Data Science Cost Estimates

How to Build a Data Science Cost Estimate with RRRrrr Studio

Jeremy Eden
ICEAA Conference - Tampa, FL
May 2018

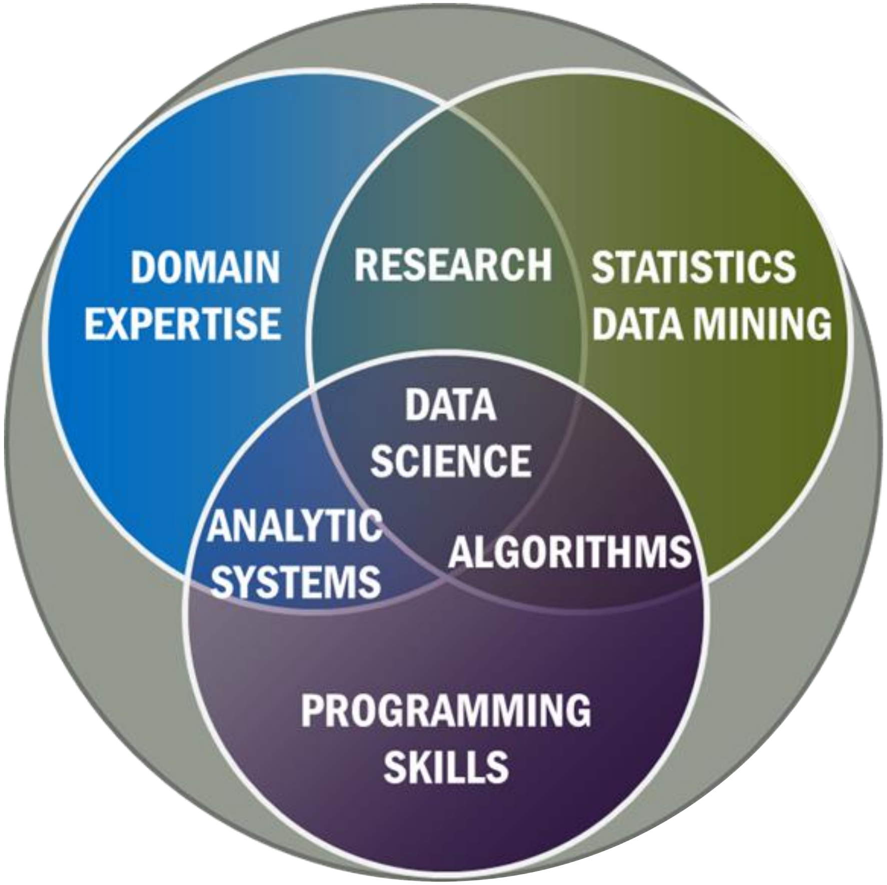
Data Science

A Refresher

Data Science- What is it?

National Institute of Standards and Technology
U.S. Department of Commerce

The extraction of actionable knowledge directly from data through a process of discovery, hypothesis, and analytical hypothesis analysis



NIST Big Data Workgroup

Data Science – How good is it really?

Stanford/s Lucile Packard Children's Hospital

Jenny Frankovich, attending physician

- Young girl had lupus and her kidneys were shutting down
- Some also developed blood clots which can be prevented with drugs, but those carry high risks
- Not sure if drugs should be administered or not, what to do?

Lucile Packard
Children's Hospital
Stanford

The Data Science

- How many lupus patients?
- How many with same symptoms as patients?
- How many of those patients had a clot?

Patients treated for clots and made a full recovery

However, program abandoned due to liability concerns

Data Science – How good can it get?

Wired Magazine- June 2008

Chris Anderson -The End of Theory: The Data Deluge Makes The Scientific Method Obsolete

Scientific Method: “Correlation is not Causation”

- No conclusions on basis of correlation between x and y
- Must understand underlying connections for a model
- Hard to test and experiment unknowns

Data Science Method: “Correlation is enough”

- Computing power and large amounts of data at the problem
- Find patterns in data
- All questions can be answered even if we don’t know why

The World's Cheapest Car | 23 Hot Summer Gadgets
Get Ready for the Google Phone
WIRED
The End of Science
The quest for knowledge used to begin with grand theories.
Now it begins with massive amounts of data. Welcome to the Petabyte Age.

1976 George Box - Statistician - Journal of the American Statistical Association

"All models are wrong, but some are useful.“

2008 Peter Norvig - Google’s research director - O'Reilly Emerging Technology Conference

"All models are wrong, and increasingly you can succeed without them."

Data Science and Cost Estimating

A Refresher

Data Science & Cost Estimating – Who is doing it?

ICEAA

Walt Disney Imagineering

Booz | Allen | Hamilton

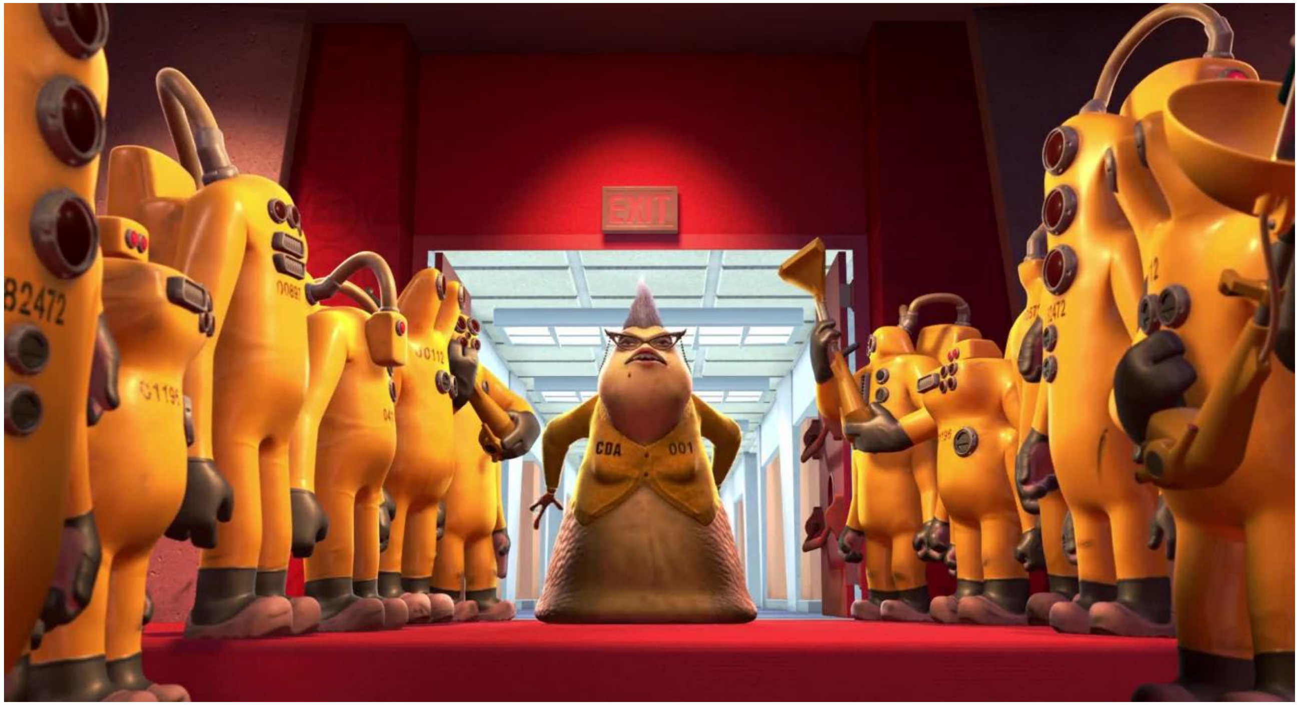
amazon

DAU

Organizations using data science for other disciplines are also using it for cost estimating. Additionally, the ICEAA community and it's members have been used some data science techniques already ….

- Construction
- Software Cost Estimation
- General Cost Estimation
- Defense Cost Estimation
- Computing and Network Structures … and more

ICEAA & Data Science – Wait, didn't ICEAA call the CDA?



Disney Monsters Inc.

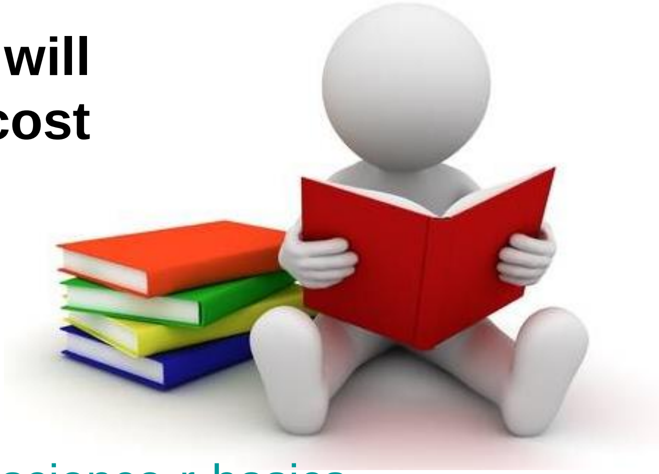
No! There was no need for ICEAA to be alarmed or to call the CDA (Cost Decontamination Unit), they agree!

Over the last few years ICEAA has continues to state it is a priority to:

- Advance the cost estimating industry by supporting new fields like Agile and Data Science
- Enable and encourage the growth of the cost estimating profession and ICEAA membership
- Revise reference material and training opportunities as needed to include the latest techniques

Ok, How Do I Get Started Today?

These are a list of resources you can go to that will enable you to start using Data Science in your cost estimates today …



Learn RStudio

HARVARD UNIVERSITY

https://online-learning.harvard.edu/course/data-science-r-basics

Capture data/favorite data sites whenever possible even if it seems unrelated to what you are doing at the time

DATA.GOV
https://www.data.gov/

amazon web services
https://aws.amazon.com/datasets/

CENTRAL INTELLIGENCE AGENCY
THE WORLD FACTBOOK
https://www.cia.gov/library/publications/the-world-factbook/

Government of Canada
https://open.canada.ca/en

EU Open Data Portal
http://data.europa.eu/euodp/en/data/

NUMBEO
https://www.numbeo.com

That’s Great! How do I get Started NOW?

Ok! We will capture data, filter/analyze that data, serve it up in a tool format, and I will provide the following…

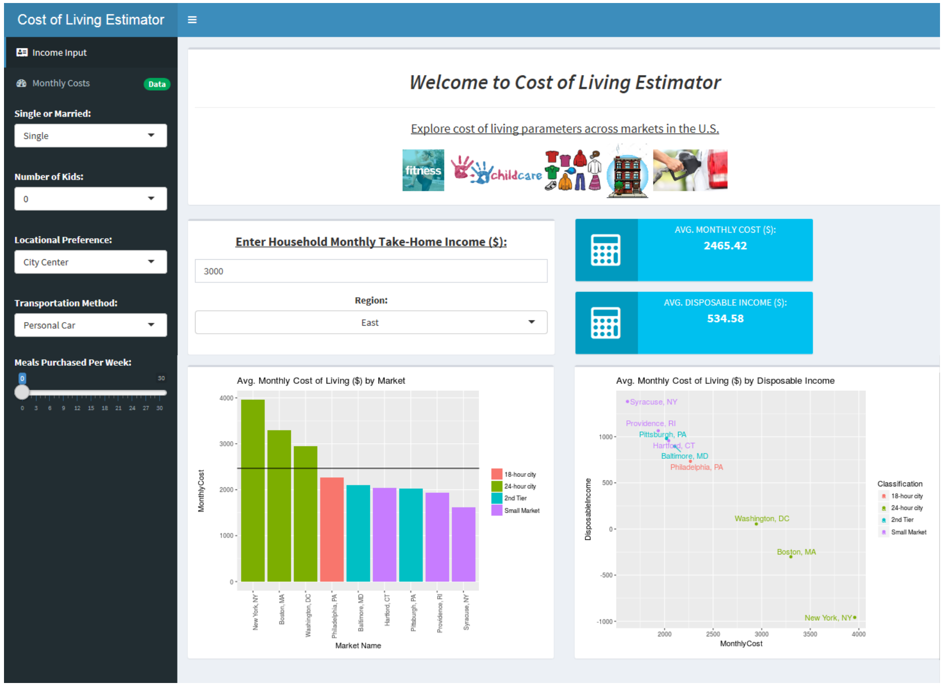
- Data capture location and technique/code for capture
- Data analysis explanation, methodology, and code
- Tool explanation, code, and final demonstration

The Cost of Living Estimator code, data, and analysis can be accessed in a GitHub Repository here: …

https://github.com/tiwariraj/CostofLivingEstimator

Cost of Living Estimator – Introduction and Scenario

We want to build a estimating tool built on actual data that will allow us to estimate how expensive it is to live in one city vs another.



Cost of Living Estimator ≡

Income Input

Monthly Costs Data

Single or Married: Single

Number of Kids: 0

Locational Preference: City Center

Transportation Method: Personal Car

Meals Purchased Per Week: 0

Welcome to Cost of Living Estimator

Explore cost of living parameters across markets in the U.S.

Enter Household Monthly Take-Home Income ($): 3000

Region: East

AVG. MONTHLY COST ($): 2465.42

AVG. DISPOSABLE INCOME ($): 534.58

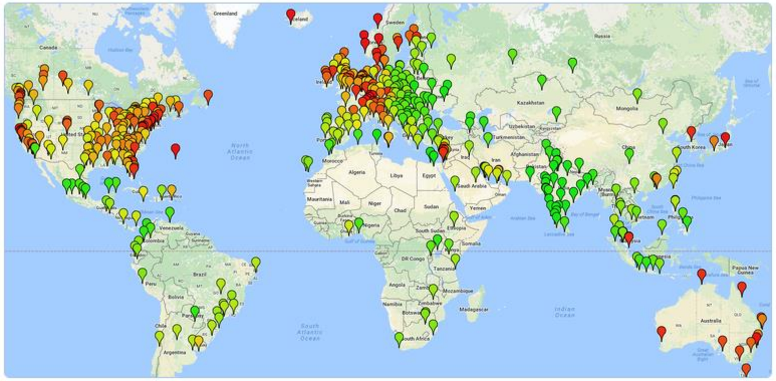
We will be using real data and popular tools for the analysis and development of the estimator interface

Cost of Living Estimator - Data Collection (1 of 2)

Like any estimate, you must start by finding and collecting data. In the analysis/tool, we will capture data for 55 major cities…

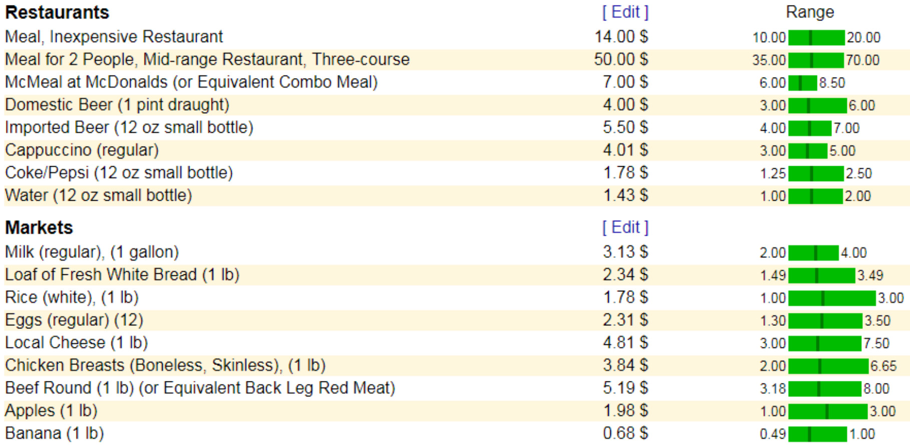
NUMBEO

https://www.numbeo.com/cost-of-living



There is data available for 28 cost types relevant to our exercise including…

- Rent costs
- Grocery costs
- Fuel costs
- Dining costs
- Costs of services (utilities, etc.)



Restaurants	[Edit]	Range
Meal, Inexpensive Restaurant	14.00 $	10.00 🟩 20.00
Meal for 2 People, Mid-range Restaurant, Three-course	50.00 $	35.00 🟩 70.00
McMeal at McDonalds (or Equivalent Combo Meal)	7.00 $	6.00 🟩 8.50
Domestic Beer (1 pint draught)	4.00 $	3.00 🟩 6.00
Imported Beer (12 oz small bottle)	5.50 $	4.00 🟩 7.00
Cappuccino (regular)	4.01 $	3.00 🟩 5.00
Coke/Pepsi (12 oz small bottle)	1.78 $	1.25 🟩 2.50
Water (12 oz small bottle)	1.43 $	1.00 🟩 2.00
Markets	[Edit]	
Milk (regular), (1 gallon)	3.13 $	2.00 🟩 4.00
Loaf of Fresh White Bread (1 lb)	2.34 $	1.49 🟩 3.49
Rice (white), (1 lb)	1.78 $	1.00 🟩 3.00
Eggs (regular) (12)	2.31 $	1.30 🟩 3.50
Local Cheese (1 lb)	4.81 $	3.00 🟩 7.50
Chicken Breasts (Boneless, Skinless), (1 lb)	3.84 $	2.00 🟩 6.65
Beef Round (1 lb) (or Equivalent Back Leg Red Meat)	5.19 $	3.18 🟩 8.00
Apples (1 lb)	1.98 $	1.00 🟩 3.00
Banana (1 lb)	0.68 $	0.49 🟩 1.00

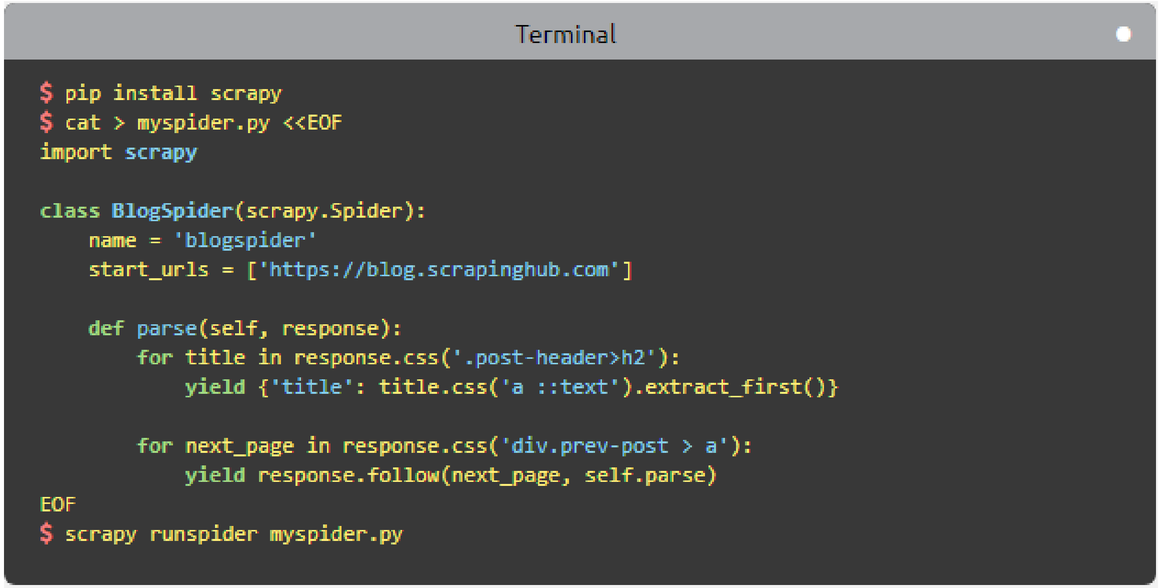
The data collected can be found here:
https://github.com/tiwariraj/CostofLivingEstimator/tree/master/Scrapy%20Crawlers/costofliving

Cost of Living Estimator - Data Collection (2 of 2)

Since the data is contained within an interactive webpage, we will use a popular tool for the capture using spiders. Scrapy allows us to “scrape” the data from the Numbeo site automatically for all 28 types of data and store them in a file in the cloud, a computer, or server for analysis…

https://scrapy.org

This is the basic Scrapy spider code template available on their website for modification.



```
$ pip install scrapy
$ cat > myspider.py <<EOF
import scrapy

class BlogSpider(scrapy.Spider):
    name = 'blogspider'
    start_urls = ['https://blog.scrapinghub.com']

    def parse(self, response):
        for title in response.css('.post-header>h2'):
            yield {'title': title.css('a ::text').extract_first()}

        for next_page in response.css('div.prev-post > a'):
            yield response.follow(next_page, self.parse)
EOF
$ scrapy runspider myspider.py
```

The modified spider code we use for capturing the Numbeo data is found here:
https://github.com/tiwariraj/CostofLivingEstimator/blob/master/Scrapy%20Crawlers/costofliving/costofliving/spiders/costofliving_spider.py

Cost of Living Estimator - Data Filtering/Cleanup

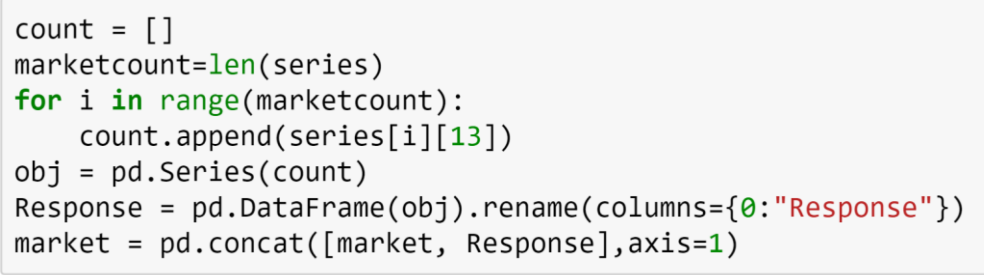
Now that we have captured the data some “clean up” is required to ensure it is complete and organized the way we want. To do this, we will use the programming language Python. There are many tutorials available for the basic code on the internet…

The code changes the data set to include:

- How many data points were in each city’s cost
- Tagging each city with a type (24 hour, 18 hour, 2nd tier market, small market) by combining it from another data source (The book 24 hour cities by Hugh Kelly) since that data wasn’t available from Numbeo

https://www.python.org/

This is the python code that appends the data set:

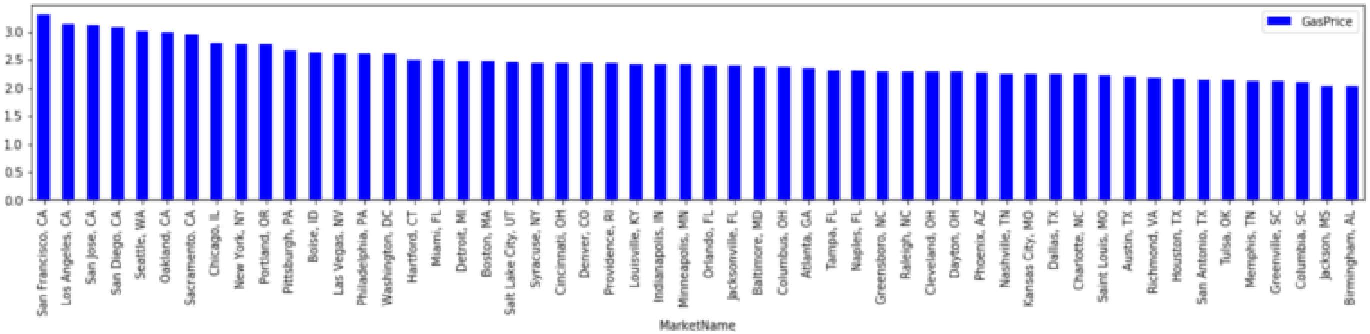


```
count = []
marketcount=len(series)
for i in range(marketcount):
    count.append(series[i][13])
obj = pd.Series(count)
Response = pd.DataFrame(obj).rename(columns={0:"Response"})
market = pd.concat([market, Response],axis=1)
```

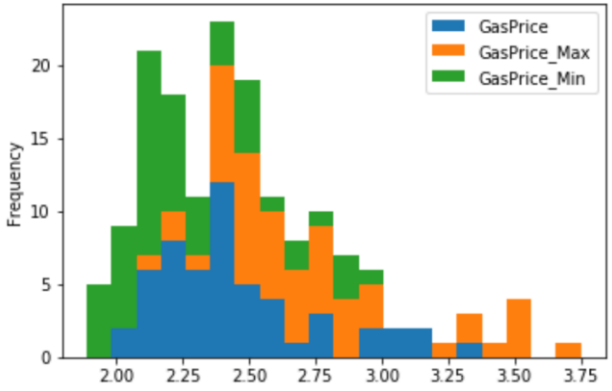
Cost of Living Estimator - Data Analysis (1 of 2)

Now that the data has been collected and it has been cleaned up, we can do some analysis which, once done, will feed the tool for end users. To do this we will examine the distribution/dispersion of the data and the frequency…

This is an example of the gas prices for the fifty markets…



… and here the frequency they occur at various prices

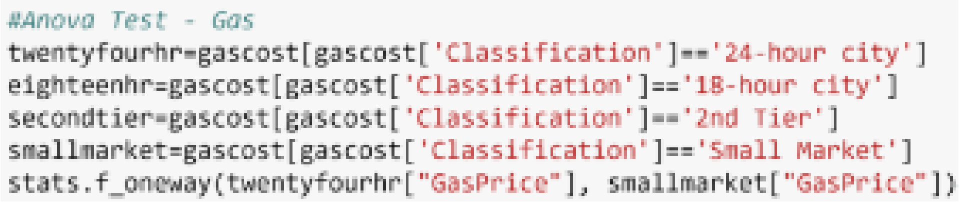


We do this analysis for other data points collected. Here is the file for the gas prices:
https://github.com/tiwariraj/CostofLivingEstimator/blob/master/Scrapy%20Crawlers/costofliving_gas/gascost.csv

Cost of Living Estimator - Data Analysis (2 of 2)

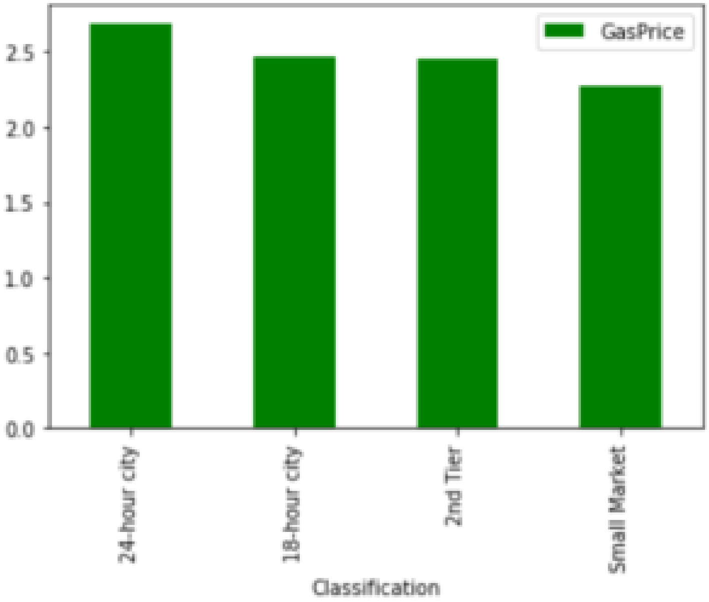
Going further, we can associate costs with other pieces of the data …

For example, the following code allows the association of gas prices with the city type…



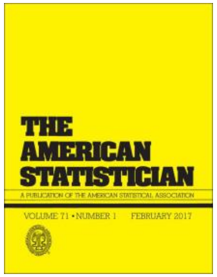
```
#Anova Test - Gas
twentyfourhr=gascost[gascost['Classification']=='24-hour city']
eighteenhr=gascost[gascost['Classification']=='18-hour city']
secondtier=gascost[gascost['Classification']=='2nd Tier']
smallmarket=gascost[gascost['Classification']=='Small Market']
stats.f_oneway(twentyfourhr["GasPrice"], smallmarket["GasPrice"])
```

There is an impact on cost as a result of the city type (i.e. 24 hour city and gas prices) on the cost of living.



Cost of Living Estimator - Data Analysis

Using Fisher's method for combining P-values we can see the impact of a city type on all the cost of living types and place into a matrix…

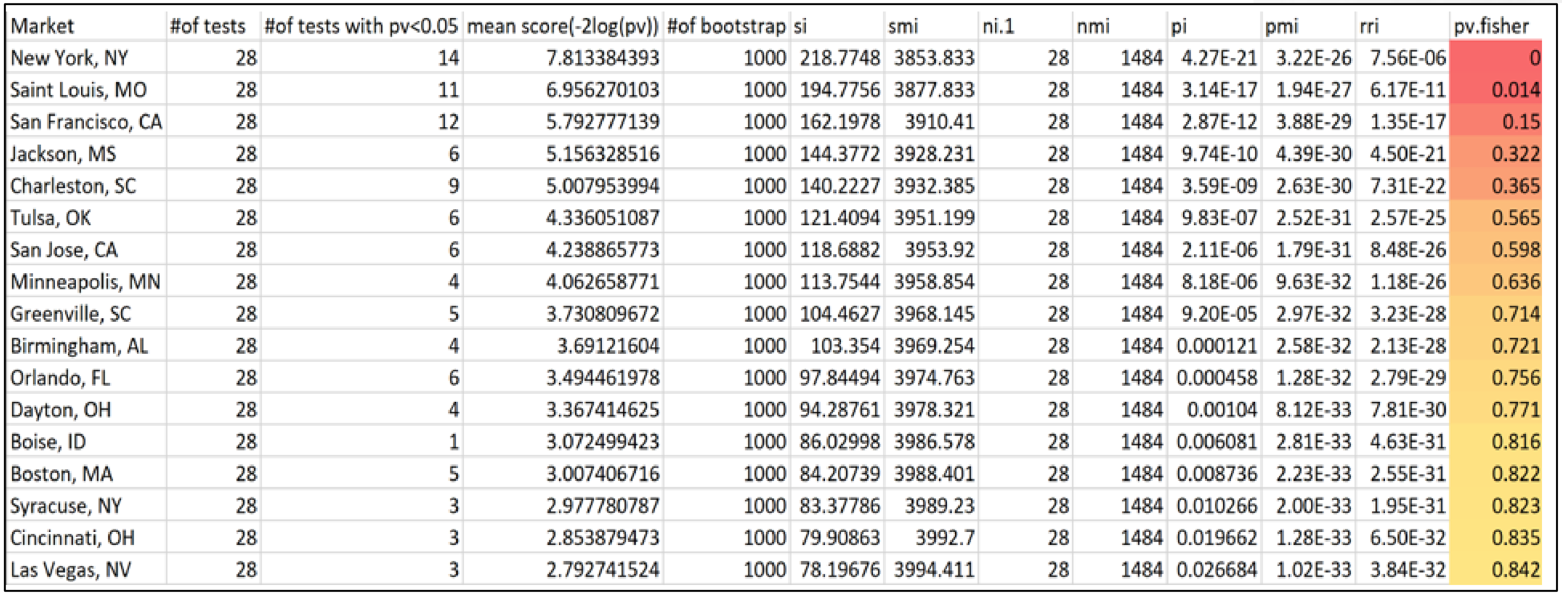


THE AMERICAN STATISTICIAN

A PUBLICATION OF THE AMERICAN STATISTICAL ASSOCIATION

VOLUME 71 • NUMBER 1 FEBRUARY 2017

$$-2\sum_{i=1}^{k}\log(p_i) \sim \chi^2_{2k}$$



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

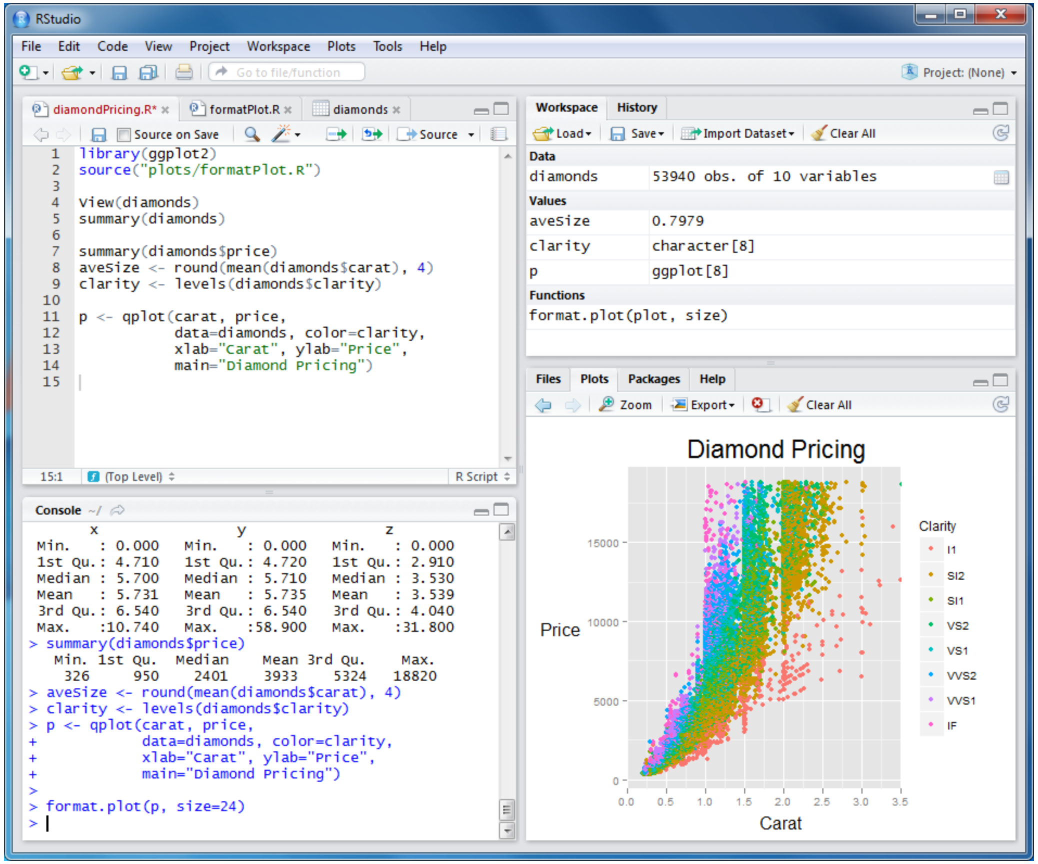
The entire matrix of all the cost of living types is available here:
https://github.com/tiwariraj/CostofLivingEstimator/tree/master/Fishers_pmatrix/market_raj

Cost of Living Estimator – Interface with RStudio

Instead of building our tool in MS Excel we will be using…

RStudio®

https://www.rstudio.com



```r
library(ggplot2)
source("plots/formatPlot.R")

View(diamonds)
summary(diamonds)

summary(diamonds$price)
aveSize <- round(mean(diamonds$carat), 4)
clarity <- levels(diamonds$clarity)

p <- qplot(carat, price,
           data=diamonds, color=clarity,
           xlab="Carat", ylab="Price",
           main="Diamond Pricing")

```

Data

diamonds	53940 obs. of 10 variables

Values

aveSize	0.7979
clarity	character[8]
p	ggplot[8]

Functions

format.plot(plot, size)

```
       x                y                z         
 Min.   : 0.000   Min.   : 0.000   Min.   : 0.000  
 1st Qu.: 4.710   1st Qu.: 4.720   1st Qu.: 2.910  
 Median : 5.700   Median : 5.710   Median : 3.530  
 Mean   : 5.731   Mean   : 5.735   Mean   : 3.539  
 3rd Qu.: 6.540   3rd Qu.: 6.540   3rd Qu.: 4.040  
 Max.   :10.740   Max.   :58.900   Max.   :31.800  
> summary(diamonds$price)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    326     950    2401    3933    5324   18820 
> aveSize <- round(mean(diamonds$carat), 4)
> clarity <- levels(diamonds$clarity)
> p <- qplot(carat, price,
+            data=diamonds, color=clarity,
+            xlab="Carat", ylab="Price",
+            main="Diamond Pricing")
> 
> format.plot(p, size=24)
> 
```

With …

Shiny

Cost of Living Estimator – Interface with R Shiny

No! All Hail Excel!

Wait! I am a huge fan of MS Excel and am the first one to say “Can I?” when I should say “Should I?”, but it DOES have it’s limitations including…

- Users per instance
- Data pipeline limitations
- Version control issues
- Limited by hardware resources

R Studio…

- Is open source and free
- Enables easier collaboration
- Is available for all platforms
- Runs on the web
- Is accepted by my commercial and government organizations (including most of DoD)

Cost of Living Estimator – Interface with R Shiny

The Cost of Living Estimator interface captures user inputs, makes living type/style assumptions, and provides the analyzed city data and the city that matches best for the user. The user inputs include the following…

- Married status
- Number of children
- Urban or suburban preference
- Transportation method preference
- Number of meals purchased
- Preferred region
- Household monthly take-home income

Welcome to Cost of Living Estimator

Explore cost of living parameters across markets in the U.S.

Cost of Living Estimator – Data Pipeline

Tools created using data science structures and tools can be designed or modified to accept data pipelines. These pipelines can provide numerous benefits over other tools without the capability including…

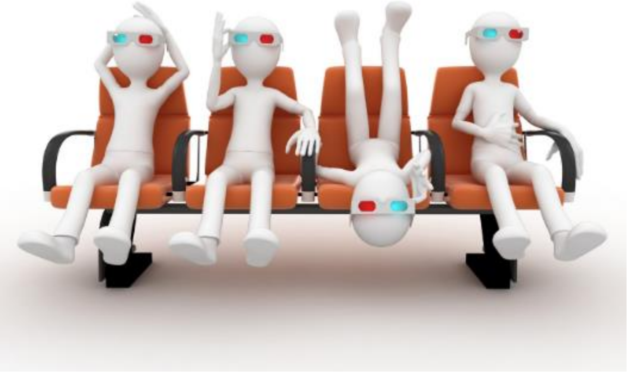
- Data sets that don’t expire
- Near real time analysis
- Ability to accommodate collaborative data sets

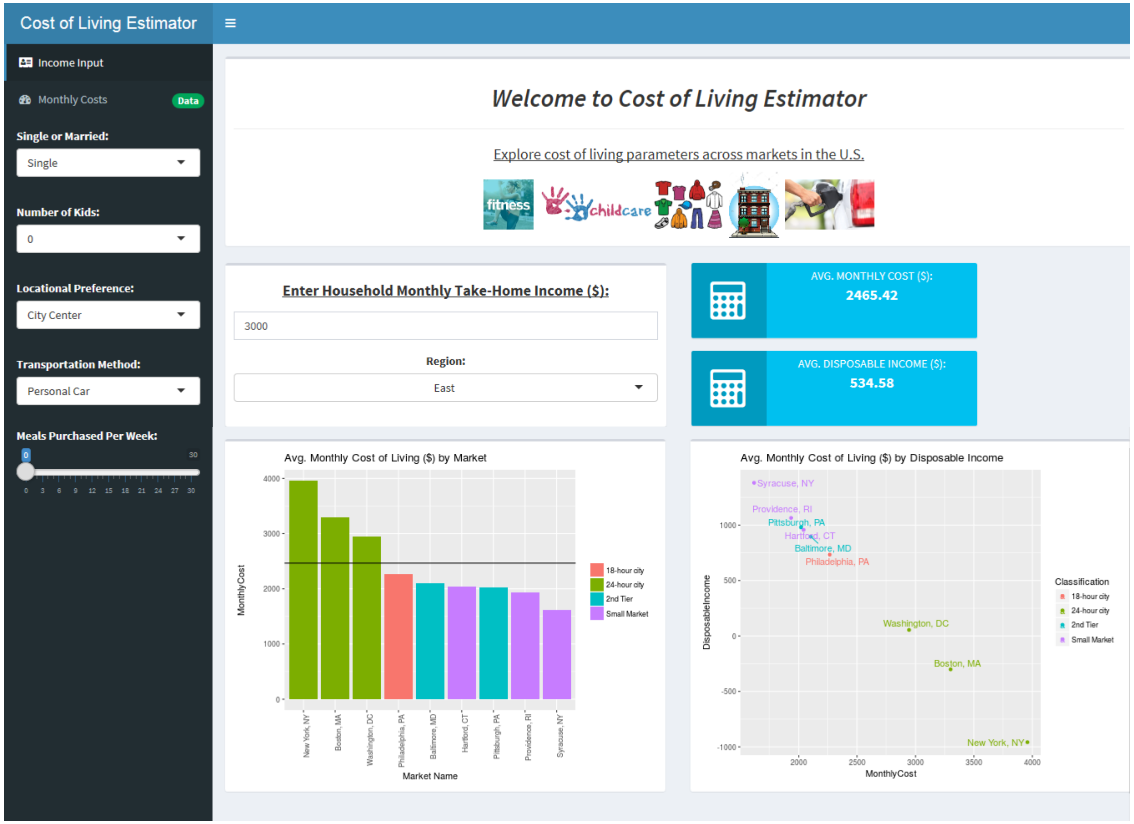
It is important to recognize that data portals controlled by third party operators have the possibility of interruption resulting from format changes, portal configuration changes, etc. beyond the control of the tool author/end user and should be planned for

Cost of Living Estimator – Live Demonstration

Live demonstration of the Cost of living Estimator

The Cost of Living Estimator is available here:
https://tiwariraj.shinyapps.io/CostofLivingEstimator/





Cost of Living Estimator ≡

- Income Input
- Monthly Costs Data

Single or Married: Single

Number of Kids: 0

Locational Preference: City Center

Transportation Method: Personal Car

Meals Purchased Per Week: 0

Welcome to Cost of Living Estimator

Explore cost of living parameters across markets in the U.S.

Enter Household Monthly Take-Home Income ($): 3000

Region: East

AVG. MONTHLY COST ($): **2465.42**

AVG. DISPOSABLE INCOME ($): **534.58**

Ok, How Do I Build MY OWN Data Science Cost Tool?

With same basic steps you can build a data science cost estimate no matter what tools you use …

1. Plan the analysis and/or tool like you would with any approach and find your data just like you would with any other estimate.
2. Using Scrapy or a similar tool to capture and document your data
3. Using Python or a similar tool, clean, sort, gap fill, and analyze the data
4. If and interface is required, use RStudio or R Shiny to create an interface and populate a user tool for the estimate

Don't try to do all the above at once! Integrate individual techniques and tools into the methods you use now and go from there.

You can do this! I have faith in you!

Sources and Special Thanks

A special thanks to Raj Tiwari for giving me permission to use his code and tool for discussion in this paper/presentation

Cost of Living Estimator, Raj Tiwari
https://tiwariraj.shinyapps.io/CostofLivingEstimator/

“The End of Theory: The Data Deluge Makes the Scientific Method Obsolete,” 2008.
https://www.wired.com/2008/06/pb-theory/

“Big Data Not a Cure All In Medicine,” 2015.
https://www.npr.org/2015/01/05/375201444/big-data-not-a-cure-all-in-medicine

“The American Statistician Vol. 2, No.5,” 1948, Frederick Mosteller and R.A. Fisher.
https://www.npr.org/2015/01/05/375201444/big-data-not-a-cure-all-in-medicine

“24 Hour Cities,” 2016, Hugh Kelly.
https://www.amazon.com/dp/B01IHQ23V2/ref=cm_sw_em_r_mt_dp_U_htbCCbFDV1ZPY

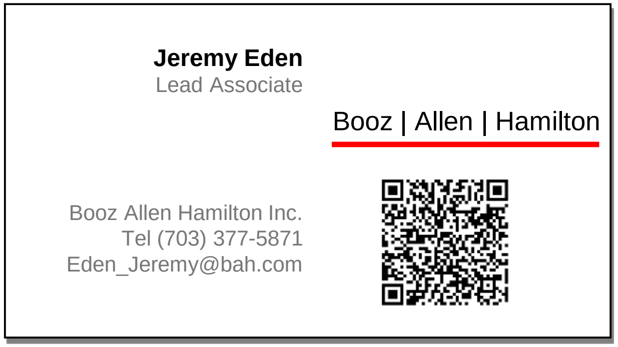
Scrapy
https://scrapy.org/

Numbeo Cost of Living
https://www.numbeo.com/cost-of-living/

Python
https://www.python.org/

R Studio
https://www.rstudio.com/

Questions



Jeremy Eden
Lead Associate

Booz | Allen | Hamilton

Booz Allen Hamilton Inc.
Tel (703) 377-5871
Eden_Jeremy@bah.com