	1165	4537	6829	159424	25183	119605	14636
2000	56	Wyoming	493782	16285	1316	12	160	70	1074	14969	2078	12318	573

potential references (place, place, gender)

attributes

origin	destination	Gender	originX	originY	destX	destY	flow
Antrim	Antrim	F	315806.73	403360.92	315806.73	403360.92	0
Antrim	Armagh	F	315806.73	403360.92	294480.97	339920.26	502
Antrim	Carlow	F	315806.73	403360.92	279397.34	165005.49	25
Antrim	Cavan	F	315806.73	403360.92	242635.95	304821.62	76
Antrim	Clare	F	315806.73	403360.92	130823.78	178333.56	19
Antrim	Cork	F	315806.73	403360.92	142281.36	74280.64	157
Antrim	Donegal	F	315806.73	403360.92	205940.43	408126.96	127
Antrim	Down	F	315806.73	403360.92	334046.09	347483.30	2804
Antrim	Dublin	F	315806.73	403360.92	314024.04	238670.95	862
Antrim	Fermanagh	F	315806.73	403360.92	223486.34	344754.55	42
Antrim	Galway	F	315806.73	403360.92	136788.42	235114.68	56
Antrim	Kerry	F	315806.73	403360.92	81866.57	97009.75	29
Antrim	Kildare	F	315806.73	403360.92	278834.09	215992.34	49
Antrim	Kilkenny	F	315806.73	403360.92	251965.45	147229.91	24
Antrim	Kings (Offally)	F	315806.73	403360.92	226500.35	217365.81	37
Antrim	Leitrim	F	315806.73	403360.92	198588.94	320839.72	9
Antrim	Limerick	F	315806.73	403360.92	149926.06	138169.03	57
Antrim	Londonderry	F	315806.73	403360.92	274071.84	409526.88	1430
Antrim	Longford	F	315806.73	403360.92	218533.48	272904.99	23
Antrim	Louth	F	315806.73	403360.92	304455.02	296454.51	138
Antrim	Mayo	F	315806.73	403360.92	108886.05	297515.50	36
Antrim	Meath	F	315806.73	403360.92	284326.19	264824.63	38

One reference

- Dataset may be:
  - Object-referenced
  - Time-referenced (time-series)
  - Space-referenced (spatial data)
- May have multiple attributes
  - Multivariate/multi-dimensional/high dimensional
  - Multi-dimensional time-series
  - Multi-dimensional spatial data

Two references

- Dataset may be
  - Object-referenced time-series
  - Spatial time-series
- May have multiple attributes
  - multidimensional spatial time series

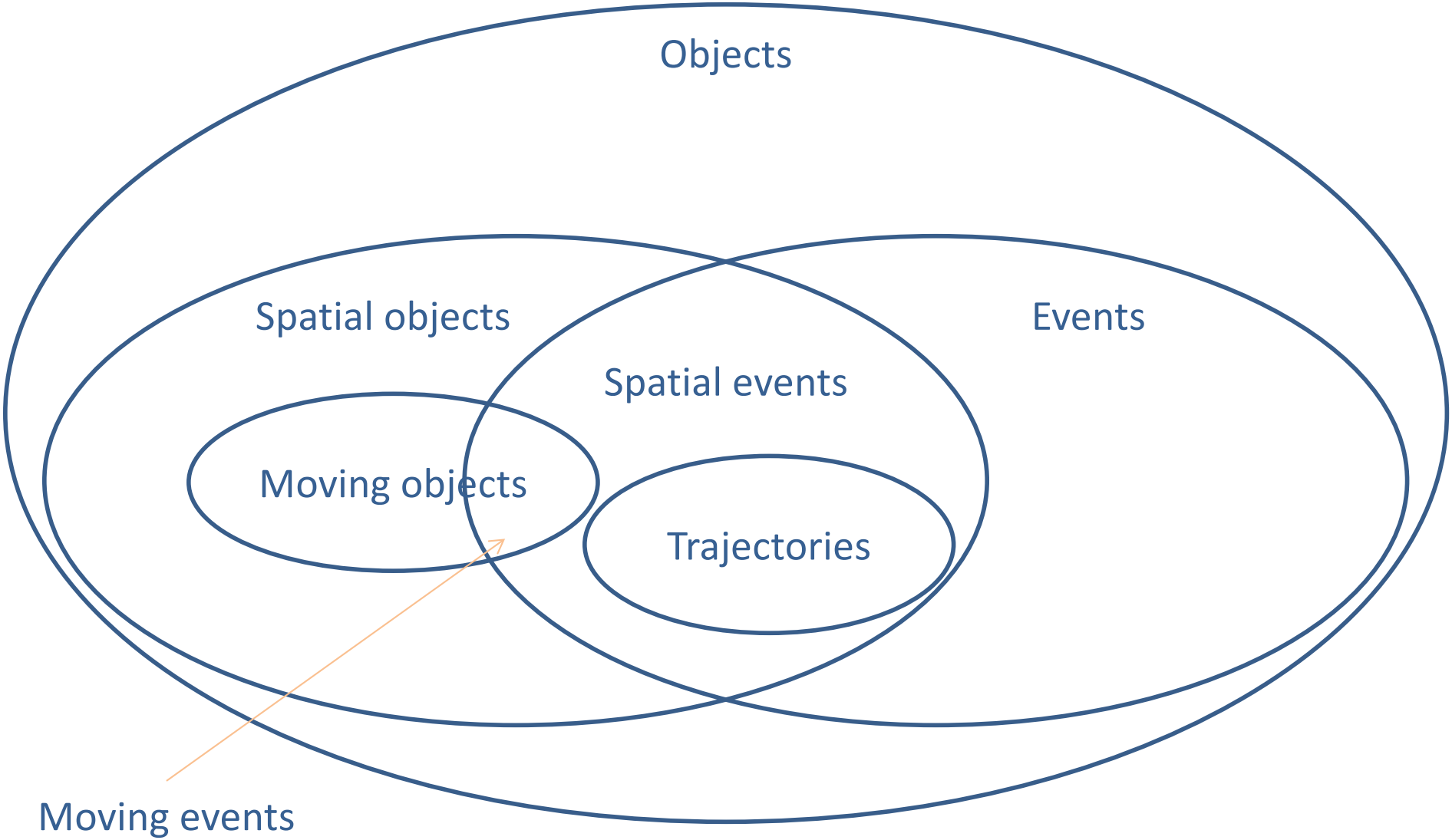
Types of Object (1)

- Generic
- Spatial objects
  - locations in space, sometimes with areal extent
- Temporal objects (events)
  - **temporal categories**: month, year, hour aggregate
  - **instant events**: e.g. tweet postings, bank transactions
  - **events with duration**: e.g. holidays, electoral campaigns…

Types of (Spatiotemporal) Object (2)

- Spatial events
  - events with location: e.g. lightning strikes, geolocated tweet posting
- Moving objects
  - object which change their location over time
  - time-series of spatial locations (trajectories)
  - E.g. people, animals, vehicles
- Trajectories
  - May have other attributes: shape, travelled distance, mean speed

Types of object



Task

Analytical tasks

- We're interested in how the attributes vary across the references
  - References: independent variables
  - Attributes: dependent variables
- Consider as a function
  - *f(indepVar)=depVar*
  - crime rates as a function of space and time
  - How do crime rates vary over space and time?

Study the behaviour of the function

- The general aim of analysis is to study the behaviour of the function:
  - **Describe**: how attributes vary
  - **Locate**: referrers and/or subsets for which particular behaviours or attribute value apply
  - **Compare:** behaviours between different **attributes** or different **subsets**
  - **Relate:** find similar behaviours
- Analysis uses specific versions of these generic tasks

Behaviour: describe

- Describe the behaviour
  - What is the distribution of values
- Examples
  - Has crime been increasing over the past decade?
  - How normal is the distribution tweets per twitter user?

Behaviour: locate

- Locate the behaviour
  - Which referrers exhibit a particular behaviour?
  - Identify
- Examples
  - Which places have both high unemployment and a high proportion of under 30s?
  - Which days of the week have low burglary?
  - Which students are above average height?

Behaviour: compare

- Compare two or more behaviours (find similarities and differences)
  - Different attributes over the same set of referrers
  - Same attributes over different subsets of referrers
- Examples
  - Does spatial distribution of pubs compare to that of craft beer bars?
  - How do grades of two modules compare?

Behaviour: relate

- Relate behaviours of two or more attributes
  - Is there a correlation between two or more attributes?
- Examples
  - Is there a relationship between density of CCTV cameras and amount of recorded crime?
  - Does Tweet frequency relate to hour of day?

Data formats and structure

Semantics are independent of structure

- Data can be structured in different ways
  - Trees: XML, JSON
  - Fields: e.g. sea surface temperature
  - Networks: e.g. social networks
  - Geometry: geographical areas
  - Tabular: CSV, Excel, tab-delimited, ASCII
- Data can be represented in different ways
- But may have the same semantics

Tabular data is common

- Most software rely on tabular data
  - Often use relational database theory to represent **1:n, n:1** or ***n:n*** relations.
  - Mondrian uses a text-based representation:
    - http://www.theusrus.de/Mondrian/Mondrian.html#ASCII
  - Tableau compiles a large table from one or more joined tables

Not all tables are the same…

- Tables can be structured in lots of different ways
- Software/libraries often requires particular structure over another.

Tidy data

- Wickham's definition, a statistical perspective on database normalisation (3NF)
  - Each variable forms a column
  - Each observation forms a row
  - Each type of *observational unit* forms a table

Wickham, H. (2014). Tidy Data. *Journal of Statistical Software, 59*(10), 1 - 23. doi:http://dx.doi.org/10.18637/jss.v059.i10

Example

Messy:

	treatmenta	treatmentb
John Smith	—	2
Jane Doe	16	11
Mary Johnson	3	1

Tidy:

person	treatment	result
John Smith	a	—
Jane Doe	a	16
Mary Johnson	a	3
John Smith	b	2
Jane Doe	b	11
Mary Johnson	b	1

Tidy data

- Cases when you want tidy data:
  - Don’t know how many *things* we’ll have
    - E.g. treatments A and B, what if we have C at some point?
  - Data transformations
    - Sorting
    - Aggregating
    - Filtering
    - Transformation

Tidy data

- When you don’t want tidy data:
  - Presentation: very long tables
  - Don’t need all the data
- Wide vs narrow (or long) data formats

Example

- UN's Human Development data
  - http://hdr.undp.org/en/data

Measure name as variable

...but different years as columns

dimension	indicator_id	indicator_name	iso3	country_name	1990	1991	1992	1993	1994	1995
Composite indices	146206	HDI rank	AFG	Afghanistan						
Composite indices	146206	HDI rank	ALB	Albania						
Composite indices	146206	HDI rank	DZA	Algeria						
Composite indices	146206	HDI rank	AND	Andorra						
Composite indices	146206	HDI rank	AGO	Angola						
Composite indices	146206	HDI rank	ATG	Antigua and Barbuda						
Composite indices	146206	HDI rank	ARG	Argentina						
Composite indices	146206	HDI rank	ARM	Armenia						
Composite indices	146206	HDI rank	AUS	Australia						
Composite indices	146206	HDI rank	AUT	Austria						
...	...	...	...	...	...	...	...	...	...	...
Demography	47906	Median age (years)	SGP	Singapore	29,3					31,8
Demography	47906	Median age (years)	SVK	Slovakia	31,2					32,4
Demography	47906	Median age (years)	SVN	Slovenia	34,1					36,2
Demography	47906	Median age (years)	SLB	Solomon Islands	16,9					17,9
Demography	47906	Median age (years)	SOM	Somalia	17,8					17,4

Messy data

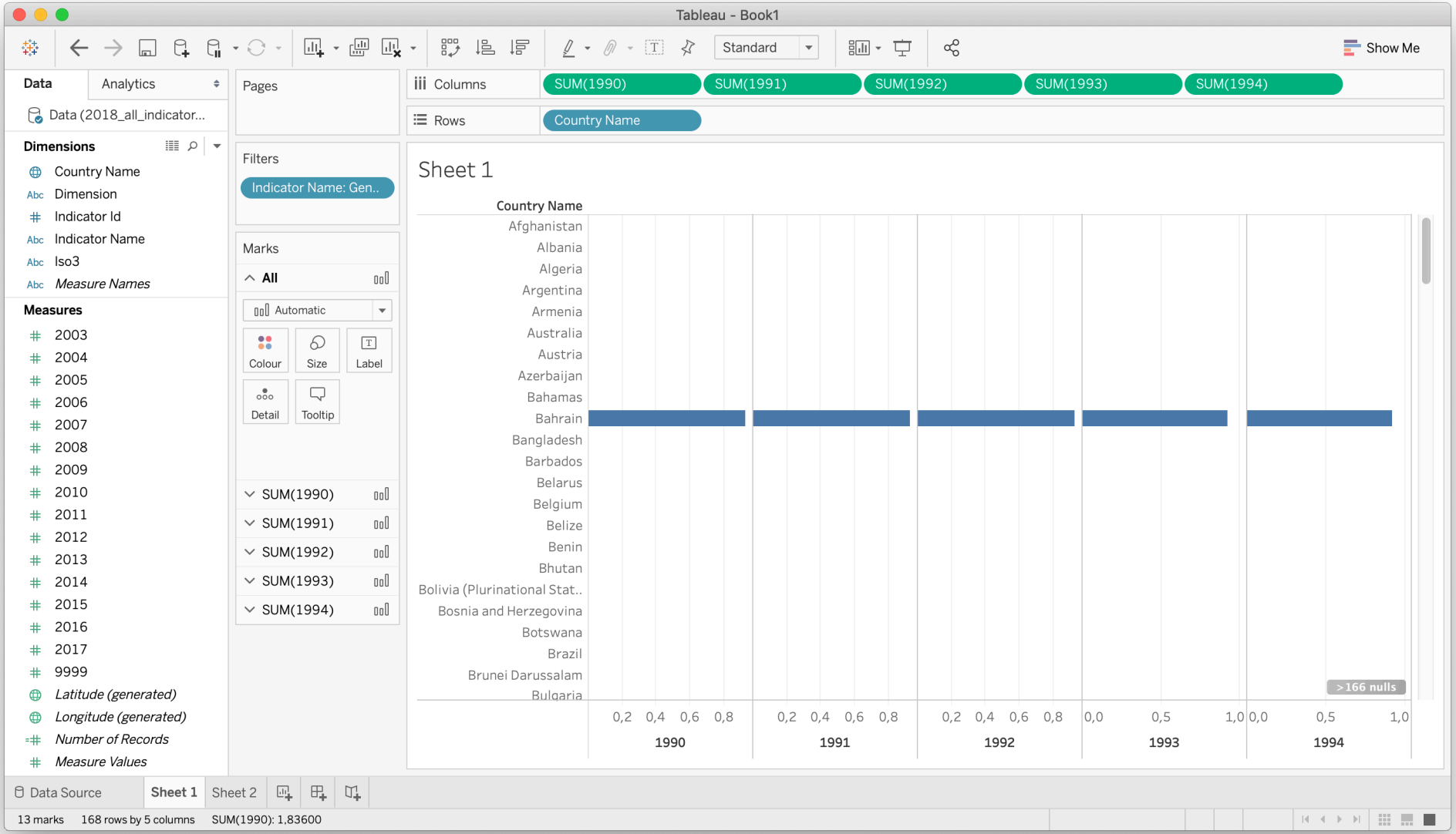


Tableau - Book1
Standard
Show Me
Data
Analytics
Data (2018_all_indicator...
Dimensions
Country Name
Dimension
Indicator Id
Indicator Name
Iso3
Measure Names
Measures
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
9999
Latitude (generated)
Longitude (generated)
Number of Records
Measure Values
Pages
Filters
Indicator Name: Gen..
Marks
All
Automatic
Colour
Size
Label
Detail
Tooltip
SUM(1990)
SUM(1991)
SUM(1992)
SUM(1993)
SUM(1994)
Columns
Rows
Country Name
Sheet 1
Afghanistan
Albania
Algeria
Argentina
Armenia
Australia
Austria
Azerbaijan
Bahamas
Bahrain
Bangladesh
Barbados
Belarus
Belgium
Belize
Benin
Bhutan
Bolivia (Plurinational Stat..
Bosnia and Herzegovina
Botswana
Brazil
Brunei Darussalam
Bulgaria
>166 nulls
0,2 0,4 0,6 0,8
1990
0,2 0,4 0,6 0,8
1991
0,2 0,4 0,6 0,8
1992
0,0 0,5 1,0
1993
0,0 0,5 1,0
1994
Data Source
Sheet 1
Sheet 2
13 marks
168 rows by 5 columns
SUM(1990): 1,83600

Tidy data

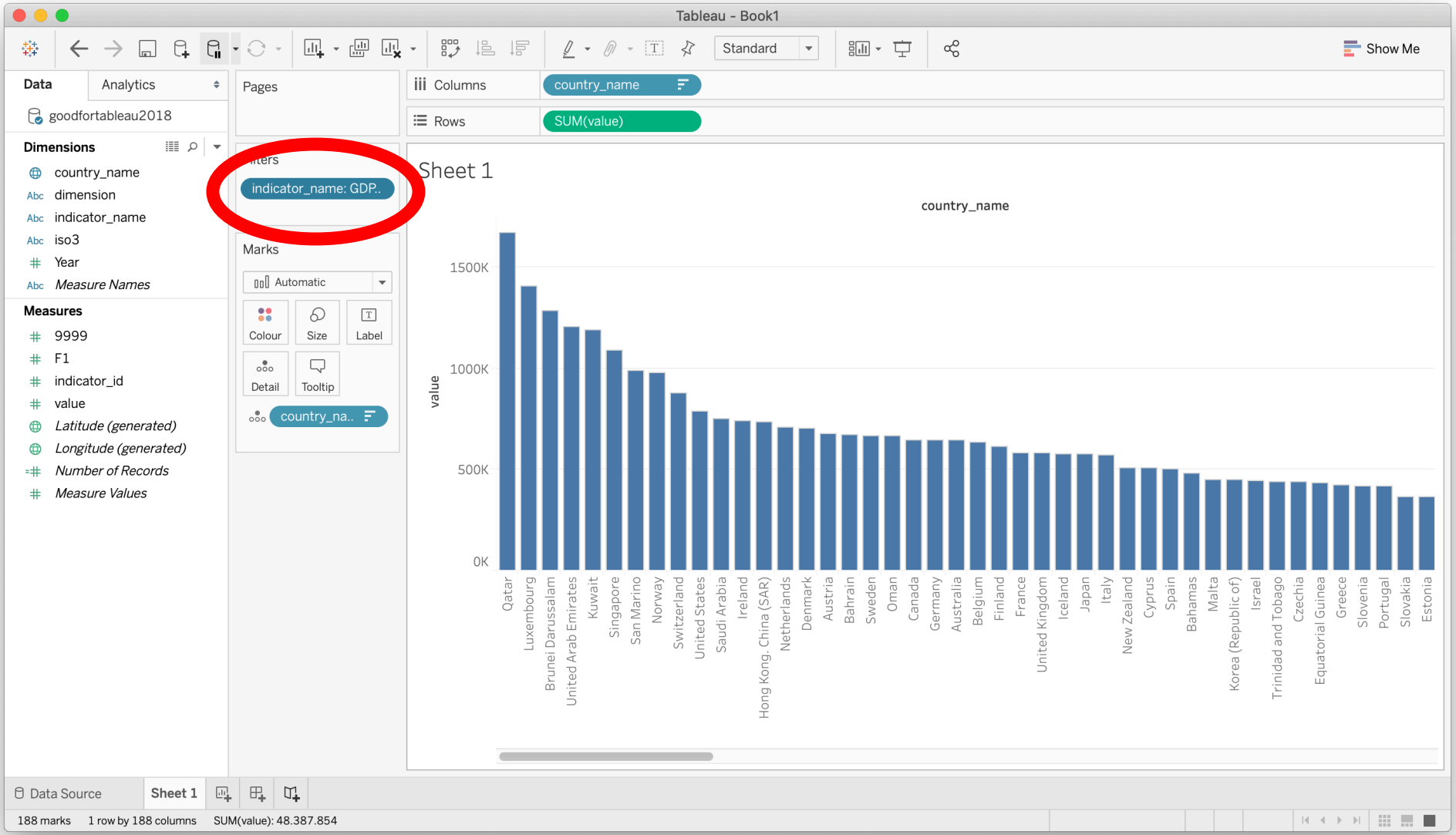
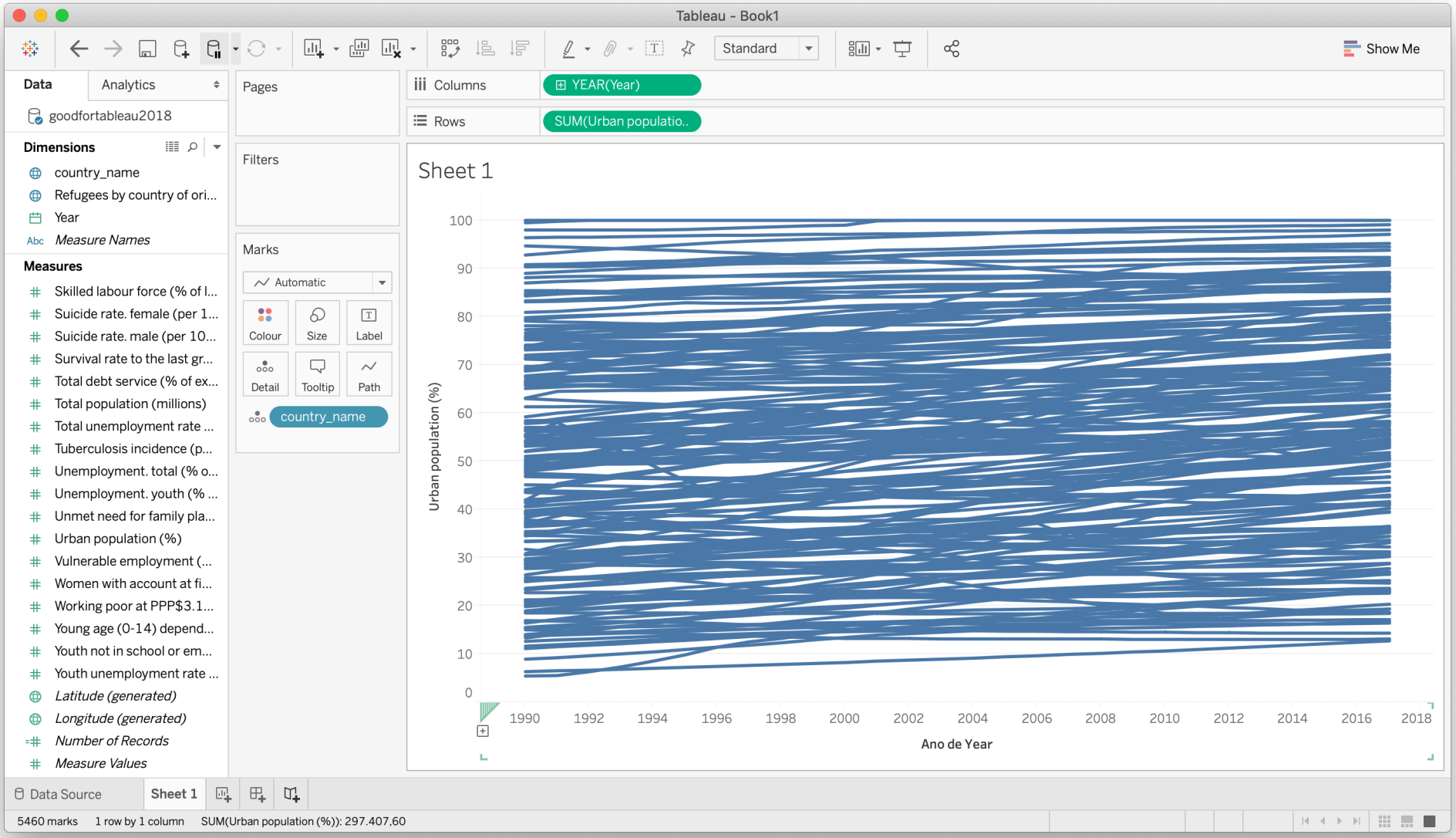


Tableau - Book1
Standard
Show Me
Data
Analytics
goodfortableau2018
Dimensions
country_name
dimension
indicator_name
iso3
Year
Measure Names
Measures
9999
F1
indicator_id
value
Latitude (generated)
Longitude (generated)
Number of Records
Measure Values
Pages
Columns
country_name
Rows
SUM(value)
indicator_name: GDP..
Marks
Automatic
Colour
Size
Label
Detail
Tooltip
country_na..
Sheet 1
country_name
value
1500K
1000K
500K
0K
Qatar
Luxembourg
Brunei Darussalam
United Arab Emirates
Kuwait
Singapore
San Marino
Norway
Switzerland
United States
Saudi Arabia
Ireland
Hong Kong, China (SAR)
Netherlands
Denmark
Austria
Bahrain
Sweden
Oman
Canada
Germany
Australia
Belgium
Finland
France
United Kingdom
Iceland
Japan
Italy
New Zealand
Cyprus
Spain
Bahamas
Malta
Korea (Republic of)
Israel
Trinidad and Tobago
Czechia
Equatorial Guinea
Greece
Slovenia
Portugal
Slovakia
Estonia
Data Source
Sheet 1
188 marks
1 row by 188 columns
SUM(value): 48.387.854

Good for Tableau



Structuring data: Wide vs long format

"Wide" (32 rows, 68 variables)

	County	X_COOR	Y_COOR	Antrim_M	Antrim_F	Armagh_M	Armagh_F	Carlow_M	Carlo
1	Antrim	315806.73	403360.92	0	0	2939	3934	33	27
2	Armagh	294480.97	339920.26	502	525	0	0	54	28
3	Carlow	279397.34	165005.49	25	20	24	13	0	0
4	Cavan	242635.95	304821.62	76	42	147	110	12	10
5	Clare	130823.78	178333.56	19	10	12	10	17	7
6	Cork	142281.36	74280.64	157	81	82	37	142	100
7	Donegal	205940.43	408126.96	127	89	51	49	6	9
8	Down	334046.09	347483.30	2804	3382	3290	3463	20	22
9	Dublin	314024.04	238670.95	862	1361	608	755	2938	3746
10	Fermanagh	223486.34	344754.55	42	35	72	67	10	6
11	Galway	136788.42	235114.68	56	36	68	45	49	28
12	Kerry	81866.57	97009.75	29	20	28	13	12	6
13	Kildare	278834.09	215992.34	49	30	48	33	648	704
14	Kilkenny	251965.45	147229.91	24	21	39	15	636	882
15	Kings	226500.35	217365.81	37	28	34	13	59	38
16	Leitrim	198588.94	320839.72	9	11	27	17	6	5
17	Limerick	149926.06	138169.03	57	37	38	21	38	29
18	Londonderry	274071.84	409526.88	1430	1610	220	145	9	5
19	Longford	218533.48	272904.99	23	10	33	22	25	11
20	Louth	304455.02	296454.51	138	121	722	969	52	33
21	Mayo	108886.05	297515.50	36	23	41	20	18	9
22	Meath	284326.19	264824.63	38	32	70	44	54	60

"Narrow/Long" (2080 rows, 6 columns)

	County	X_COOR	Y_COOR	flow	destination	gender
1	Antrim	315806.73	403360.92	0	Antrim	M
2	Armagh	294480.97	339920.26	502	Antrim	M
3	Carlow	279397.34	165005.49	25	Antrim	M
4	Cavan	242635.95	304821.62	76	Antrim	M
5	Clare	130823.78	178333.56	19	Antrim	M
6	Cork	142281.36	74280.64	157	Antrim	M
7	Donegal	205940.43	408126.96	127	Antrim	M
8	Down	334046.09	347483.30	2804	Antrim	M
9	Dublin	314024.04	238670.95	862	Antrim	M
10	Fermanagh	223486.34	344754.55	42	Antrim	M
11	Galway	136788.42	235114.68	56	Antrim	M
12	Kerry	81866.57	97009.75	29	Antrim	M
13	Kildare	278834.09	215992.34	49	Antrim	M
14	Kilkenny	251965.45	147229.91	24	Antrim	M
15	Kings	226500.35	217365.81	37	Antrim	M
16	Leitrim	198588.94	320839.72	9	Antrim	M
17	Limerick	149926.06	138169.03	57	Antrim	M
18	Londonderry	274071.84	409526.88	1430	Antrim	M
19	Longford	218533.48	272904.99	23	Antrim	M
20	Louth	304455.02	296454.51	138	Antrim	M
21	Mayo	108886.05	297515.50	36	Antrim	M
22	Meath	284326.19	264824.63	38	Antrim	M

Structuring data: Wide vs long format

- Pandas: melt, pivot_table, stack/unstack
- Human Development data
  - Wide format: each year is a different column
    - But indicator is tidy
  - Tidy: each year is a value of a new column
    - Indicator still tidy – not good for Tableau

Visually-supported data wrangling

- Excel
  - Easy to use and can be quite powerful
- TriFacta
  - https://www.trifacta.com/products/
- Open Refine
  - http://openrefine.org/
  - Used to be “Google Refine”

Summary

- Each data source will be different
- Attributes of the same type can have different semantic roles in the task
- Re-structure is needed:
  - Depending on tasks/questions we're addressing
  - Depending on the software/library we're using

Analysing multi-dimensional data with one reference

One reference

- Dataset may be:
  - Object-referenced
  - Time-referenced (time-series)
  - Space-referenced (spatial data)
- May have multiple attributes
  - Multivariate/multi-dimensional/high dimensional
  - Multi-dimensional time-series
  - Multi-dimensional spatial data

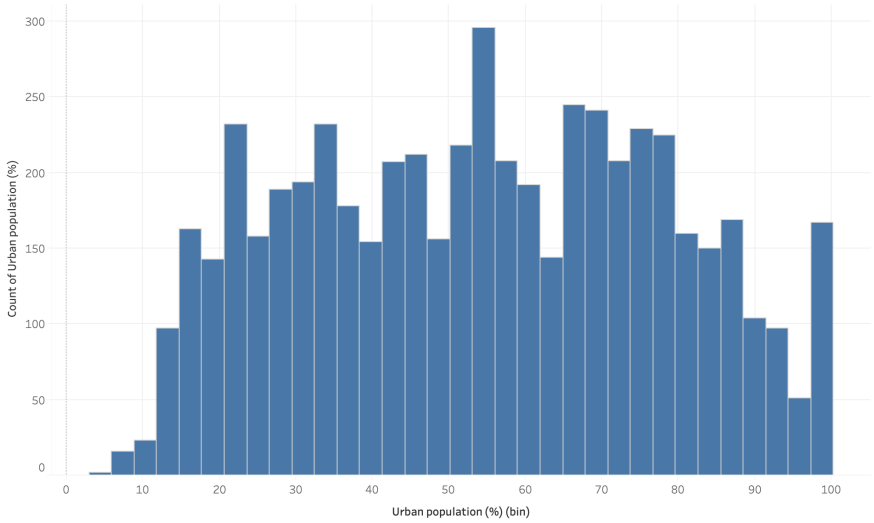
Object-referenced data

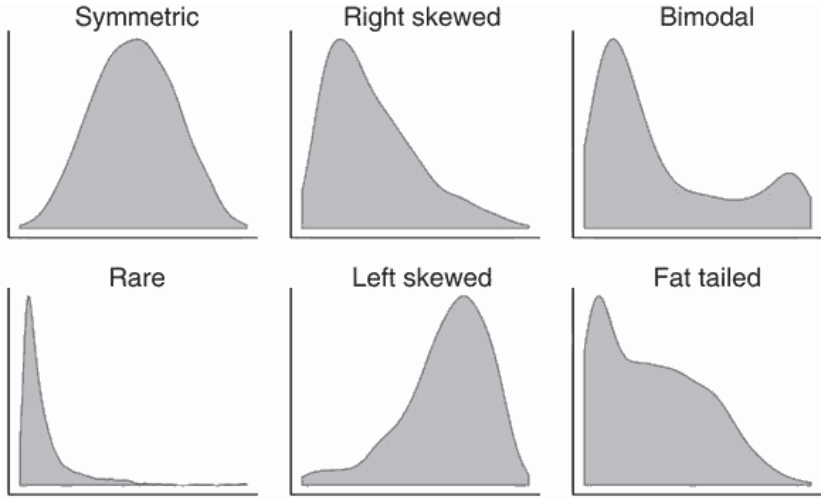
Aim

- Study the distribution of the attribute values over the set of objects
  - describe
  - locate
  - compare
  - relate

Task: Describe

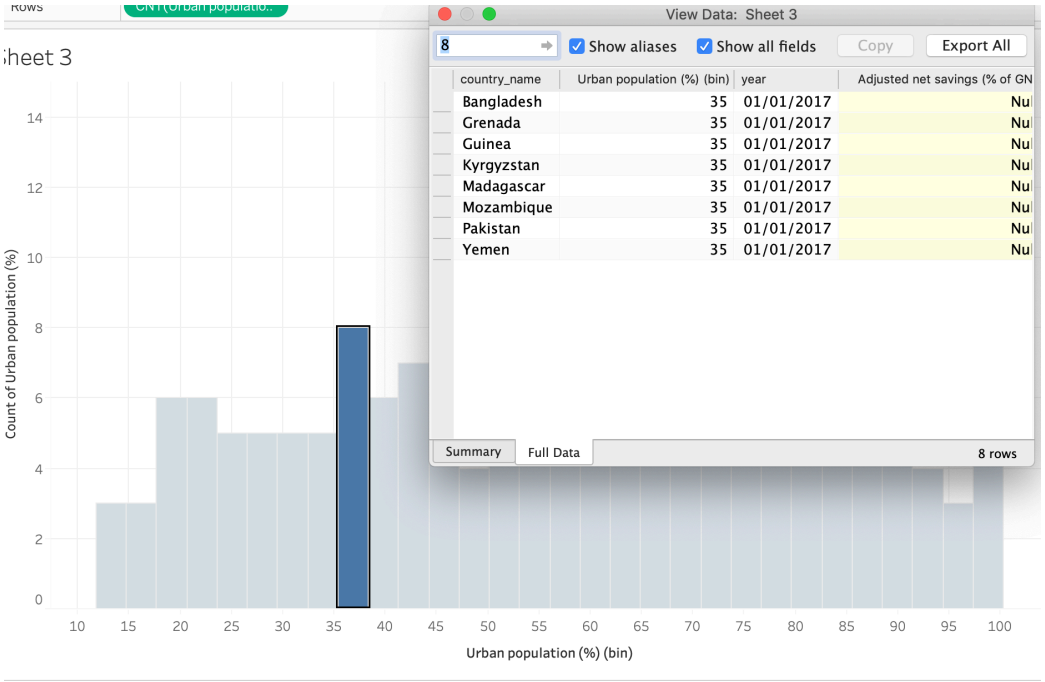
- **Task:** Describe the value distribution of a single numeric attribute
  - Use a frequency histogram to look at the shape of the distribution and any outliers





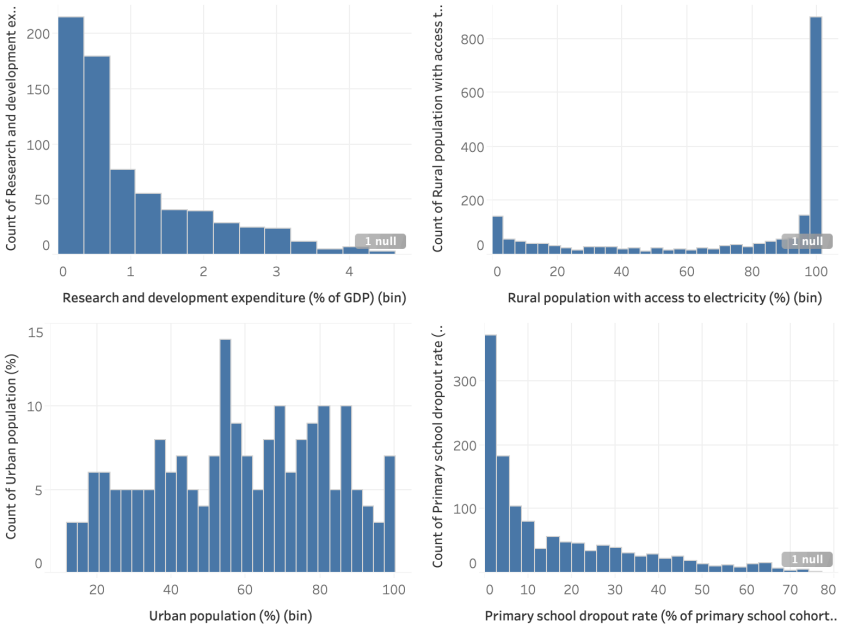
Task: Locate

- **Task:** Locate referrers with attributes of various values
  - Use interaction on a frequency histogram



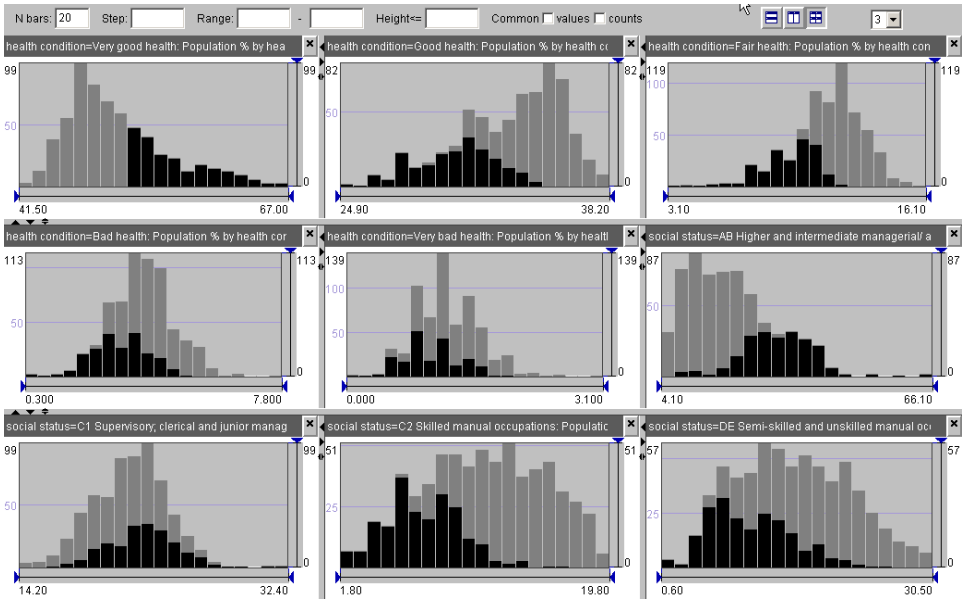
Task: Compare

- **Task:** Compare value distributions of several attributes
  - Juxtapose multiple distributions in histograms
    - make sure they scaled appropriately!



Task: Relate

- **Task:** Relate value distributions of several attributes
  - Relate a subset frequency to the whole dataset, juxtaposing those for different attributes



Task: Relate

- **Task:** Relate value distributions of two attributes
  - Use a scatterplot for pairwise comparison looking for apparent correlations, clusters and outliers

Where objects are geographic

Geographical distribution

- We might (but not always) be interested in the geographical distribution
  - Describe, locate, compare, relate
- Geographical concepts include
  - Location, distance between items, neighbourhoods, spatial distribution

Tobler's first law of geography

- “Everything is related to everything else, but near things are more related than distant things”
  - Spatial dependence
  - neighbouring objects or locations expected to have similar attribute values
  - outliers are values that deviate from that
  - But these may not hold once we spatially aggregate data (e.g., by district)

Tobler, W. (1970). "A computer movie simulating urban growth in the Detroit region”. **Economic Geography**, 46(2): 234-240.

Geographical space

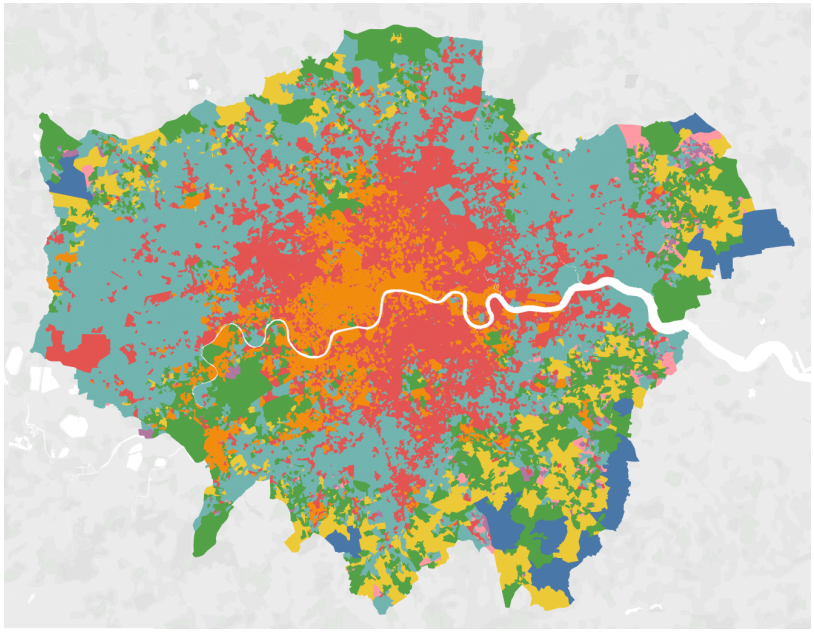
- Geographical space contains physical features
  - rivers, motorways, coastlines, land use
  - this interferes with the First Law of Geography
- So, taking geographical context into account is often important
  - distance to coastline
  - altitude
  - sources of noise/pollution

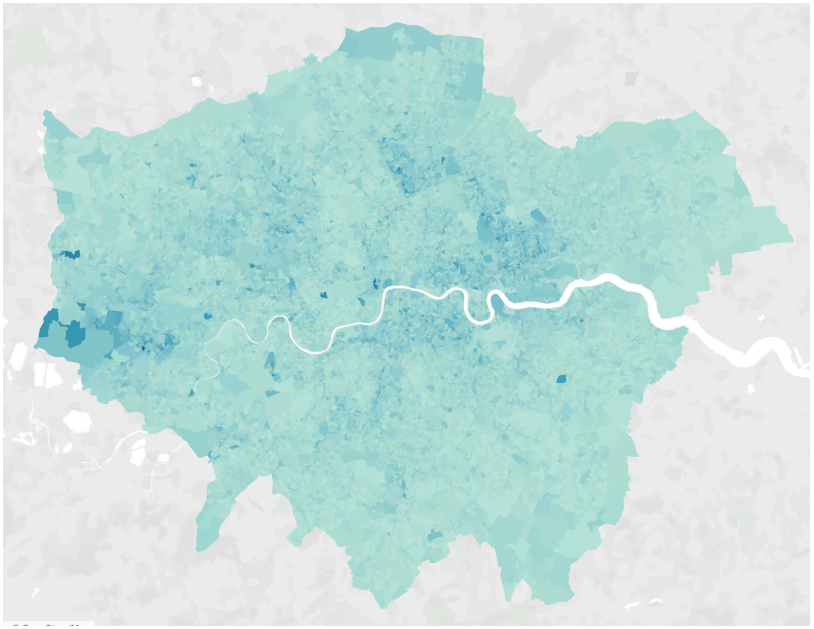
Displaying geography

- Maps
  - **use:** both dimensions of **position** (*x* and *y* visual variables)
  - **show:** geographical distributions objects and attributes, distance/neighbourhood relations, geographical context (same coordinate system)
  - **problems**: clutter/occlusion, cartographic distortion, little space to show other data
    - **solutions**: aggregate, interactive filtering, cartograms and other projections… or don't use maps

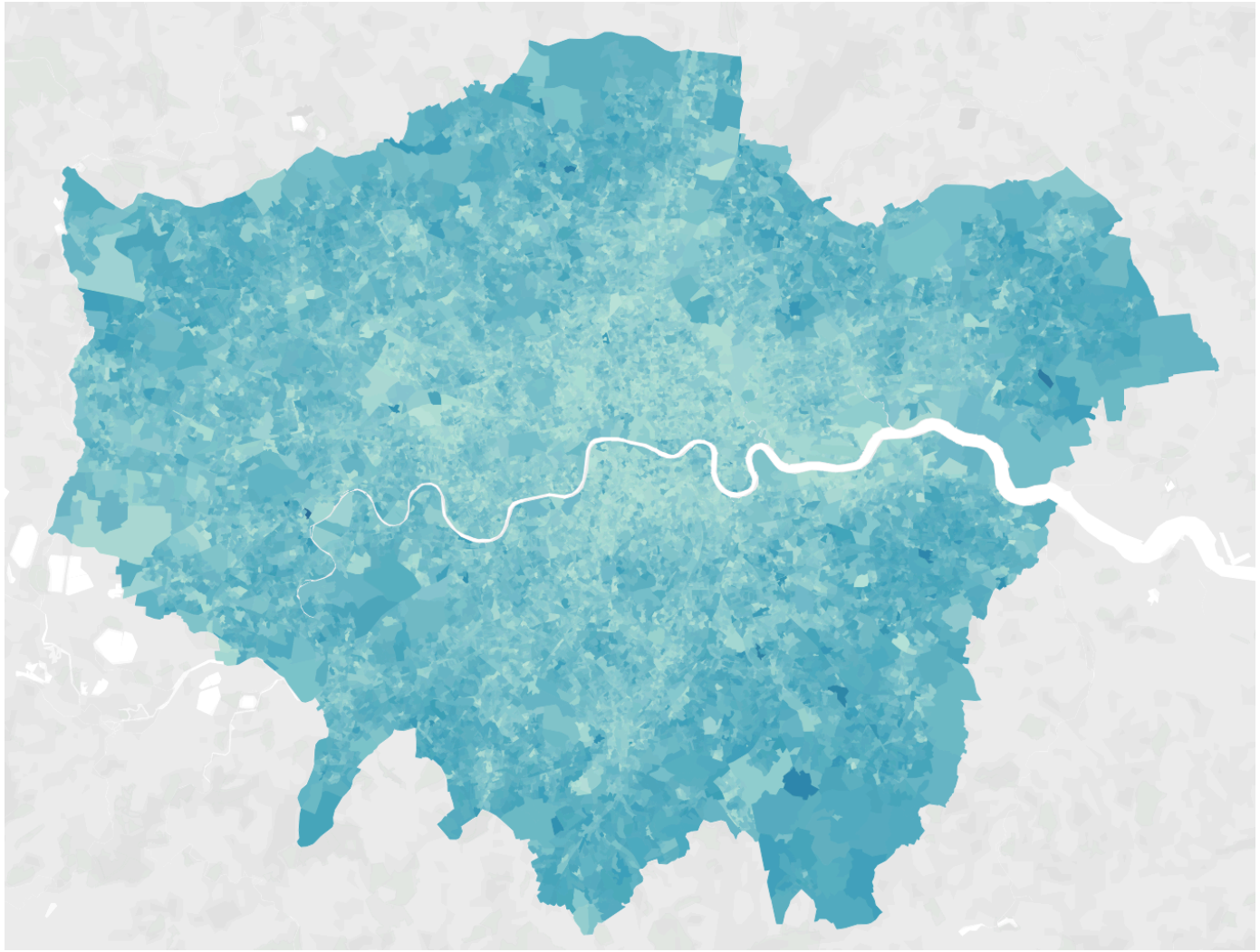
Choropleth maps

- Good for maps of spatial distributions of attribute values

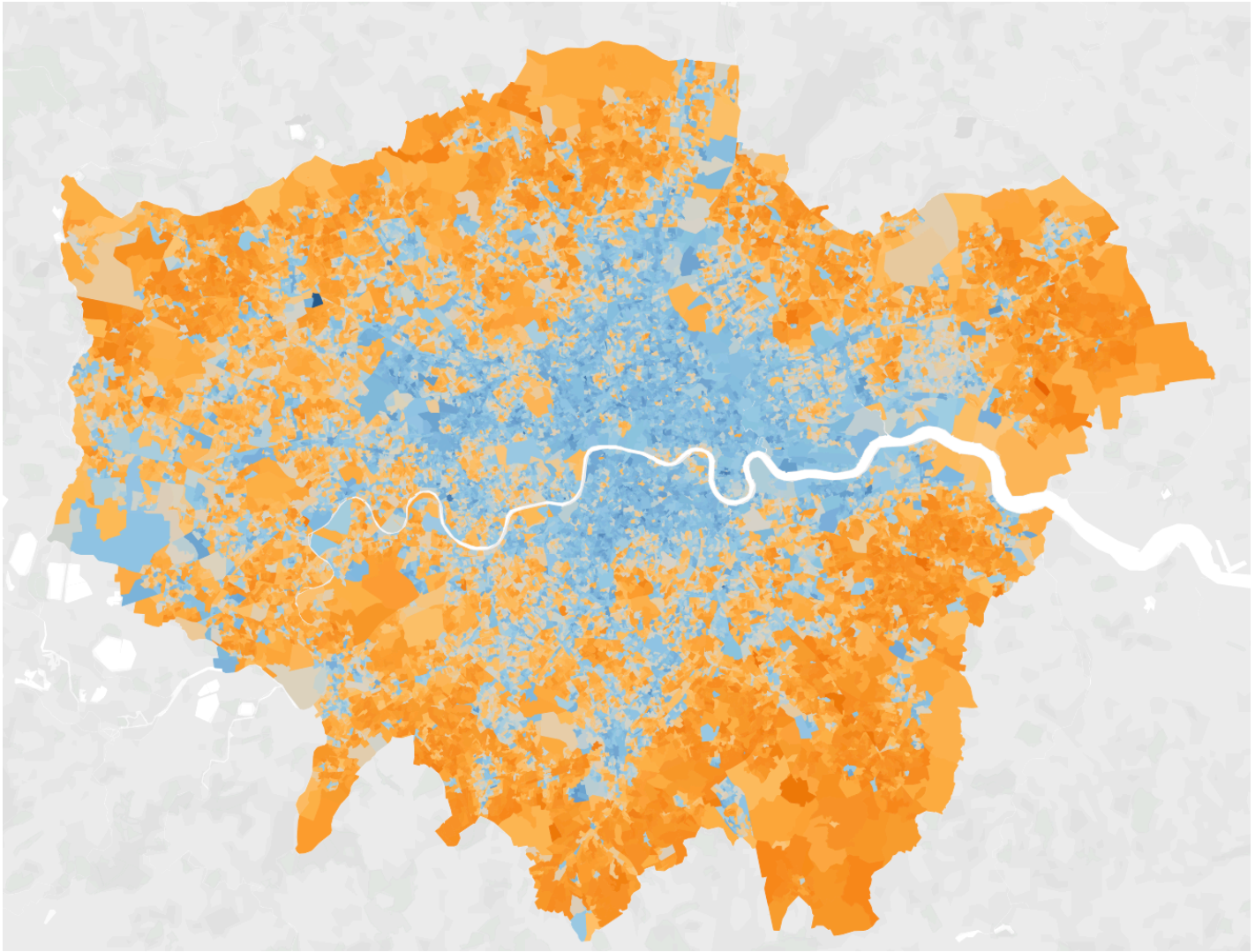




Choropleth maps: sequential single-hue

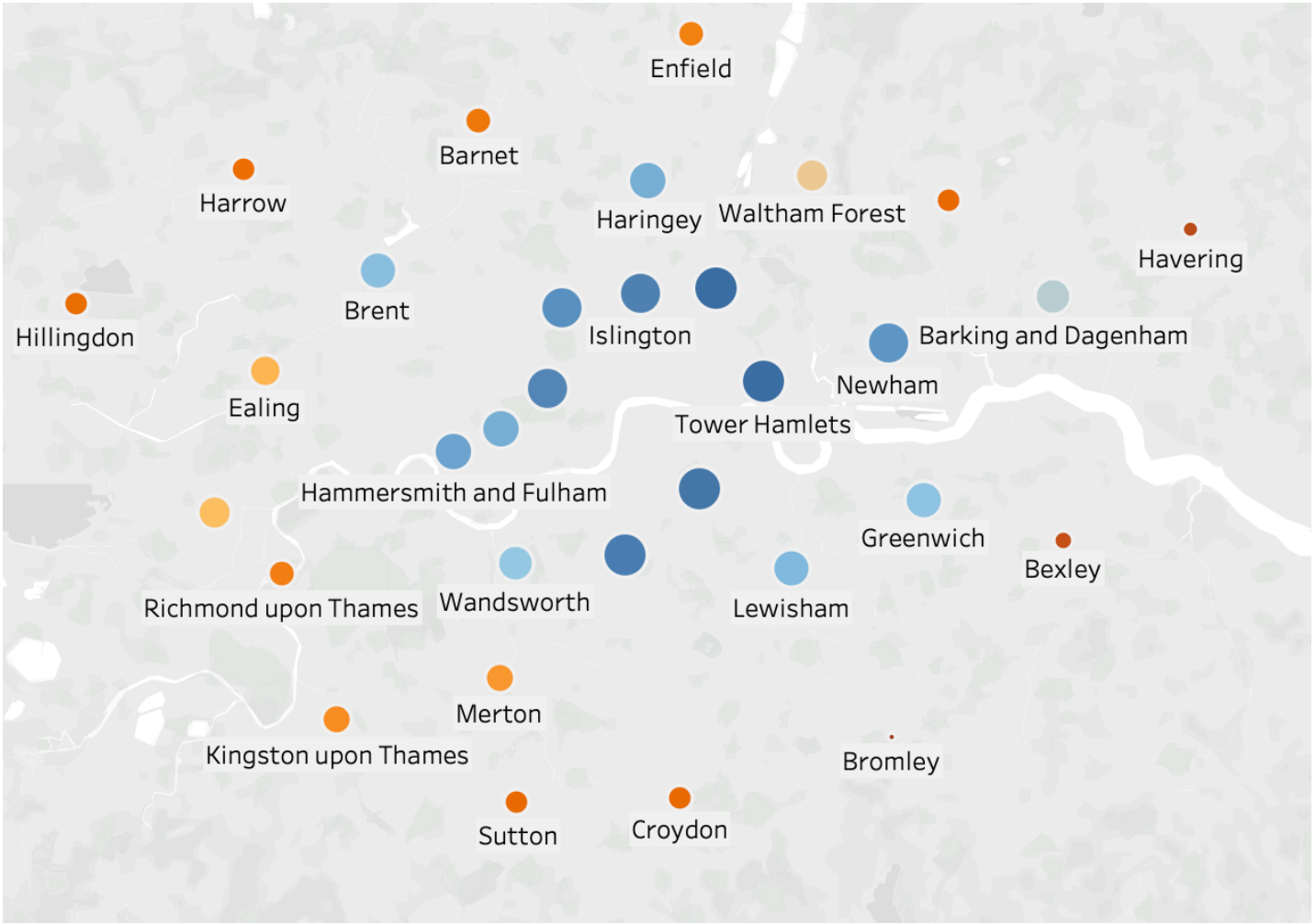


Choropleth maps: diverging multi-hues

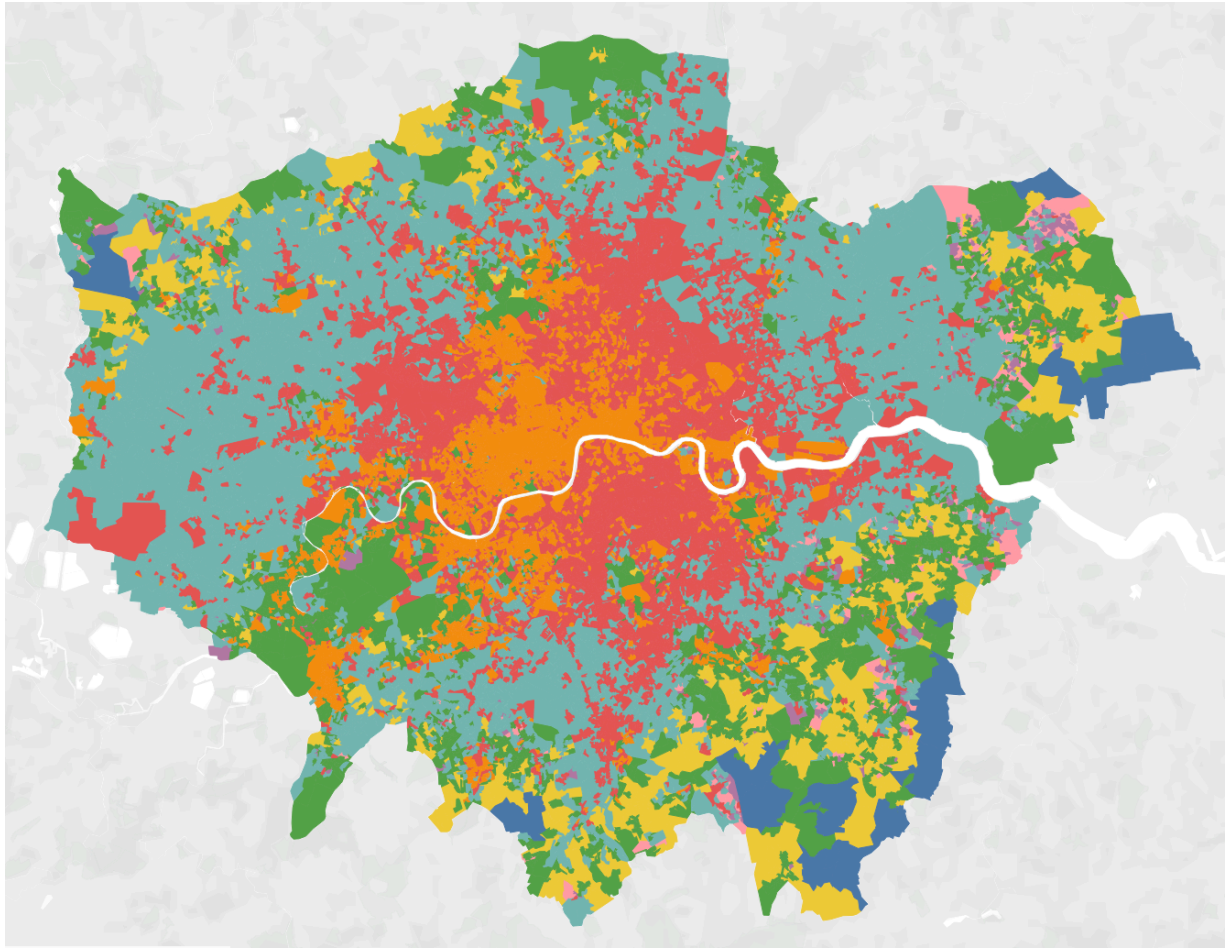


Proportional symbol map

Proportional symbol map: diverging



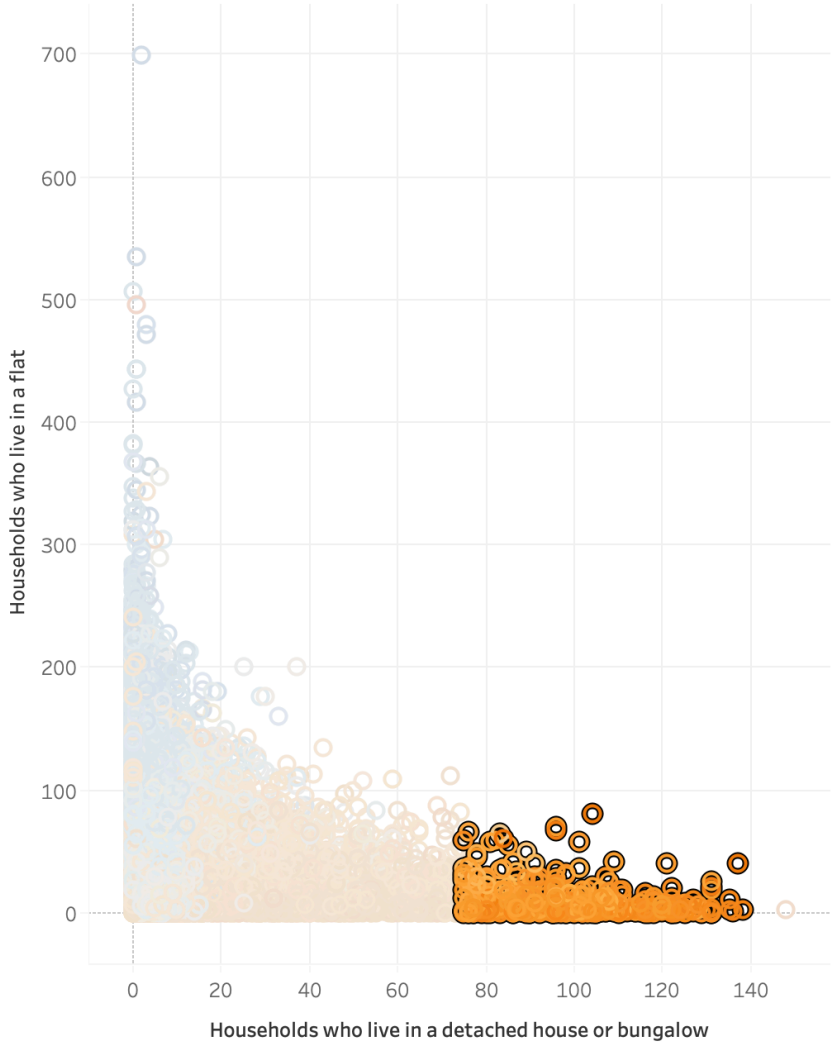
Choropleth map: qualitative

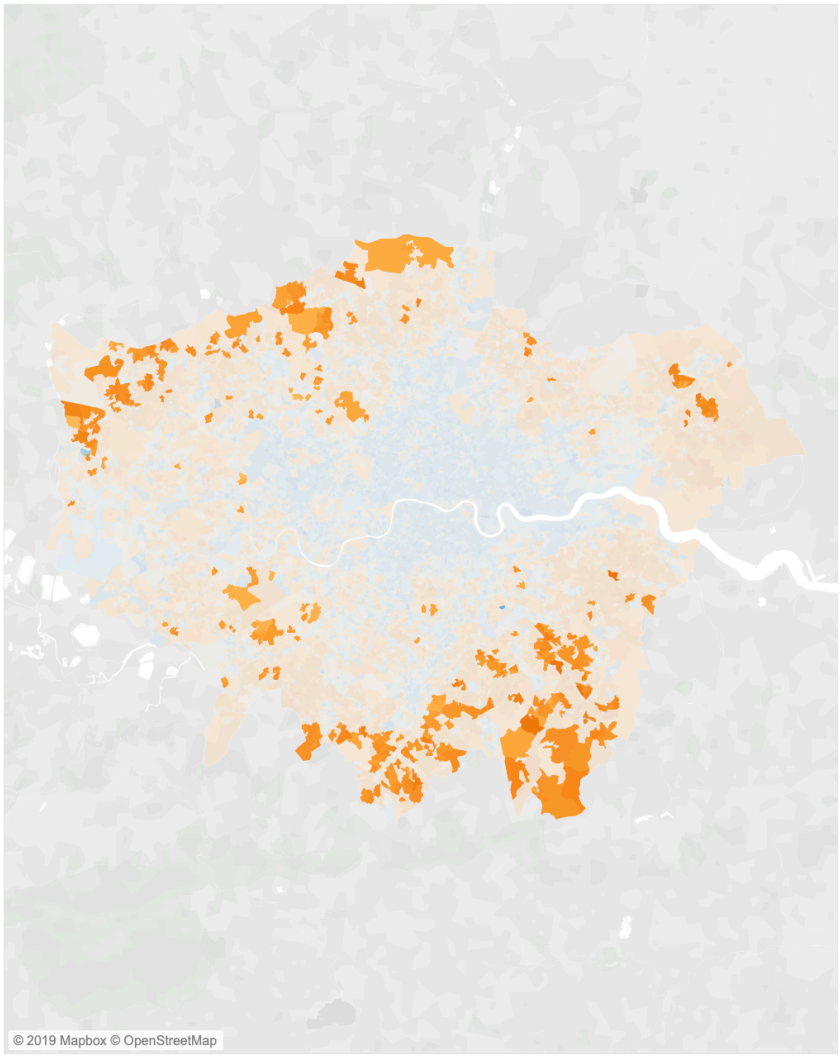


Relate geographic and non-geographic data

- Brushing and interactive linking
- Select objects on a plot, diagram, or histogram
  - see the spatial distribution of the selected objects on a map
  - are there any spatial patterns?
- Select spatial objects on a map
  - See those objects in the attribute displays
  - which values and value combinations occur in the selected part of space?

Attribute displays <-> map



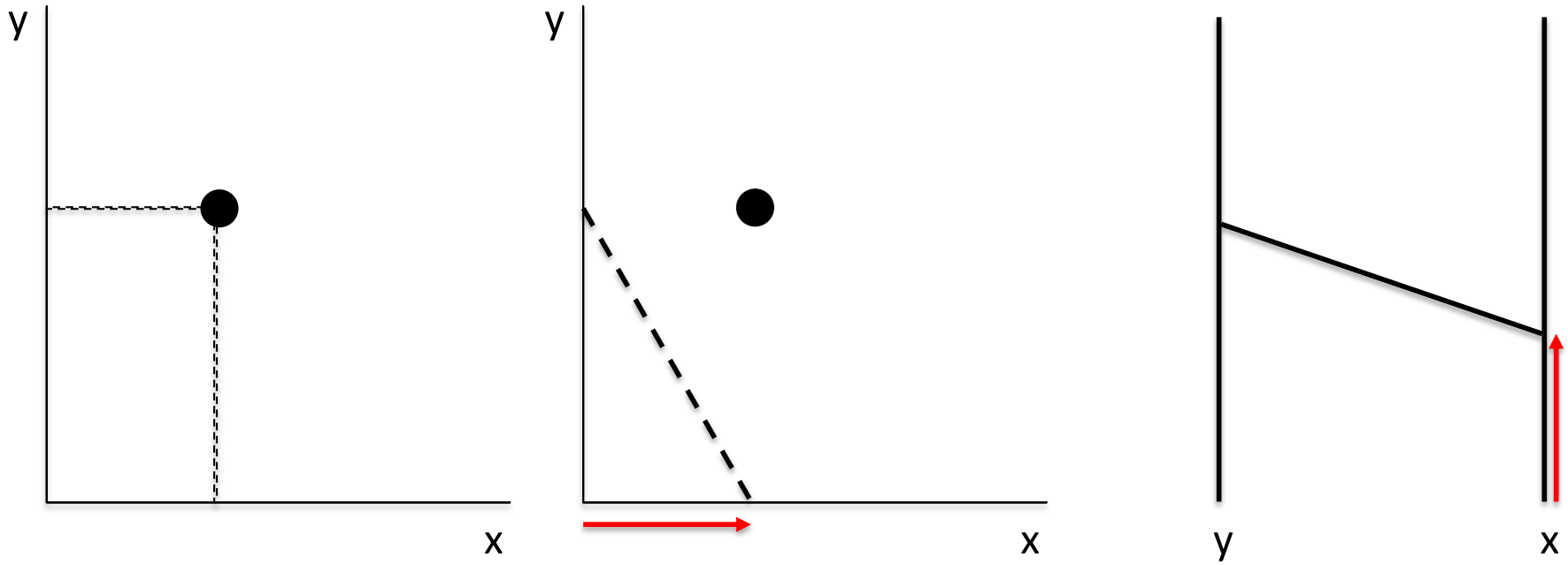


Part 2: Using analytics to simplify data

Task: Describe **with lots of data**

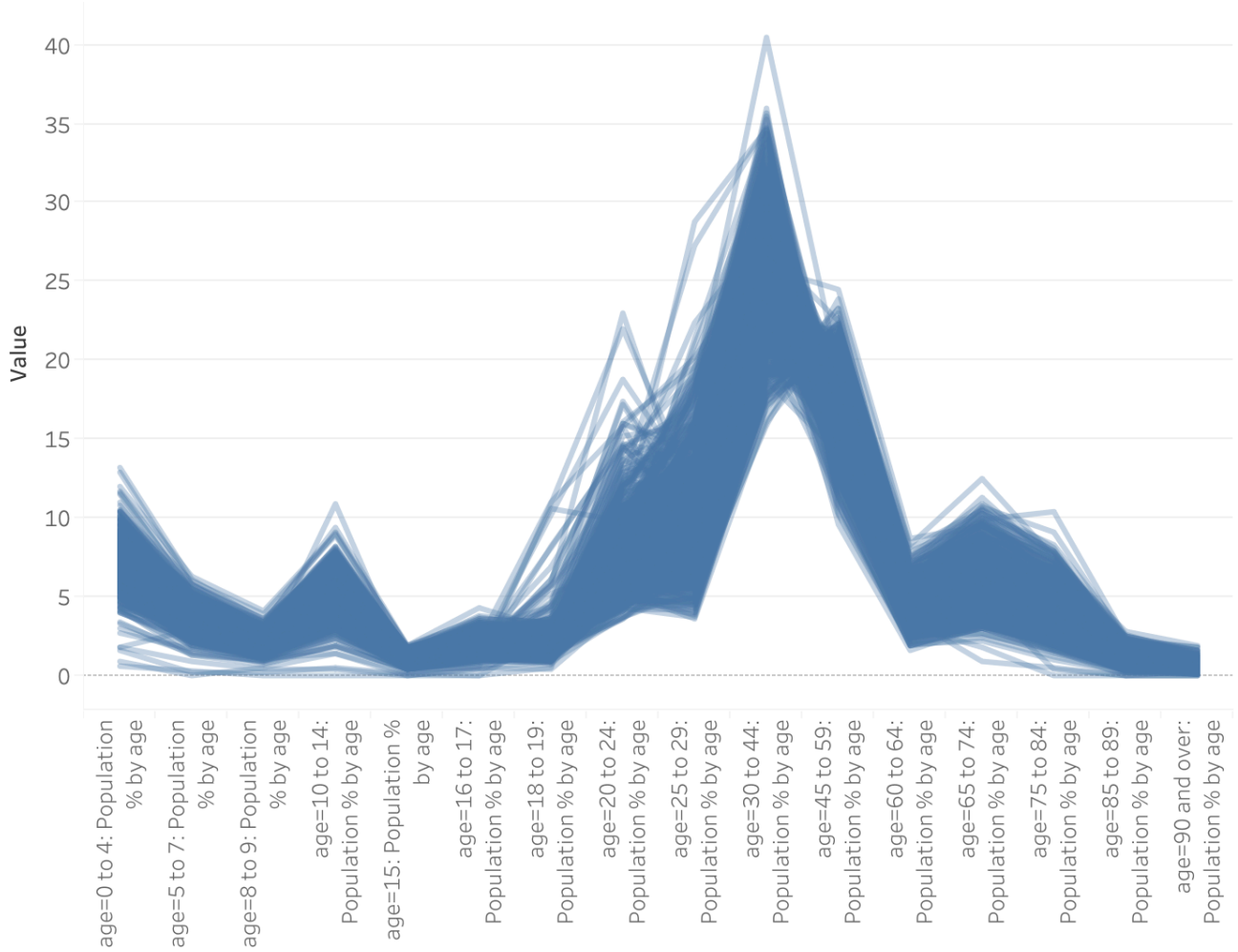
- **Task:** Describe the joint value distribution of multiple attributes
  - Tricky where there are lots of variables
    - Scatterplot matrices are pairwise
  - Try using summary statistics (deciles, interquartiles)
  - Try visually: transparency, binning

Parallel coordinates



Task: Describe with **lots of data**

~650 wards as lines, 16 variables



Simplify the data

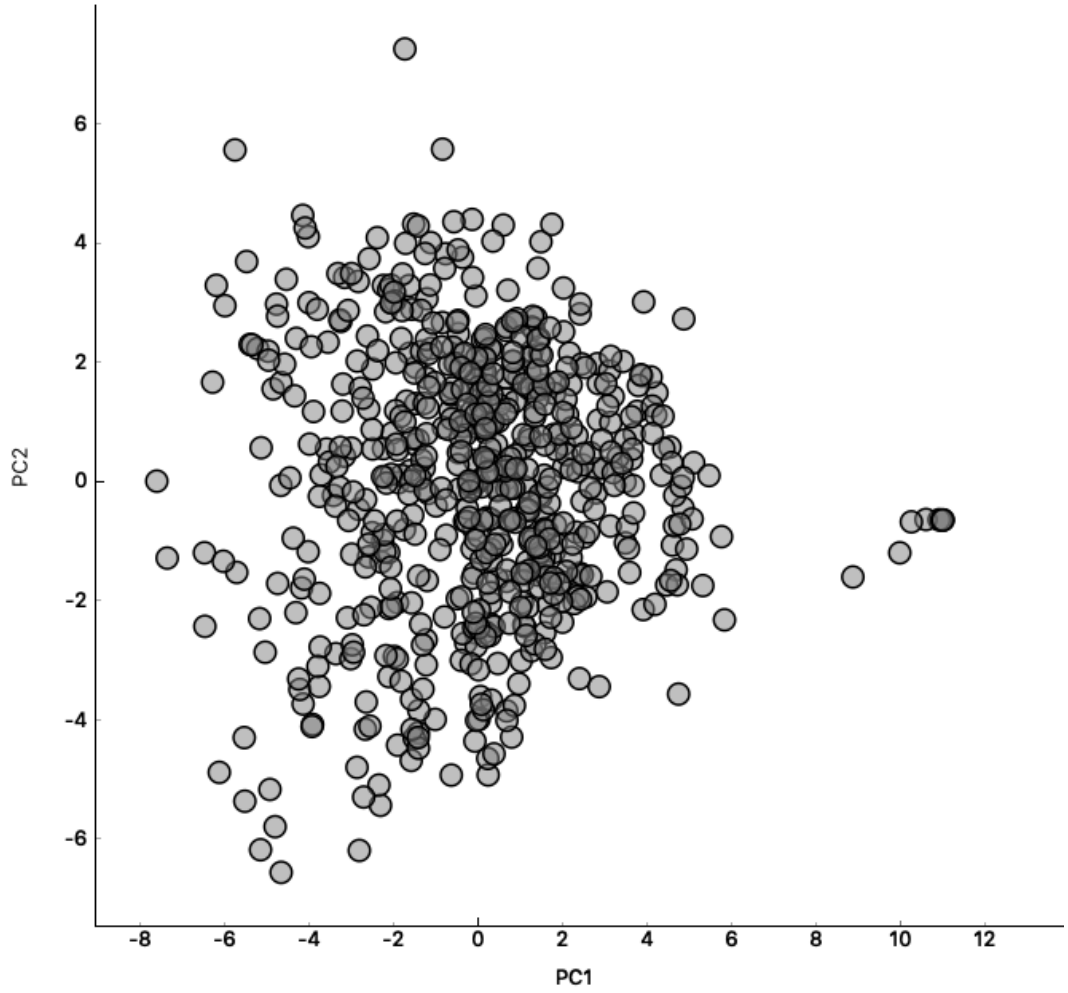
- When there are too many data points, simplify.
- Two approaches:
- 1. Reduce the number of **attributes**
  - Dimensionality reduction: create two synthetic variables that try to represent the variation
- 2. Reduce the number of objects (referrers)
  - **Group** the objects into a (much) smaller set of *representative* groups
    - low within variation and high between variation
  - **Summarise** the attributes by group

Approach 1: Dimension-reduction

Approach 1: Dimension-reduction

- (Reducing the number of attributes)
- There are many approaches to reduce many dimensions to two (so they can be plotted)
  - multidimensional scaling (MDS), principal component analysis (PCA), Sammon's mapping

Approach 1: Dimension-reduction

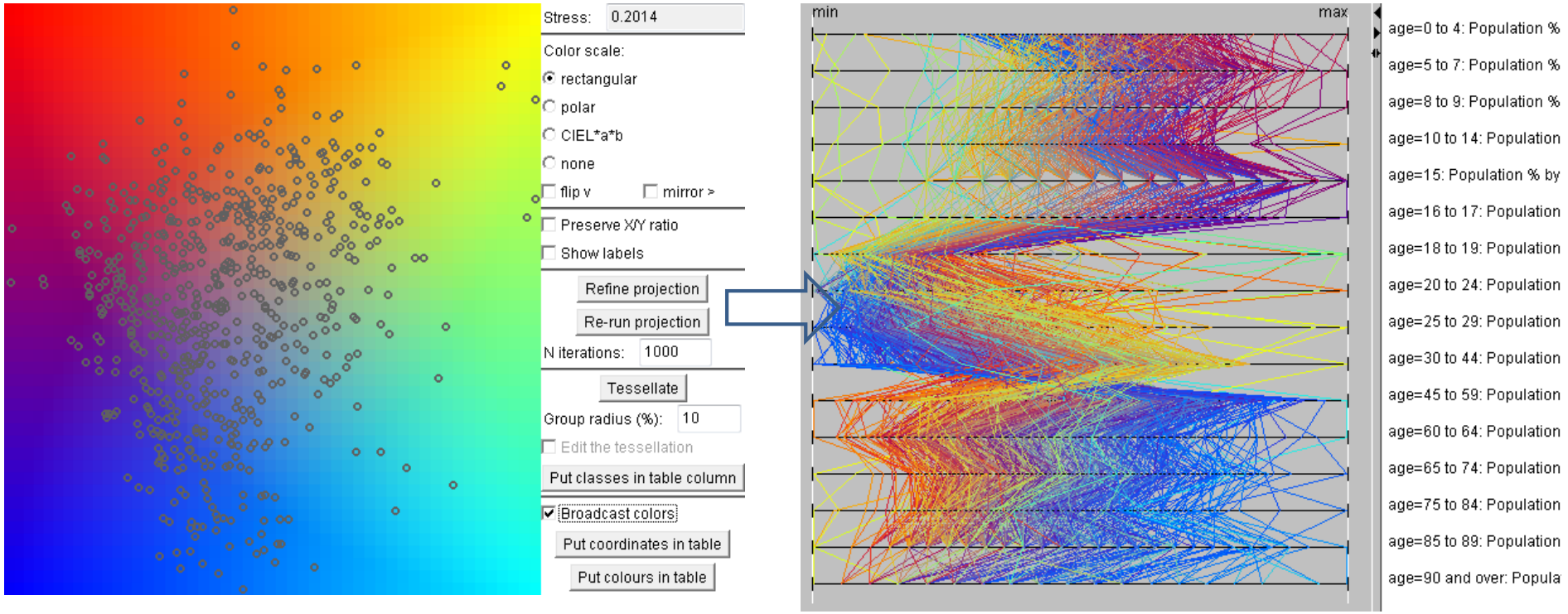


Approach 1: Dimension-reduction

- Is it useful?
  - Less clutter
  - Shows objects that (likely) have **similar** characteristics (close to each other)
    - similar = similar combinations of attribute values
  - Shows (likely) **outlier** objects
  - Shows **groups** of objects that are (likely) similar
- But…
  - These two dimensions are not interpretable
  - Need to link them back to the original data

Approach 1: Dimension-reduction

- We can use hue!
  - "Sammon's mapping" provides a 2D-varying hue
  - We can relate object in two projections with hue



Approach 1: Dimension-reduction

- Problems
  - Inherent distortions
    - Close means similar over many attributes
  - The projection itself only gives an approximate understanding of the data distribution.
  - Try different methods of simplification
    - consistent results can be trusted
    - discrepancies require detailed investigation

Approach 2: Object-based clustering

Approach 2: Object-based clustering

- Reducing the number of objects
- Two major types of clustering
  - **Partition-based clustering**: all object allocated
  - **Density-based clustering**: not every object allocated
- We'll use **partition-based clustering** to group our objects with similar attribute "signatures". Aim:
  - objects **within** a group should be similar.
  - objects should be dissimilar **between** groups.
- How does it work?
  - https://educlust.dbvis.de
  - Principles of Data Science module ☺

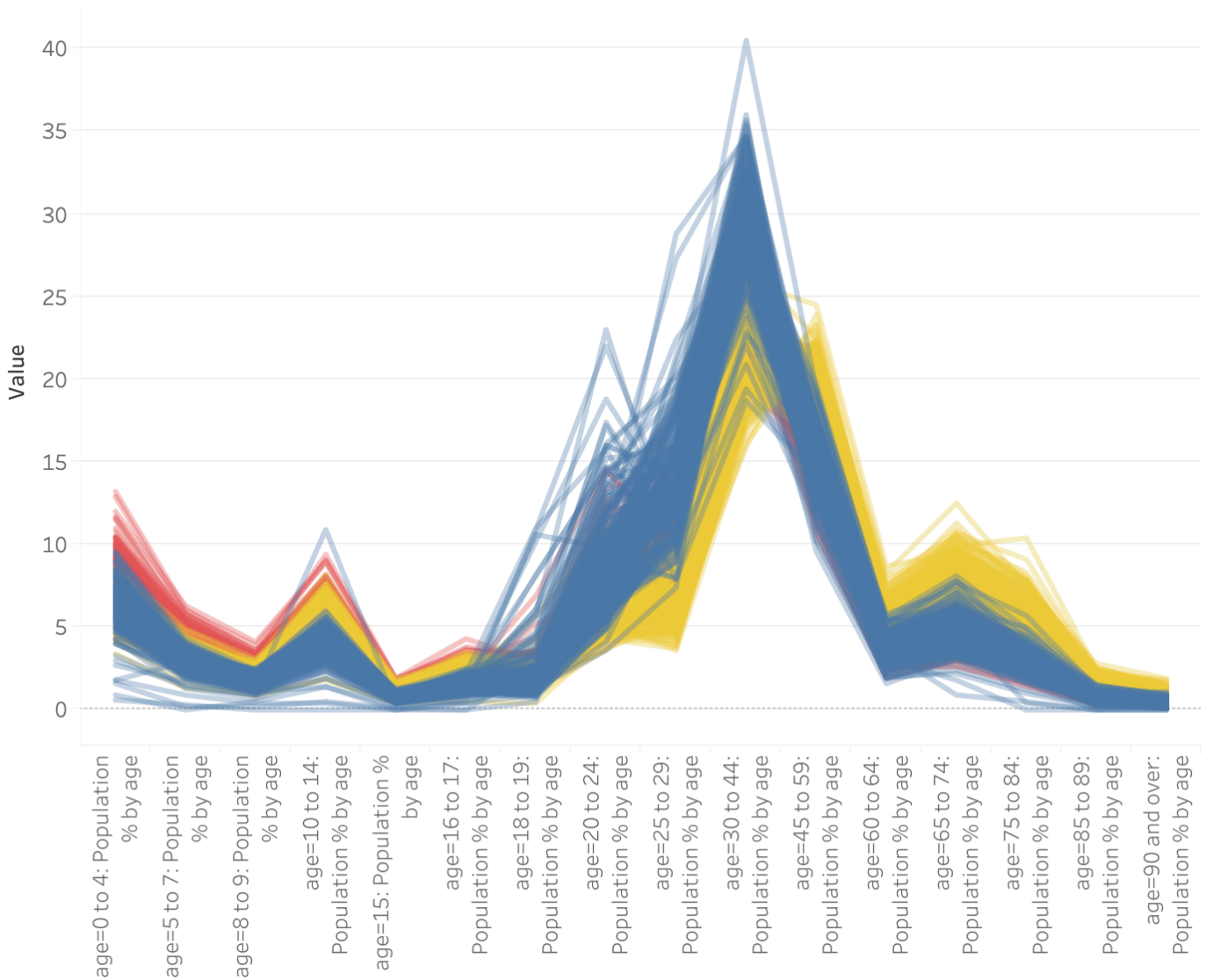
Partition-based clustering

- Most methods ask for:
  - the desired number of clusters
  - the features with which to cluster (and weighting)
  - sometime other parameters
- It's hard to choose good ones, so try a few:
  - you need to discriminate objects based on characteristics that are of interest to your analysis
  - **don't forget the purpose is to help your analysis!**
  - use interactive graphics to compare

Partition-based clustering

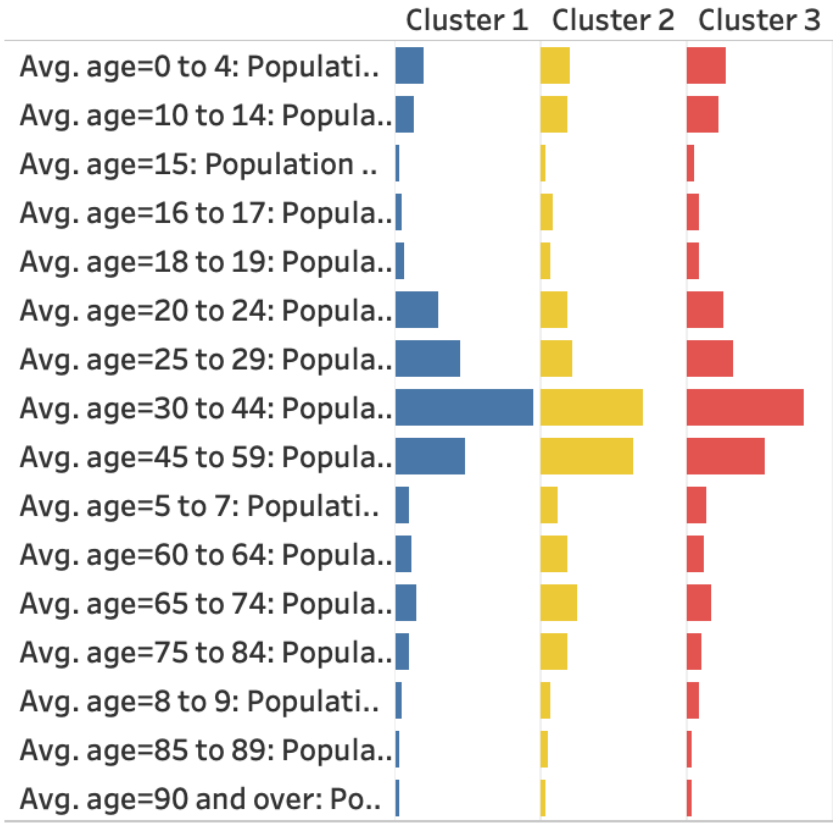
- Output
  - A cluster id assigned to each object
- Interpret
  - How do attributes vary within & between clusters
- Note that:
  - There's often a stochastic element, so slightly different solutions each time (e.g. initial state, seeds)
  - Hue is a good way to show cluster (why?)
  - Often hard to compare alternative cluster solutions

A 3-cluster solution



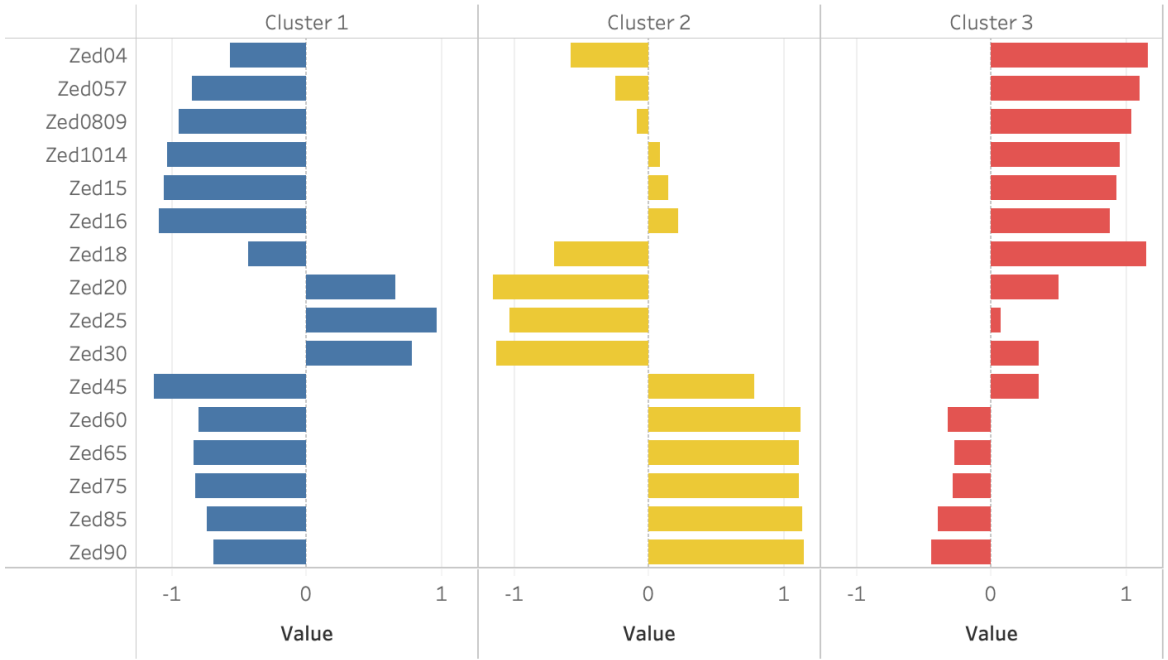
Attribute values by cluster

- But large magnitude differences make differences between clusters hard to see

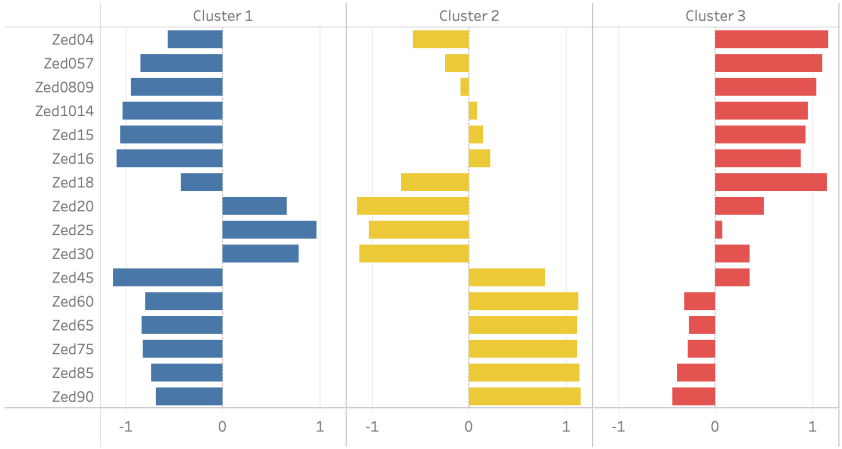


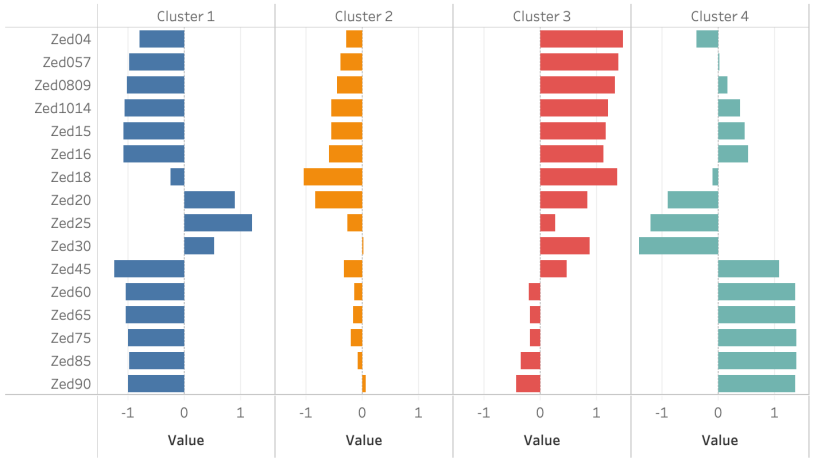
Transformation to z-scores

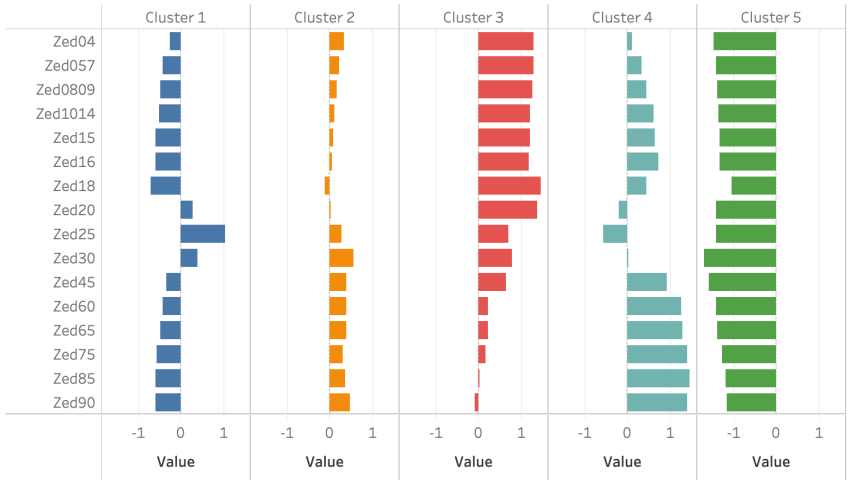
- z-score is the deviation from the mean divided by the standard deviation



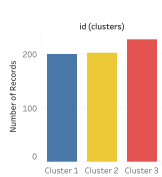
Impact of k


Cluster 1
Cluster 2
Cluster 3
Zed04
Zed057
Zed0809
Zed1014
Zed15
Zed16
Zed18
Zed20
Zed25
Zed30
Zed45
Zed60
Zed65
Zed75
Zed85
Zed90
-1
0
1


Cluster 1
Cluster 2
Cluster 3
Cluster 4
Zed04
Zed057
Zed0809
Zed1014
Zed15
Zed16
Zed18
Zed20
Zed25
Zed30
Zed45
Zed60
Zed65
Zed75
Zed85
Zed90
-1
0
1
Value

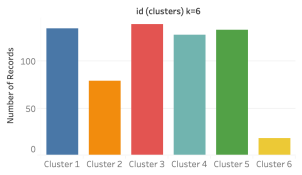

Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Zed04
Zed057
Zed0809
Zed1014
Zed15
Zed16
Zed18
Zed20
Zed25
Zed30
Zed45
Zed60
Zed65
Zed75
Zed85
Zed90
-1
0
1
Value


Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Zed04
Zed057
Zed0809
Zed1014
Zed15
Zed16
Zed18
Zed20
Zed25
Zed30
Zed45
Zed60
Zed65
Zed75
Zed85
Zed90
-2
-1
0
1
Value


id (clusters)
Number of Records
200
100
0
Cluster 1
Cluster 2
Cluster 3


id (clusters) (1)
Number of Records
150
100
50
0
Cluster 1
Cluster 2
Cluster 3
Cluster 4


id (clusters) (2)
Number of Records
150
100
50
0
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5

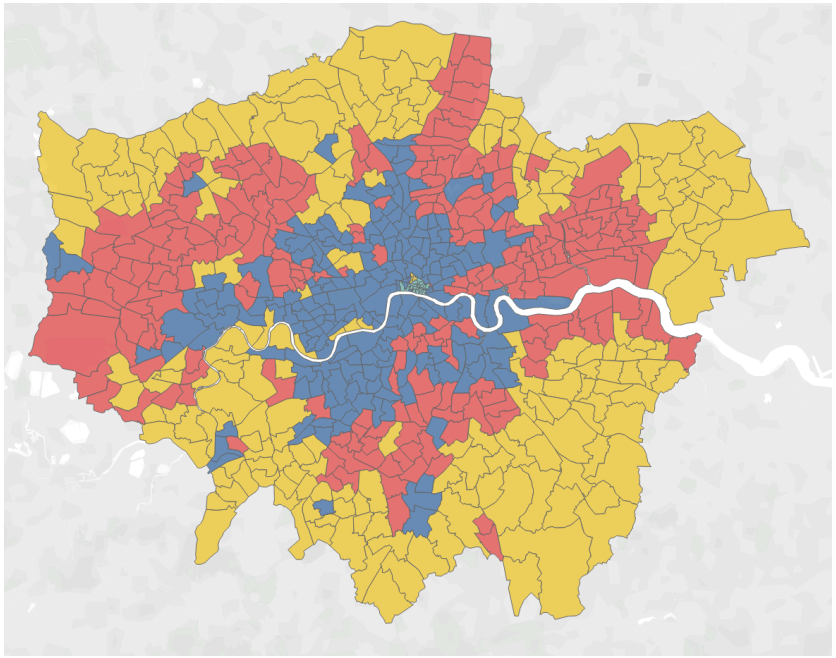

id (clusters) k=6
Number of Records
100
50
0
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6

No “right” and “wrong” groupings!

- Choose a solution that supports your analysis
- All groupings are a (huge) simplification
  - and information loss
- Try different solutions and explore these with interactive graphics

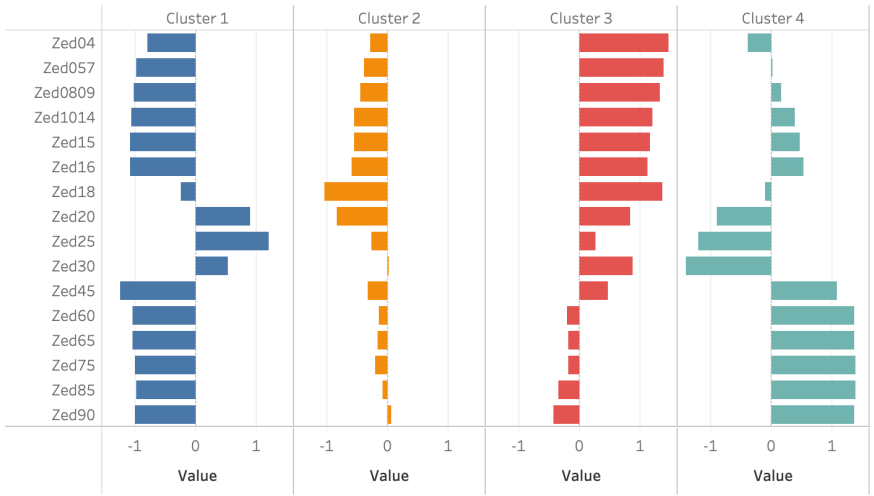
Partition-based clustering for space-referenced multidimensional data

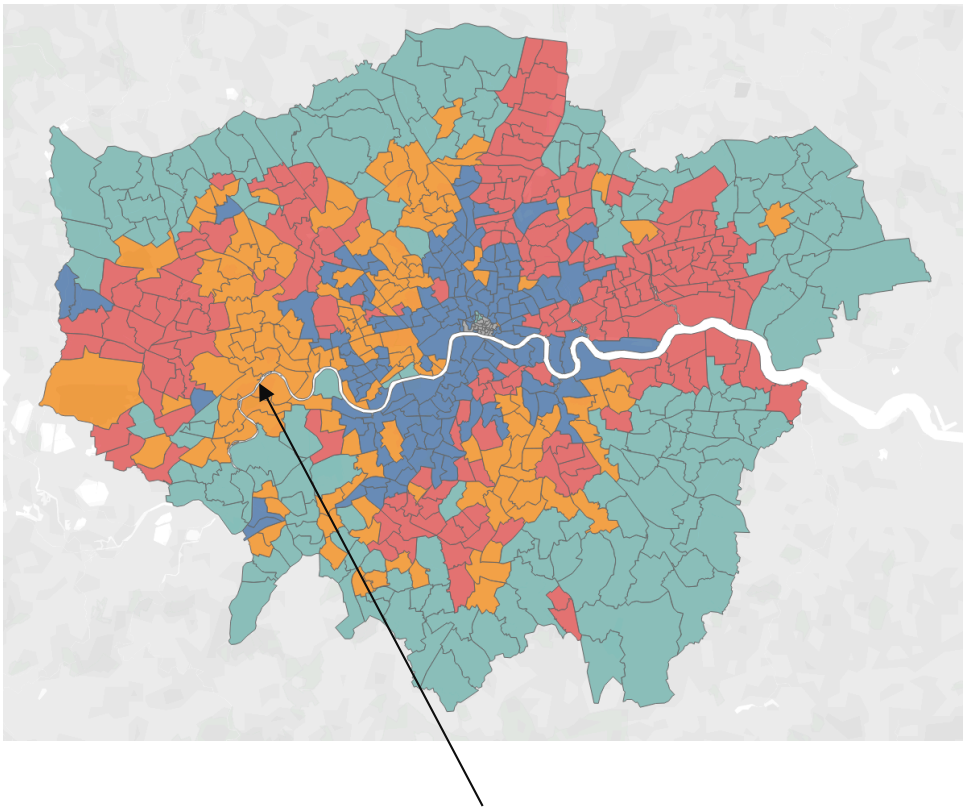
Partition-based clustering of space-referenced data



proportions of the respective age groups are higher than on the average

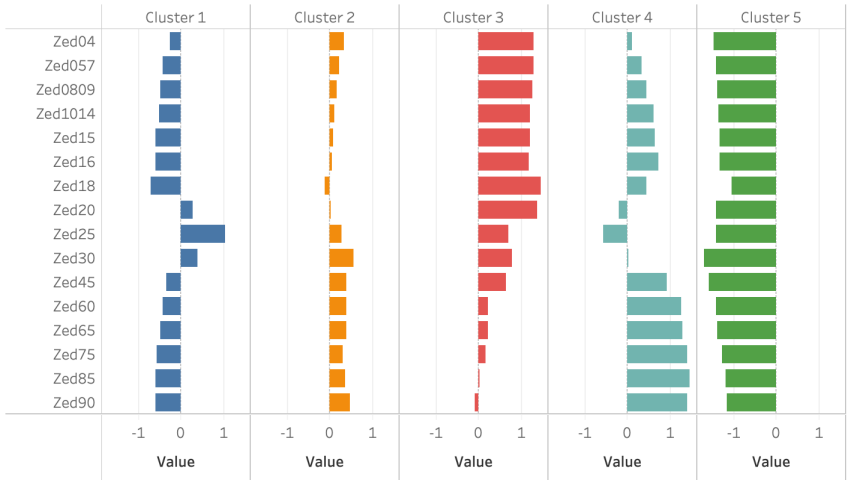
Clustering with different k's





Refinement: separation of a group of wards with attribute values close to average.

Clustering with different k’s



Clustering

- When to stop:
  - When it reveals only subtle differences between groups and does not bring significant new information.
  - We use clustering to simplify data; too many clusters mean high complexity.

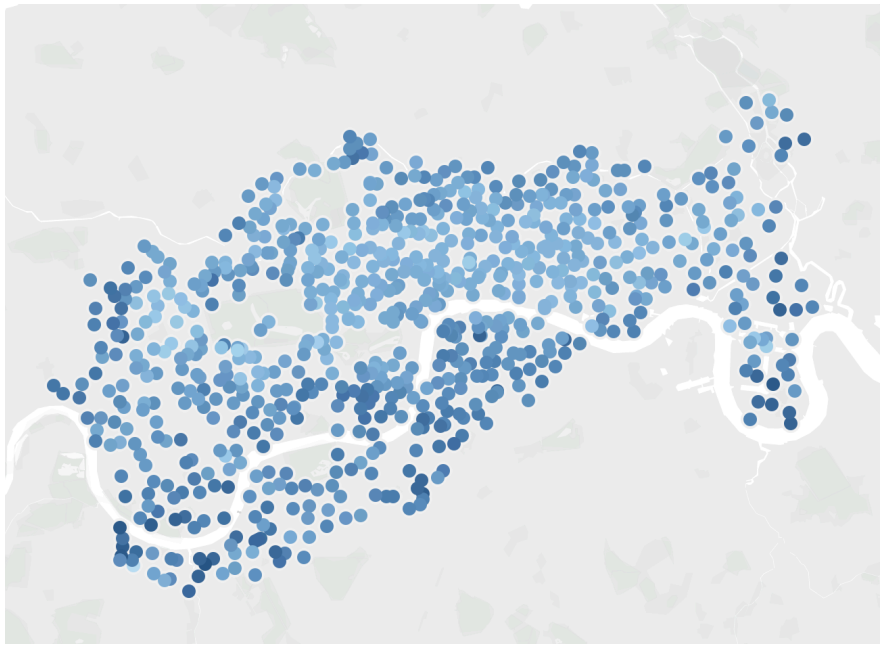
How to choose suitable parameter settings?

- Run clustering with different settings and investigate how the results change
- Select the settings bringing the “best” results:
  - makes sense? (e.g., understandable spatial patterns)
  - internal variance within the clusters is sufficiently low
  - fit to the purpose (e.g., the intended analysis scale may require coarser or finer division)
- Use **progressive clustering** for targeted refinement of clusters with high internal variance.

Partition-based clustering for space-referenced time series data

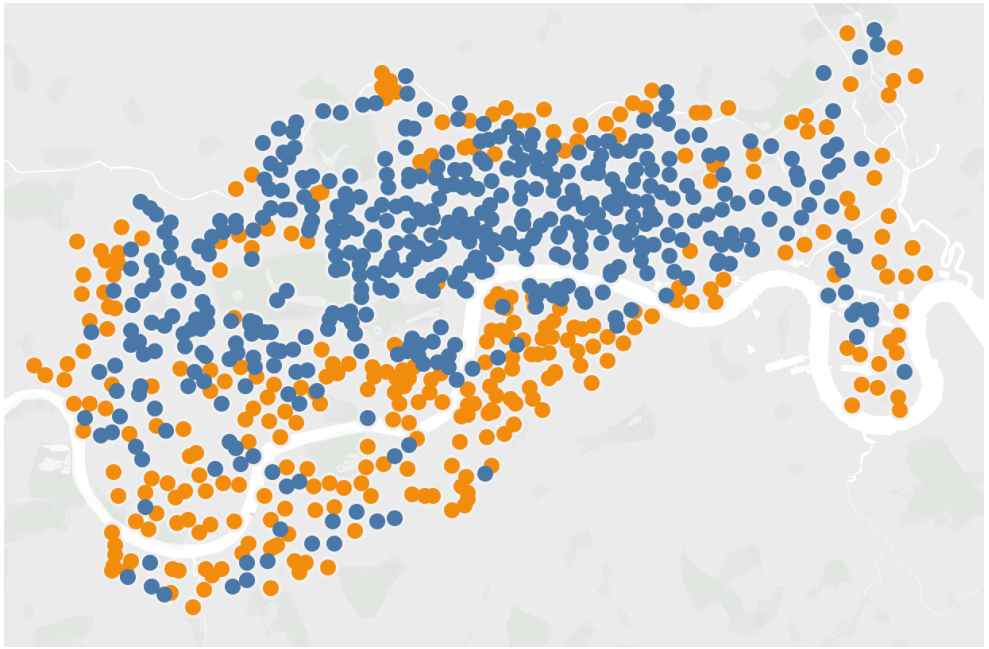
London Cycle Hire Scheme

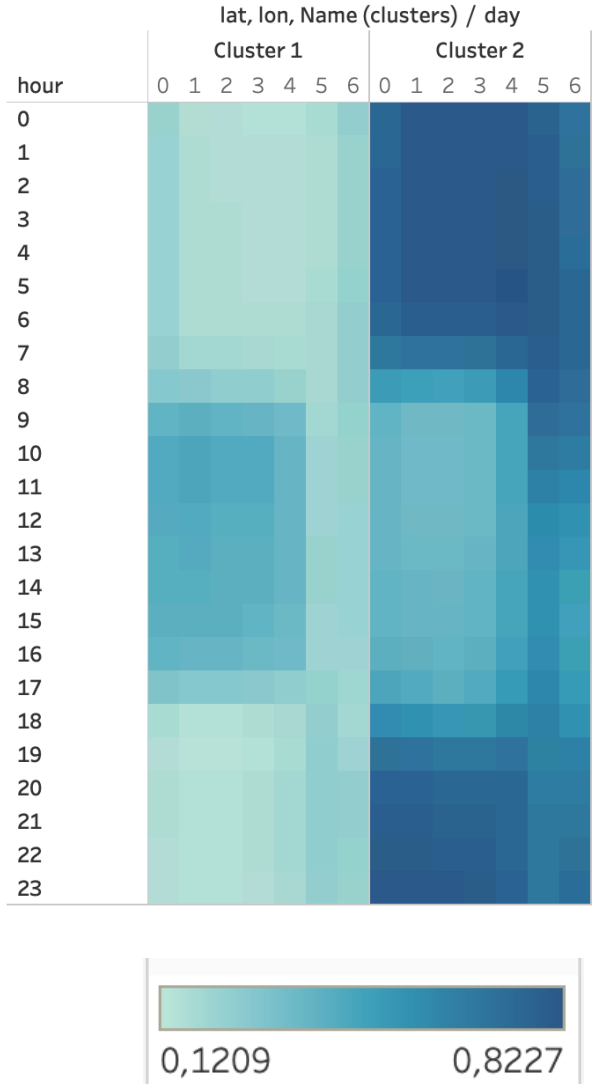
- Bike stations
  - Spatial
  - Aggregate usage over time (bikes/max bikes per hour/day of week)



Cluster by average proportion of available bikes

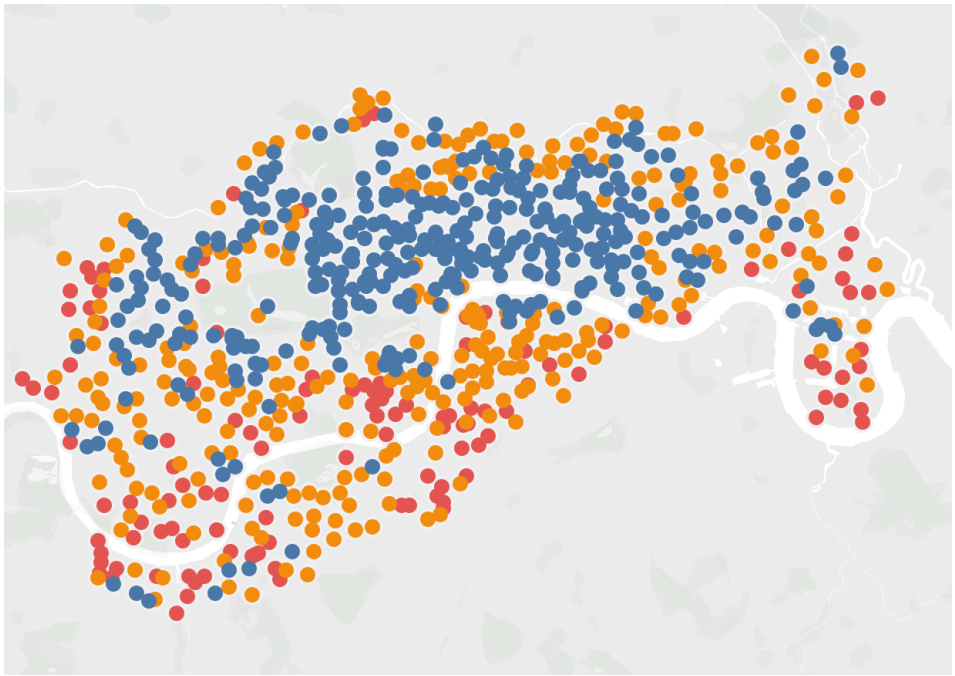
k=2



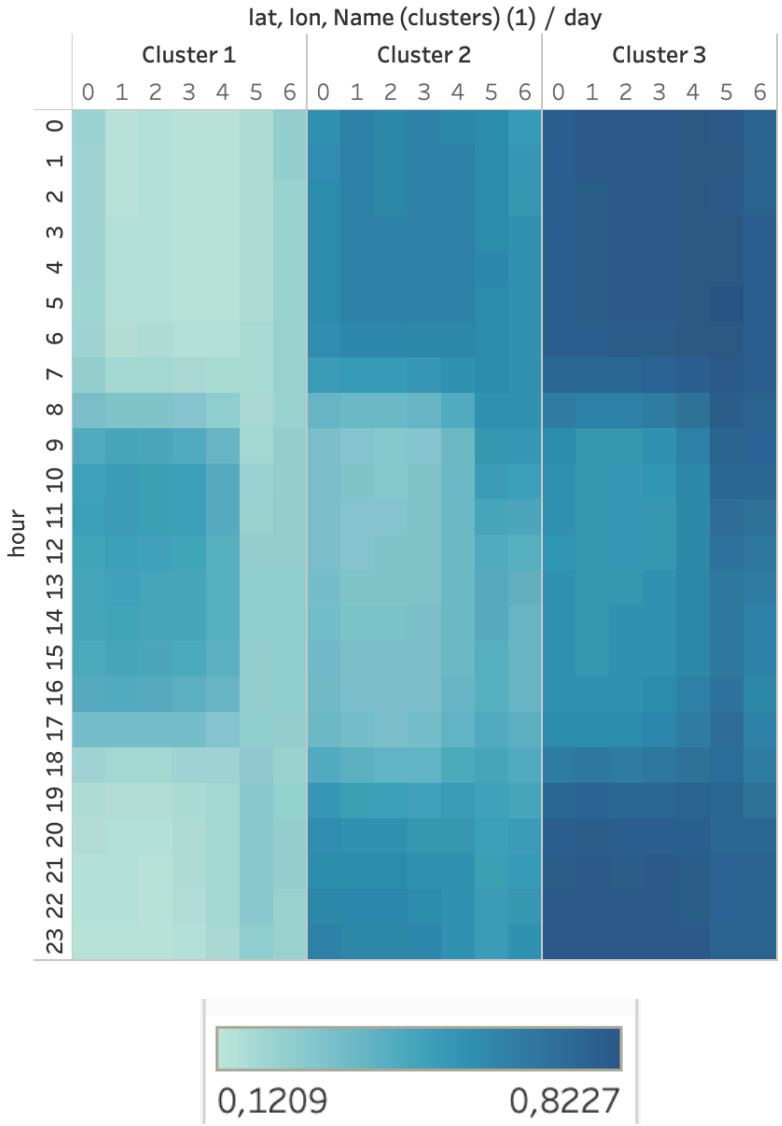


Cluster by average proportion of bikes

k=3

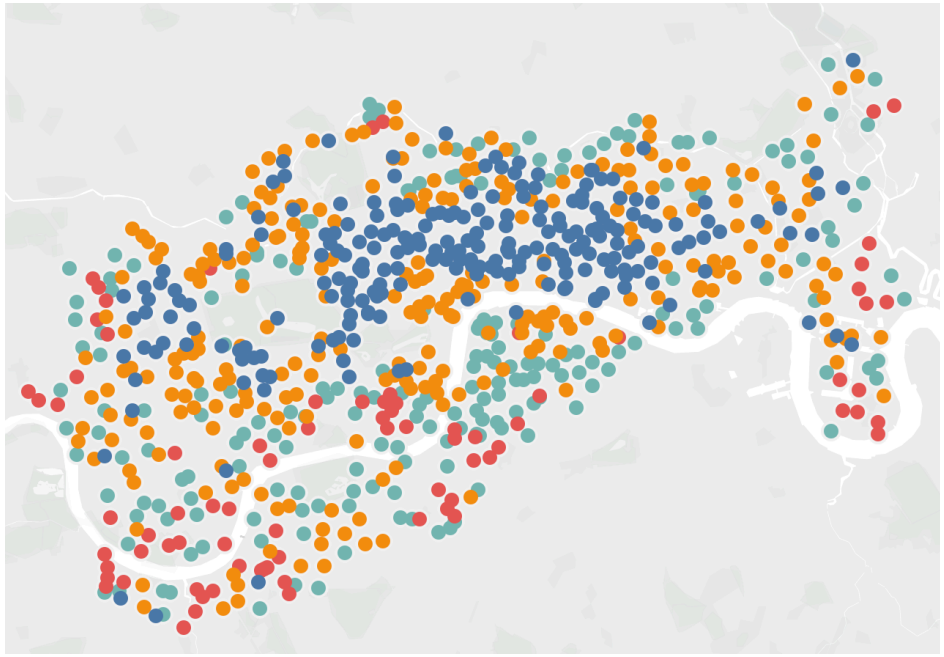


Cluster 1
Cluster 2
Cluster 3

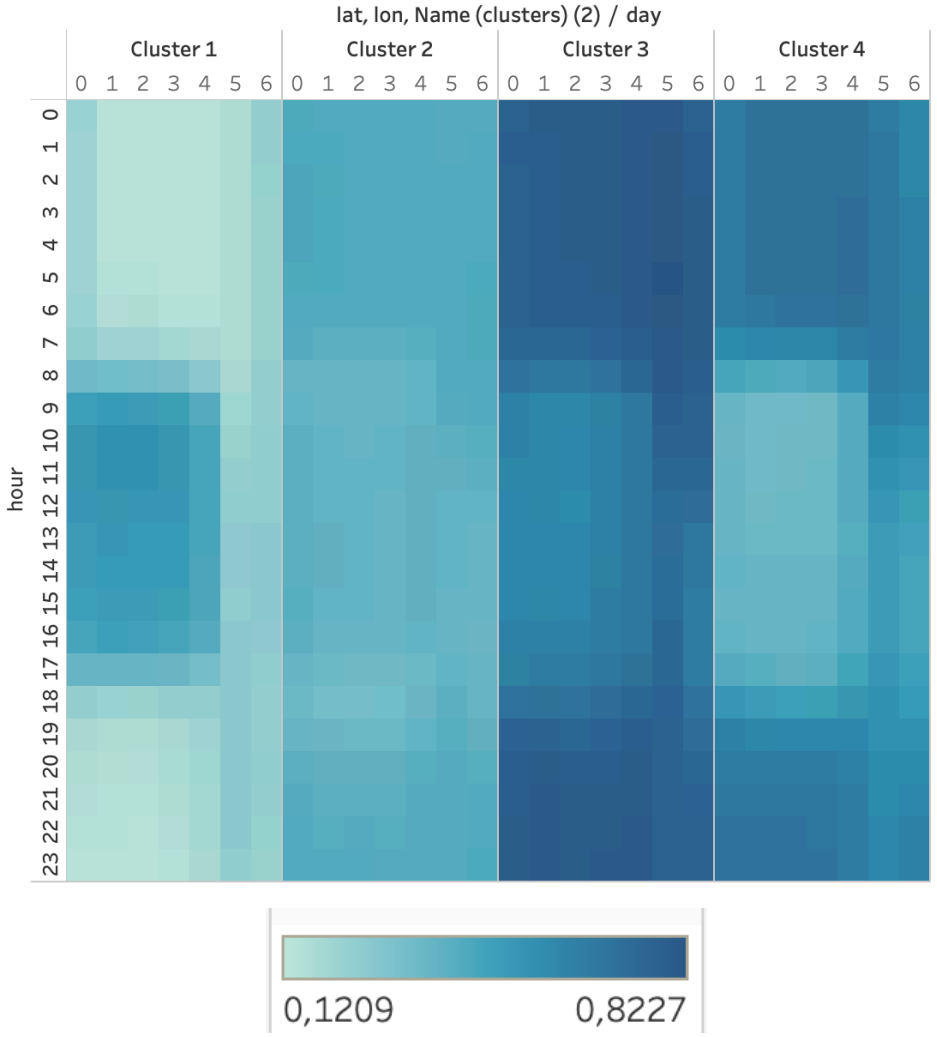


Cluster by average proportion of bikes

k=4

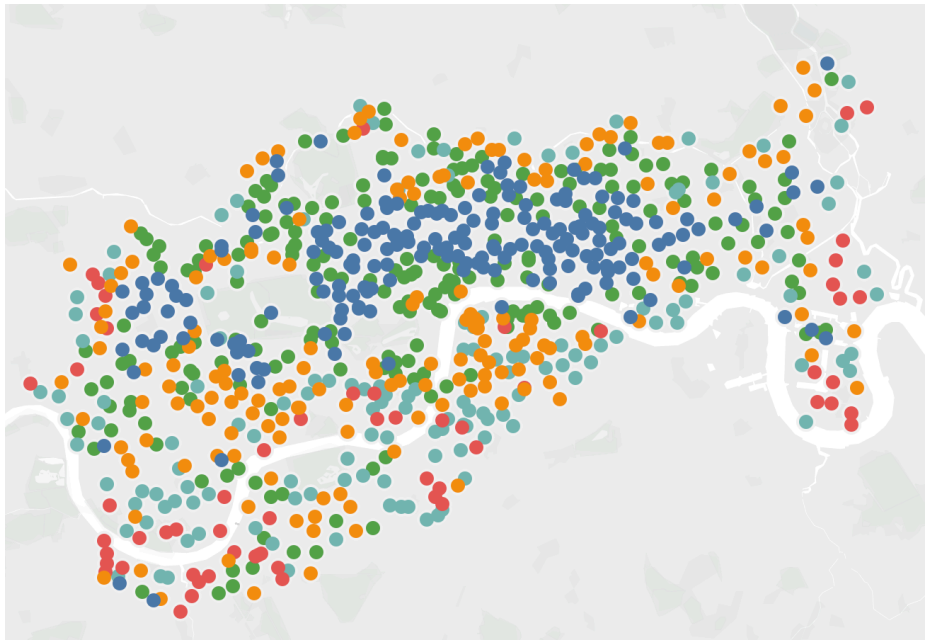


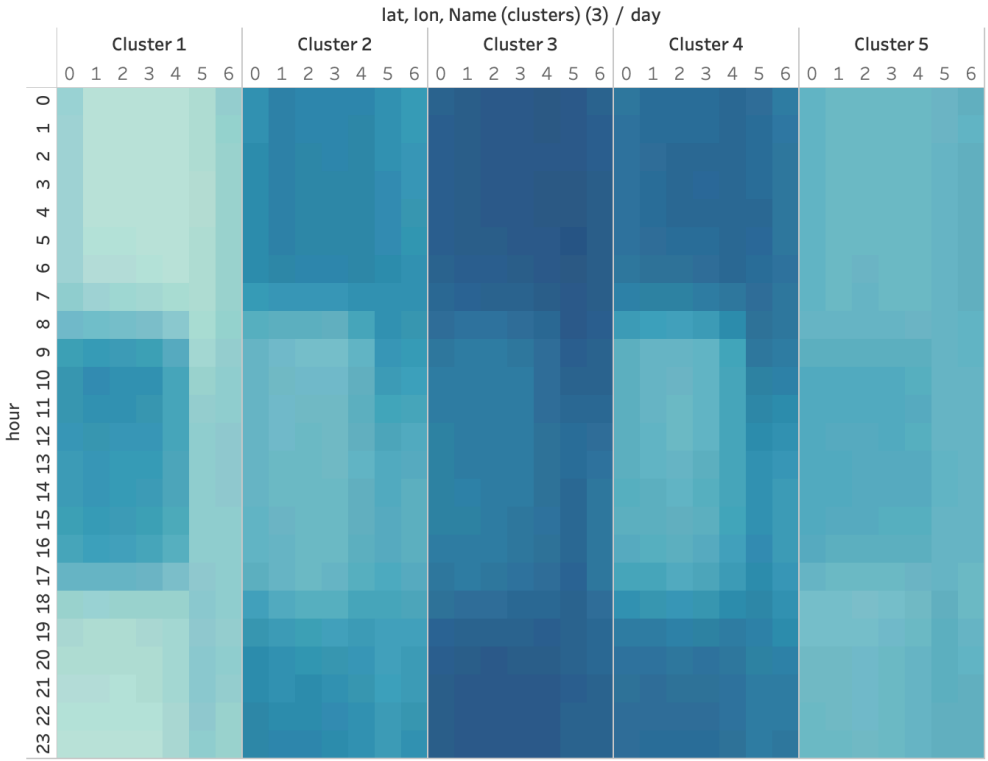
Cluster 1
Cluster 2
Cluster 3
Cluster 4



Cluster by average proportion of bikes

k=5

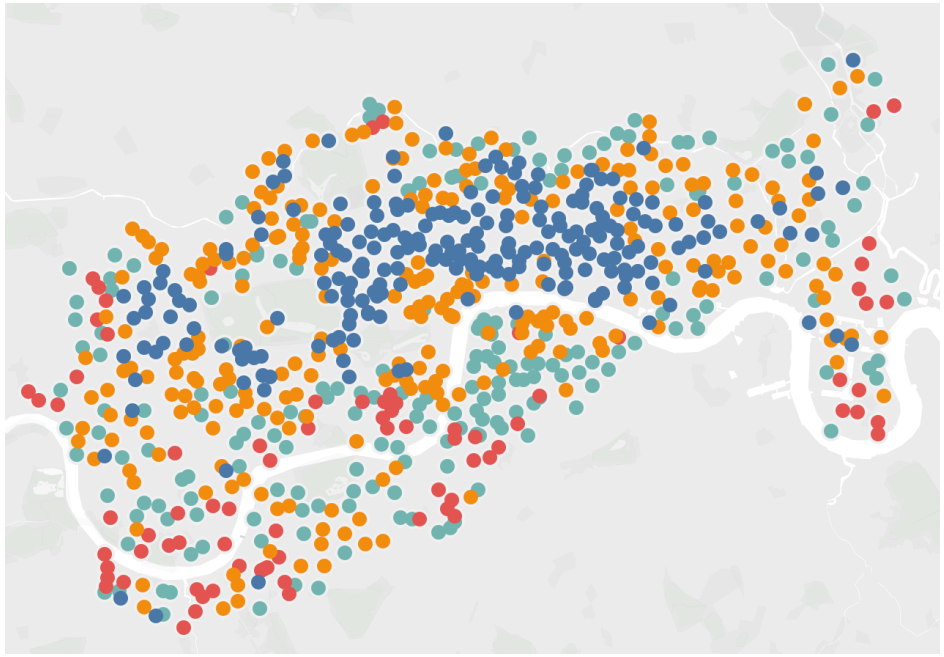




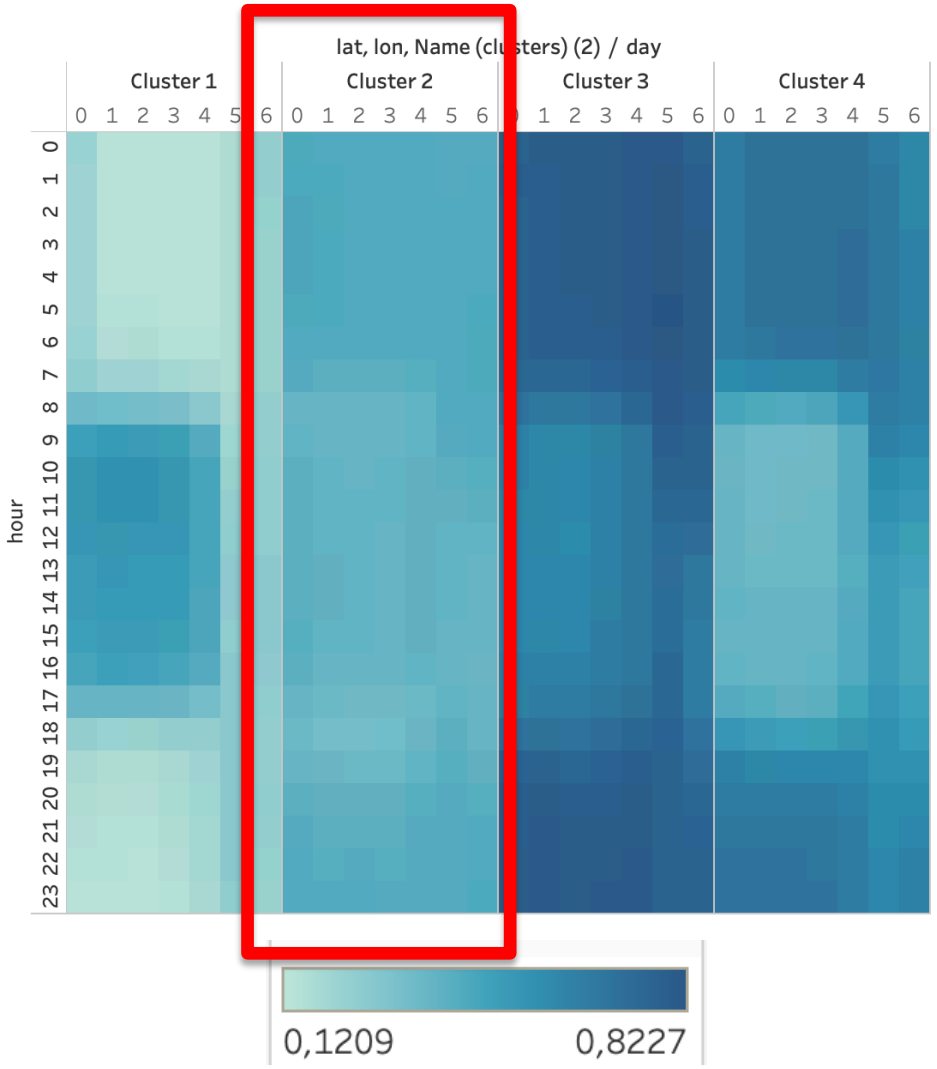
2 and 4 overlap too much?

Cluster by average proportion of bikes

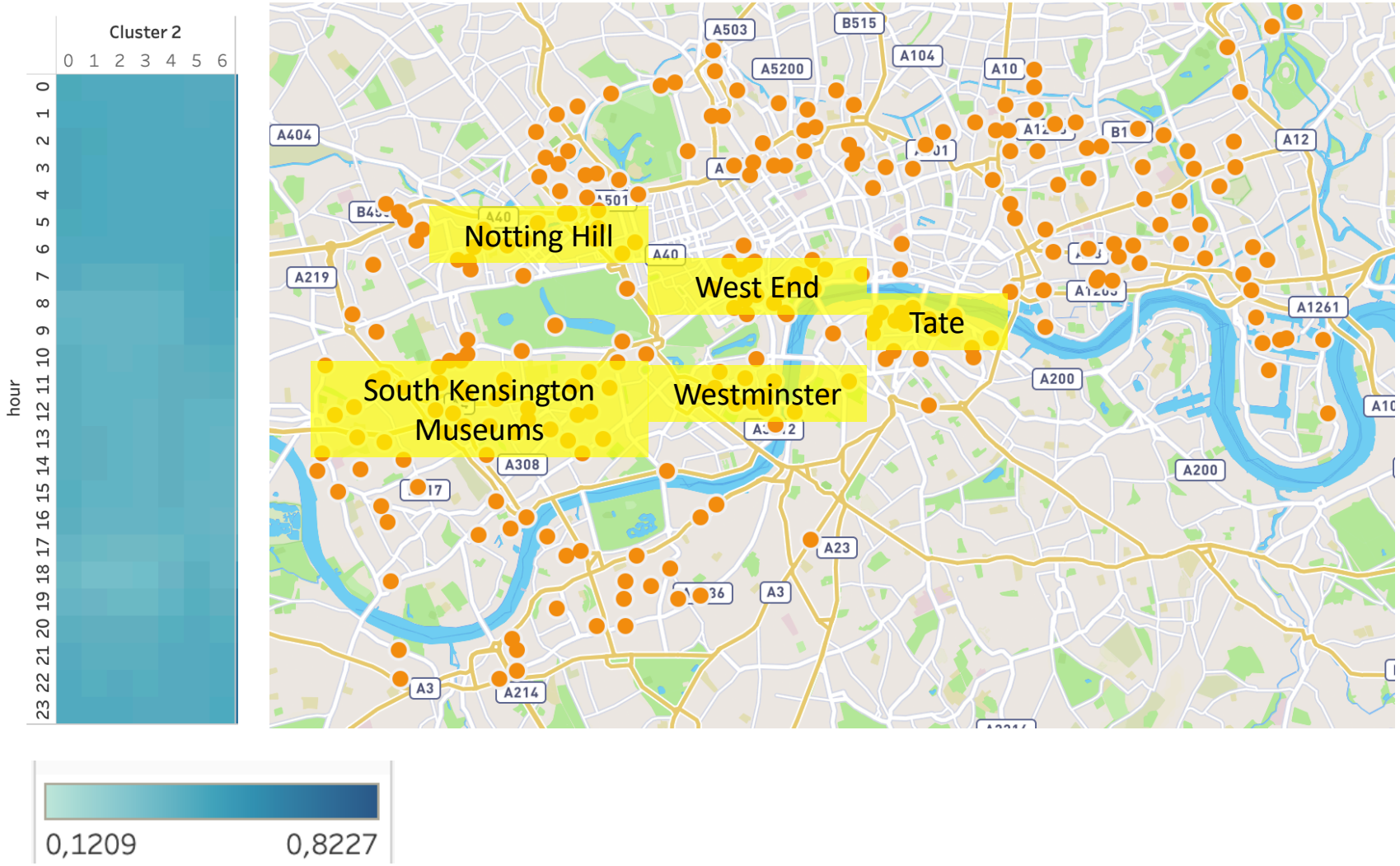
k=4



Cluster 1
Cluster 2
Cluster 3
Cluster 4

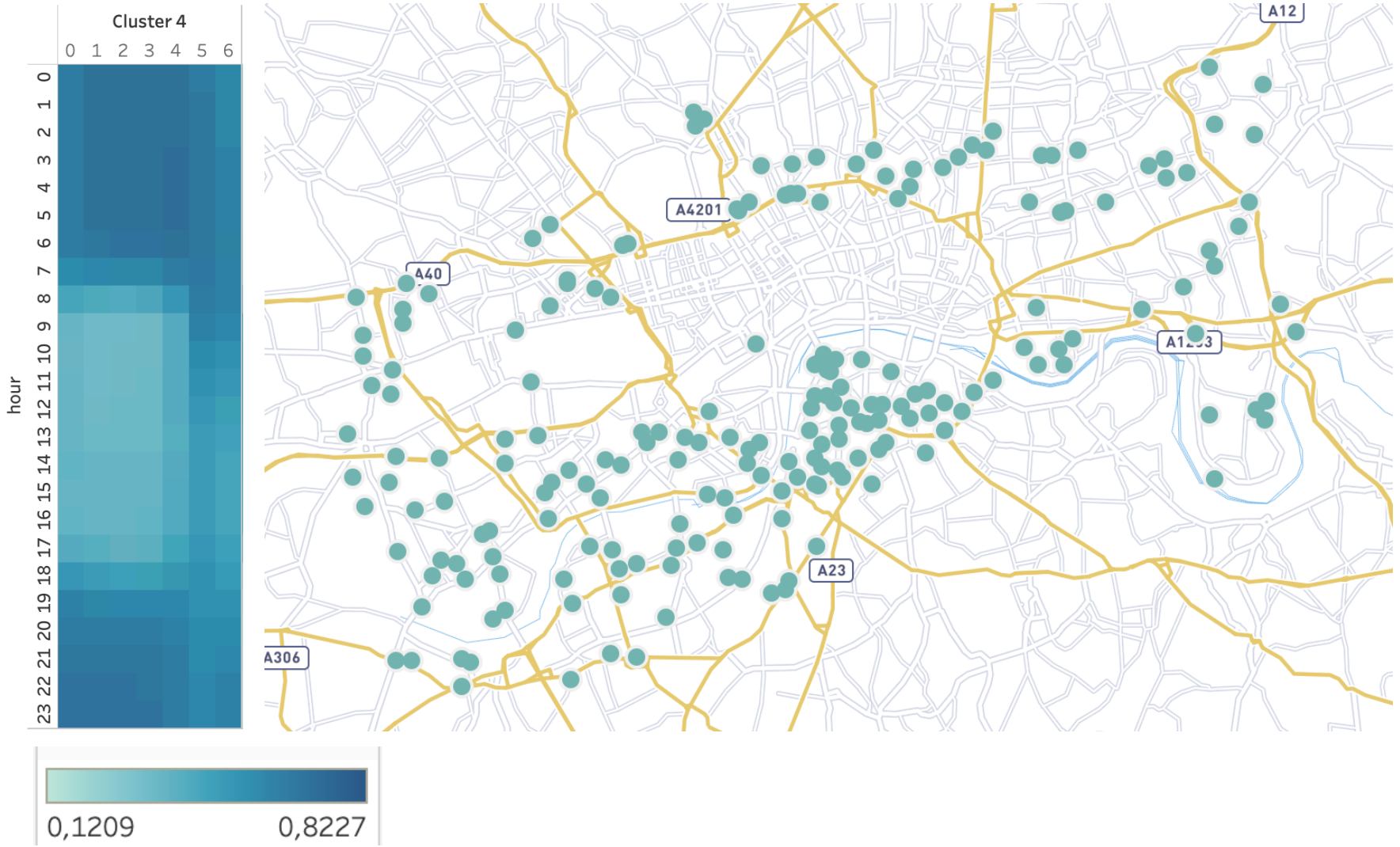


Analysing clusters

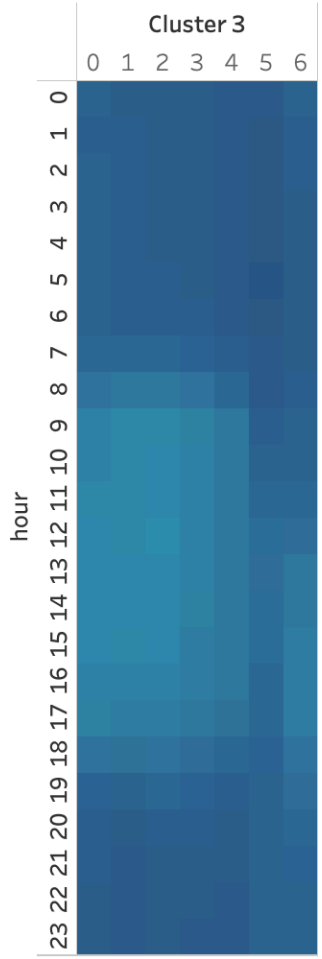


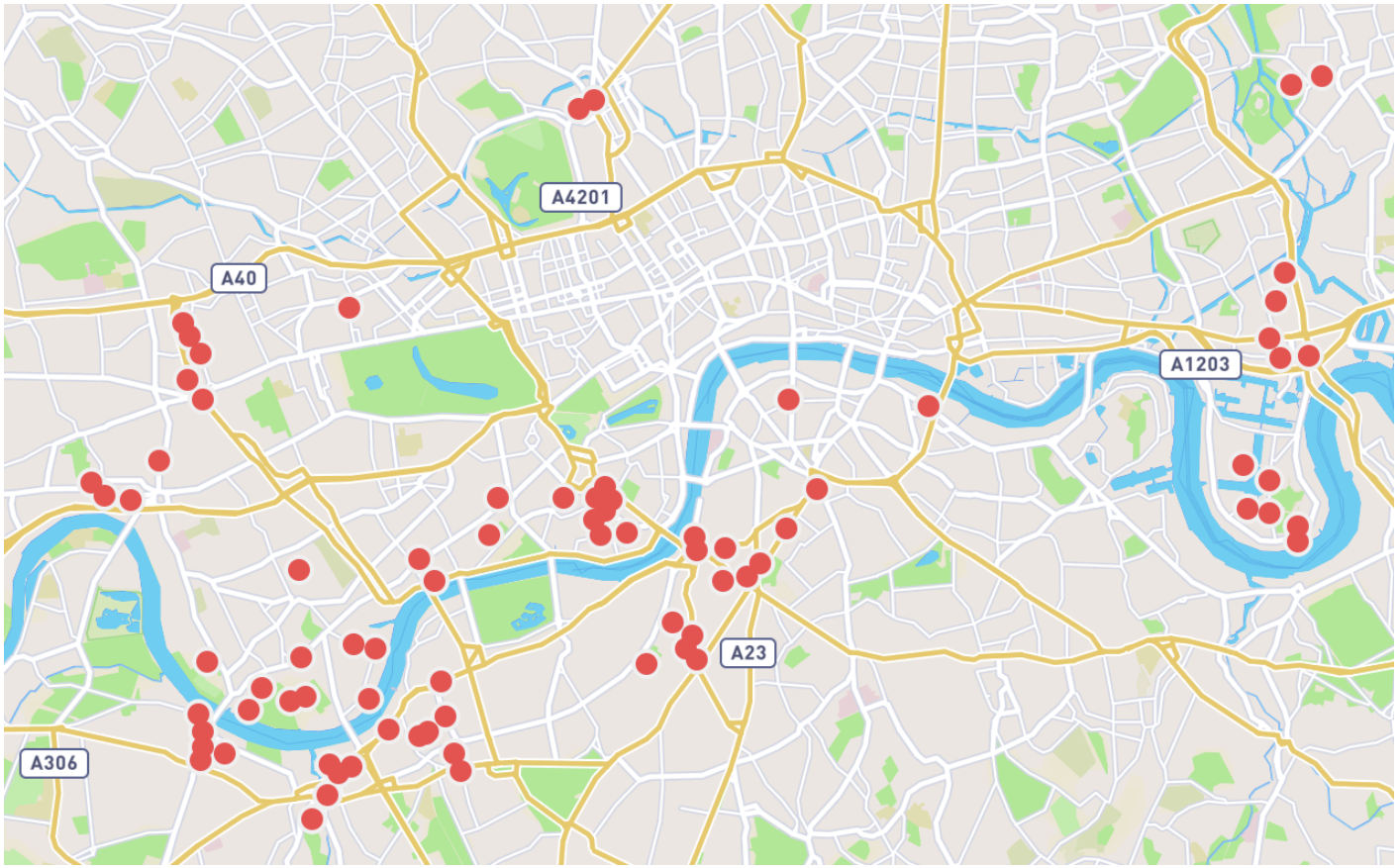
Analysing clusters

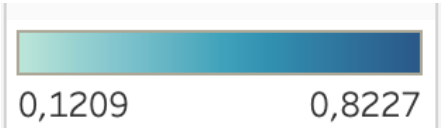
Transport hubs? Commuter areas within coverage area?



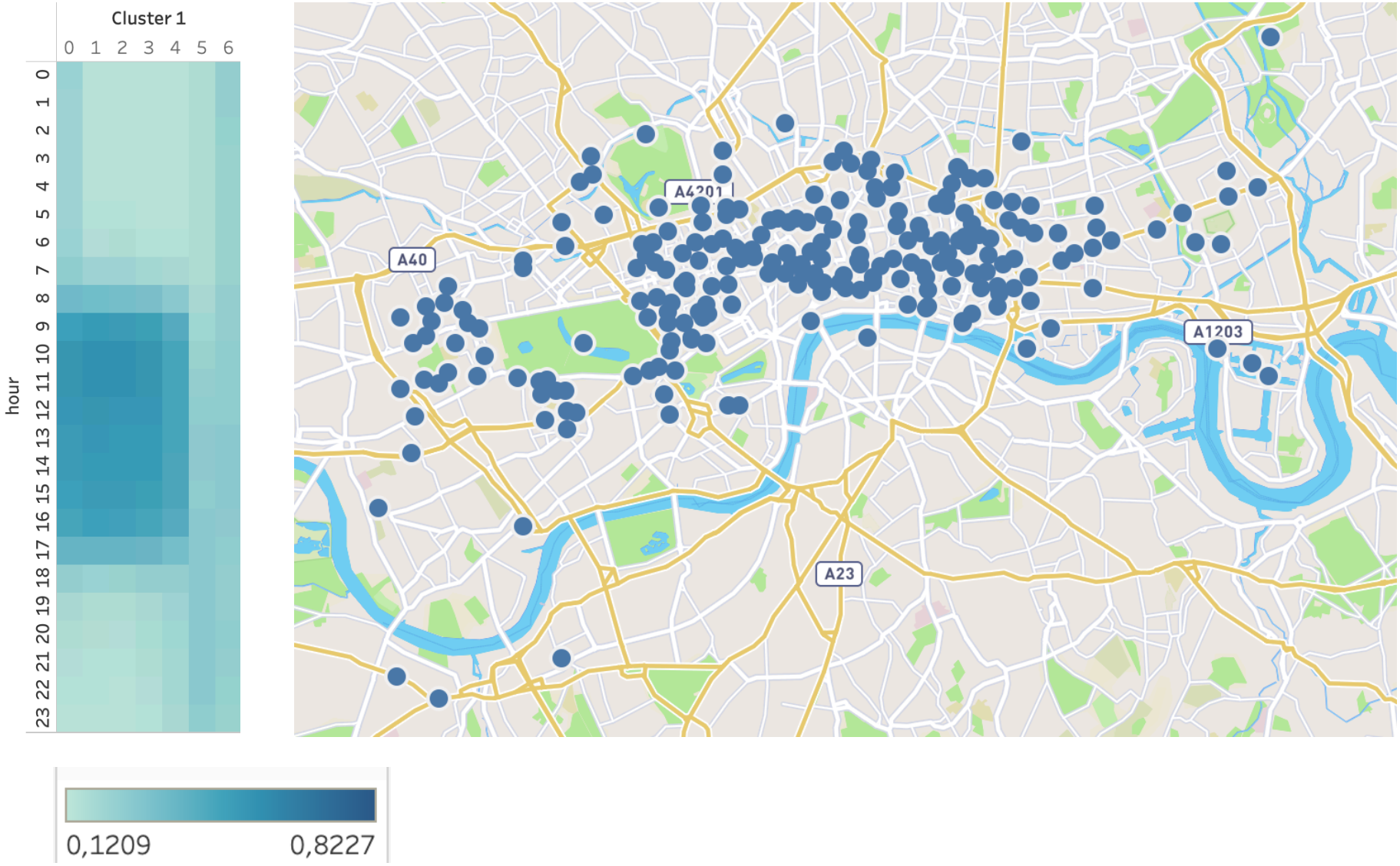
Analysing clusters







Analysing clusters



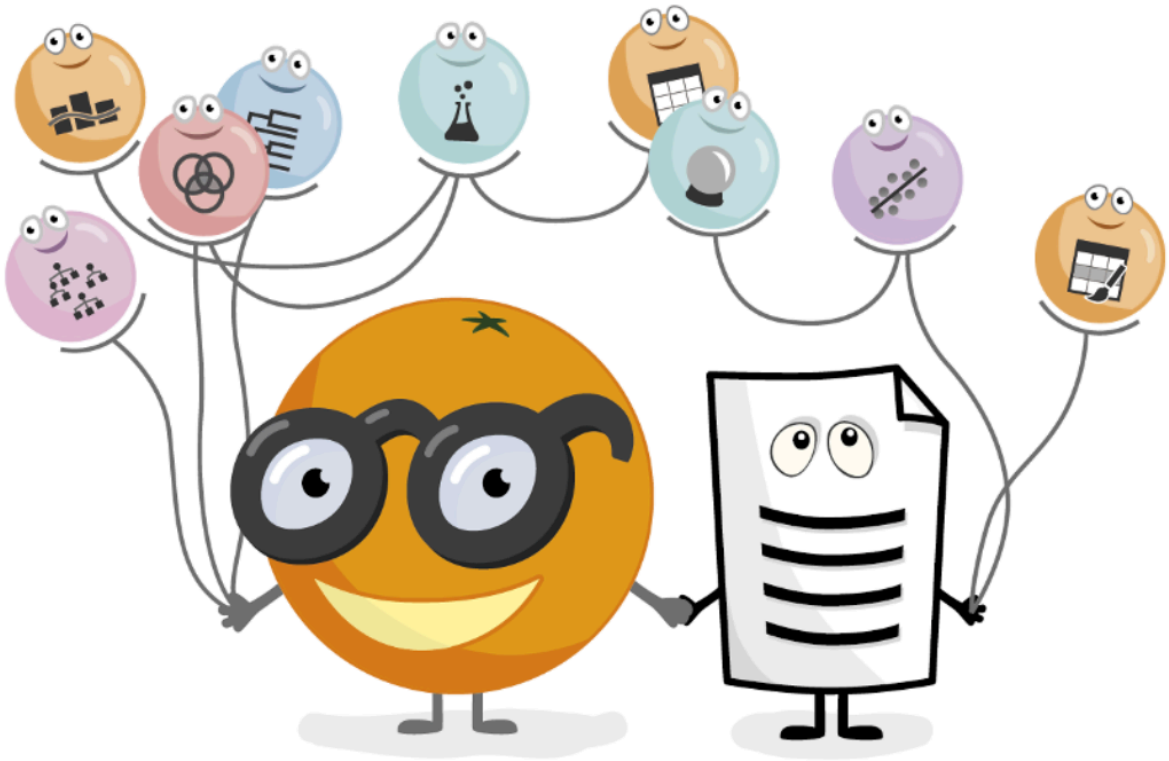
Software

Software

- Standard analytical methods found in most data analysis software
  - **Specialist visual analytics GUI software**: highly interactive, good visual support, perhaps less flexible
    - V-analytics, Orange
  - **Command-line statistics software**: most flexible, can use in a visual analytics way, but not always interactive
    - R, Python, etc.
  - Tableau: built-in clustering (limited)

Orange

- https://orange.biolab.si/
- Interesting and worth a look
- GUI
- Modular



Wrap up

Partition-based clustering

- Groups objects into clusters by similarity of attribute values
  - ✓reduces and simplifies the data to analyse
  - ✓facilitates abstraction
  - ☹but involves large information losses
- To decrease the information loss, interact:
  - Vary parameter settings & compare different groups
  - Examine internal variance and refine clusters by progressive clustering

Visualisation of clustering results

- Colour objects by cluster colour across multiple views
- Summarise and visualise attribute values by cluster
- Link back to original data

More clustering to come!

- How clustering algorithms work
  - PDS and Machine Learning
- Two-way application of partition-based clustering to multiple time series
  - this module
- Density-based clustering
  - in this module.
- Progressive clustering with different distance functions
  - in this module.

Wider context

- Not only about clustering
- A good example of the general principle of visual analytics
- Principles
  - Iteratively vary parameters and refine your results
  - Visualise **all** your results!

Intended learning outcomes

- Data types and structure
  - And how these affect analysis and interpretation
- How **partition-based clustering** combined with interactive visualisation can help deal with large complex datasets
  - Density-based is the other type of clustering that will be covered later