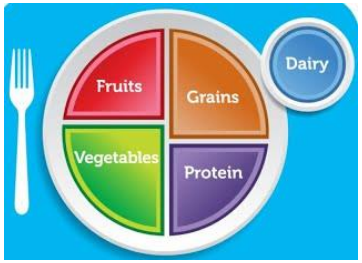
Cost of Living Estimator via Web Scraping

Raj Tiwari
February 14, 2018



Goal: Benchmark estimated average monthly costs across U.S. cities.

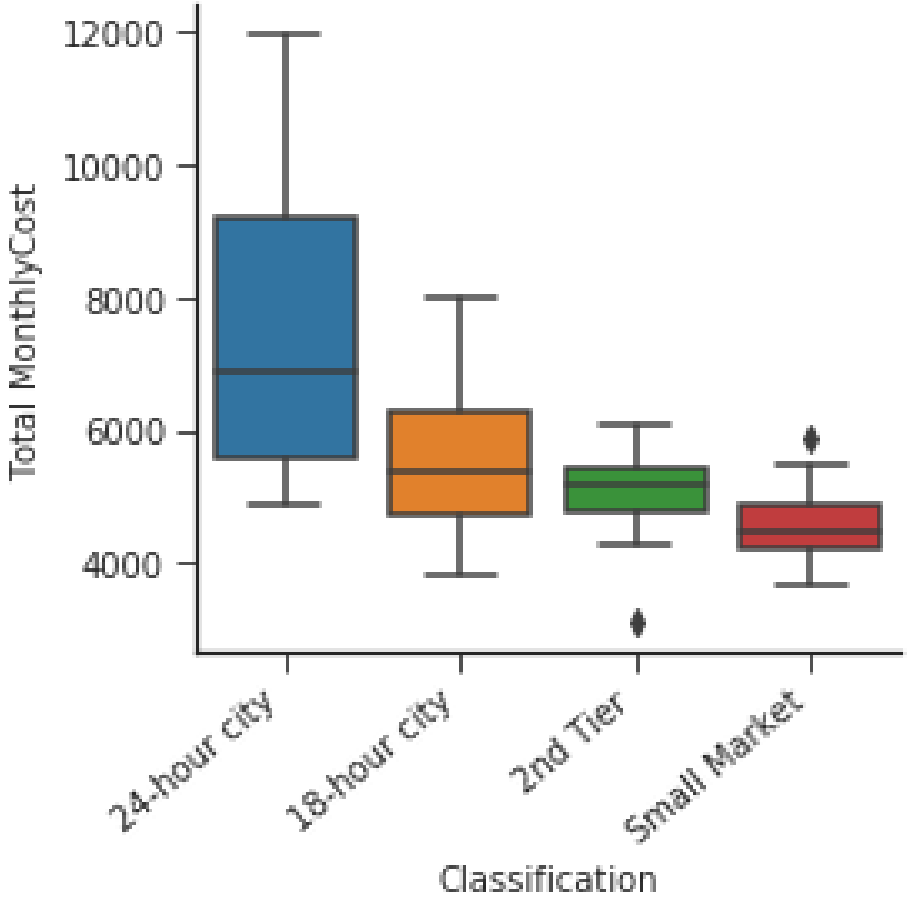
Fields scrapped across 55 U.S. Cities:

- Food Prices (Fruits, Vegetables, Grains, etc.)
- Clothing & Meal Costs
- Household Expenses (Utilities, Internet)
- Apartment Rental Costs
- Childcare and Fitness Club Costs

Cost of Living Analysis Plan

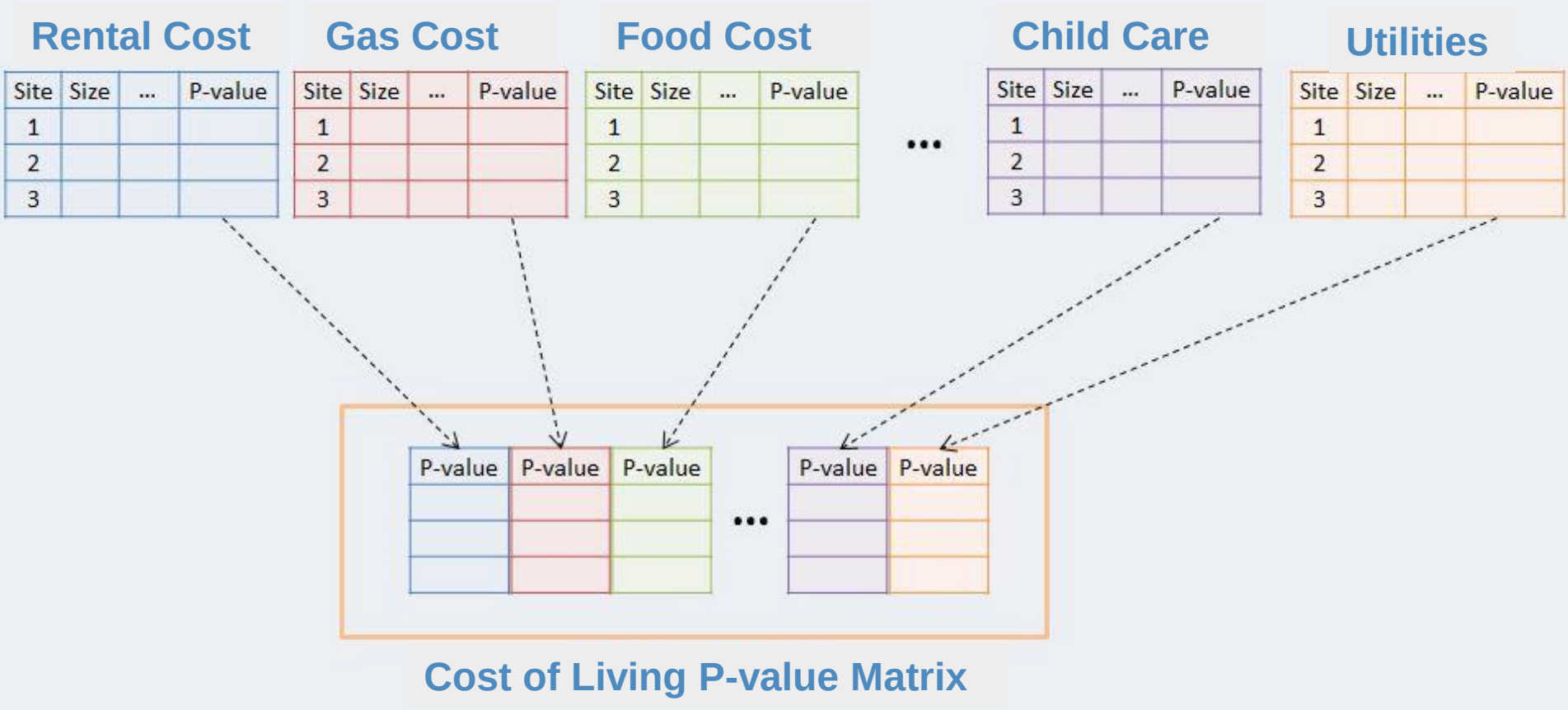
Build Web Scrapping Algorithm	Data Wrangling	Compute P-values and Apply Fisher's Method	Build Cost of Living Generator
Leverage Scrapy crawler to extract cost of living measures across all major U.S. cities.	Clean scrapped data for analysis purposes. Parse text, strip data fields, append additional indicators.	Compute the P-value with respect to all 28 variables. Employ Fisher's Method to measure dispersion for all cost of living indicators.	Leverage R-Shiny to build a cost of living generator application.

Explore Cost of Living Parameters



- Visually, we observe a variance in overall monthly costs by city profile.
- Therefore, we'd like to examine the statistical variance for the cost of living measures, as well as the statistical variance for all markets.

Construct an overall P-value matrix



The first step in Fisher's combined probability test is to build a P-value matrix for all measures. Fisher's Method will systematically control the Type I Error.

Fishers Method Results

Market	#of tests	#of tests with pv<0.05	mean score(-2log(pv))	#of bootstrap	si	smi	ni.1	nmi	pi	pmi	rri	pv.fisher
New York, NY	28	14	7.813384393	1000	218.7748	3853.833	28	1484	4.27E-21	3.22E-26	7.56E-06	0
Saint Louis, MO	28	11	6.956270103	1000	194.7756	3877.833	28	1484	3.14E-17	1.94E-27	6.17E-11	0.014
San Francisco, CA	28	12	5.792777139	1000	162.1978	3910.41	28	1484	2.87E-12	3.88E-29	1.35E-17	0.15
Jackson, MS	28	6	5.156328516	1000	144.3772	3928.231	28	1484	9.74E-10	4.39E-30	4.50E-21	0.322
Charleston, SC	28	9	5.007953994	1000	140.2227	3932.385	28	1484	3.59E-09	2.63E-30	7.31E-22	0.365
Tulsa, OK	28	6	4.336051087	1000	121.4094	3951.199	28	1484	9.83E-07	2.52E-31	2.57E-25	0.565
San Jose, CA	28	6	4.238865773	1000	118.6882	3953.92	28	1484	2.11E-06	1.79E-31	8.48E-26	0.598
Minneapolis, MN	28	4	4.062658771	1000	113.7544	3958.854	28	1484	8.18E-06	9.63E-32	1.18E-26	0.636
Greenville, SC	28	5	3.730809672	1000	104.4627	3968.145	28	1484	9.20E-05	2.97E-32	3.23E-28	0.714
Birmingham, AL	28	4	3.69121604	1000	103.354	3969.254	28	1484	0.000121	2.58E-32	2.13E-28	0.721
Orlando, FL	28	6	3.494461978	1000	97.84494	3974.763	28	1484	0.000458	1.28E-32	2.79E-29	0.756
Dayton, OH	28	4	3.367414625	1000	94.28761	3978.321	28	1484	0.00104	8.12E-33	7.81E-30	0.771
Boise, ID	28	1	3.072499423	1000	86.02998	3986.578	28	1484	0.006081	2.81E-33	4.63E-31	0.816
Boston, MA	28	5	3.007406716	1000	84.20739	3988.401	28	1484	0.008736	2.23E-33	2.55E-31	0.822
Syracuse, NY	28	3	2.977780787	1000	83.37786	3989.23	28	1484	0.010266	2.00E-33	1.95E-31	0.823
Cincinnati, OH	28	3	2.853879473	1000	79.90863	3992.7	28	1484	0.019662	1.28E-33	6.50E-32	0.835
Las Vegas, NV	28	3	2.792741524	1000	78.19676	3994.411	28	1484	0.026684	1.02E-33	3.84E-32	0.842

- The Fishers Method is an application for combining independent tests, forming an overall test statistic, which has a Chi Squared Distribution.

Global Null Hypothesis: Each market's 'relative risk' is the same.

Market's with a low P-Value are considered different (e.g. New York, Saint Louis, San Francisco).

Resources

- NUMBEO (www.numbeo.com)
- Bureau of Labor Statistics
- Publications:
  - Analytical methods for identifying study site outliers/signals for inspection or monitoring (Ram Tiwari, Jianjin Xu, Lan Huang)
  - 24 Hour Cities, Hugh F. Kelly