Quality of Life During the Chinese Cultural Revolution

An Exploratory Visual Analysis of the World Development Indicators Dataset

November 2nd, 2025

Henry Shi

Henry Shi November 2nd, 2025

Introduction

In the following project, I perform an exploratory visual analysis of the World Development Indicators using Python for data processing, analysis, and most of the visualizations. The final time series graphs were created using Tableau Desktop. This writeup will profile the data, detail the iterative process of finding and testing my hypothesis through visual analysis, and end with a reflection on the process.

Data Profile

The World Development Indicators (WDI) is a dataset compiled by The World Bank consisting of numerous "internationally comparable statistics" that serve as indicators for varying country's development and well-being. The dataset was downloaded from The World Bank website as a zip file containing multiple csv's. The primary dataset contains the time series values of each indicator for each country or region and is nearly 200MB. Meanwhile, the other csv files are much smaller in size and explain details such as what each indicator code corresponds to.

For this project, I focused on the primary csv file, which had 403,256 rows and 69 columns. Four of the columns corresponded to the categorical variables of "Country Name," "Country Code," "Indicator Name," and "Indicator Code," with the remaining 65 columns each dedicated to a year from 1960-2024. Each row represented the quantitative time series data for each indicator for each country or region. There are 266 unique country names, with some of those referring to a region rather than a specific country, and 1516 unique indicators for economy, politics, society, and more.

Aside from the staggering number of null values taking up much of the dataset, the WDI dataset can be described as rich, though overwhelmingly so, and hence arguably impractical for many users. This is mainly because of the incredible number of indicators, which provide both variety and confusion at the same time. There is no feasible way to assess data quality across the board, and it is also near impossible for someone to choose indicators if they do not have strong pre-conceived ideas and domain knowledge in the subject (e.g. understanding of economics when attempting to narrow down which economic indicators should be included in an analysis). Even then, there may be too many null values in the desired variables. In other words, the dataset offers lots of potential but can be clunky and requires lots of refining for exploratory analysis.

Question Exploration

FINAL HYPOTHESIS: I predict that key indicators for quality-of-life in China will worsen and conflict with the overall trend from 1960-2024 during the Cultural Revolution of 1966-1976.

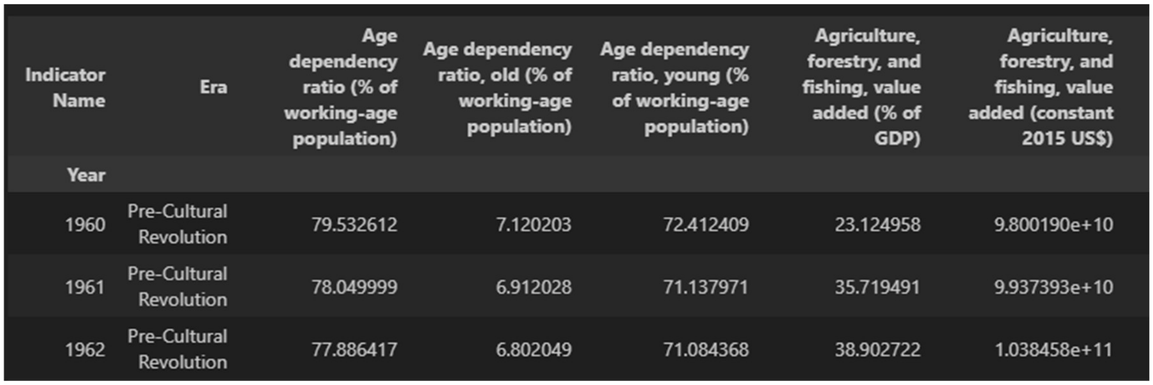
Initial question:
My first idea was inspired by a recent trip to Iceland, where I was shocked by prices even in non-touristy places. I soon realized that Iceland is a lot wealthier than I had expected for such a small island nation. My initial intent was to explore the factors that led to Iceland's prosperity. However, this idea was immediately discarded upon downloading the WDI data and realizing the earliest year was 1960 (well after WWII, a pivotal part of Iceland's story).

Formulating a new question:
The next idea was, broadly speaking, to explore the impact of China's Cultural Revolution (CR) on development indicators. Does the data agree with the sentiments I hear from my elders who lived through it, specifically that life was "miserable"? I had my reservations regarding the quality or accuracy of data from China, especially during the relevant period, but nonetheless hoped to see a meaningful pattern.

I started by filtering the dataset to include only China, which allowed me to drop the "Country Name" and "Country Code" columns, as well as "Indicator Code" (redundant with "Indicator Name"). I then transposed the dataframe to follow the convention of having variables (indicators) as columns and observations (year) as rows. I set the year as an index and added an "Era" column denoting if the year corresponded to before, during, or after the CR (Table 1).

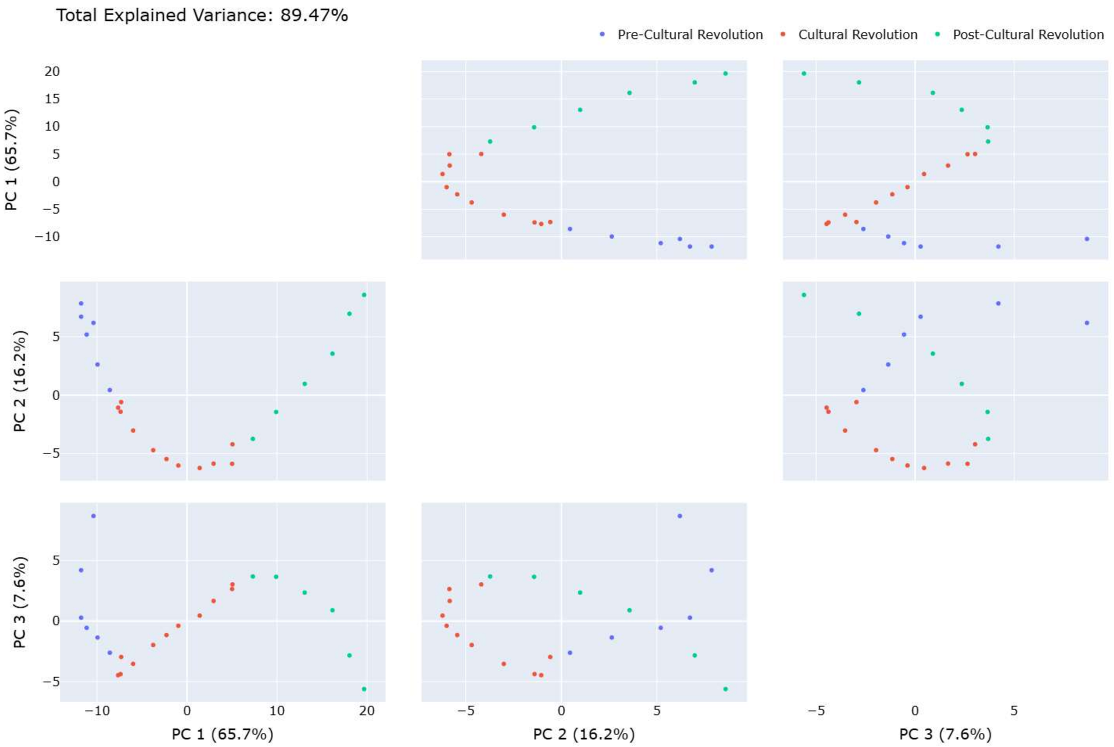
PCA:
China's data contained many nulls, especially before the year 2000. I decided to remove any indicators that contained nulls to allow for PCA, drastically shortening the list of indicators from 1516 to 111. My hope was to use PCA to determine influential variables that I could then plot in a classic time series graph to analyze the CR's impact.

Table 1: First 3 rows of cleaned data for China (only showing the first 5 of 111 indicators).



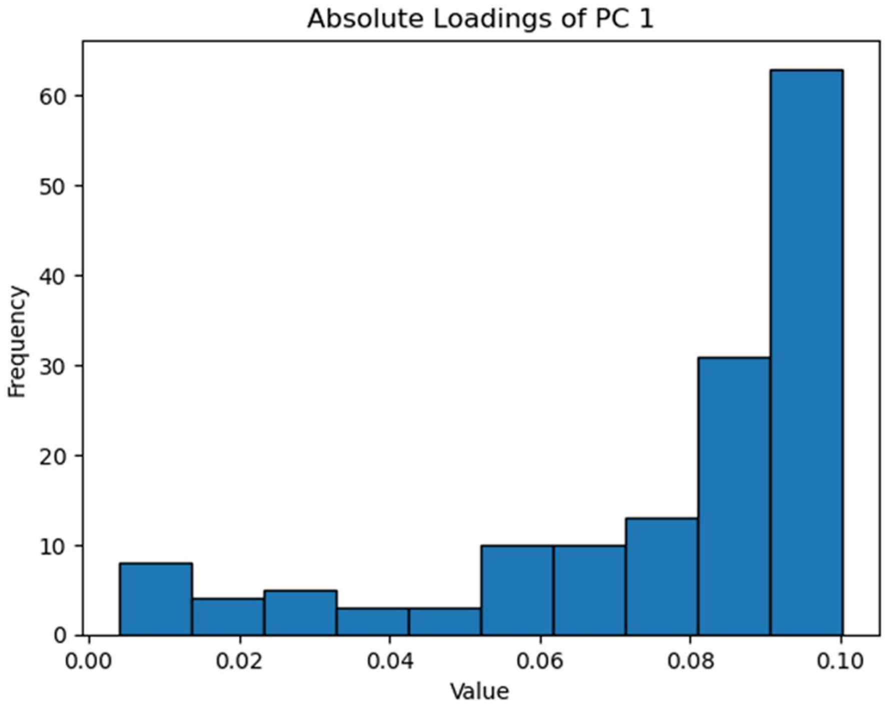
Indicator Name	Era	Age dependency ratio (% of working-age population)	Age dependency ratio, old (% of working-age population)	Age dependency ratio, young (% of working-age population)	Agriculture, forestry, and fishing, value added (% of GDP)	Agriculture, forestry, and fishing, value added (constant 2015 US$)
Year						
1960	Pre-Cultural Revolution	79.532612	7.120203	72.412409	23.124958	9.800190e+10
1961	Pre-Cultural Revolution	78.049999	6.912028	71.137971	35.719491	9.937393e+10
1962	Pre-Cultural Revolution	77.886417	6.802049	71.084368	38.902722	1.038458e+11

However, I felt that this was too harsh and hence narrowed the data to be between 1960 to 1982 (to include 5 years of data before and after the CR). Dropping indicators with null values for this narrowed dataset resulted in 150 indicators, 39 more than when looking at 1960-2024. Most importantly, this new list included commonly cited indicators such as life expectancy.



Graph 1: Scatterplot matrix for first 3 principal components, accounting for 89.47% of explained variance.

After normalizing all numerical variables (i.e. not era), I conducted PCA. Three principal components account for nearly 90% of total explained variance. While there was clear clustering based on the era (pre, during, and post CR), none of the clusters were noticeably separated (Graph 1). It was difficult to interpret anything about the CR based on the above PCA plots alone. It was also impractical to plot all the loadings (original indicators) onto the graphs due to sheer number of them. I thought to only plot the most significant ones (and select those for a time series analysis later) but realized quickly that there was no meaningful cutoff to how many given the distribution of absolute loadings. Specifically, there were no notable variables that were particularly influential to any of the principal components (Graph 2). Ultimately, I moved on from PCA to find other methods of selecting indicators for “quality-of-life.”



Graph 2: Histogram of absolute loadings for PC 1, with more than half of them being between 0.08-0.1.

Manual selection of indicators:

I looked to regain some of the data lost when I removed indicators with missing values, given that outside of PCA, it was potentially acceptable to keep some nulls. Of course, it would be meaningless to include indicators consisting of primarily missing values, particularly during the period of study. Hence, I dropped indicators again, this time based on if they included nulls for more than 1/3 of the years from 1960-1982. Out of the resulting 183 indicators, I spent quite some time deciding how to choose the most meaningful ones for a time series graph. There were two main issues:

1. There is no singular “list" of indicators for the quality-of-life for a country. While many resources online may point to broad indicators (e.g. wealth, life expectancy, education, unemployment, access to electricity, air and water quality, etc.), the WDI dataset contains too many detailed indicators that did not necessarily match with the broad ones.
2. It became clear that researching relevant indicators would be time-consuming and out of scope for this project, with a very high chance of finding them missing in the 183 anyways (e.g. Gini index and education-related indices).

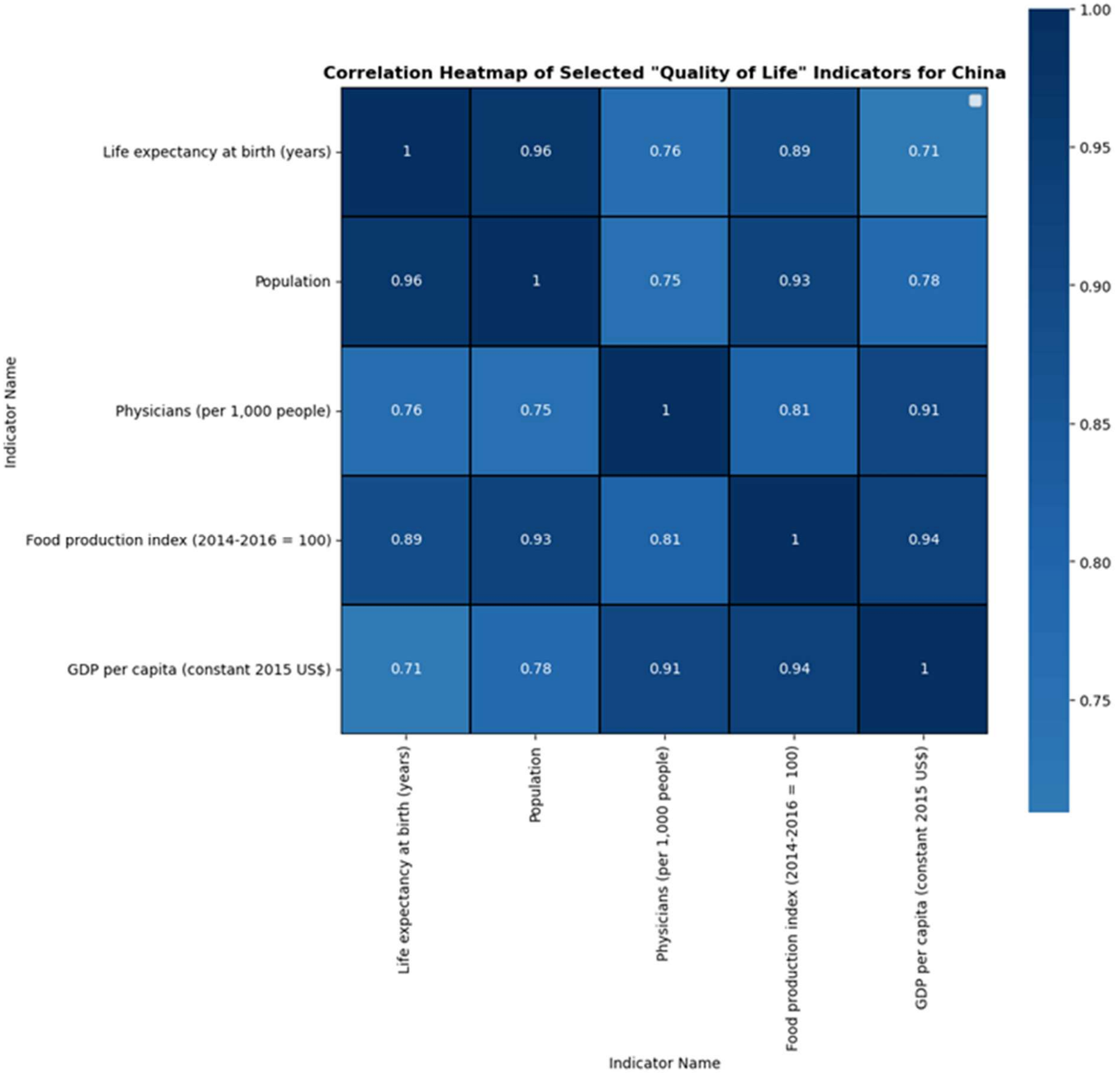
In the end, I opted for a more straightforward approach: manual selection. While it certainly felt like an inelegant and non-scalable method, it saved me time from the second point above. And to the first point, there will always be subjectivity in what constitutes "the best indicators" for quality-of-life. My belief is that I followed the most efficient method given the circumstances, and picked based on intuition or how they matched with online resources regarding quality-of-life indicators[1-3].

Narrowing down the question to a hypothesis:

As I neared an answer to my broad question of what quality-of-life in China during the CR was like, I decided to change it to a more specific hypothesis. This became the final version: I predict that key indicators for quality-of-life in China will worsen and conflict with the overall trend from 1960-2024 during the Cultural Revolution. I expected to see the data agree with the near unanimous anecdotal evidence I've heard from elders that life was overwhelmingly difficult.

Testing my hypothesis:

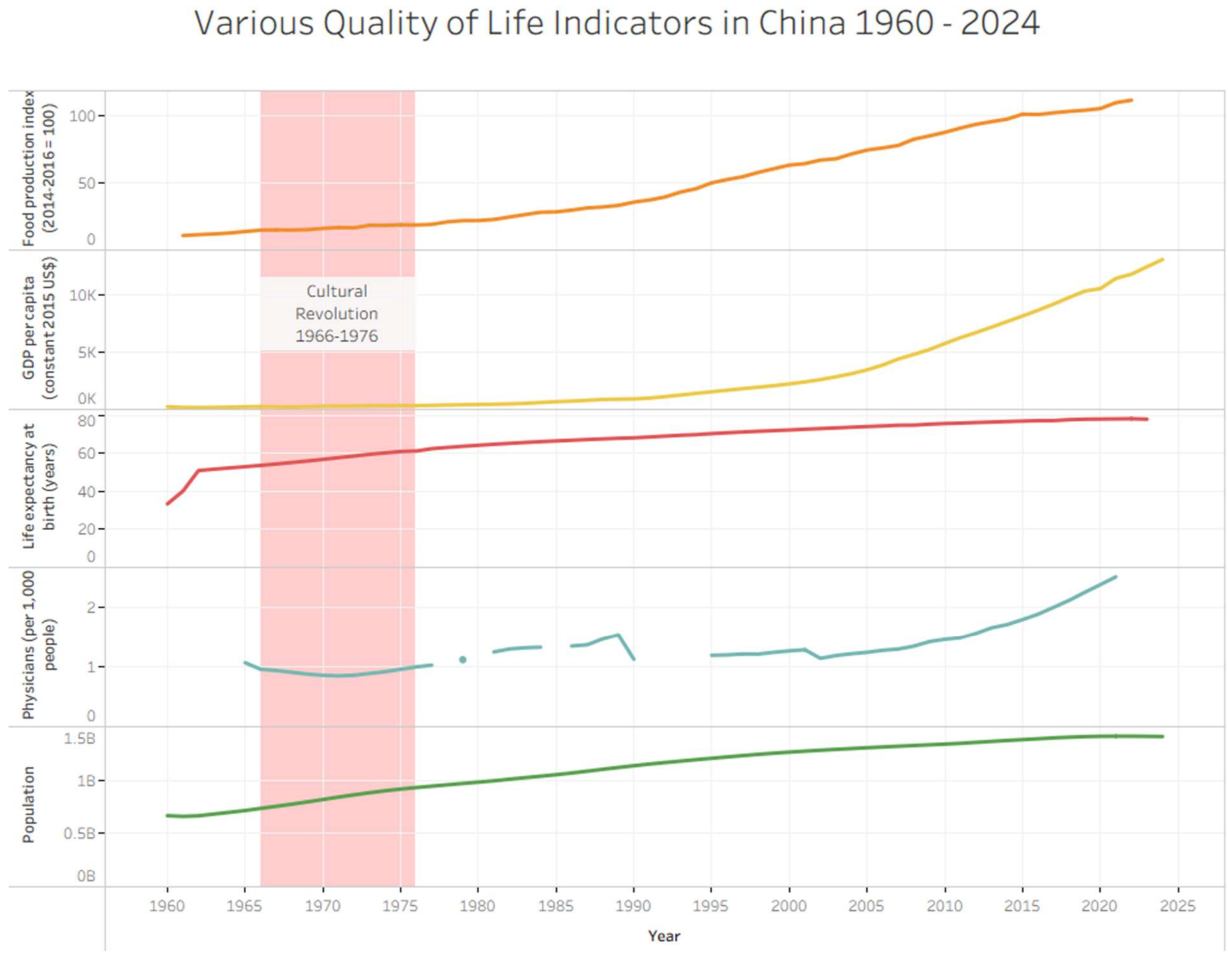
Five indicators were chosen in the end to represent quality-of-life. The correlation heatmap for them was generally very positive (ranging from 0.71-0.96), suggesting consistent growth for all variables and potentially conflicting with my hypothesis (Graph 3). However, I was hoping that the CR would see contradictions to the overall trend and be the cause for some of the correlation values being closer to 0.7 rather than 0.9.



	Life expectancy at birth (years)	Population	Physicians (per 1,000 people)	Food production index (2014-2016 = 100)	GDP per capita (constant 2015 US$)
Life expectancy at birth (years)	1	0.96	0.76	0.89	0.71
Population	0.96	1	0.75	0.93	0.78
Physicians (per 1,000 people)	0.76	0.75	1	0.81	0.91
Food production index (2014-2016 = 100)	0.89	0.93	0.81	1	0.94
GDP per capita (constant 2015 US$)	0.71	0.78	0.91	0.94	1

Graph 3: Correlation heatmap of manually selected quality-of-life indicators for China (1960-2024), all being positive, suggesting overall growth for all indicators.

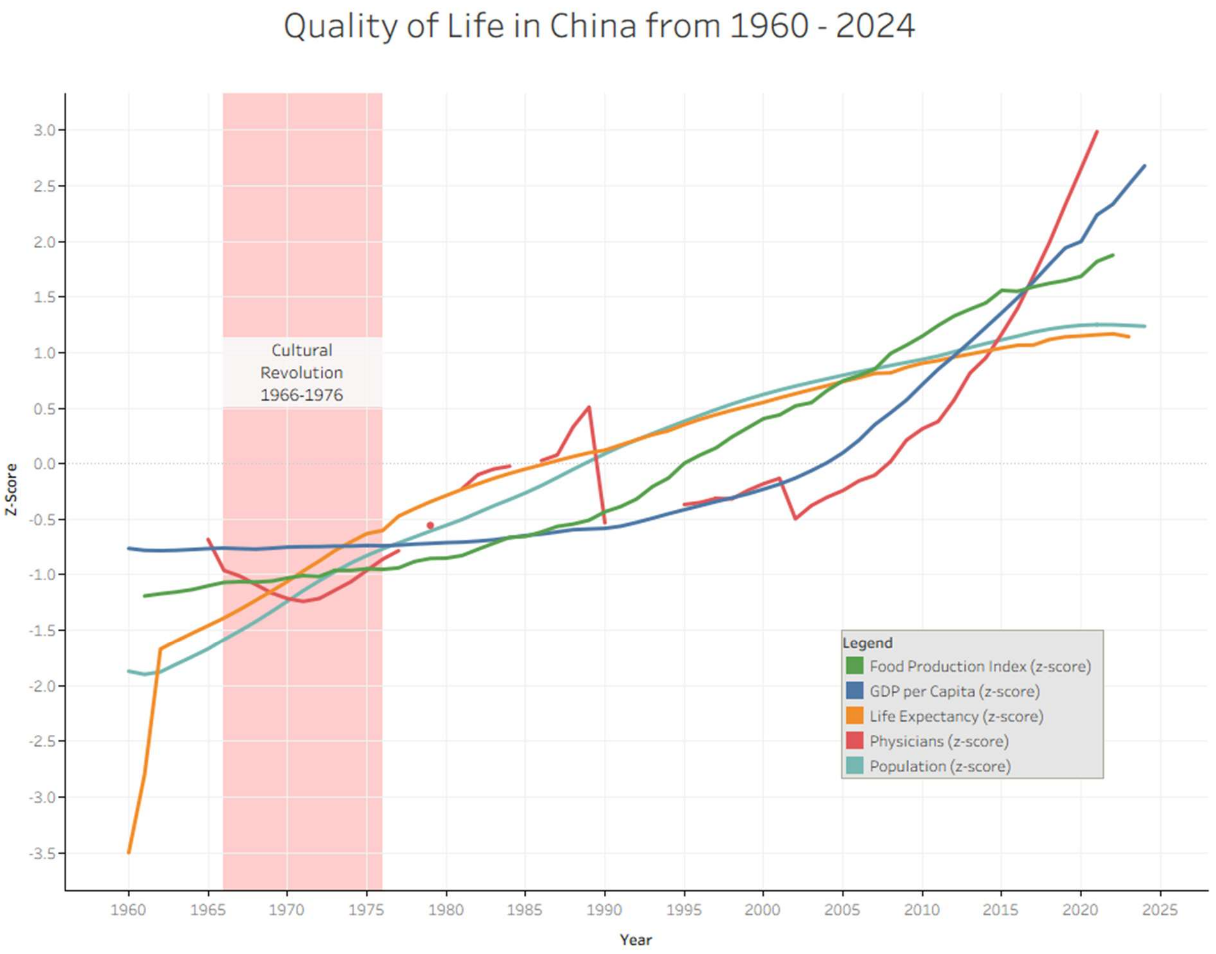
Plotting each indicator onto a time series graph, there appeared to be no noticeable contradiction to the overall trend during the CR years (Graph 4). Given that there are no distinctly differing regimes in the time series, and because specific values don't matter (quality-of-life is defined by the overall pattern amongst these indicators, rather than if certain values are reached), I then normalized all values to z-scores.



Graph 4: Time series graphs of quality-of-life indicators for China from 1960-2024 showing no apparent significant contradiction to the overall trend during the CR.

Final visualization and rejection of hypothesis:

The final visualization features all indicators (encoded by color) plotted together onto one multivariate graph with a shared y-axis representing z-score (Graph 5). All indicators show consistent growth, and there is no significant deviation to the overall trend during the CR, except for the number of physicians lowering. Understanding why the number of physicians decreased may be a whole other non-data science project in itself, though I found that this interestingly coincides with my father's accounts of how good physicians were in shortage (as part of the mass persecution of the rich and educated), and how he was given the wrong antibiotic upon birth, resulting in his permanently discolored teeth. In any case, ultimately, aside from physicians, all indicators in the final visualization point to a rejection of my hypothesis.



Graph 5: Time series graph of normalized quality-of-life indicators for China from 1960-2024 showing a consistent positive trend even during the CR, except for number of physicians dipping during it.

Discussion and Conclusion

In this exploratory analysis of the World Development Indicators, I set out to better understand if data agreed with my iteratively created final hypothesis that the CR saw a marked decrease in the quality-of-life for Chinese citizens. The selection of indicators for "quality-of-life" was made difficult by the large amounts of null values, and the sheer specificity of most of the provided indicators. In the end, five indicators were selected and plotted on a multivariate time series graph. Except for the number of physicians, the CR did not appear to create any deviations in the growth trends of the indicators, much less a strong decrease in any of them.

As mentioned at the beginning, I do have some doubt towards the accuracy of the data for China, especially in the 1960s and 70s. For the scope of this project though, my hypothesis has been rejected, suggesting that the CR was not as disastrous as people make it out to be. It reminds me a little of traditional research, where you often don't get expected results.

I am very content with my choice of tools. Python was an excellent tool for both data manipulation and creating visuals. It provided the much-needed flexibility for an exploratory analysis where I wasn't even sure what to look for when I downloaded the dataset. On the other hand, Tableau was helpful in creating a visually appealing final graph faster than customizing one in Python.

I think this project was a great example of the iterative nature of exploratory visual analysis. What questions can be asked depends heavily on the dataset, and there will be constant obstacles to how you may have initially planned to approach answering a question. As such, both questions and methods must be refined before reaching a final visualization or answer.

References

1. Henderson, H., Lickerman, J., Flynn, P., & Calvert Group. (2000). *Calvert-Henderson quality of life indicators*. Calvert Group.

2. Eurostat. (n.d.). *Quality of life indicators – measuring quality of life*. Retrieved [October 31, 2025], from https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Quality_of_life_indicators_-_measuring_quality_of_life

3. Guliyeva, A. (2022). Measuring quality of life: A system of indicators. *Economic and Political Studies, 10*(4), 476–491. https://doi.org/10.1080/20954816.2021.1996939