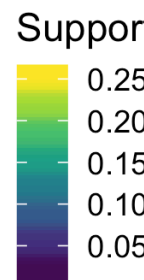
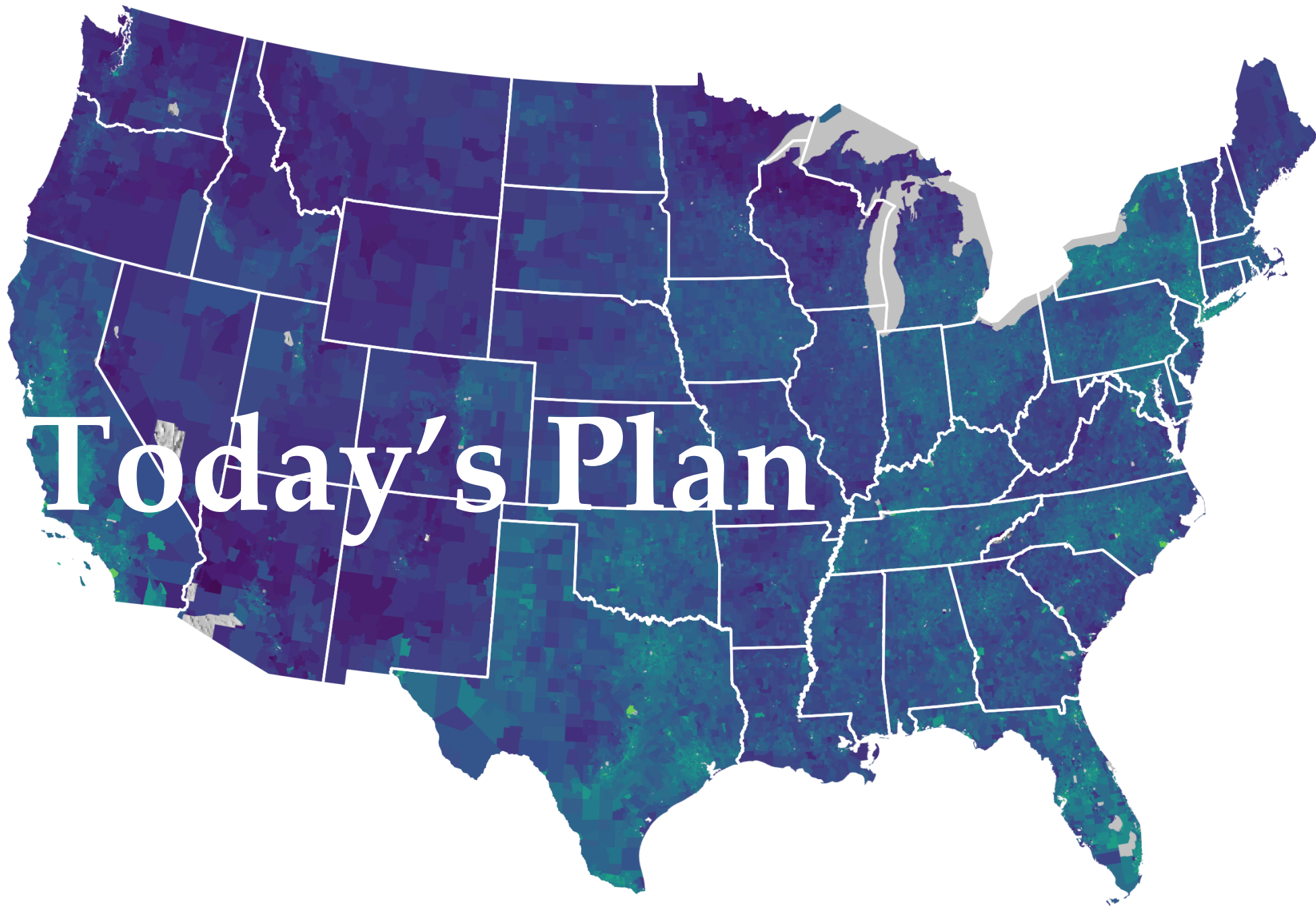


Combining Tabular and Spatial Data

HES 505 Fall 2024: Session 14

Carolyn Koehn



Objectives

By the end of today, you should be able to:

- Define *spatial analysis*
- Describe the steps in planning a spatial analysis
- Understand the structure of relational databases
- Use attributes and topology to subset data
- Generate new features using geographic data
- Join data based on attributes and location

What is spatial analysis?

What is spatial analysis?

“The process of examining the locations, attributes, and relationships of features in spatial data through overlay and other analytical techniques in order to address a question or gain useful knowledge. Spatial analysis extracts or creates new information from spatial data”.

— ESRI Dictionary

What is spatial analysis?

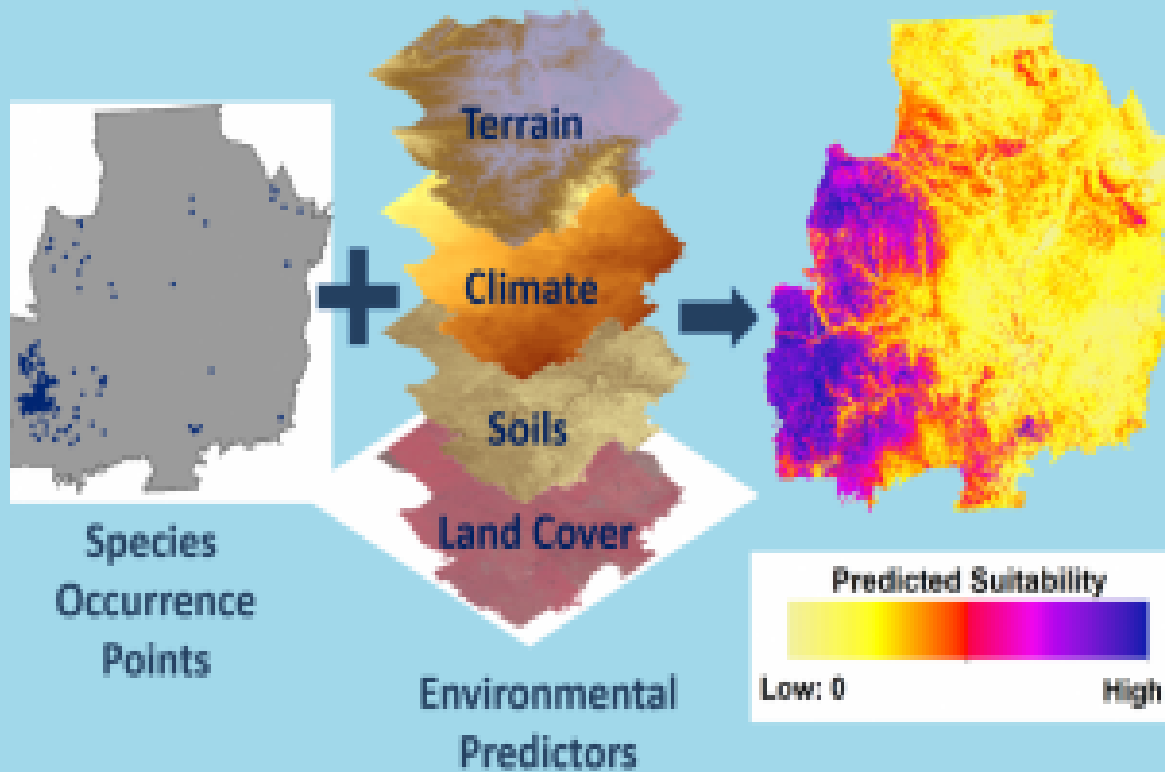
- The process of turning maps into information
- Any- or everything we do with GIS
- The use of computational and statistical algorithms to understand the relations between things that co-occur in space.



John Snow's cholera outbreak map

Common goals for spatial analysis

Building a Model



- Describe and visualize locations or events
- Quantify patterns
- Characterize 'suitability'
- Determine (statistical) relations

courtesy of NatureServe

Common pitfalls of spatial analysis

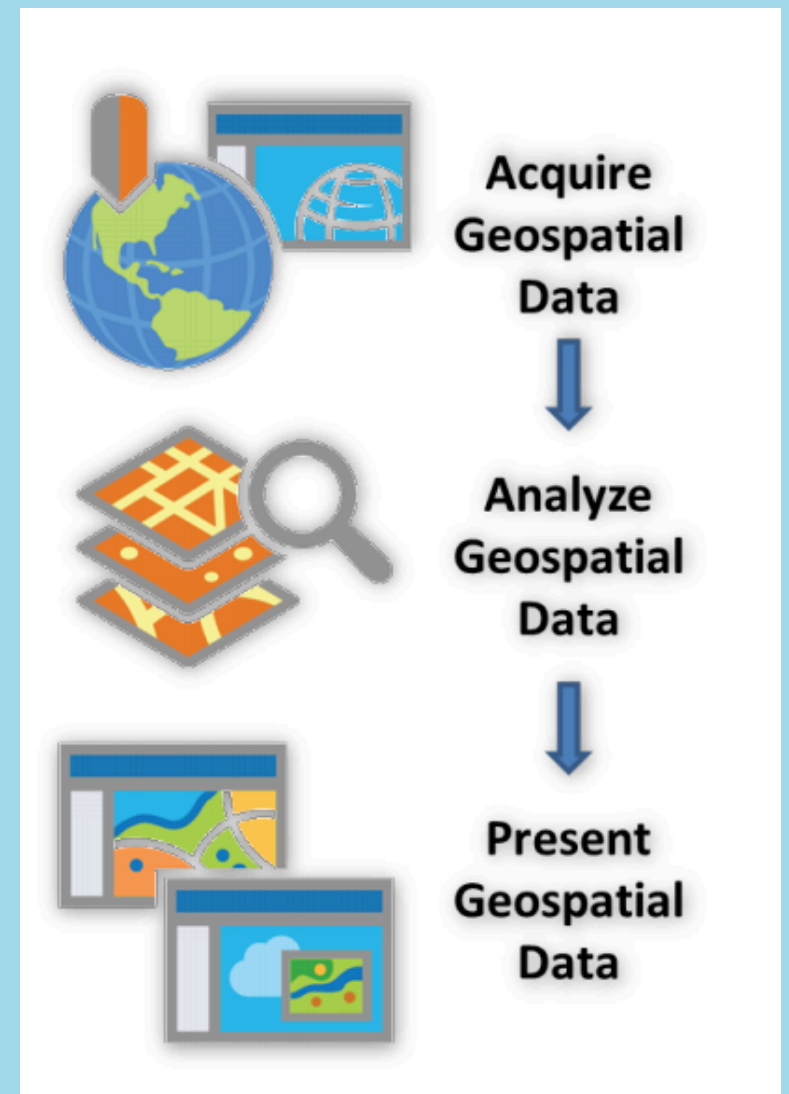
- **Locational Fallacy:** Error due to the spatial characterization chosen for elements of study
- **Atomic Fallacy:** Applying conclusions from individuals to entire spatial units
- **Ecological Fallacy:** Applying conclusions from aggregated information to individuals

Spatial analysis is an inherently complex endeavor and one that is advancing rapidly. So-called “best practices” for addressing many of these issues are still being developed and debated. This doesn’t mean you shouldn’t do spatial analysis, but you should keep these things in mind as you design, implement, and interpret your analyses

Workflows for spatial analysis

Workflows for spatial analysis

- Acquisition (not really a focus, but see **Resources**)
- Geoprocessing
- Analysis
- Visualization



courtesy of **University of Illinois**

Geoprocessing

Manipulation of data for subsequent use

- Alignment
- Data cleaning and transformation
- Combination of multiple datasets
- Selection and subsetting

Databases and Attributes

Databases and Attributes

	A	B	C	D	E	F
	AREA	PERIMETER	APN	LANDUSE	LOT_SIZE	NEBRHC
1						
2	6474154.35276	10145.96973	20100400020000	HMAJCG	6795360.000000	M0000
3	7076794.10172	10644.47635	20100400010000	HFAJAG	6969600.000000	M0000
4	12993367.28984	15307.50117	20100300200000	HFAJAG	13229172.000000	M0000
5	2942042.70203	7688.52193	20100400030000	HPAJAG	2744260.000000	M0000
6	102725.86950	5216.30257	20100300190000	WBAC0A	187308.000000	M0000
7	78659.06631	3743.20269	20101000140000	WCAC0A	262656.000000	M0000
8	208715.26064	5238.44336	20101000140000	MROADA	219106.800000	M0000
9	12711269.20939	15000.44010	20100300100000	HFAJAG	13009700.000000	M0000
10	8530649.18776	11583.31722	20100200150000	HFAJAG	8819157.600000	M0000
11	2534604.48019	7728.17055	20100200200000	HFAJAG	2800472.400000	M0000
12	2459663.50513	7083.54911	20100200190000	HFAJAG	2090008.800000	M0000
13	4389060.54201	8923.10466	20100200180000	WCAC0A	4420468.800000	M0000
14	385170.08143	13827.90196	20100100450000	WCAC0A	402930.000000	M0000
15	8702821.65378	5361.67716	20100100150000	WCAC0A	8887962.400000	M0000
16	1488916.88618	2496.00860	20100100190000	WCAC0A	1494108.000000	M0000
17	229970.91558	4569.26427	20100100170000	WCAC0A	1153468.800000	M0000
18	1368014.23169	5911.54636	20100100110000	WCAC0A	1594296.000000	M0000
19	1615128.08861	752.44578	20100100140000	WCAC0A	35142.000000	M0070
20	32486.36388	3274.84752	20100510010000	A1E00A	600692.400000	M0000
21	595458.06886	28027.24631	20101000060000	WGAC0A	4194392.400000	M0000
22	3450710.31760	2466.20972	20100520010000	WGAC0A	236095.200000	E0000
23	210706.26281	15248.86526	20101000080000	WHAC0A	20037.600000	E0000
24	796657.93179	3775.17969	20100530060000	IAGAAB	235224.000000	E0000
25	269178.57938	808.95637	20100100130000	WCAC0A	30608.000000	M0070
26	37129.21791	2205.72911	20100530060000	A1D00A	161172.000000	E0000
27	158741.85422					

- Attributes: Information that further describes a spatial feature
- Attributes → predictors for analysis
- Last week focus on thematic relations between datasets
 - Shared 'keys' help define linkages between objects
- Sometimes we are interested in attributes that describe location (overlaps, contains, distance)
- Sometimes we want to join based on location rather than thematic connections
 - Must have the same CRS

courtesy of [Giscommons](#)

Databases and attributes

	A	B		D	E	F
	AREA	PERIMETER	APN	LANDUSE	LOT_SIZE	NEBRHC
1						
2	6474154.35276	10145.96973	20100400020000	HMAJCG	6795360.000000	M0000
3	7076794.10172	10644.47635	20100400010000	HFAJAG	6969600.000000	M0000
4	12993367.28984	15307.50117	20100300200000	HFAJAG	13229172.000000	M0000
5	2942042.70203	7688.52193	20100400030000	HFAJAG	2744260.000000	M0000
6	102725.86950	5216.30257	20100300190000	WBAC0A	187308.000000	M0000
7	78659.06631	3743.20269	20101000140000	WCAC0A	262666.000000	M0000
8	208715.26064	5238.44336	20101000140000	MROADA	219106.800000	M0000
9	12711269.20939	15000.44010	20100300100000	HFAJAG	13009700.000000	M0000
10	8530649.18776	11583.31722	20100200150000	HFAJAG	8819157.600000	M0000
11	2534604.48019	7728.17055	20100200200000	HFAJAG	2800472.400000	M0000
12	2459663.50513	7083.54911	20100200190000	HFAJAG	2090008.800000	M0000
13	4389060.54201	8923.10466	20100200180000	WCAC0A	4420468.800000	M0000
14	385170.08143	30329.90196	20100100450000	WGAC0A	402930.000000	M0000
15	8702821.65378	13827.90196	20100100150000	WCAC0A	8887962.400000	M0000
16	1488916.88618	5361.67716	20100100190000	WCAC0A	1494108.000000	M0000
17	229970.91558	2496.00860	20100100160000	WCAC0A	217364.400000	M0000
18	1368014.23169	4569.26427	20100100170000	WCAC0A	1153468.800000	M0000
19	1615128.08861	5911.54636	20100100110000	WCAC0A	1594296.000000	M0000
20	32486.36388	752.44578	20100100140000	WCAC0A	35142.000000	M0070
21	595458.06886	3274.84752	20100510010000	A1E00A	600692.400000	M0000
22	3450710.31760	28027.24631	20101000060000	WGAC0A	4194392.400000	M0000
23	210706.26281	2466.20972	20100520010000	WGAC0A	236095.200000	E0000
24	796657.93179	15248.85526	20101000080000	WHAC0A	20037.600000	E0000
25	269178.57938	3775.17969	20100530060000	IAGAAB	235224.000000	E0000
26	37129.21791	808.95637	20100100130000	WCAC0A	30608.000000	M0070
27	158741.85422	2205.72911	20100530060000	A1D00A	161172.000000	E0000

- Previous focus has been largely on *location*
- Geographic data often also includes non-spatial data
- Attributes: Non-spatial information that further describes a spatial feature
- Typically stored in tables where each row represents a spatial feature
 - Wide vs. long format

courtesy of [Giscommons](#)

Common attribute operations

- `sf` designed to work with `tidyverse`
- Allows use of `dplyr` data manipulation verbs (e.g. `filter`, `select`, `slice`)
- Can use `scales` package for units
- Also allows `%>%` to chain together multiple steps
- geometries are “sticky”

Subsetting by Field

Subsetting by Features

- Features refer to the individual observations in the dataset
- Selecting features

```
1 head(world)[1:3, 1:3] %>%  
2   st_drop_geometry()
```

```
# A tibble: 3 × 3  
  iso_a2 name_long      continent  
* <chr>  <chr>         <chr>  
1 FJ     Fiji         Oceania  
2 TZ     Tanzania       Africa  
3 EH     Western Sahara Africa
```

```
1 world %>%  
2   filter(continent == "Asia")  
3   dplyr::select(name_long, continent)  
4   st_drop_geometry() %>%  
5   head(.)
```

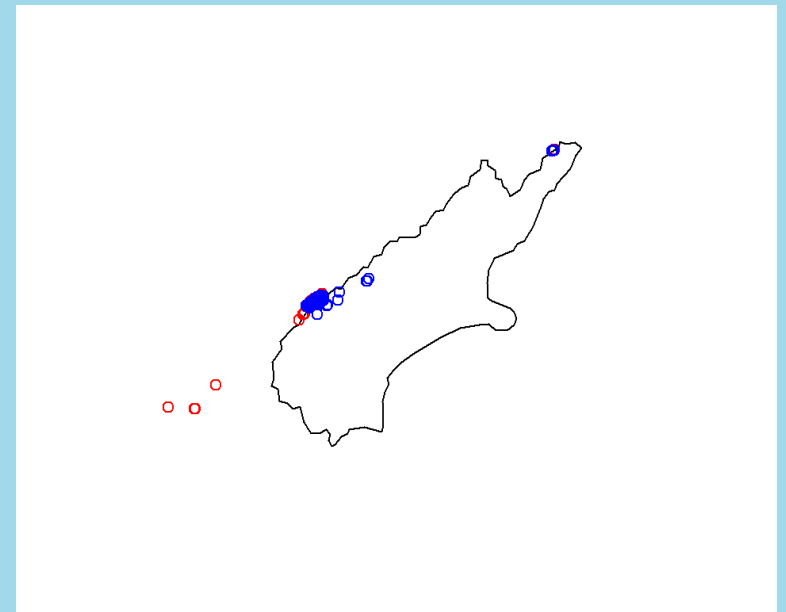
```
# A tibble: 6 × 2  
  name_long      continent  
  <chr>         <chr>  
1 Kazakhstan Asia  
2 Uzbekistan Asia  
3 Indonesia Asia  
4 Timor-Leste Asia  
5 Israel Asia  
6 Lebanon Asia
```

Spatial Subsetting

Topological Subsetting

- Topological relations describe the spatial relationships between objects
- We can use the overlap (or not) of vector data to subset the data based on topology
- Need *valid* geometries
- Easiest way is to use `[` notation, but also most restrictive

```
1 canterbury = nz %>% filter(Name = "Canterbury")
2 canterbury_height = nz_height[canterbury]
```



Topological Subsetting

- Lots of verbs in **sf** for doing this (e.g., **st_intersects**, **st_contains**, **st_touches**)
- see **?geos_binary_pred** for a full list
- Creates an **implicit** attribute (the *records* in **x** that are “in” **y**)

Using **sparse=TRUE**

```
1 co = filter(nz, grepl("Car", nz_name))
2 st_intersects(nz_height, co,
3               sparse = TRUE)
```

```
[[1]]
integer(0)
```

```
[[2]]
[1] 2
```

```
[[3]]
[1] 2
```

```
1 lengths(st_intersects(nz_height, co,
2                       sparse = TRUE))
```

```
[1] FALSE TRUE TRUE
```

Topological Subsetting

- The **sparse** option controls how the results are returned
- We can then find out if one or more elements satisfies the criteria

Using **sparse=FALSE**

```
1 st_intersects(nz_height, co, sparse = FALSE)[1:3,]
```

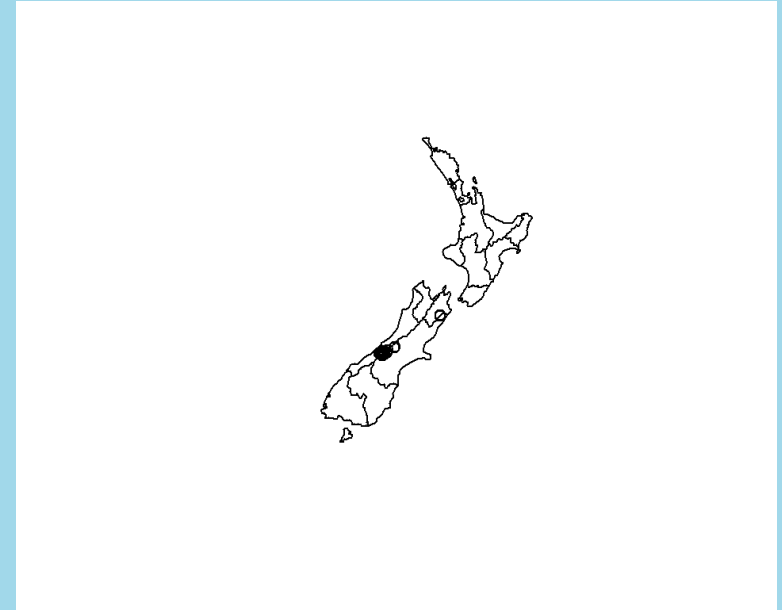
```
      [,1] [,2]  
[1,] FALSE FALSE  
[2,] FALSE  TRUE  
[3,] FALSE  TRUE
```

```
1 apply(st_intersects(nz_height, co, sparse = FALSE), 1, any)[1:3]
```

```
[1] FALSE  TRUE  TRUE
```

Topological Subsetting

```
1 canterbury_height3 = nz_height %>%  
2   filter(as.vector(st_intersects(2
```



New Attributes from Existing Fields

Revisiting the tidyverse

- Creating new fields

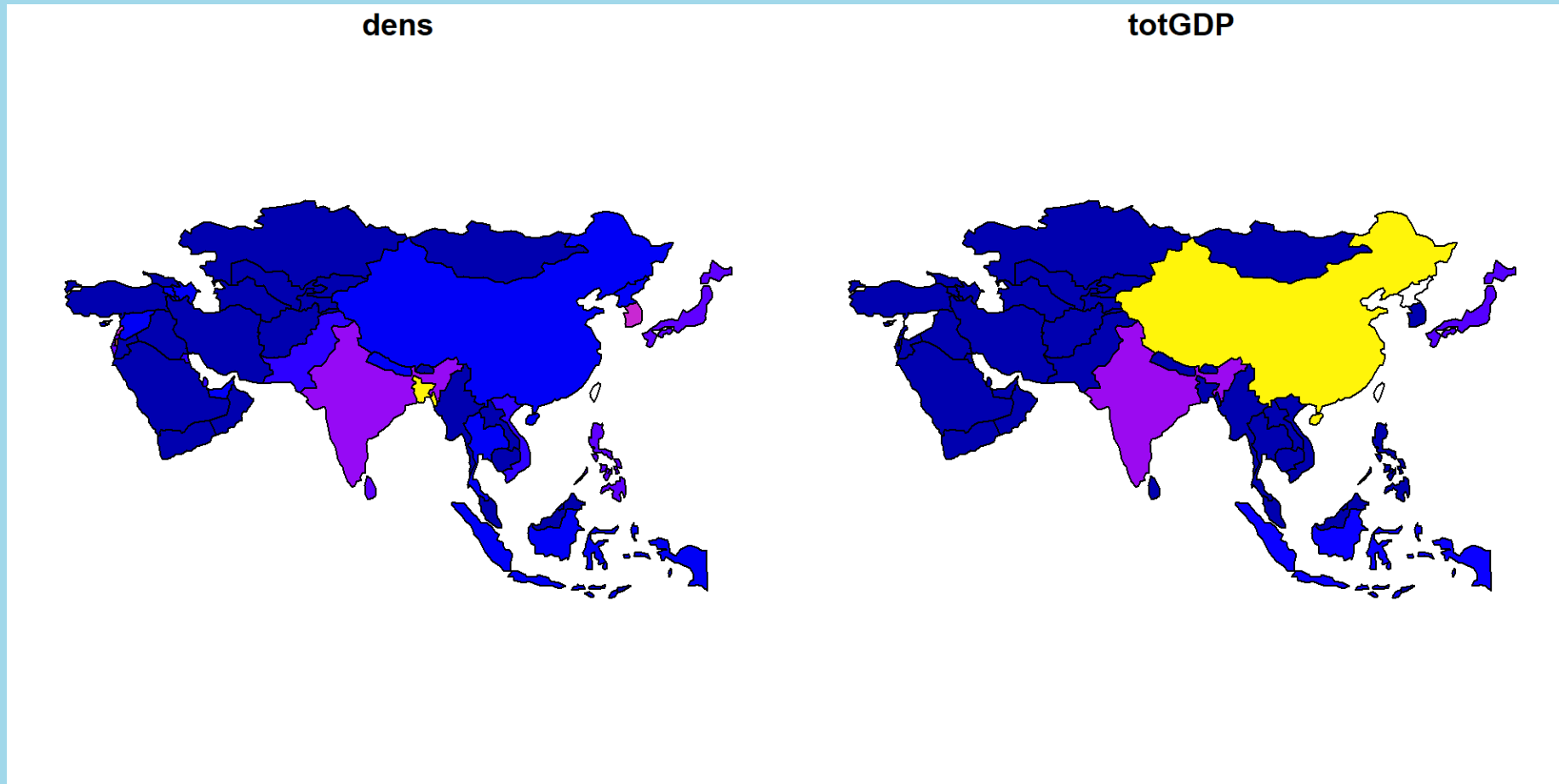
```
1 world %>%
2   filter(continent == "Asia") %>%
3     dplyr::select(name_long, continent, pop, gdpPercap ,area_km2) %>%
4     mutate(., dens = pop/area_km2,
5            totGDP = gdpPercap * pop) %>%
6     st_drop_geometry() %>%
7     head(.)
```

A tibble: 6 × 7

	name_long	continent	pop	gdpPercap	area_km2	dens	totGDP
	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	Kazakhstan	Asia	17288285	23587.	2729811.	6.33	4.08e11
2	Uzbekistan	Asia	30757700	5371.	461410.	66.7	1.65e11
3	Indonesia	Asia	255131116	10003.	1819251.	140.	2.55e12
4	Timor-Leste	Asia	1212814	6263.	14715.	82.4	7.60e 9
5	Israel	Asia	8215700	31702.	22991.	357.	2.60e11
6	Lebanon	Asia	5603279	13831.	10099.	555.	7.75e10

Revisiting the tidyverse

- Creating new fields



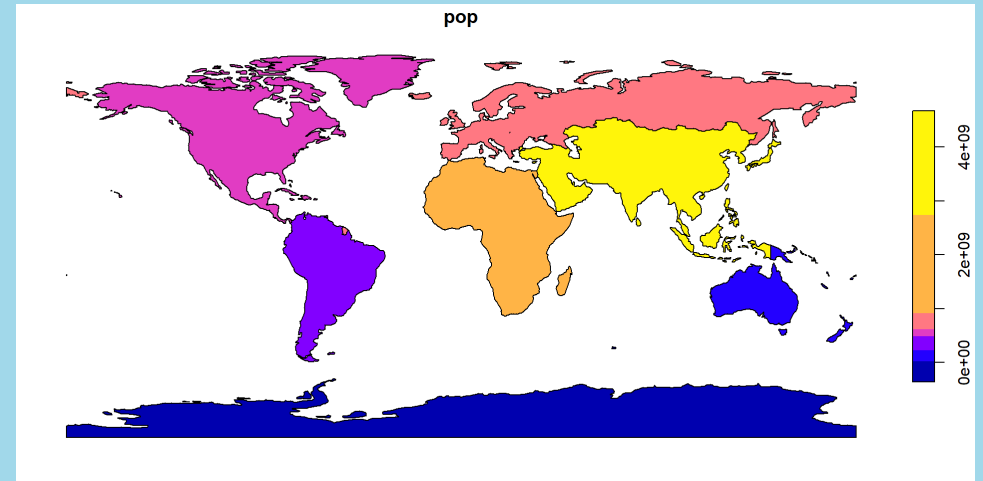
Revisiting the tidyverse

- Aggregating data

```
1 world %>%  
2   st_drop_geometry(.) %>%  
3   group_by(continent) %>%  
4   summarize(pop = sum(pop,
```

```
# A tibble: 8 × 2
```

	continent	pop
	<chr>	<dbl>
1	Africa	1154946633
2	Antarctica	0
3	Asia	4311408059
4	Europe	669036256
5	North America	565028684
6	Oceania	37757833
7	Seven seas (open ocean)	0
8	South America	412060811



New Attributes from Topology

Attributes based on geometry and location (**measures**)

- Attributes like area and length can be useful for a number of analyses
 - Estimates of 'effort' in sampling designs
 - Offsets for modeling rates (e.g., Poisson regression)
- Need to assign the result of the function to a column in data frame (e.g., **\$**, **mutate**, and **summarize**)
- Often useful to test before assigning

Estimating area

- **sf** bases area (and length) calculations on the map units of the CRS
- the **units** library allows conversion into a variety of units

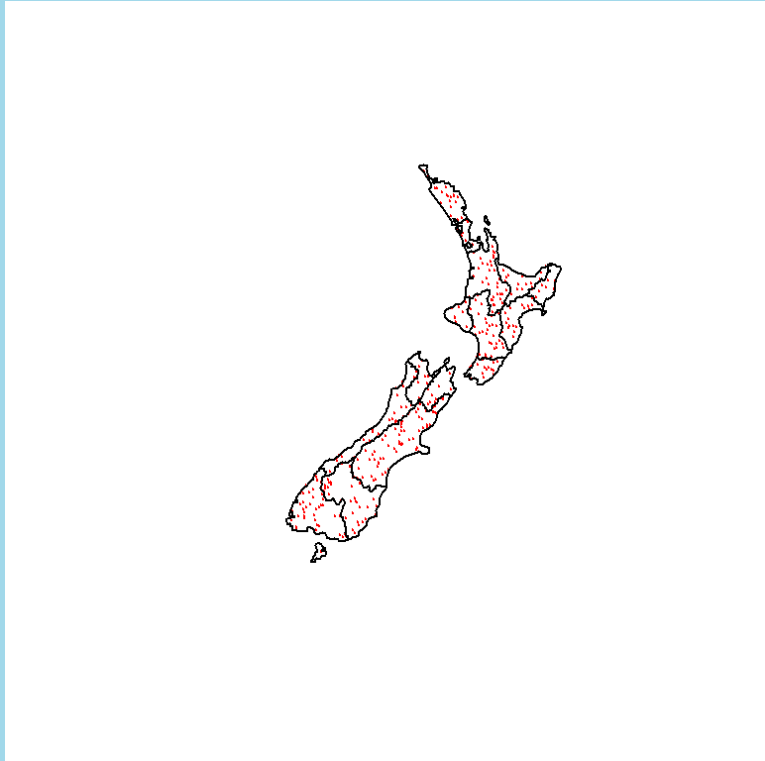
```
1 nz.sf <- nz %>%  
2   mutate(area = st  
3   head(nz.sf$area, 3
```

```
Units: [m^2]  
[1] 12890576439  4911565037  
24588819863
```

```
1 nz.sf$areakm <- un  
2 head(nz.sf$areakm,
```

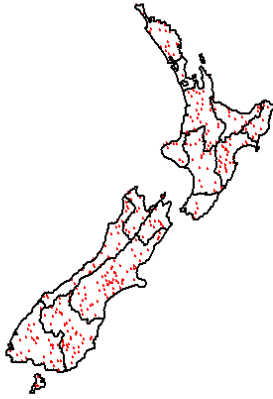
```
Units: [km^2]  
[1] 12890.576  4911.565  
24588.820
```

Estimating Density in Polygons



- Creating new features based on the frequency of occurrence
- Clarifying graphics
- Underlies quadrat sampling for point patterns
- Two steps: count and area

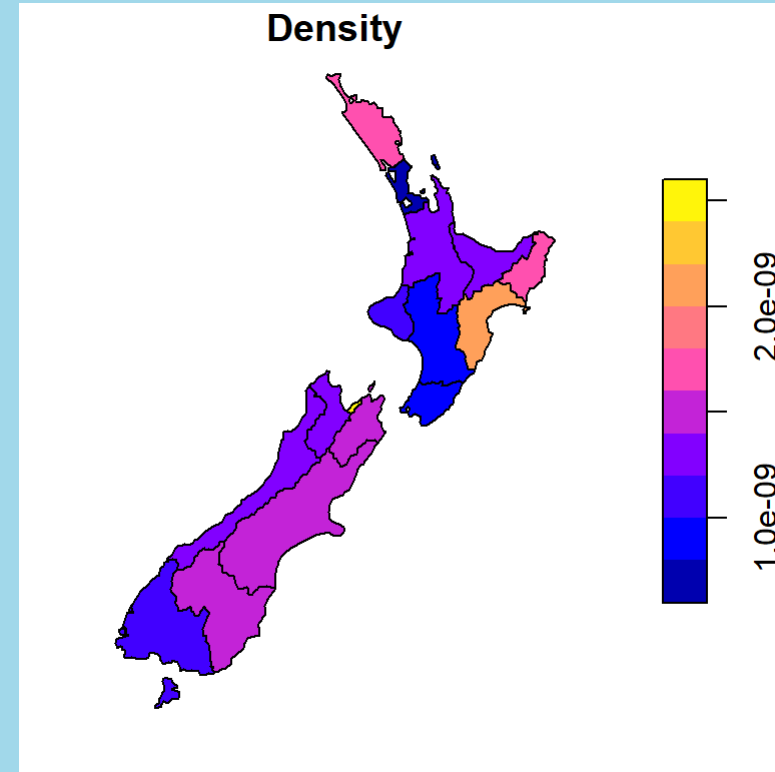
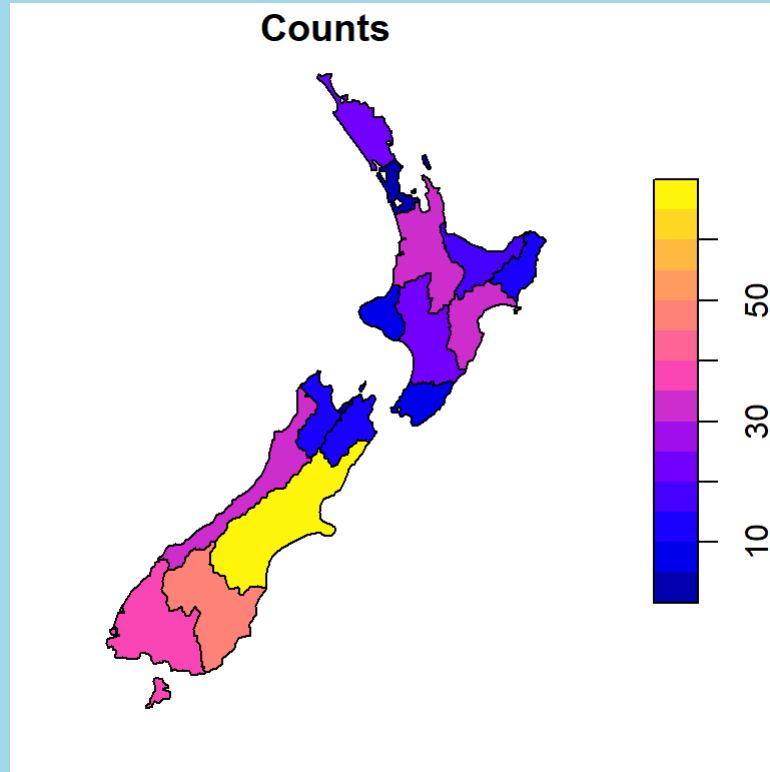
Estimating Density in Polygons



```
1  nz.df <- nz %>%  
2  mutate(counts = lengths(st_interse  
3           area = st_area(nz),  
4           density = counts/area)  
5  head(st_drop_geometry(nz.df[,7:10])
```

	counts	area
density		
1	22 12890576439 [m^2]	1.706673e-09
	[1/m^2]	
2	3 4911565037 [m^2]	6.108033e-10
	[1/m^2]	
3	33 24588819863 [m^2]	1.342073e-09
	[1/m^2]	
4	16 12271015945 [m^2]	1.303886e-09
	[1/m^2]	
5	14 8364554416 [m^2]	1.673729e-09
	[1/m^2]	
6	31 14242517871 [m^2]	2.176581e-09
	[1/m^2]	

Estimating Density in Polygons

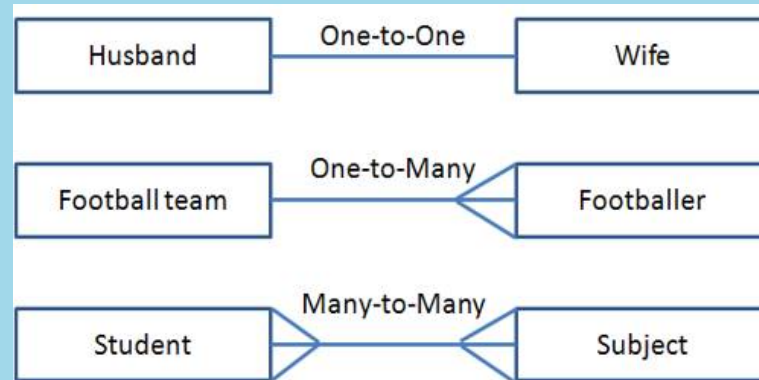


Estimating Distance

- As a covariate
- For use in covariance matrices
- As a means of assigning connections in networks

Estimating Single Point Distance

- **st_distance**
returns distances
between all features
in **x** and all features
in **y**
- One-to-One
relationship requires
choosing a single
point for **y**



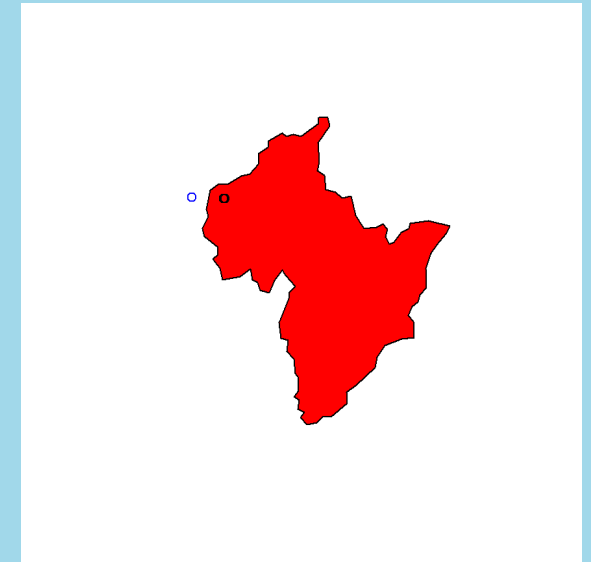
Estimating Single Point Distance

- Subsetting **y** into a single feature

```
1 canterbury = nz %>% filter(Name == "Canterbury")
2 canterbury_height = nz_height[canterbury, ]
3 co = filter(nz, grepl("Canter|Otago", Name))
4 st_distance(nz_height[1:3, ], co)
```

Units: [m]

	[,1]	[,2]
[1,]	123537.16	15497.72
[2,]	94282.77	0.00
[3,]	93018.56	0.00



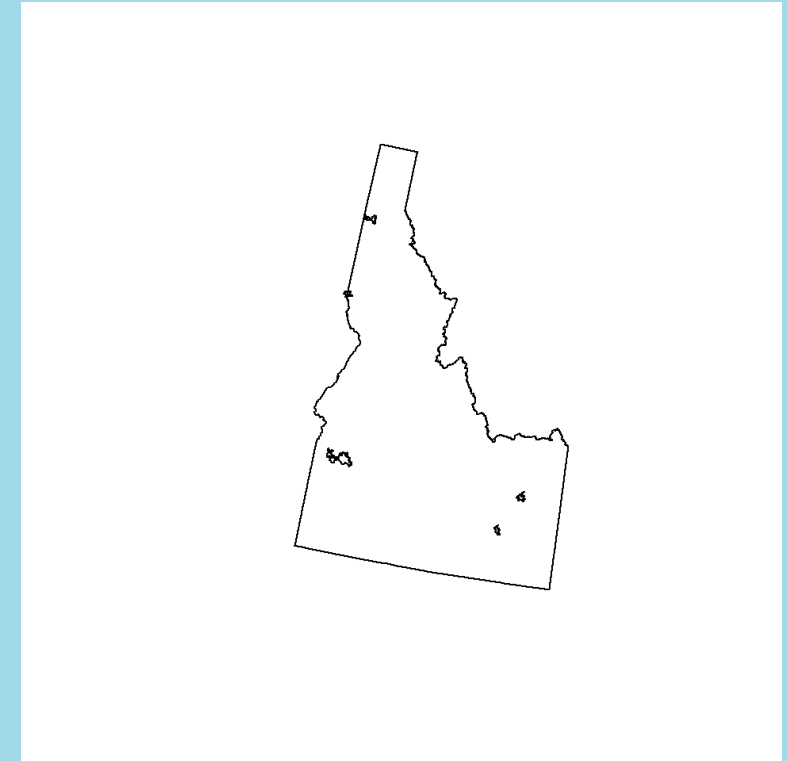
Estimating Single Point Distance

- Using nearest neighbor distances

```
1 ua <- urban_areas(cb = FALSE, proc
2   filter(., UATYP10 == "U") %>%
3   filter(., str_detect(NAME10, "ID
4   st_transform(., crs=2163)
5
6 #get index of nearest ID city
7 nearest <- st_nearest_feature(ua
8 #estimate distance
9 (dist = st_distance(ua, ua[nearest
```

Units: [m]

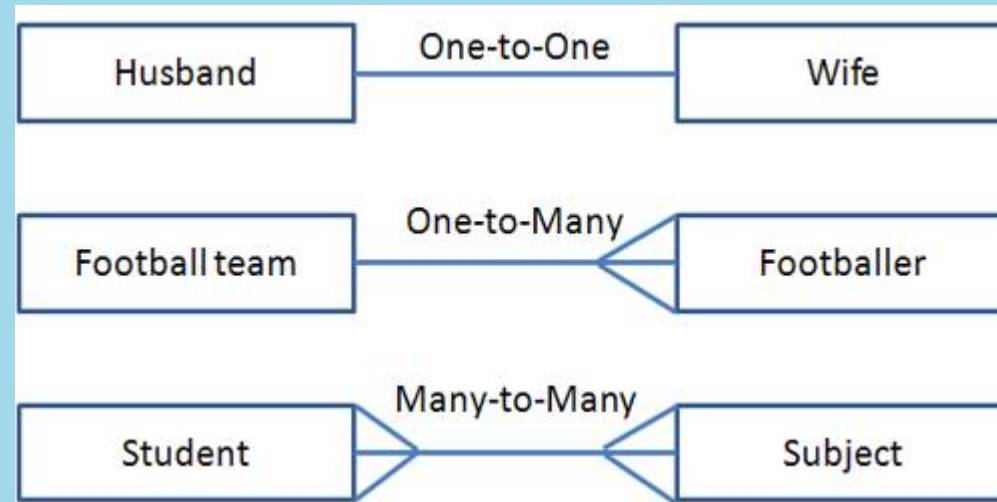
```
[1] 61373.575 61373.575 1647.128
1647.128 136917.546 136917.546
```



Joining (a)spatial data

Joining (a)spatial data

- Requires a “key” field
- Multiple outcomes possible
- Think about your final data form

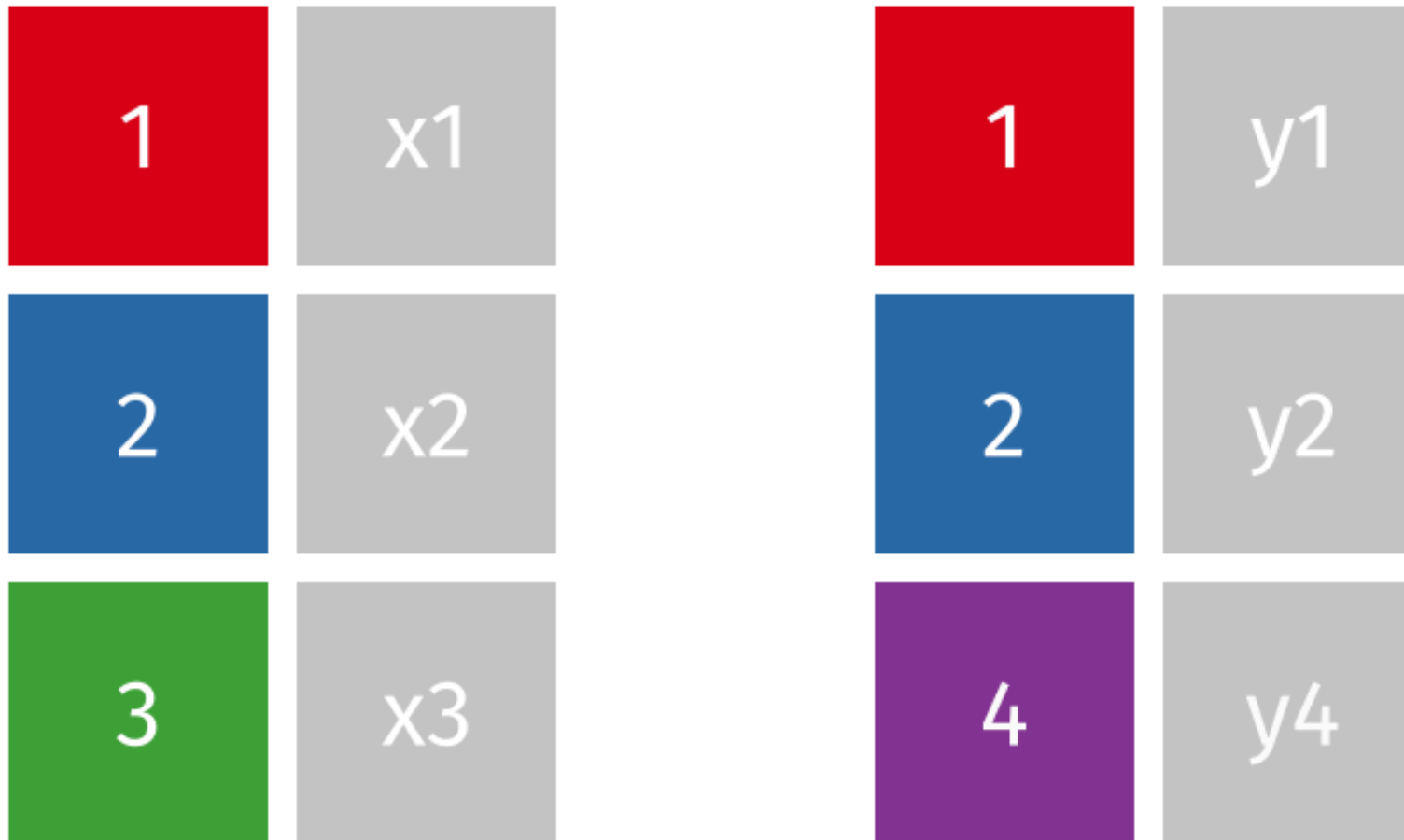


Left Join

- Useful for adding other attributes not in your spatial data
- Returns all of the records in **x** attributed with **y**
- Pay attention to the number of rows!

Left Join

`left_join(x, y)`



Left Join

```
1 head(coffee_data)
```

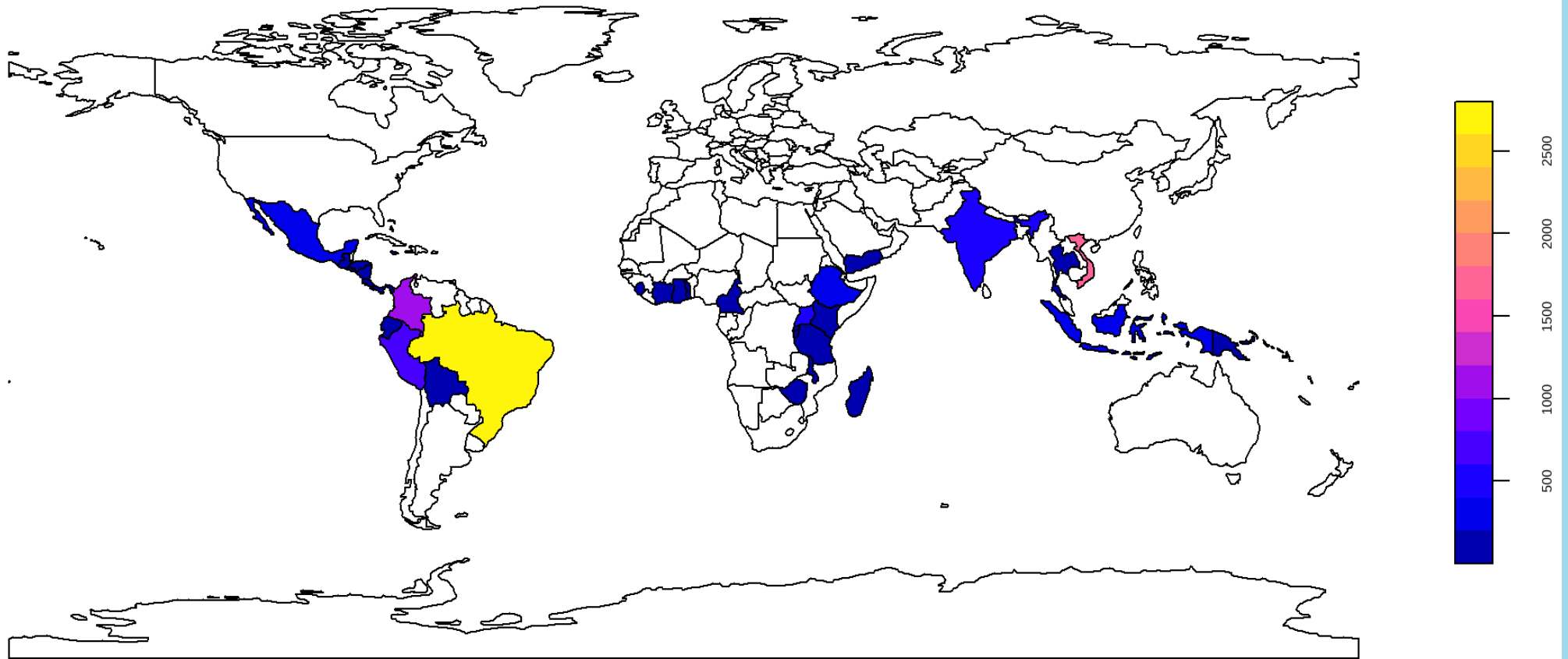
```
# A tibble: 6 × 3
  name_long
coffee_production_2016
coffee_production_2017
  <chr>
<int>          <int>
1 Angola
NA            NA
2 Bolivia
3          4
3 Brazil
3277         2786
4 Burundi
37           38
5 Cameroon
~            ~
```

```
1 world_coffee = left_join(v
2 nrow(world_coffee)
```

```
[1] 177
```

Left Join

coffee_production_2017



Inner Join

- Useful for subsetting to “complete” records
- Returns all of the records in **x** with matching **y**
- Pay attention to the number of rows!

Inner Join

`inner_join(x, y)`

1	x1	1	y1
2	x2	2	y2
3	x3	4	y4

Inner Join

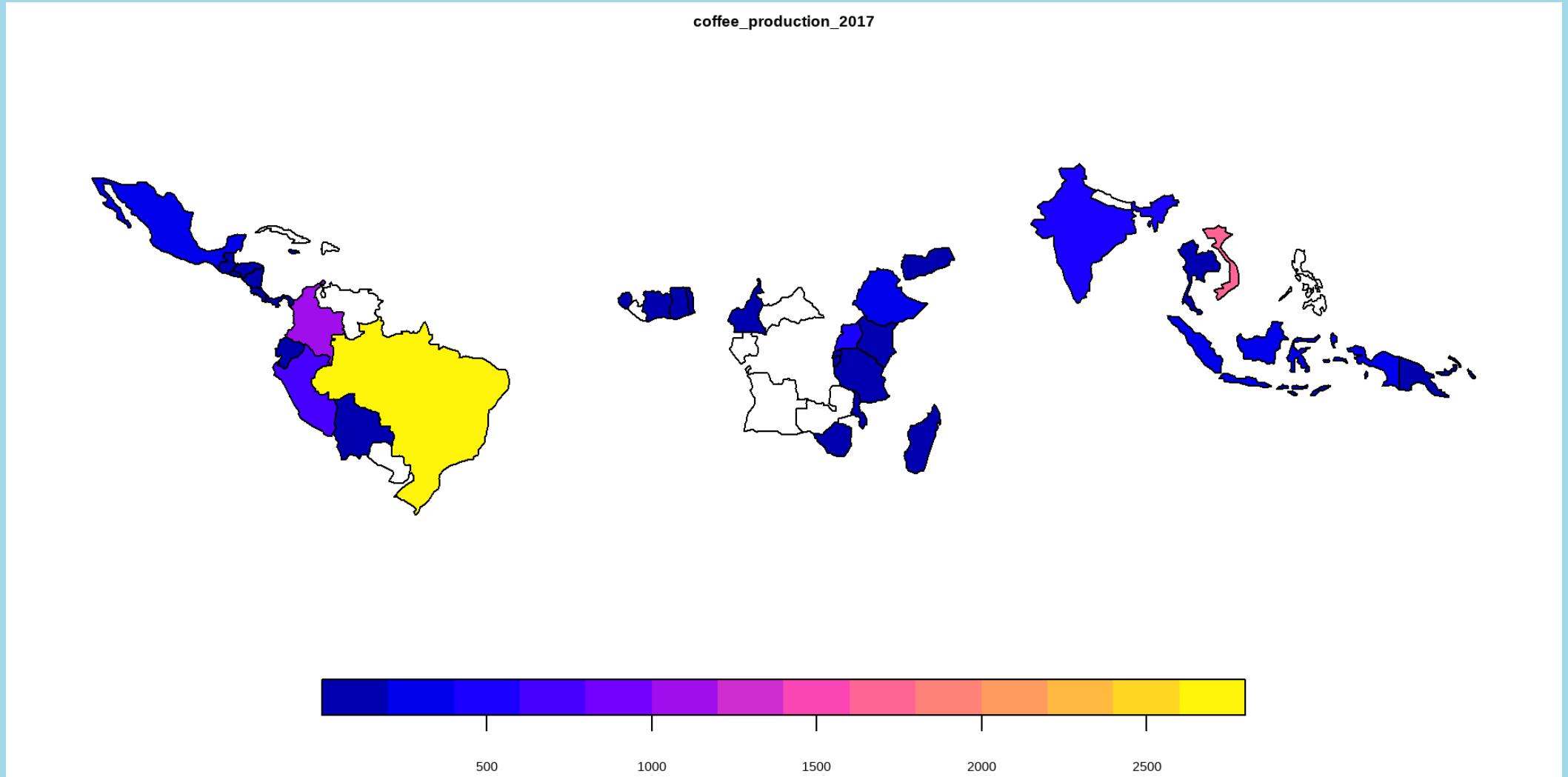
```
1 world_coffee_inner = inner_join(world_coffee, world_coffee_2013)
2 nrow(world_coffee_inner)
```

```
[1] 45
```

```
1 setdiff(coffee_data$name_2013, world_coffee_inner$name_2013)
```

```
[1] "Congo, Dem. Rep. of" "Others"
```

Inner Join



Spatial Joins

Spatial Joins

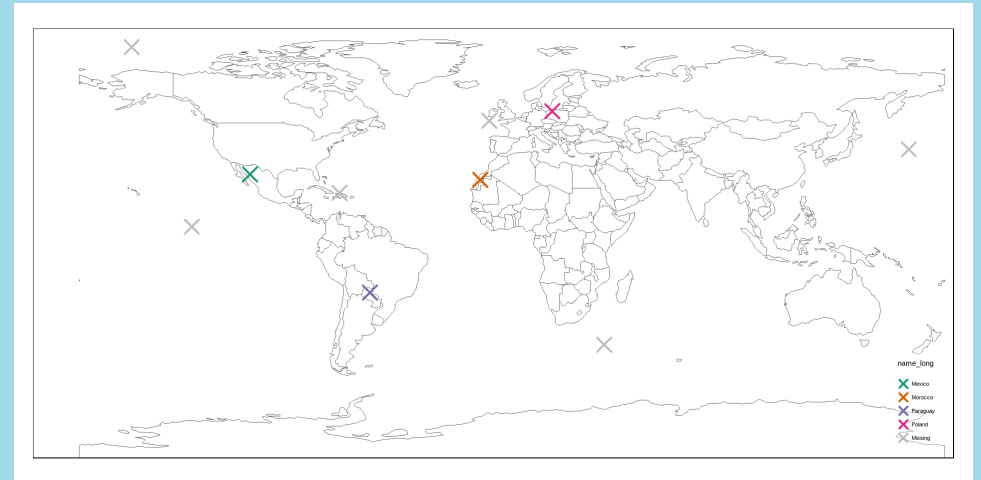
- `sf` package provides `st_join` for vectors
- Allows joins based on the predicates (`st_intersects`, `st_touches`, `st_within_distance`, etc.)
- Default is a left join

Spatial Joins

```
1 set.seed(2018)
2 (bb = st_bbox(world)) # tl
```

```
      xmin      ymin      xmax
ymax
-180.00000 -89.90000  179.99999
83.64513
```

```
1 #>      xmin      ymin      xmax
2 #> -180.0    -89.9    180.0
3 random_df = data.frame(
4   x = runif(n = 10, min =
5   y = runif(n = 10, min =
6 )
7 random_points = random_df
8   st_as_sf(coords = c("x",
9   st_set_crs("EPSG:4326")
10
11 random_joined = st_join(ra
```



Spatial Joins

- Sometimes we may want to be less restrictive
- Just because objects don't touch doesn't mean they don't relate to each other
- Can use `predicates` in `st_join`
- Remember that default is `left_join` (so the number of records can grow if multiple matches)

Spatial Joins

```
1 any(st_touches(cycle_hire, cycle_hire_osm
```

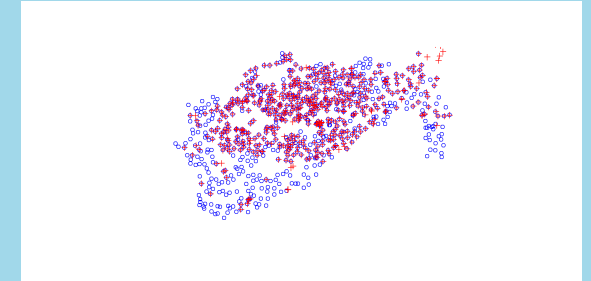
```
[1] FALSE
```

```
1 z = st_join(cycle_hire, cycle_hire_osm, s  
2 nrow(cycle_hire)
```

```
[1] 742
```

```
1 nrow(z)
```

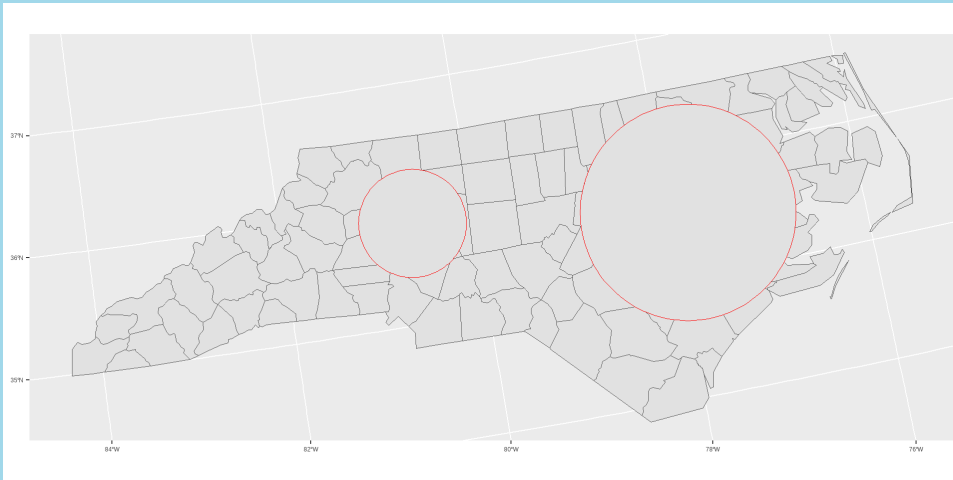
```
[1] 762
```



Extending Joins

Extending Joins

- Sometimes we are interested in analyzing locations that contain the overlap between two vectors
 - How much of home range a occurs on soil type b
 - How much of each Census tract is contained within a service provision area?
- `st_intersection`, `st_union`, and `st_difference` return new geometries that we can use as records in our spatial database



```
1 intersect_pct <- st_interse
2   mutate(intersect_area =
3     dplyr::select(NAME, int
4     st_drop_geometry())
5
6 nc <- mutate(nc, county_a
7
8 # Merge by county name
9 nc <- merge(nc, intersect_
10
11 # Calculate coverage
12 nc <- nc %>%
13   mutate(coverage = as.nu
```

Extending Joins

