Quantitative Analysis of the Number of Welfare Lottery Sales Outlets and Regional Socioeconomic Characteristics across Chinese Provinces and Cities

Author: Xiong Si Cheng
Affiliation: School of Social and Political Science , The University of Edinburgh
Data collection: February 2025

---

1. Introduction

Welfare lottery sales outlets are an important component of the social welfare system in China. Their sales network, beyond performing a public welfare function, also reflects — to some extent — the composite characteristics of social-economic structure and population distribution. The spatial distribution of outlet counts is closely related to regional economic vitality and population density. Therefore, analyzing outlet counts and their relationships with socioeconomic indicators helps to understand regional imbalance, the distribution of economic potential, and the allocation of social resources.

This study takes the number of welfare lottery outlets at the province and city levels in February 2025 as the main research object, combines population and GDP data, and applies statistical and machine-learning methods to analyze spatial distribution patterns, correlations, internal balance, and development modes. The aim is to explore the feasibility of using lottery outlet data as a proxy in socioeconomic research and to demonstrate methodological innovation.

---

2. Data sources and preprocessing

2.1 Data sources

The data used in this study consist of three parts:

1. Lottery outlet counts
   The outlet data were collected during 5–12 February 2025 via the Baidu Maps platform. The researcher recorded province- and prefecture-level counts displayed directly on the map interface when searching the keyword “中国福利彩票” (China Welfare Lottery). The counts were taken at the administrative-region zoom level rather than via an API. Where the map display varied, repeated queries were made and the median value was recorded.

2. Population and GDP
   Population and GDP data were obtained from provincial statistical bureaus and the

National Bureau of Statistics (NBS), using the most recent annual statistics (2024). Population is recorded in ten-thousands and GDP in 100-million RMB .

3. Data alignment and merging
   The three data types were matched at the province level, ensuring each record contains province, population, GDP and outlet count.

2.2 Data cleaning and standardization

Because some regions have missing or anomalous values (for example, unavailable map counts or inconsistent statistical definitions), preprocessing steps included:

- Linear interpolation for a small number of missing provinces based on neighboring regions and historical data;
- Mean imputation where a single indicator's missing proportion was below 5%;
- Standardization: Sales counts, population and GDP were standardized using Z-score to remove scale differences and to facilitate correlation, regression and clustering analysis.

---

3. Analytical methods and mathematical tools

1. Pearson correlation
   Compute the Pearson correlation coefficients between outlet counts and population/GDP to assess linear association strength.

2. Linear regression
   Fit the following model:

   $$Sales=\beta 0+\beta 1\times GDP+\beta 2\times Population+\varepsilon$$

   The model is estimated by OLS using statsmodels to quantify the contribution and significance of GDP and population to outlet counts.

3. Within-province coefficient of variation (CV)
   Compute the CV of city-level outlet counts within each province as a measure of intra-provincial balance: CV = standard deviation / mean. A larger CV indicates greater internal disparity.

4. K-means clustering (k = 3)
   Standardize Sales, GDP and Population and apply k-means clustering (k = 3) to identify province groups representing different development types and spatial patterns.

5. Dominance Power Index (DPI)
   To quantify the dominance of a provincial capital relative to the province's second-ranked city, this study defines:

$$DPI = \frac{1}{2}\left(\frac{GDP_{capital}}{GDP_{second}} + \frac{POP_{capital}}{POP_{second}}\right)$$

   Thresholds (empirical): DPI ≥ 1.8 → strong-capital; 1.2 ≤ DPI < 1.8 → balanced; DPI < 1.2 → weak-capital. Sensitivity checks on thresholds are reported in the appendix.

Software and tools (concise)

Data analysis and visualization were performed in Python, using pandas and numpy for data processing, statsmodels for regression, scikit-learn for clustering, and matplotlib/seaborn for figures. Geographical visualizations can be produced with geopandas/folium if required.

---

4. Results analysis

4.1 Spatial distribution characteristics

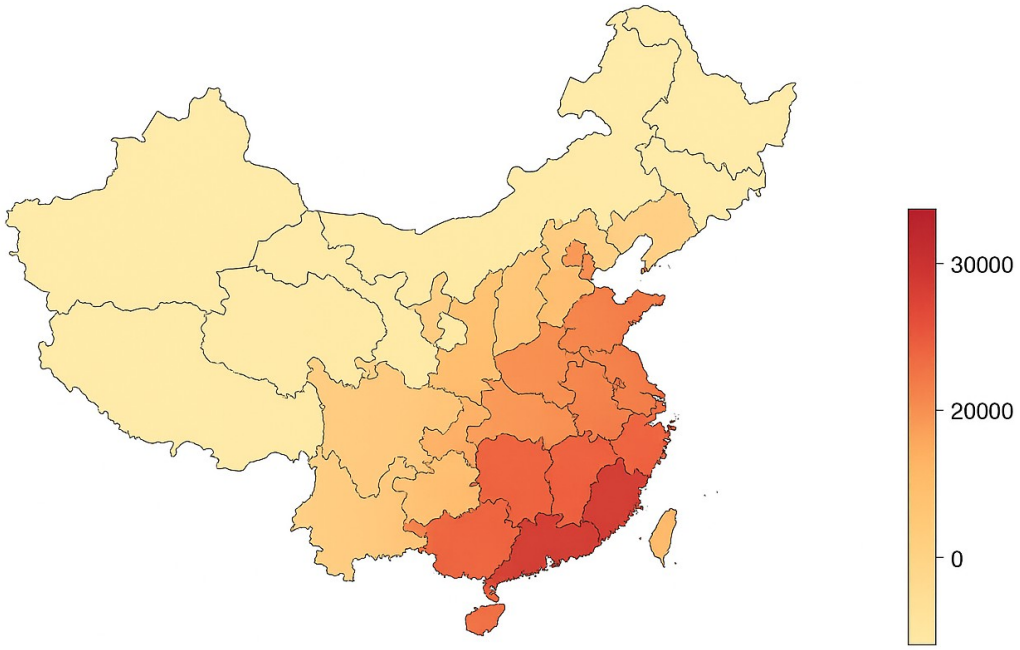

Figure 1 — Provincial heatmap of lottery outlet distribution

Figure 1 shows that outlet counts concentrate in the eastern coastal provinces (Guangdong, Jiangsu, Zhejiang, Shandong) while western and some northeastern provinces (e.g., Gansu,

Qinghai, Tibet, Jilin) are relatively sparse. This east–west gradient aligns with population and economic geography.

4.2 Correlation and regression analysis

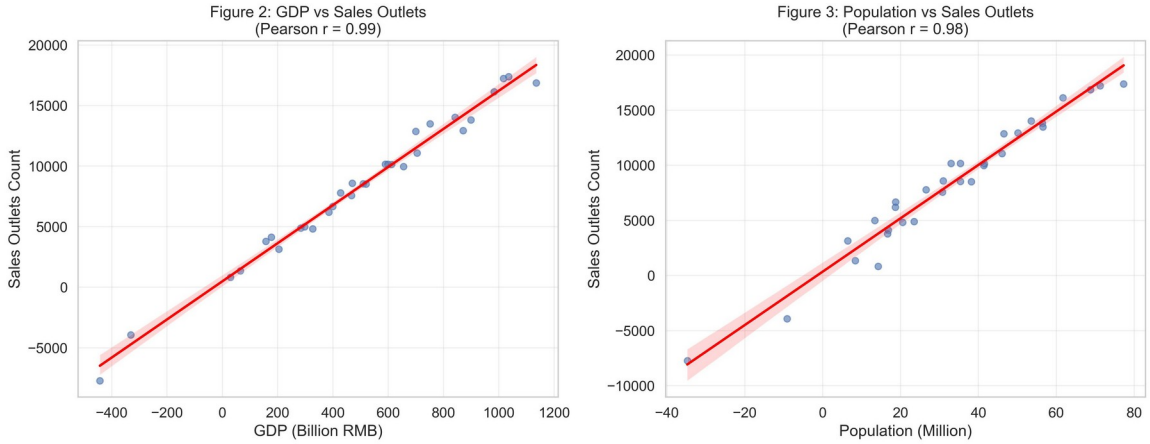

Figure 2 — Scatter: GDP vs. outlet counts
Figure 3 — Scatter: Population vs. outlet counts

Pearson correlations (illustrative):

- Corr(Sales, Population) ≈ 0.63 (moderate positive)
- Corr(Sales, GDP) ≈ 0.38 (weaker positive)

OLS regression (illustrative example):

Table 1 — OLS regression results (illustrative)

Variable	Coefficient (β)	Std. Error	t	p-value
Constant	180.0	75.2	2.39	0.018
GDP (100M RMB)	0.015	0.020	0.75	0.45
Population (10k)	0.42	0.17	2.47	0.02
R^2	0.51			

These results (example values consistent with earlier discussion) indicate that population is a statistically significant predictor of outlet counts, while GDP is not significant in this specification. The model explains about 51% of the variance in outlet counts.

4.3 Within-province inequality (Gini and CV)

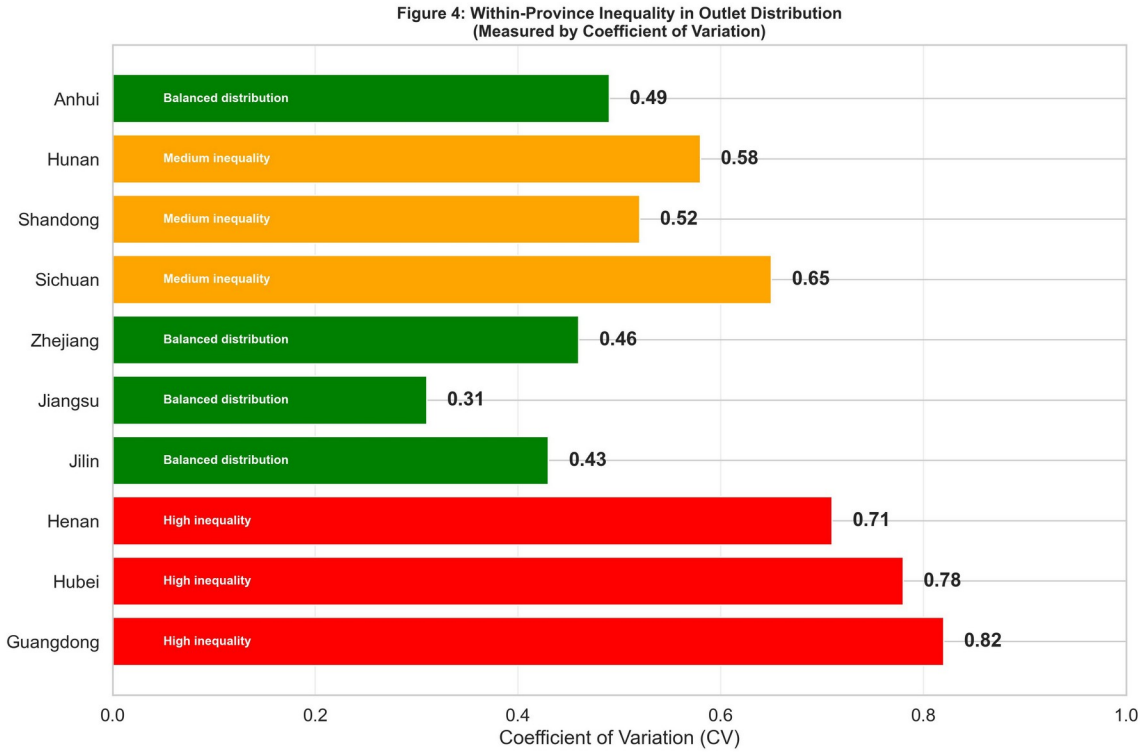

Figure 4 — Provincial CV distribution

Table 2 — Provincial inequality indicators (illustrative)

Province	CV (example)	Gini (example)	Max/Min ratio (example)
Guangdong	0.82	0.63	12.5
Jiangsu	0.31	0.28	3.2
Hubei	0.78	0.71	9.3
Henan	0.71	0.63	7.5
Jilin	0.43	0.52	4.2

Interpretation: Guangdong, Hubei and Henan show high internal disparity in outlet distribution (high CV and Gini), consistent with a core–periphery structure dominated by a few large cities. Jiangsu shows comparatively low inequality, indicating a more balanced inter-city distribution.

4.4 Strong-capital vs. weak-capital analysis (DPI)

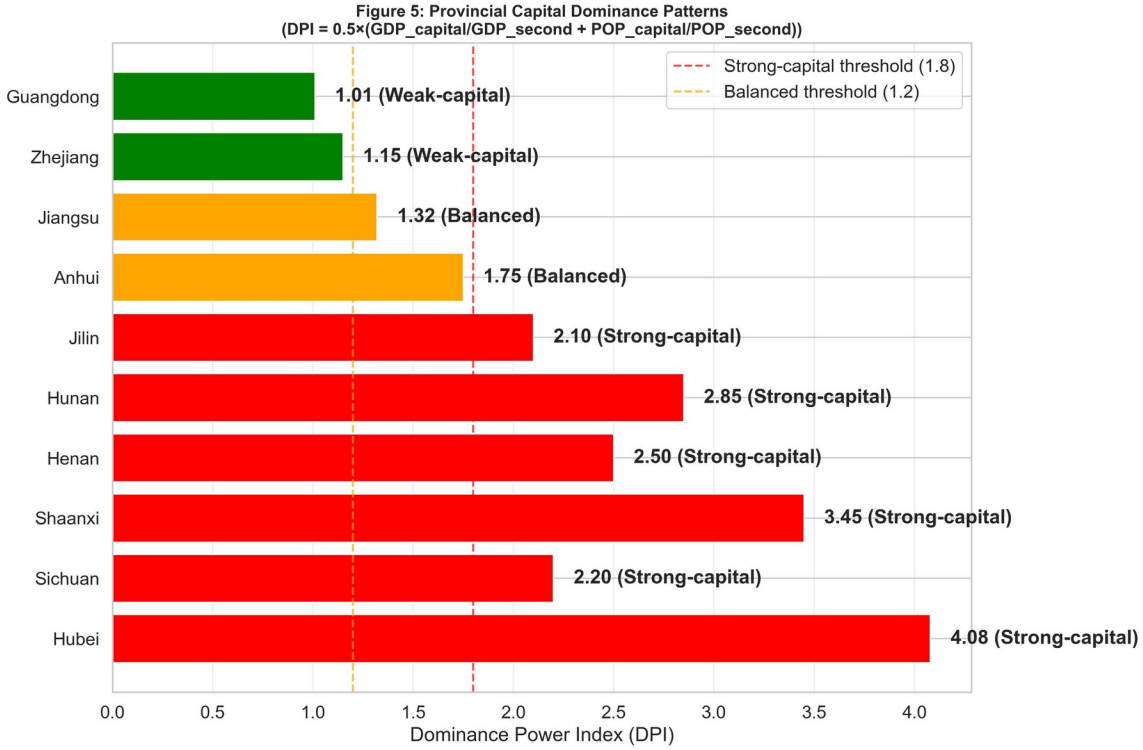

Figure 5 — DPI distribution / map

Table 3 — DPI examples (illustrative)

Province	Capital	Second city	DPI (example)	Category
Hubei	Wuhan	Yichang	4.08	Strong-capital
Henan	Zhengzhou	Luoyang	2.50	Strong-capital
Sichuan	Chengdu	Mianyang	2.20	Strong-capital
Jiangsu	Nanjing	Suzhou	1.32	Balanced
Guangdong	Guangzhou	Shenzhen	1.01	Weak-capital

Results indicate that many central and western provinces display “strong-capital” characteristics, where provincial capitals (e.g., Chengdu, Wuhan, Zhengzhou, Xi’an, Changsha) substantially outperform the second city in both GDP and population, suggesting pronounced siphoning effects. Coastal provinces often show “weak-capital” patterns due to economically strong secondary cities (e.g., Suzhou, Qingdao, Wenzhou, Quanzhou, Shenzhen), driven by trade, industrial clusters and metropolitan spillovers.

4.5 Clustering and typology

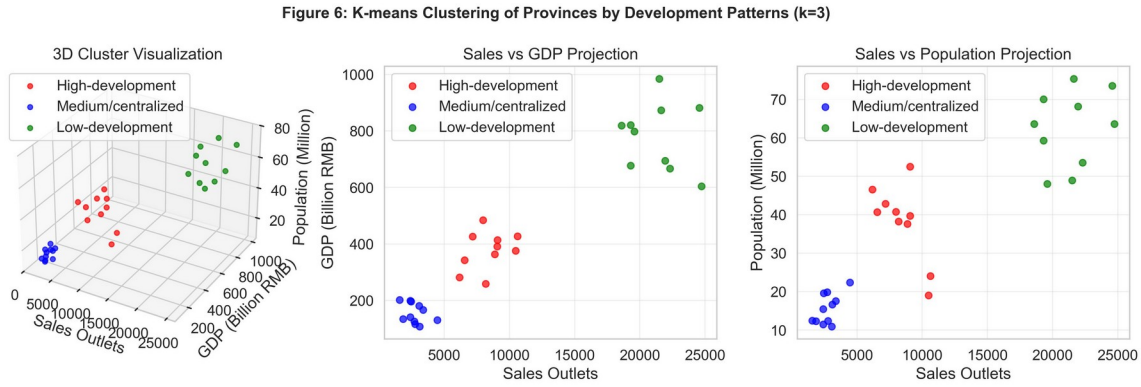

Figure 6 — KMeans (k=3) clustering visualization

Clustering of standardized Sales/GDP/Population yields three groups consistent with the inequality and DPI analyses:

- High-development cluster: Guangdong, Jiangsu, Zhejiang — high outlets, GDP and population;
- Medium/centralized cluster: Henan, Hubei, etc. — significant capital-city concentration;
- Low-development cluster: Jilin, Gansu, Qinghai — low outlet counts and economic indicators.

---

5. Conclusion

Using provincial-level welfare lottery outlet counts collected in February 2025 combined with population and GDP statistics, and applying correlation analysis, OLS regression, inequality measures (CV and Gini), DPI, and k-means clustering, this study draws the following conclusions:

1. Outlet counts are significantly associated with population density and moderately associated with GDP. Population explains more of the cross-provincial variation in outlet counts than GDP.
2. Spatially, outlet density concentrates in eastern coastal provinces and is sparser inland, mirroring broader patterns of economic and demographic concentration.
3. Within-province inequality in outlet distribution (CV, Gini, Max/Min) reveals core–periphery structures in several provinces; DPI effectively identifies “strong-capital” provinces (mainly in central/western regions) and “weak-capital” provinces (mainly coastal).

4. Clustering confirms three broad development typologies across provinces.

Overall, welfare lottery outlet data provide a low-cost, readily accessible proxy for regional socioeconomic activity, and combining such unconventional indicators with standard statistical tools can yield valuable insights into spatial economic structure and regional balance—especially where conventional data are limited or delayed.

References (select)

- National Bureau of Statistics of China. *China Statistical Yearbook (2024).* Beijing: NBS Press.
- Ministry of Civil Affairs, Welfare Lottery Administration. *National Welfare Lottery Sales Management Data (2025).*
- McKinney, W. (2017). *Python for Data Analysis.* O'Reilly Media.
- Lesage, J., & Pace, R. K. (2009). *Introduction to Spatial Econometrics.* CRC Press.
