




# County-level HIV Prevalence and Risk Analysis

Presented by:

Christine Gao & Heidi Sipes



# Outline

- Introduction
  - HIV prevalence and infection in the U.S. may be more severe in certain geographic locations and within demographics
  - Certain behaviors are likely to result in more cases in certain locations
- Methods
  - Unsupervised algorithm to determine clusters within and across different populations
- Data Analysis
  - R
- Conclusions/Takeaways
  - Do we see a trend in HIV prevalence according to various demographic groups, and is a k-means algorithm suitable for partitioning the data into distinct, meaningful risk clusters?

# United States HIV Epidemic

1.2 million HIV cases in U.S.

## Highest Risk Groups

- Injecting Drug Users (IDUs)
- Male-to-Male Sex (MSM)
- African Americans
- Younger age groups

\*The Centers for Disease Control and Prevention

# Research Question

- Do HIV cases display different clustering across geographic regions (county specific by state) based on age, race, sexual orientation, and injecting drug use?
- Can an algorithm be created to assess the risk level of these highest risk groups and detect any overlapping within the groups?

# Data & Methods

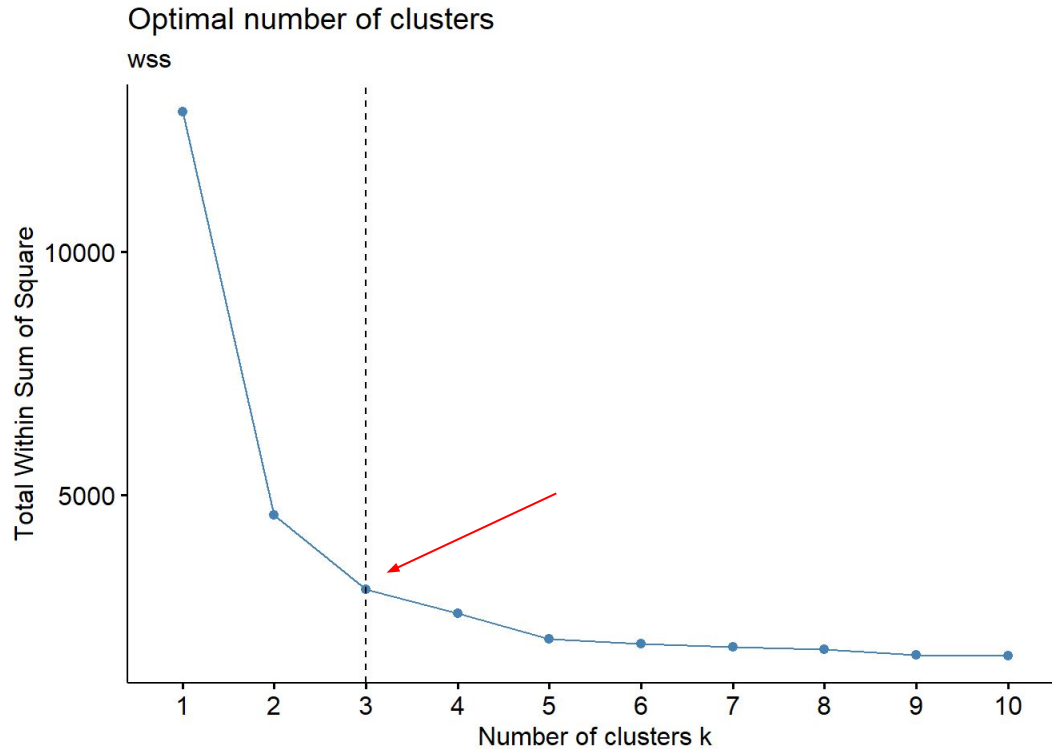
“2018 HIV infection rates in the US”

- Obtained via Sakai
- Determined the optimal number of clusters to use for cluster analysis
  - Function **fviz\_nbclust** from package **factoextra**
- k-means algorithm to visualize clusters
  - Function **kmeans** from package **stats**
  - Plot using function **fviz\_cluster** from package **factoextra**

# Analysis

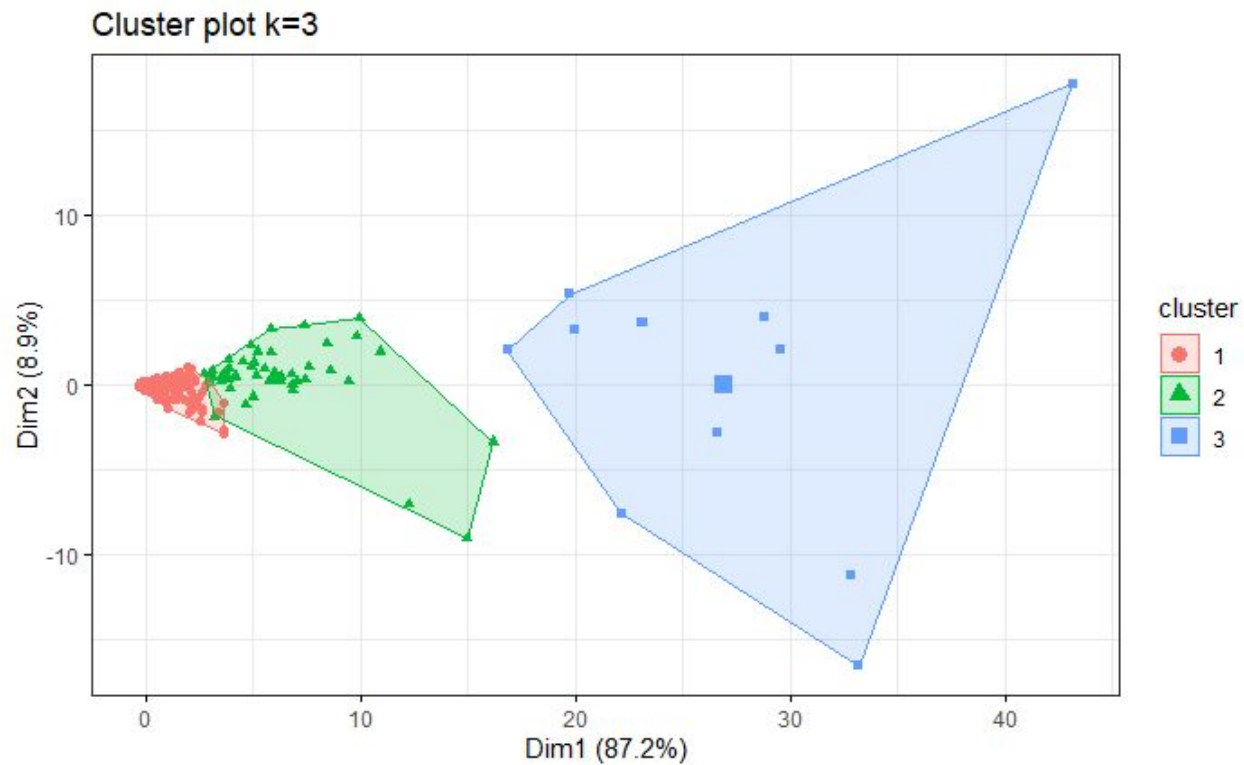
Goal: use k-means unsupervised algorithm to determine clusters or spatial patterns of HIV infections.

- Elbow method heuristic determines optimal number of meaningful clusters from the data to assign to subgroups varying in risk
- Label assigned reflect a “risk score” to create a discrete scale for different at-risk groups
  - Example: 1 to 3 (low to high risk)
- Sizes of the cluster points plotted on map reflect the risk value assigned
  - Larger clusters indicate “hotspots”



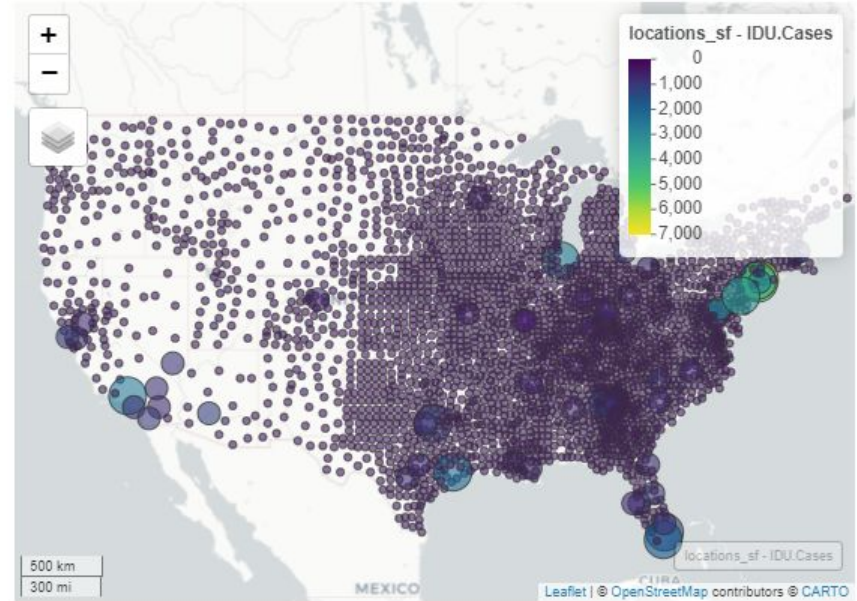
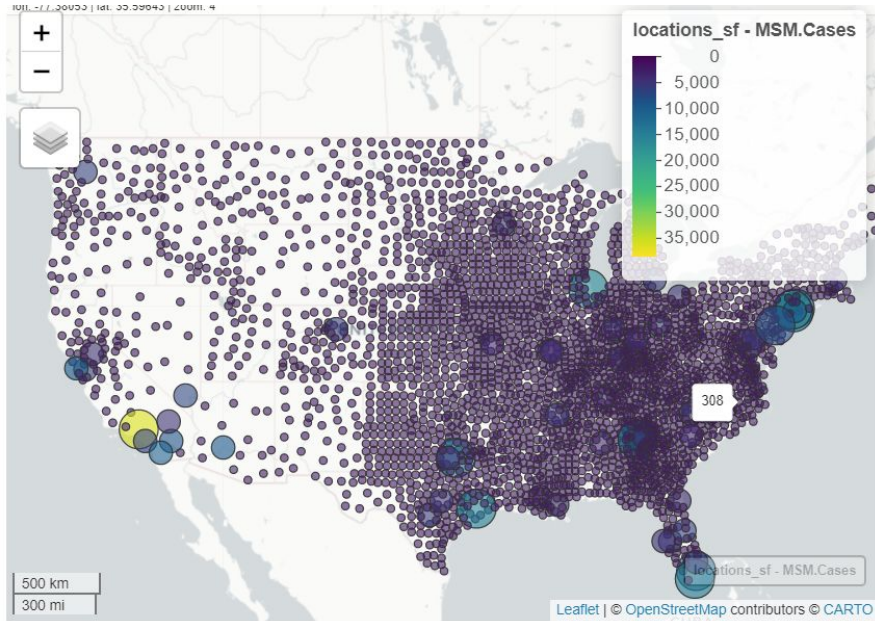
- WSS (within cluster sum of squared errors) or elbow method
- Amount of explained variation changes rapidly for small number of clusters, then decreases toward the “elbow”, which is the optimal number of clusters to use for a clustering algorithm

# Findings

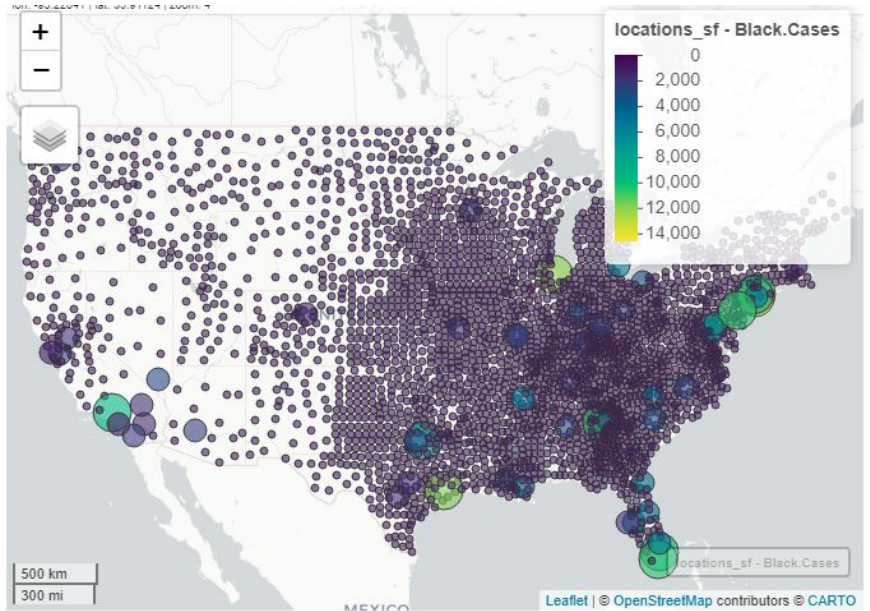
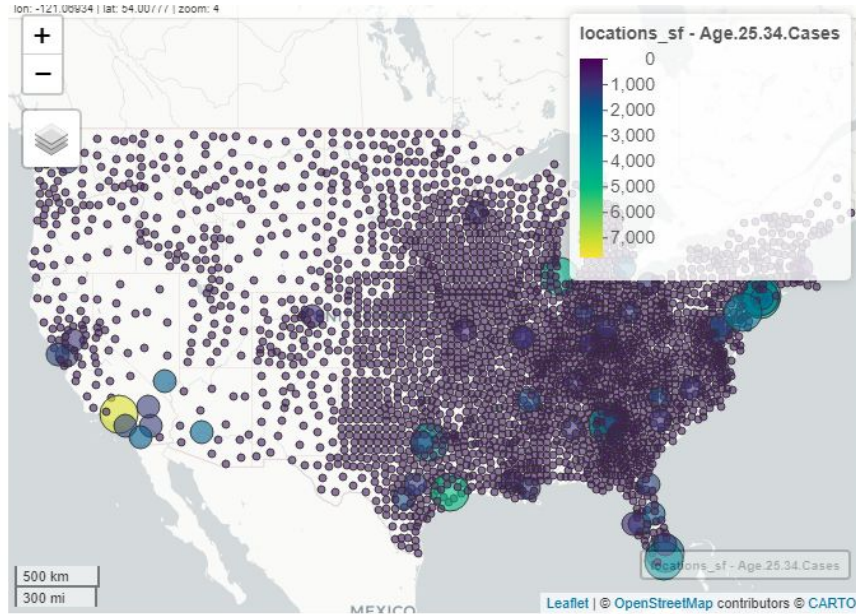


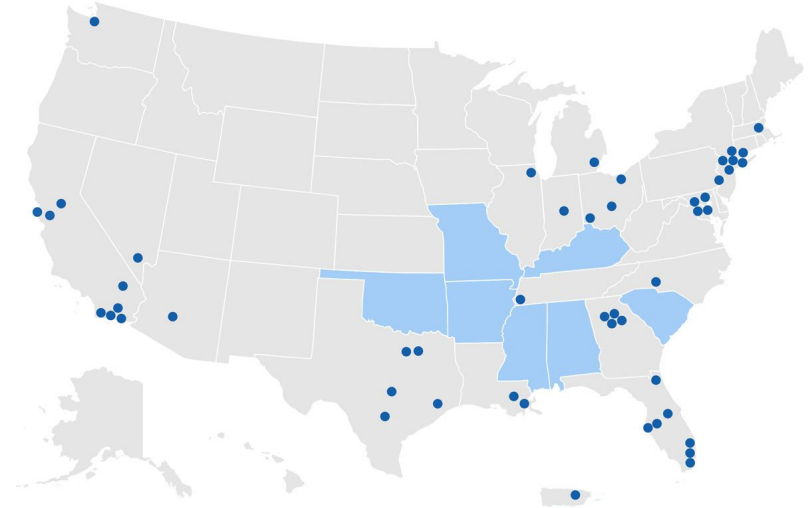
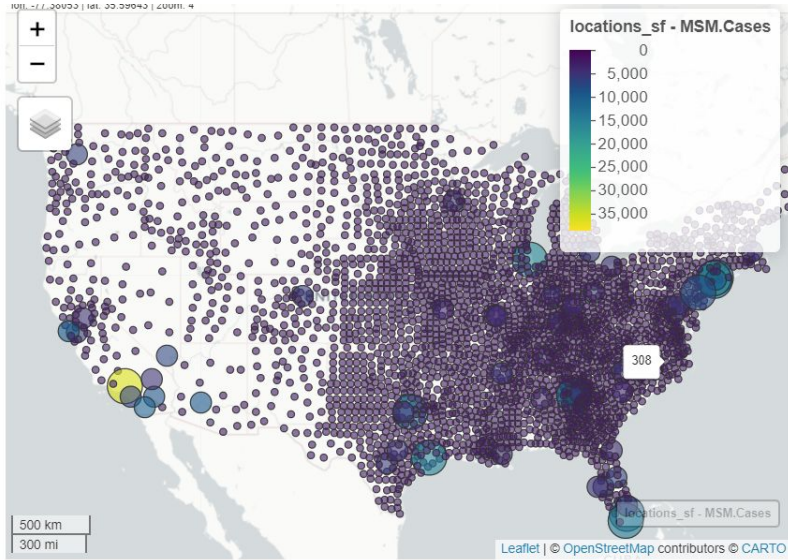


# Findings (cont.)



# Findings (cont.)





**Geographic Hotspots:** The 48 counties, plus Washington, DC, and San Juan, PR, where >50% of HIV diagnoses occurred in 2016 and 2017, and an additional seven states with a substantial number of HIV diagnoses in rural areas

<https://www.hiv.gov/federal-response/ending-the-hiv-epidemic/overview>



	county_fips	State.Abbreviation.x	County.Name	lon	lat	MSM.Cases	IDU.Cases	Age.25.34.Cases	Black.Cases	clusters
176	06037	CA	Los Angeles County	-118.21274	34.36996	38397	2515	7733	9538	3
1823	36061	NY	New York County	-73.97427	40.77022	17328	4075	3033	7434	3
577	17031	IL	Cook County	-87.82118	41.84296	16074	2552	4678	12615	3
2587	48201	TX	Harris County	-95.40106	29.87217	14916	1997	5277	12290	3
334	12086	FL	Miami-Dade County	-80.58750	25.60789	14370	1490	3449	10874	3
2543	48113	TX	Dallas County	-96.77946	32.76743	12531	1048	3761	7632	3
1816	36047	NY	Kings County	-73.95465	40.64058	11081	5412	3811	14522	3
418	13121	GA	Fulton County	-84.46753	33.79484	10636	964	3257	10812	3
297	12011	FL	Broward County	-80.49869	26.14256	10240	1046	2221	9103	3
1795	36005	NY	Bronx County	-73.86494	40.85618	9343	7006	3737	11587	3
2259	42101	PA	Philadelphia County	-75.13523	40.00487	6397	3457	2697	10841	3

	2017	2018	2019	2020	2021 Q1 Prelim.	GOAL 2025	GOAL 2030
Los Angeles County, CA, Total	1,799	1,685	1,482	1,145	132	450	180
Miami-Dade County, FL, Total	1,141	1,168	1,151	834	235	285	114
Harris County, TX, Total	1,100	1,206	1,195	647	20	275	110
Dallas County, TX, Total	815	794	733	608	68	204	82
Cook County, IL, Total	978	983	881	592	72	245	98
Maricopa County, AZ, Total	494	530	513	491	61	124	49
Broward County, FL, Total	671	616	594	489	124	168	67
Fulton County, GA, Total	618	592	537	481	67	155	62
Kings County, NY, Total	630	547	466	426	58	158	63
Orange County (FL), FL, Total	461	461	466	398	98	115	46
Bronx County, NY, Total	506	456	499	334	37	127	51
Clark County, NV, Total	444	445	449	323	53	111	44
Philadelphia County, PA, Total	498	439	446	321	54	125	50
Queens County, NY, Total	431	413	354	306	27	108	43
New York County, NY, Total	396	374	338	289	34	99	40

# Findings (cont.)

Table 1. K Means algorithm unsupervised model performance

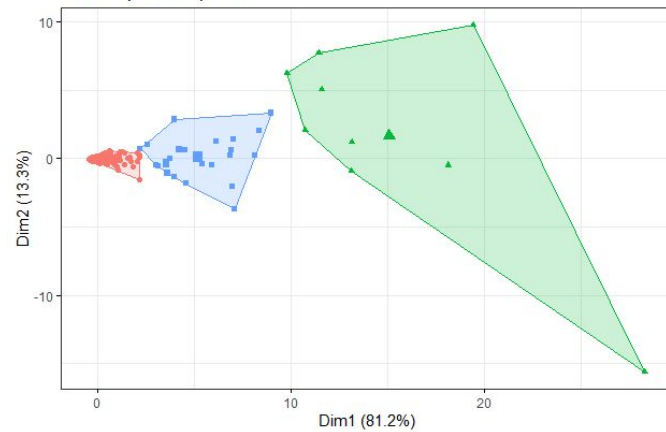
Performance Measures		Sample 1 n=1000 set.seed=21093	Sample 2 n=1000 set.seed=21045	Sample 3 n=2000 set.seed=21093
<i>Accuracy</i>		0.976	0.991	0.998
<i>Recall</i>				
	<i>Cluster 1</i>	0.984	1	1
	<i>Cluster 2</i>	0.529	0	0.875
	<i>Cluster 3</i>	0.984	1	1
<i>Precision</i>				
	<i>Cluster 1</i>	1	0.994	0.998
	<i>Cluster 2</i>	1	0	0.966
	<i>Cluster 3</i>	0	0.5	1
<i>F1</i>				
	<i>Cluster 1</i>	0.992	0.997	0.999
	<i>Cluster 2</i>	0.692	N/A	0.918
	<i>Cluster 3</i>	0	0.667	1

# Results

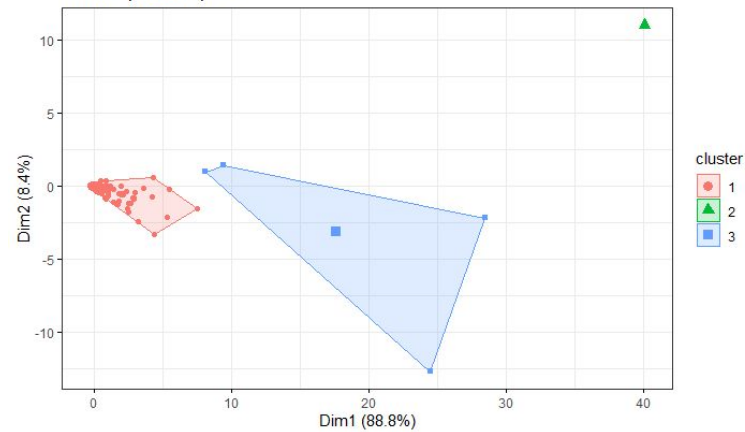
Repeated random sampling indicates the performance of the unsupervised model is very good for the group assigned a value of 1.

- This is likely because most counties in the U.S. from the dataset are considered low risk and have low HIV infection rates → **greater proportion of sample is accurately assigned**
- Cluster 2 and 3 are more ambiguous → **less accuracy surrounding the assignment of medium-to-high risk counties**
  - Room for reinvestigation on how assignments are made and how many clusters should be used for more granular distribution

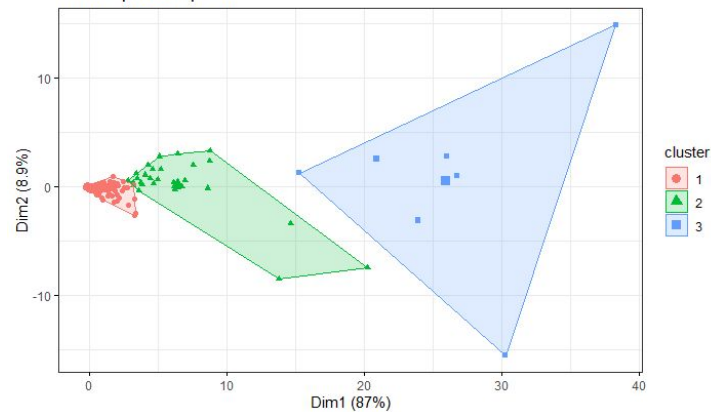
Cluster plot sample 1



Cluster plot sample 2



Cluster plot sample 3



# Results (cont.)

- The clusters assigned by the k-means algorithm appear to be robust, with consistent performance across varying random samples
  - The algorithm consistently assigns counties with lower number of cases to a risk category of 1, and counties with the highest HIV prevalence to a risk category of 3.
- The highest risk counties (counties with largest cluster size on maps) most consistently the locations with highest number of HIV cases
  - Consistent with HIV.gov list of counties with the highest number of HIV diagnoses



# Conclusion

- The K means unsupervised algorithm is pretty good, but not great for all of our subgroups when assigning risk scores based on HIV prevalence for different counties
  - Seems to perform well for only one subgroup, which is the largest subset from the data
- However, there is value in the geographical mapping of HIV cases, revealing potential hotspots
- Policy implications:
  - Regular diagnoses in regions considered to be HIV hotspots
  - Prioritize treatment and prevention in high-risk counties
  - Socioeconomic differences and marginalized communities → focus on education and increase funding to those regions

# Thank you!

Any questions?