Global Cost of Living & Affordability

Our EDA

Authors: Ksenia and Shamma
Course: Data Bootcamp, Section 002
Date: October 23, 2025

We built a small, reproducible project to compare cost of living across countries using the Big Mac Index plus World Bank data. The goal wasn't to "solve" economics, it was to show a clean data pipeline, make readable charts, and explain what they mean. We worked together on everything; I focused more on the data pulling/cleanup and modeling, and my partner fixed the nasty environment errors and helped design and sanity-check the figures.

What We Wanted to Know

1. Which countries look most/least expensive if you just compare the Big Mac price in USD?
2. How affordable is that burger once you scale by income (PPP)? A $5 burger means different things in different places.
3. What explains the differences? Is it mainly income, or do inflation, unemployment, or urbanization matter too?
4. How do countries compare to the U.S. on the same date (so the ratio for the U.S. itself is exactly 1.00)?

Data We Used (public + reproducible)

- The Economist , Big Mac Index (GitHub data). This is a long-running dataset of Big Mac prices in local currency, exchange rates, and USD conversions, by country and date. We use it as a simple, comparable price proxy.
    - Repo: The Economist, *big-mac-data* (public GitHub).
- World Bank , World Development Indicators (WDI) via the wbgapi Python package:
    - GDP per capita, PPP (current international $)
    - Inflation, CPI (annual %)
    - Unemployment (% of labor force)
    - Urban population (% of total)
    - Population (total)

- Gini index (inequality)
- Plus World Bank region and income group metadata.

We assembled these into one dataset that easily clears the assignment's minimum size (12+ features; 300+ rows).

How We Built the Dataset

1. Pull Big Mac data directly from The Economist's GitHub CSV.
2. Standardize keys: country ISO3 codes, date, and year.
3. Make sure 'usd_price' exists: if the file already has it, great; if not, compute it from local_price/exchange rate.
4. Drop impossible outliers: remove any row where usd_price > 40 (we printed how many rows this removed so it's auditable).
5. Pull WDI indicators across the Big Mac years and merge on (iso3, year). If an indicator misses a year for a country, we forward/back fill within that country just so we can draw charts without losing entire countries.
6. Engineer a few features:
   - daily_gdp_ppp_pc = gdp_ppp_pc / 365
   - Affordability ratio: bigmac_afford_ratio = usd_price / daily_gdp_ppp_pc
   - Logs: log_usd_price, log_gdp_ppp_pc
   - Relative to the U.S. at one snapshot: pick a single dense date that includes the U.S.; compute rel_to_us = usd_price_i / usd_price_US. We hard-set the U.S. to 1.00.

Everything runs end-to-end in the notebook. The merged dataset is saved as data/processed_cost_of_living.csv. Figures are saved as both PNG and HTML (interactive).

What We Looked At

- Figure 1. *Top 15 Most Expensive Big Mac Prices (Latest Snapshot)*
  Horizontal bars, labeled in USD.
- Figure 2. *Top 15 Least Expensive Big Mac Prices (Latest Snapshot)*
  Same format as Figure 1.
- Figure 3. *Affordability Ratio by World Bank Income Group (Latest Snapshot)*
  Violin + box plot of usd_price / (GDP PPP per capita / 365) grouped by income category.
- Figure 4. *GDP per capita (PPP) vs Big Mac Price (Latest Snapshot)*
  Scatter with OLS trendline (we kept it simple: log price on log income with a few controls).

- Figure 5. *Big Mac Price Relative to the United States (Snapshot)*
  A choropleth map where the value is country price divided by U.S. price, computed on the same date so comparisons are fair and the U.S. shows 1.00.

All titles, axes, and legends are labeled. For PNGs, we used Plotly+Kaleido; the HTML files are interactive and nice for hover tooltips.

Results (what stood out to us)

1) Nominal prices vary, but not wildly

Looking at Figures 1 and 2, you do see a spread: some countries are clearly pricier in USD, others cheaper. But it's not a "10x" spread; it's a few multiples. That's already interesting because inflation headlines can make it feel like prices should be all over the place.

2) Affordability gaps are bigger than price gaps

Figure 3 (the violin by income group) was our favorite. Even when the sticker prices look similar, affordability diverges a lot because incomes diverge a lot. In lower-income countries, the affordability ratio can get close to 1,meaning a burger can cost around a day of average per-capita income. In high-income countries, the ratio is much smaller, and the spread narrows: the burger is just not a big expense for the average person. This is a simple but powerful reminder: talking only about prices hides the real story.

3) Income explains a lot of the price differences

Figure 4 plots price against GDP per capita (PPP). The cloud has a clear upward slope; richer economies tend to have higher prices for this product. That lines up with the standard Balassa–Samuelson idea: non-tradable parts of a good (rent, local labor, services) get more expensive as productivity and wages rise. We ran a small OLS with robust errors: log price on log income, plus inflation, unemployment, and urbanization. Income stays positive and significant. The other variables matter less in the cross-section once income is in the model.

4) Relative-to-U.S. map works only if you anchor it right

For Figure 5, we chose a single snapshot date with lots of countries and the U.S. present. Then we divided each country's USD price by the U.S. price on that same date, and we set the U.S. exactly to 1.00. This solves the common bug where the U.S. shows something like "0.77" because the denominator price came from a different day. We also binned the legend carefully to avoid "NaN"

ranges at the top edge. The map makes it easy to eyeball clusters of countries that are relatively more or less expensive than the U.S. for this specific item.

What This Does Not Mean

- The Big Mac is one product, not a whole consumer basket. It's a useful proxy because it bakes in things like wages and rent, but it's still just one item.
- Our income measure is GDP per capita (PPP),that's not the same as median wages. Affordability for a typical worker might look different.
- Some of the World Bank indicators are thin or laggy for certain countries. We used in-country forward/back fills so the charts wouldn't drop countries entirely, but that's still an approximation.
- Taxes, subsidies, and lots of local policy details also play a role (e.g., VAT, food standards, franchise rules). We don't model those.
- Exchange rates move, PPP updates happen, and prices can be sticky. Cross-sectional snapshots are clean for comparisons, but time-series stories need more care.

What We'd Do Next (if we had more time)

- Add a rent proxy (or a small menu of items) for a broader "cost of living" view beyond a single burger.
- Replace GDP per capita with wage data (even for a subset) to see affordability for the median worker.
- Try a panel model with country fixed effects to separate time variation (inflation, FX) from long-run differences.
- Cluster countries on the standardized features to see if distinct "price-income" groups emerge (could be a nice extra figure).

How to Reproduce Everything

1. Create a virtual environment and install the requirements (we included a requirements.txt).
2. Open the notebook cost_of_living_eda.ipynb and Run All.
3. The notebook pulls data from the web, merges it, engineers features, and saves:
    - Processed CSV: data/processed_cost_of_living.csv
    - Figures (PNG + HTML): in the figures/ folder
    - OLS text report: data/ols_summary.txt

4. We added checks so you don't accidentally end up with empty charts. Also, we assert that the final dataset has ≥ 300 rows and ≥ 12 features.

What Each of Us Did

- Shamma: I set up the data pull and merge, added the affordability metric, and wrote the main notebook scaffolding plus helper functions (e.g., figure saving and snapshot logic). I handled basic cleaning (ISO3/date standardization), removed outliers (usd_price > 40), and ran the sanity checks for row/feature counts so the dataset is reproducible. I also put together the simple OLS section and drafted the first pass of the write-up, keeping filenames/labels consistent so the figures and README link up cleanly.
- Ksenia: I fixed the environment/package conflicts, cleaned up plotting defaults, made the maplook good (legends, titles, bins), and double-checked that the U.S. ratio hit 1.00 on the map. I cleaned up some of the code so that it was able to be run. I also worked on the scale of the map so that it was able to be different colors as prior it wasn't able to be divided by color due to scale. We both iterated on the figures and interpretation.

References

1. The Economist , *Big Mac Index* (source and raw data via GitHub). https://github.com/TheEconomist/big-mac-data
2. World Bank , *World Development Indicators (WDI)*. Accessed via wbgapi.
   WDI portal: https://databank.worldbank.org/source/world-development-indicators
   wbgapi docs: https://github.com/worldbank/wbgapi
3. Plotly , *Interactive Python charts*. https://plotly.com/python/
4. Kaleido , *Static image export for Plotly*. https://github.com/plotly/Kaleido
5. Balassa, B. (1964). *The Purchasing-Power Parity Doctrine: A Reappraisal*. *Journal of Political Economy*, 72(6), 584–596.
   Samuelson, P. A. (1964). *Theoretical Notes on Trade Problems*. *Review of Economics and Statistics*, 46(2), 145–154.

Conclusion

We combined the Big Mac Index and World Bank indicators to study global cost of living in a way that's easy to reproduce and easy to read. Prices are higher in richer countries (no shock there), but the

affordability differences are the real eye-opener: a similar USD price can be a tiny purchase in one place and a big bite of daily income in another. The map relative to the U.S. works cleanly once you compute it on one date and pin the U.S. to 1.00. Everything is labeled, saved, and scripted, and we kept the tone human because this is a class project, not a journal article. We hope it's useful and easy to re-run