Integrating Chat GPT in the Creation of Data Products for Consumer Data Research

Alex Singleton
Professor of Geographic Information Science
University of Liverpool

www.cdrc.ac.uk
www.alex-singleton.com
geographicdatascience.com
@alexsingleton

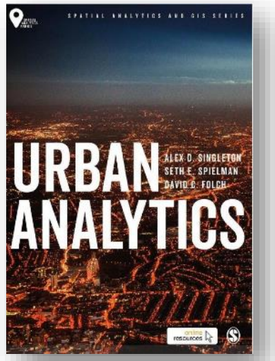
UNIVERSITY OF LIVERPOOL

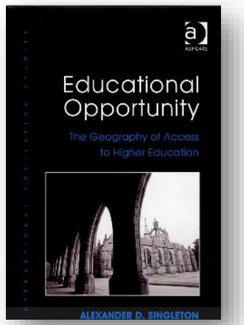
Consumer Data Research Centre

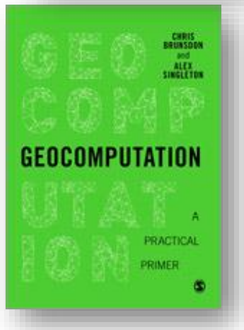
Geographic Data Science Lab

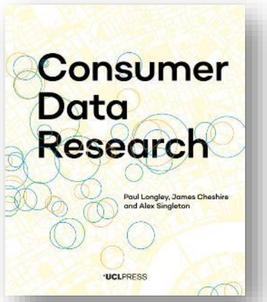
UKRI Economic and Social Research Council

About Me









Geographic Data Science Lab

Alex Singleton

PhD

2007

2000-2003 — 2005 — 2010 — 2023

MANCHESTER 1824 The University of Manchester

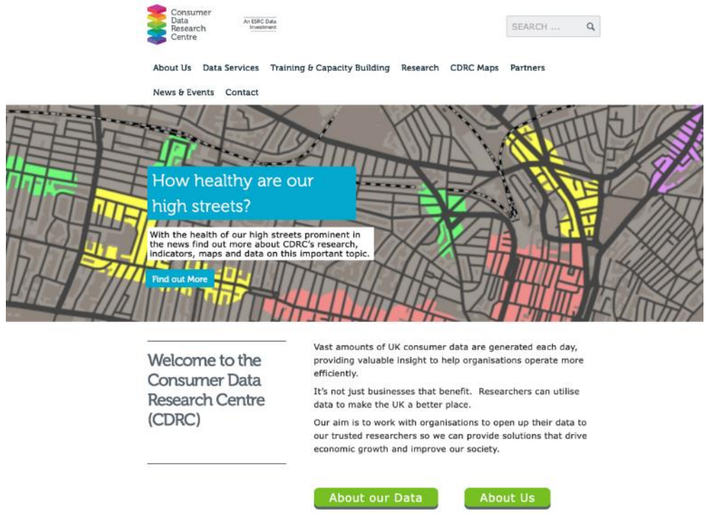
UG — *KTP* UCAS — *Researcher* — *Lecturer / Reader / Chair*

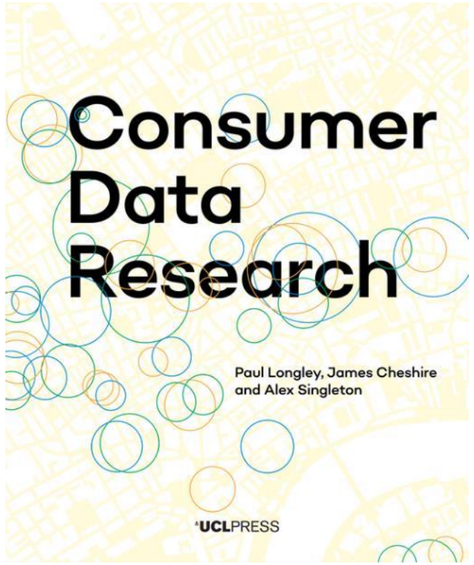
PhD pt. UCL

UNIVERSITY OF LIVERPOOL

Consumer Data Research Centre

Consumer Data Research Centre





Since 2014

- Funding emerged as a research council response to “Big Data”
  - Big Data Network II
- Pitch – “Business Data”
- Now – Digital Footprints Phase 1 Centre

University of Liverpool
UCL
University of Oxford
University of Leeds
UKRI Economic and Social Research Council

Consumer Data Research Centre

Model

Data Owners

- Trust
  - Takes time…
- Motivation for engagement
- Assurance
  - Legal
  - Data Security (Service)

Potential Research Users

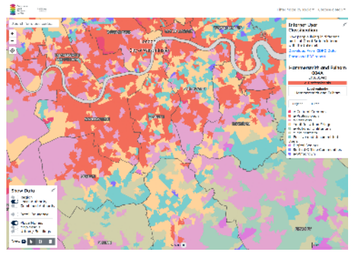
- Data service
  - Secure
  - Efficient
- Data Discoverability
  - Platforms
  - Analysis Ready Data

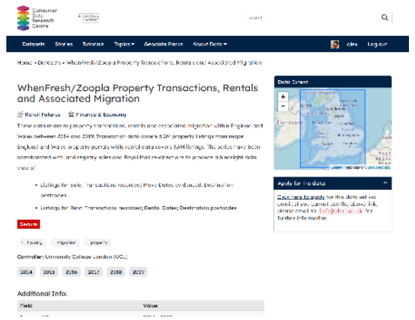
UNIVERSITY OF LIVERPOOL

Consumer Data Research Centre

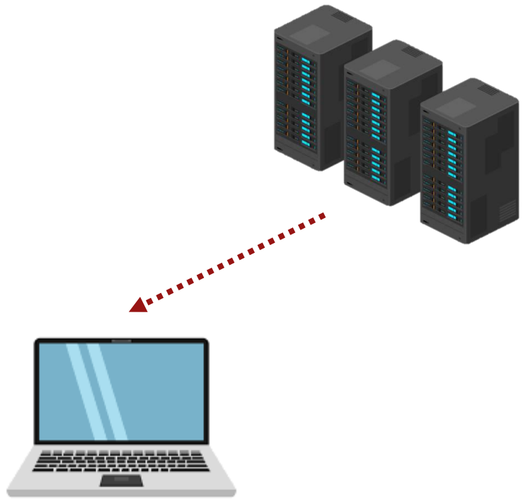
Data Service

Open





Safeguarded



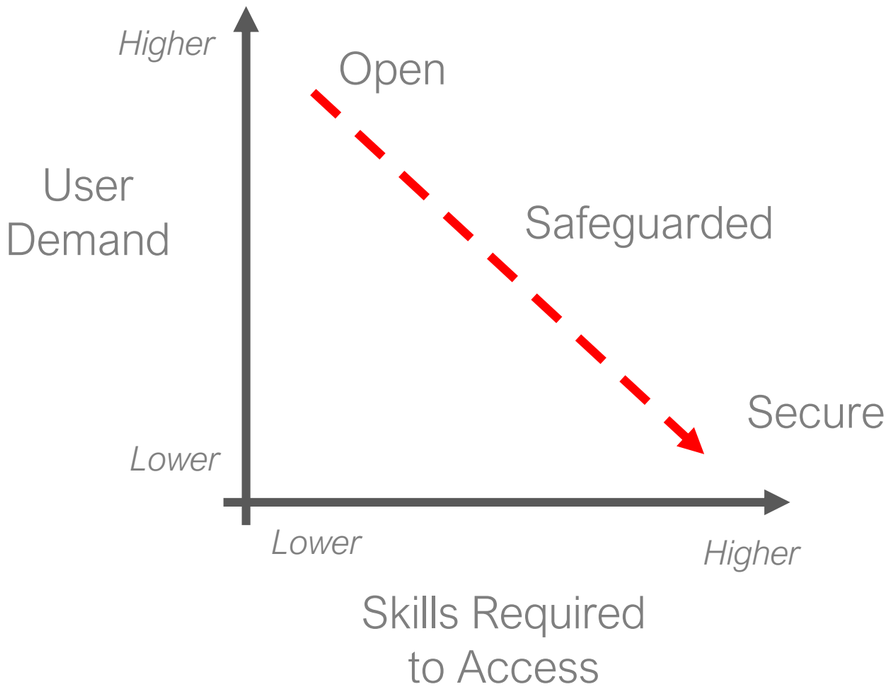
Secure

ISO27001

Research Approvals Process

UNIVERSITY OF LIVERPOOL

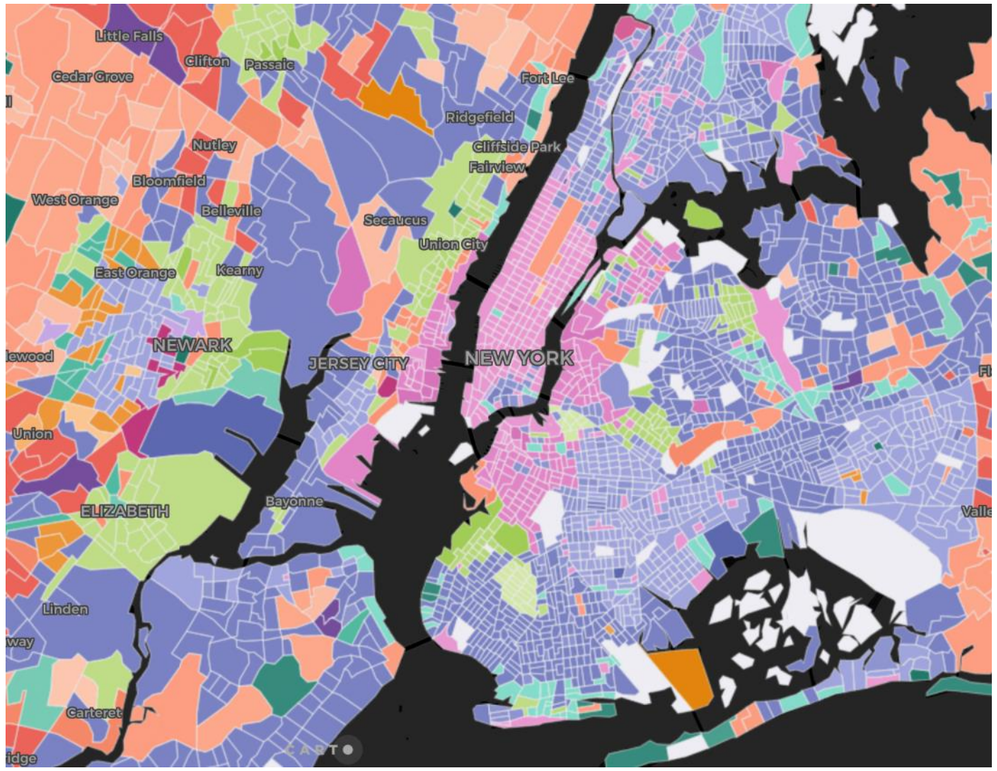
Analysis Ready Data Products



UNIVERSITY OF LIVERPOOL

US Geodemographic

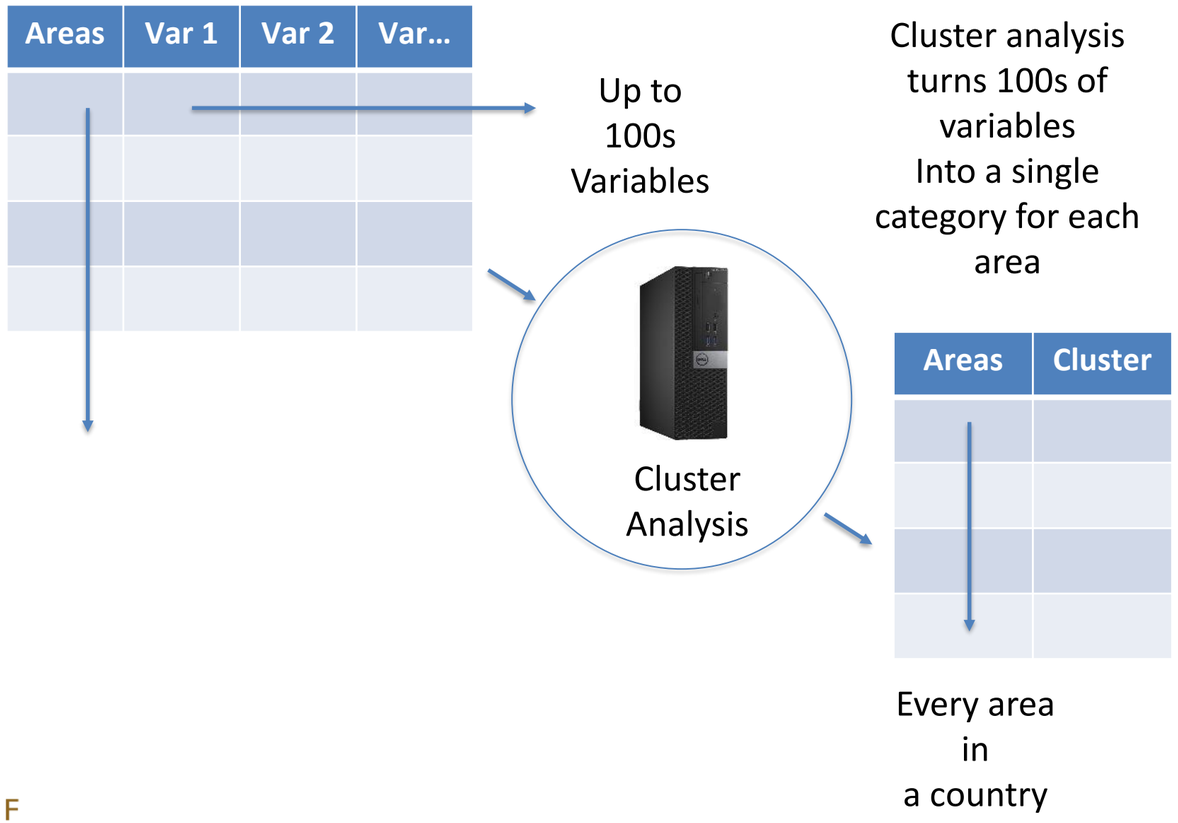
- Create an updated model for the US
- Move from Tract (1-8k people) to Block Group (600-3k people)
- Evaluate the use of Chat GPT for the Creation of Descriptions / Labels



~240k Block Groups
~85k Tracts

UNIVERSITY OF LIVERPOOL

US Geodemographics



UNIVERSITY OF LIVERPOOL

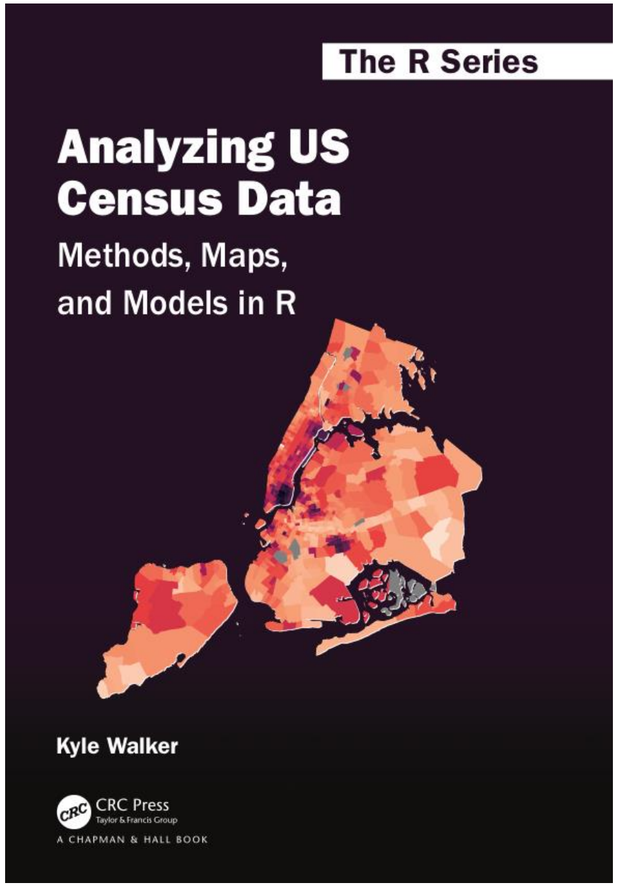
US Geodemographics

- Commercial models
  - Cost $$
  - Not open

UNIVERSITY OF LIVERPOOL

American Community Survey

- Ongoing annual survey
- Around 3.5m responses
- Replaced the long form US Census
- Small area estimates



UNIVERSITY OF LIVERPOOL

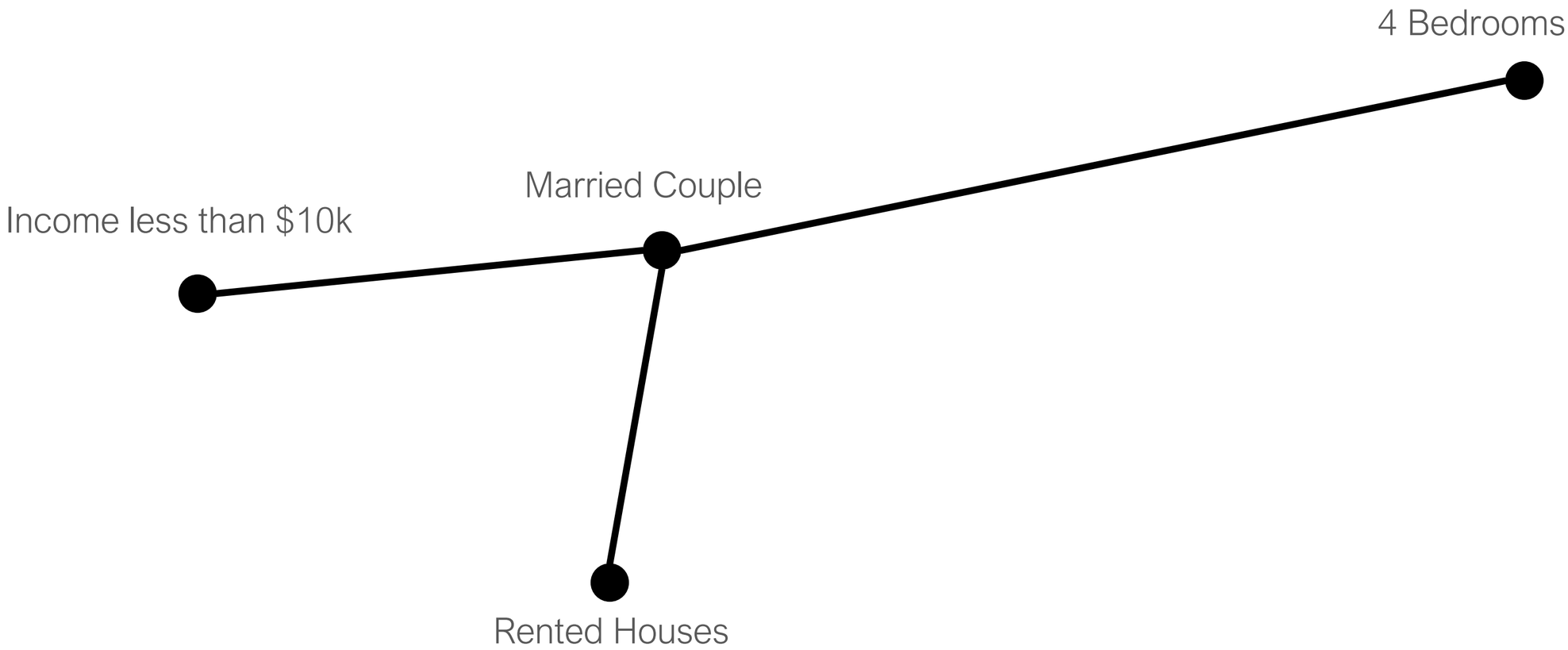
walker-data.com/census-r

Variable Selection

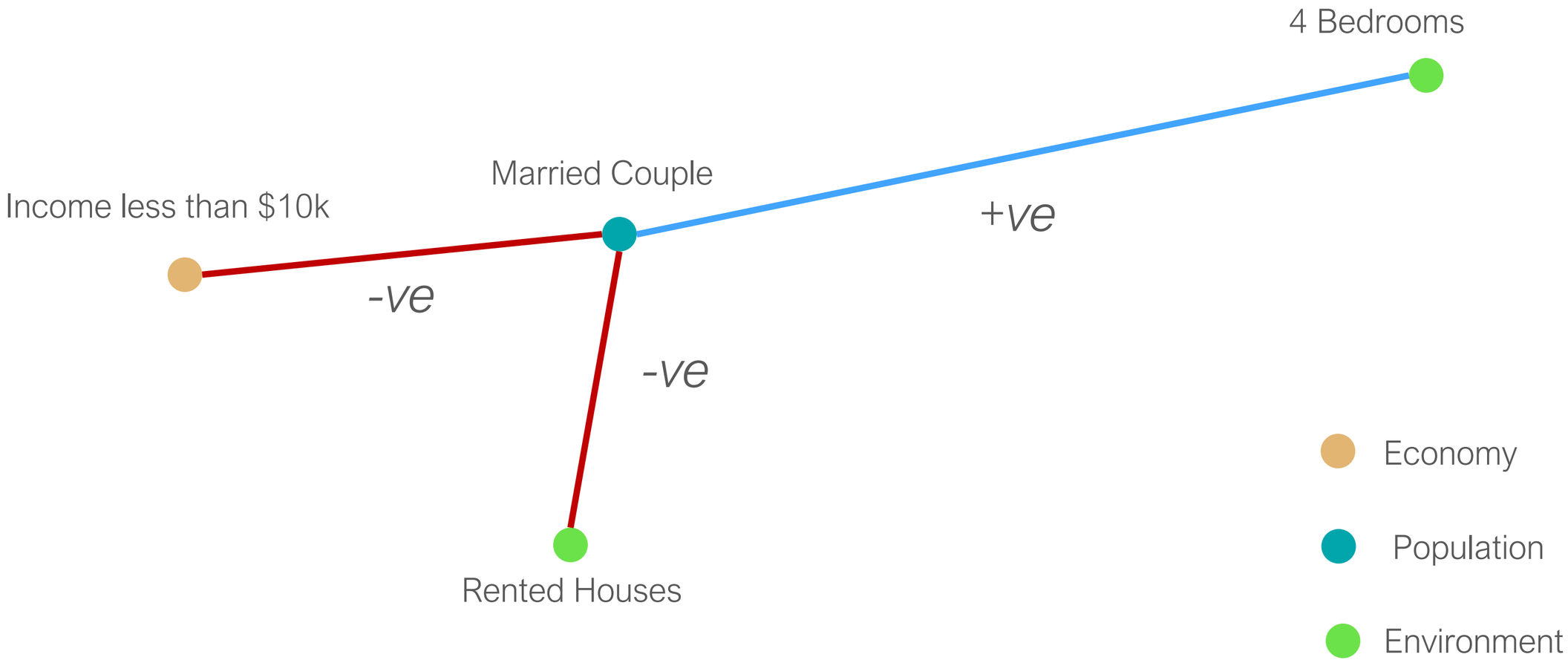
- Theoretical Framework
  - Concepts; Domains; Measures
- Past Variables
- “Good” Variables
  - Coverage; Correlation; Variance

Concept	Domain
Economy	Digital Connectivity
Economy	Digital Devices
Economy	Employment
Economy	Mobility
Economy	Start Work Time
Economy	Transport
Economy	Wealth
Environment	Accommodation size
Environment	Building Age
Environment	Housing Cost
Environment	Occupancy
Environment	Rent
Environment	Structure Type
Environment	Tenure
Environment	Unit Size
Environment	Vacant Units
Population	Children
Population	Education
Population	Family
Population	Family Structure
Population	Household
Population	Housing
Population	Language
Population	Marriage
Population	Race / Ancestry / Citizenship

Correlation

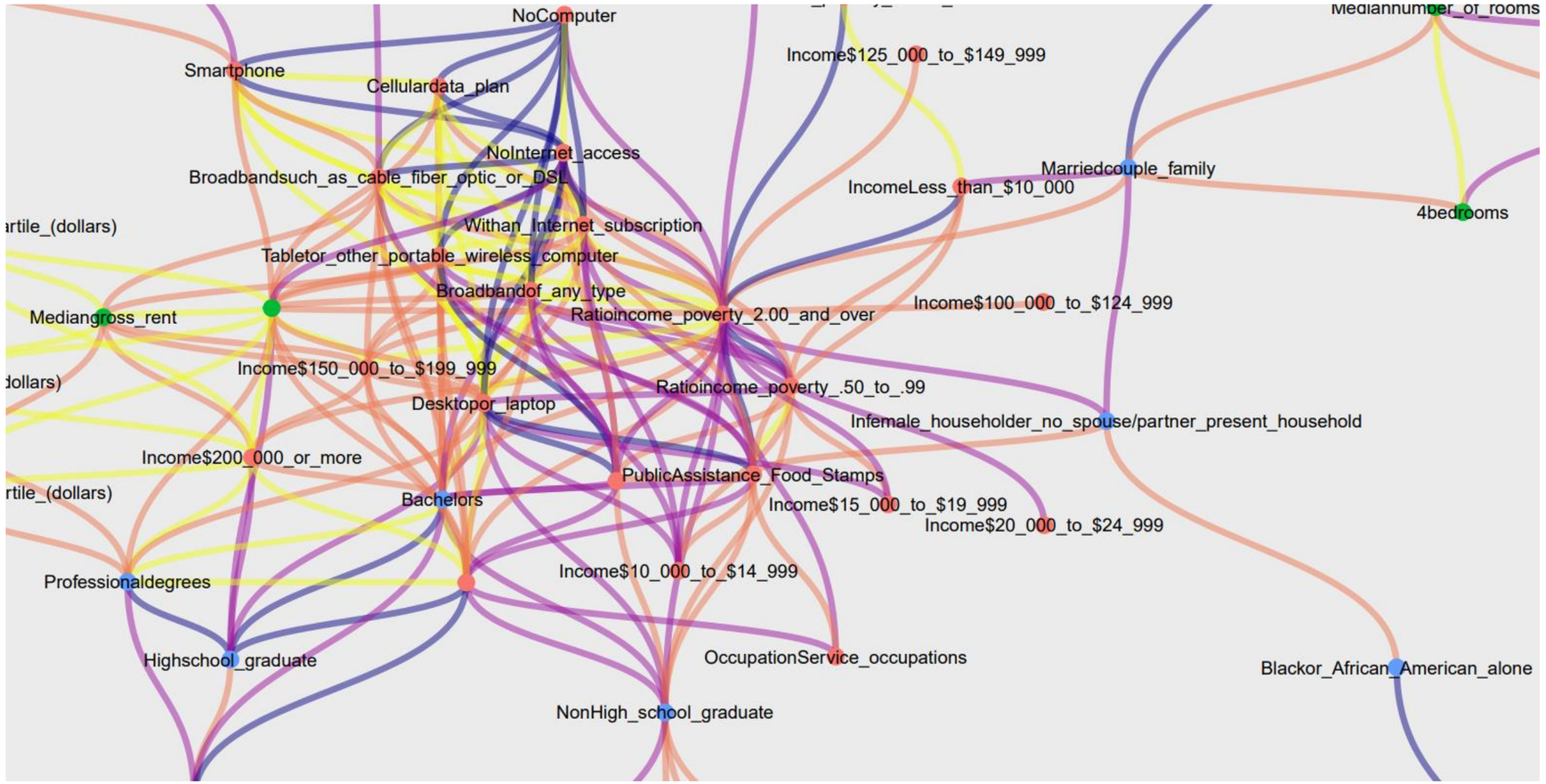


Correlation



UNIVERSITY OF LIVERPOOL

Correlation



CONCEPT

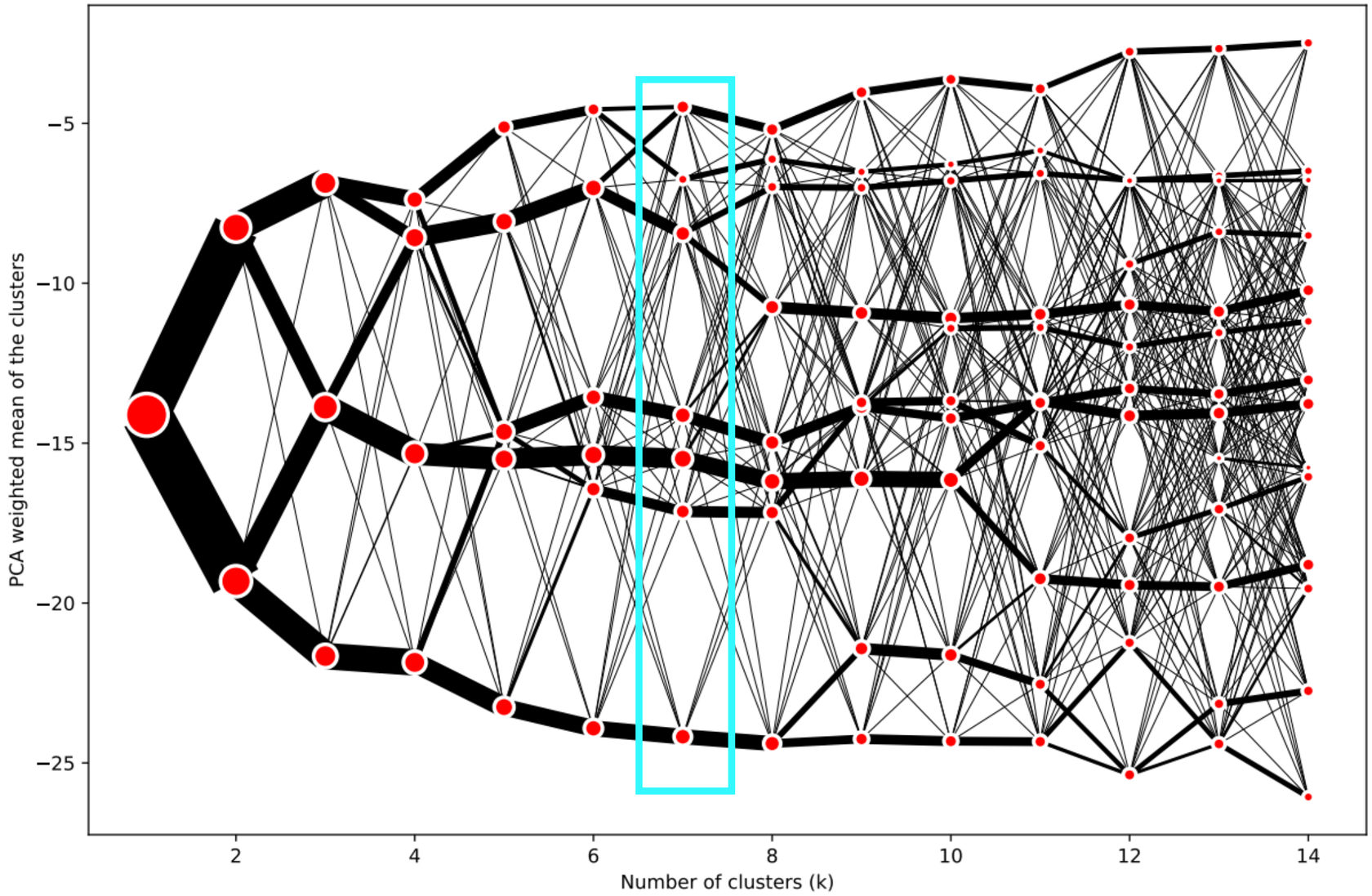
- ECONOMY
- ENVIRONMENT
- POP

band

- (−Inf,−0.7]
- (−0.7,−0.6]
- (0.6,0.7]
- (0.7, Inf]

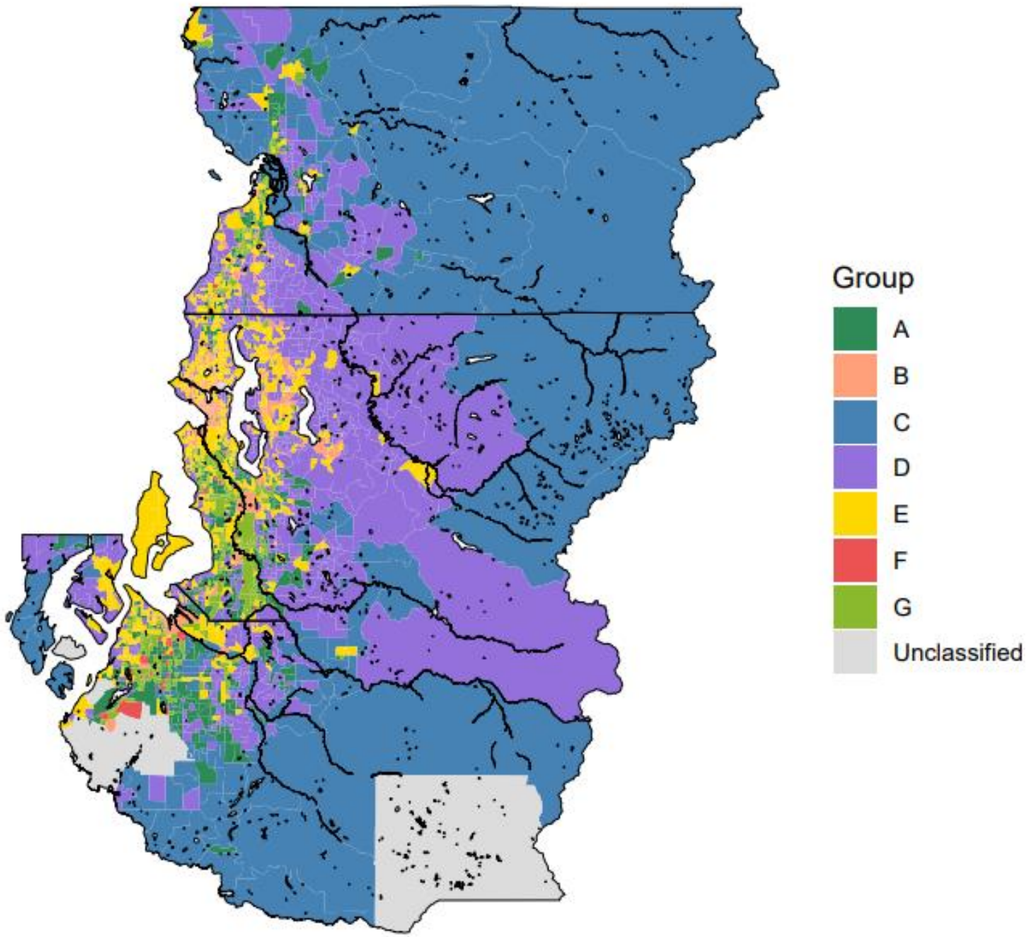
UNIVERSITY OF LIVERPOOL

Cluster Selection



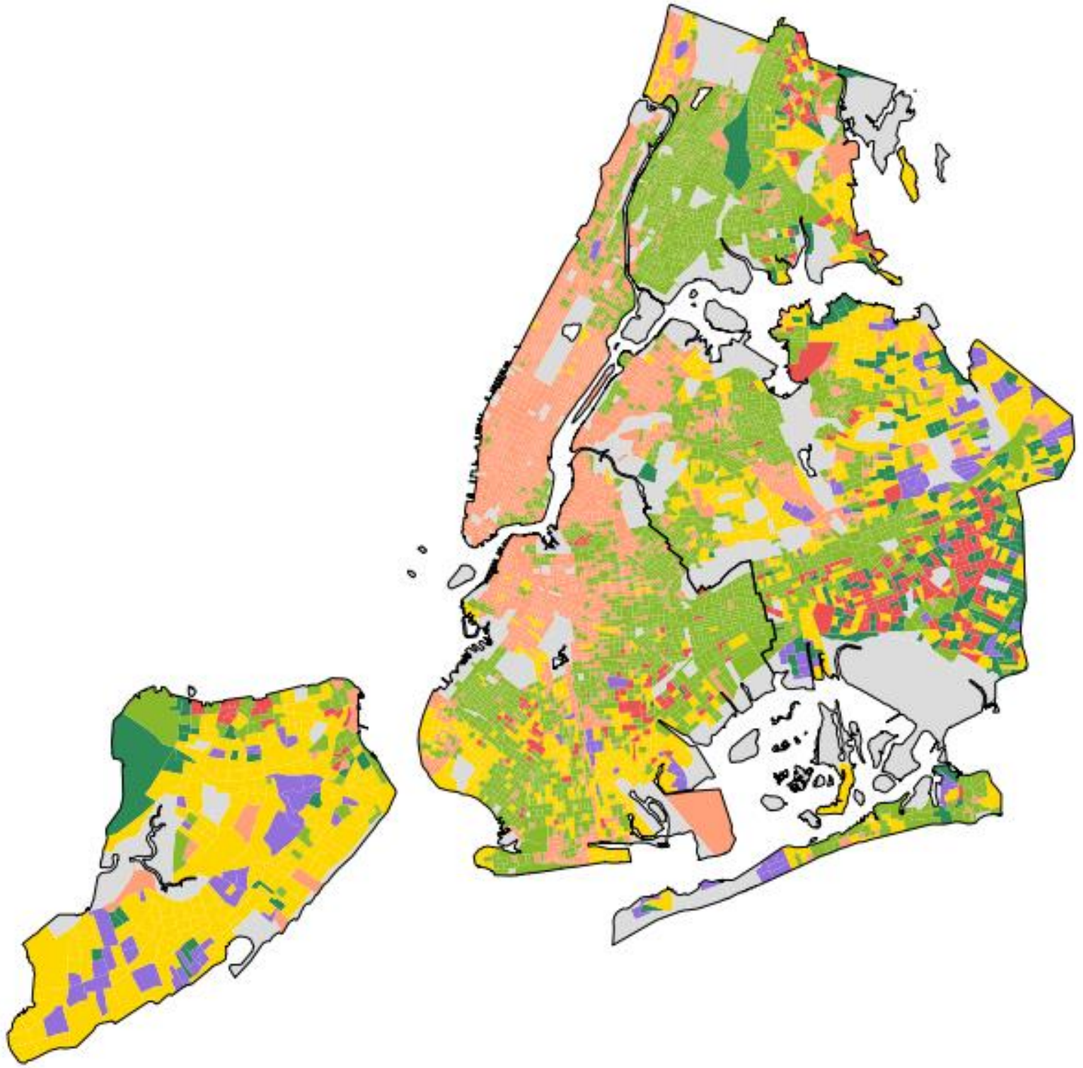
- 7 Groups
- 39 Types

Seattle



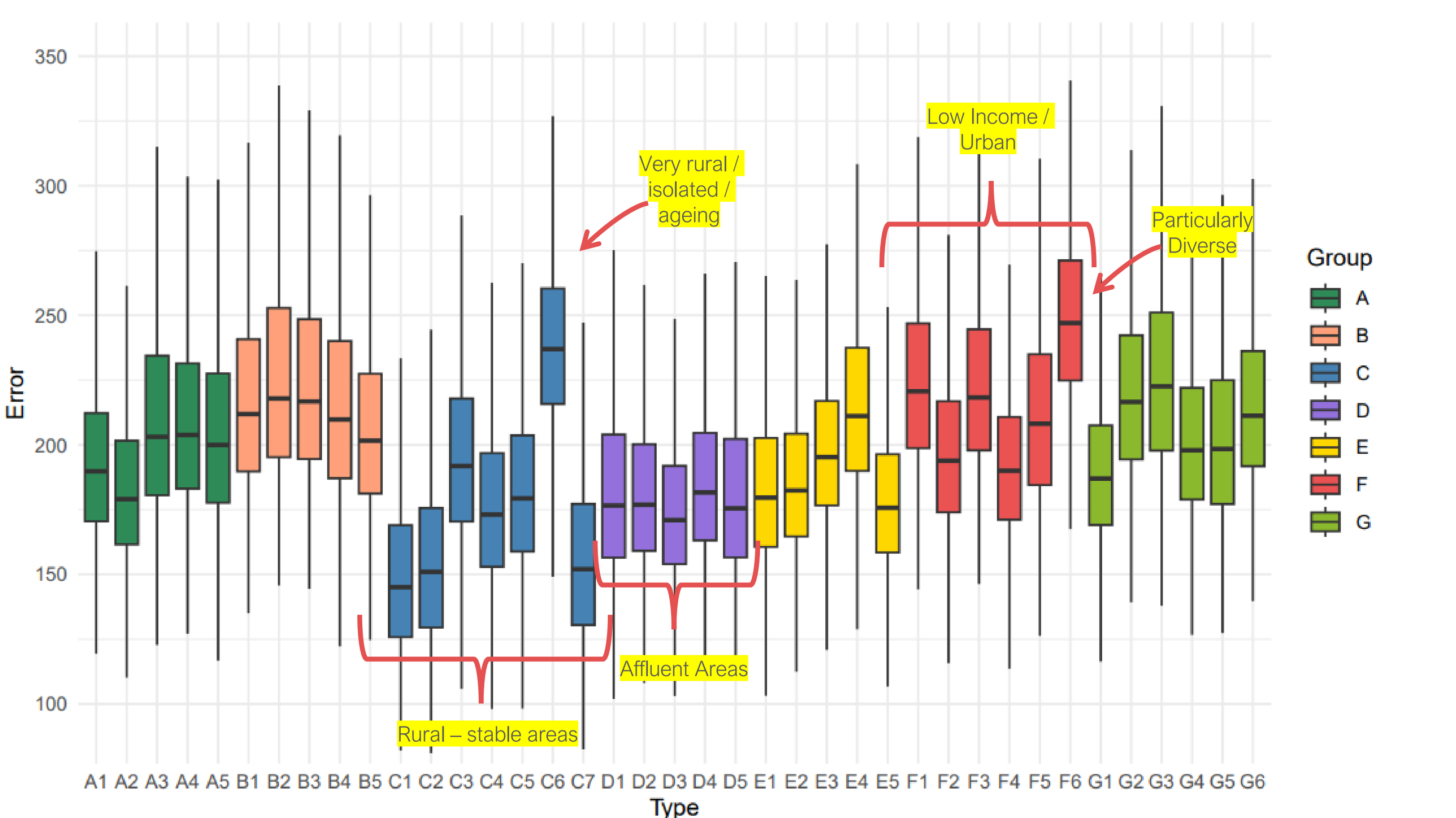
Group
A
B
C
D
E
F
G
Unclassified

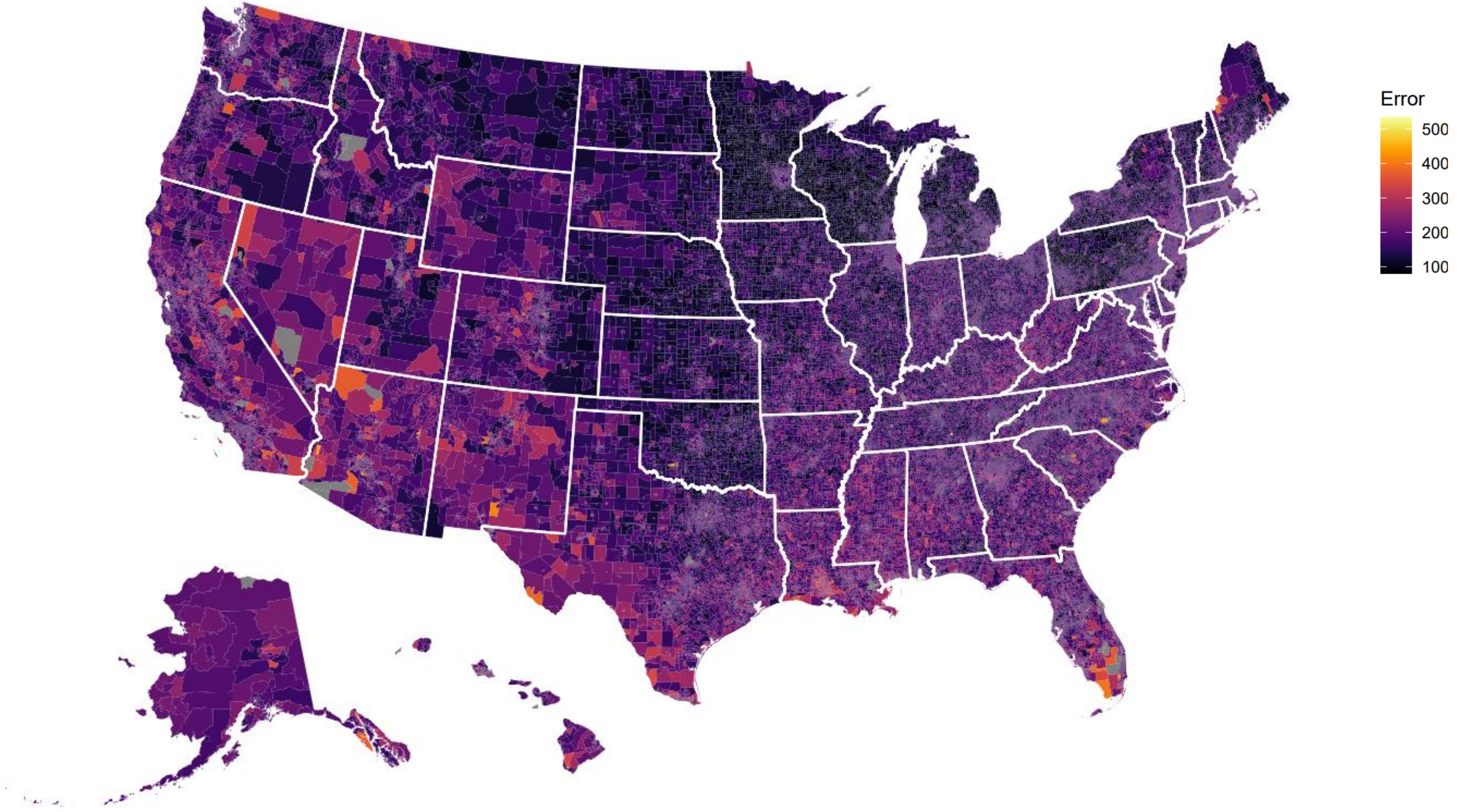
New York



Group
A
B
D
E
F
G
Unclassified

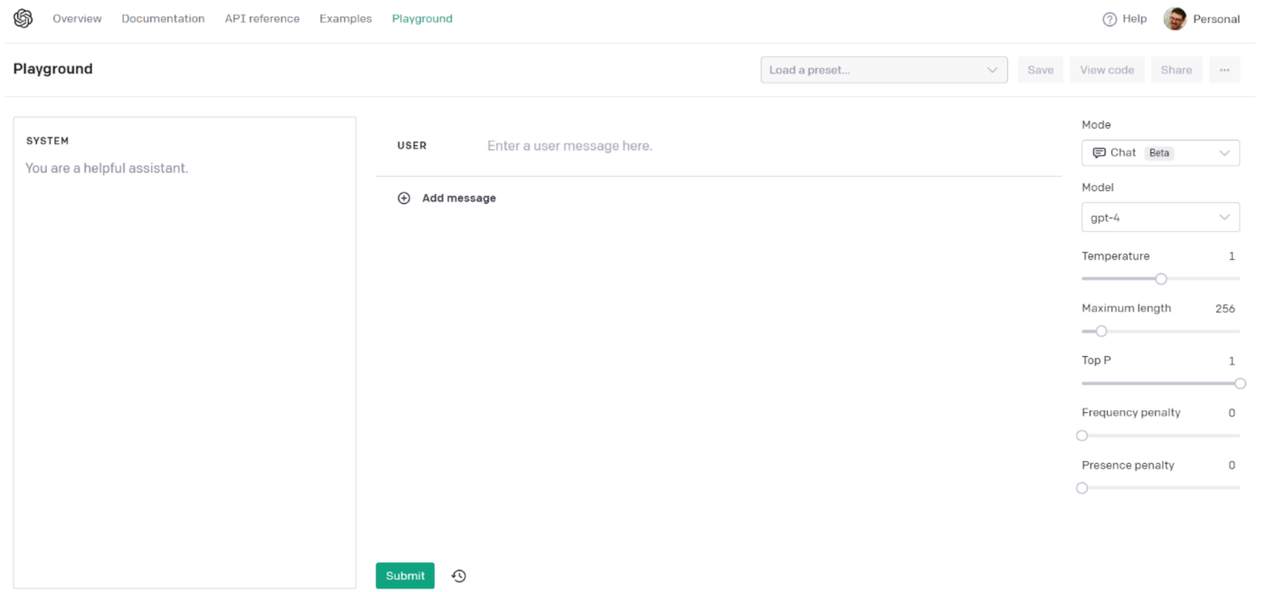
UNIVERSITY OF LIVERPOOL


Very rural / isolated / ageing
Low Income / Urban
Particularly Diverse
Affluent Areas
Rural – stable areas
Error
Type
Group
A
B
C
D
E
F
G
350
300
250
200
150
100
A1 A2 A3 A4 A5 B1 B2 B3 B4 B5 C1 C2 C3 C4 C5 C6 C7 D1 D2 D3 D4 D5 E1 E2 E3 E4 E5 F1 F2 F3 F4 F5 F6 G1 G2 G3 G4 G5 G6


Error
500
400
300
200
100

Cluster Descriptions

- Chat GPT 4 (March 23)
- Prompt Engineering –
  - Zero shot approach
- Evaluation



platform.openai.com/playground?mode=chat&model=gpt-4

UNIVERSITY OF LIVERPOOL

Cluster Descriptions

- Specify the context of the task
  - *A geodemographics company is trying to explain the characteristics of a neighborhood to a new customer. They present data comparing this neighborhood to the national average.*

Cluster Descriptions

- Explain how to interpret the scores
  - *A score of 100 means the neighborhood is equivalent to the national average, a score of 150 means the neighborhood one and a half times the national average, a score of 200 means the neighborhood is twice the national average, a score of 50 means the neighborhood is half of the national average, a score of 300 means the neighborhood is three times the national average.*

Cluster Descriptions

- *Explain the structure of the data table*
  - *Their neighborhood has the following characteristics, described in #DATA# below. Data are presented for each characteristic followed by a colon, and then a score.*

Cluster Descriptions

- *Additional focus for the task and any adjustments*
  - *The description of the neighborhood should focus on characteristics that have scores which are greater than 120 or less than 80. If 'housing units that are Renter occupied' have scores that are less than 80, then don't describe any details about the rent prices.*

Cluster Descriptions

- *Present the data*

#DATA#
male aged Under 5 years:96
male aged 5 to 9 years:94
male aged 10 to 14 years:98
male aged 15 to 17 years:101
…

Cluster Descriptions

- Explain what I wanted

In the third person, write a description of the neighborhood in no more than 500 words. Don't mention the specific scores from the #DATA#, instead use descriptive words to illustrate rates that are above or below the national average.

A1 *This neighborhood is characterized by a diverse and multicultural population, with a significantly higher proportion of Black or African American, Asian, Hispanic or Latino, and residents of two or more races. It is also home to a large number of residents who speak Asian and Pacific Island languages and other Indo-European languages. These linguistic and cultural differences make the community vibrant and unique.*

The area has a distinct pattern of commuting and work schedules. A notable number of residents travel to work by public transportation, with longer commuting times being more common in the neighborhood. Many people start their workday early, between midnight and 5:59 a.m., and a sizable portion have a commute of 30 minutes or more. This suggests that the neighborhood might be home to a considerable number of shift workers or professionals with non-traditional work hours.

The residents of this neighborhood also exhibit distinctive family and living arrangements. There is a high prevalence of grandparents living with their grandchildren and a lower rate of males and females living alone compared to the national average. This might indicate a strong emphasis on family values and multi-generational households in the area.

In terms of education, there is a notably higher percentage of residents who have not completed any schooling or have only completed up to the 12th grade without obtaining a diploma. However, there is also a higher rate of attendance in nursery school, suggesting a potential focus on early childhood education in the community. Income distribution in the area is skewed towards the middle and upper-middle range, with a higher proportion of residents earning between $125,000 and $199,999 annually.

Housing in the neighborhood is predominantly older, with a significant majority of homes built between 1940 and 1969. There is a lower rate of renter-occupied housing units, with most residents being homeowners. The area has a low vacancy rate for houses.

Regarding employment, the neighborhood has a higher proportion of men working in transportation and warehousing, utilities, and administrative support and waste management services, while women are more likely to be employed in transportation and warehousing and utilities.

In summary, this neighborhood is a diverse and multicultural community with unique family structures, a focus on early childhood education, and distinct commuting patterns. The area has a mix of older housing stock and a higher proportion of owner-occupied homes.

Cluster Names

- *In the style of a commercial geodemographic classification; create a cluster name that would summarize the following geodemographic pen portrait. The name should capture as many different characteristics contained within the description as possible. The cluster name should be no more than 3 words.*

- *TYPE DESCRIPTION…*

- 39 Type Descriptions
- Used as the prompts for Groups…
- *The following text represent a series of cluster descriptions from a geodemographic classification. Create an aggregate description that summarizes the key characteristics of all the clusters. This should describe the age profile, racial composition and language diversity, types of housing, commuting characteristics and modes of transport, and the types of employment.*
- Total 18k words of description…

Group		Type
A: Commuting Families	A1	Multicultural Commuter Communities
	A2	Diverse Family Commuters
	A3	Diverse Early-Riser Families
	A4	Multicultural Sunrise Laborers
	A5	Senior Suburban Stability
B: Young Professionals	B1	Multicultural Urban Professionals
	B2	Diverse Multilingual Commuters
	B3	Youthful Academic Diversity
	B4	Young Urbanites
	B5	Multicultural Urban Achievers
C: Aging Agricultural Settlers	C1	Elderly Agricultural Enclave
	C2	Aging Rural Haven
	C3	Diverse Elders' Enclave
	C4	Agricultural Employment Hub
	C5	Modest Agricultural Settlers
	C6	Mature Rural Dwellers
	C7	Mature, Rural Homesteaders
D: Prosperous Professionals	D1	Affluent Mature Families
	D2	Diverse Family Professionals
	D3	Diverse Aging Affluents
	D4	Prosperous Multicultural Professionals
	D5	Multilingual Professional Community
E: Culturally Rich Achievers	E1	Affluent Educated Elders
	E2	Diverse Professional Enclave
	E3	Educated Commuting Professionals
	E4	Affluent Elderly Enclave
	E5	Mature Multicultural Settlement
F: Economic Adversity	F1	Lower-Income Agricultural Community
	F2	Young Multilingual Commuters
	F3	Lower-Income Ethnic Melting Pot
	F4	Diverse Cultural Mosaic
	F5	Struggling Service Workers
	F6	Urban Economic Strugglers
G: Urban Melting Pot	G1	Diverse Young Renters
	G2	Diverse Commuter Community
	G3	Low-Income Multilingual Renters
	G4	Economically Challenged Commuters
	G5	Multigenerational Latino Enclaves
	G6	Diverse Low-Income Community

Evaluation

Factual Inaccuracy	Supposition	Repetition	Missing Features	Not Notable Feature	Attributes N
1.05	0.10	0.18	0.92	0.05	24.62

Many Thanks

UNIVERSITY OF LIVERPOOL