LBA Report: The Price of Basic Goods

Marcela Radilla Deloya

CS146

Fall 2020

The Price of Basic Goods

Introduction

With this project, we attempted to model the cost of groceries around the world. Minerva students around the world collected grocery price data from different grocery stores. This report includes details on the process of data collection, data preprocessing, building a statistical model, running the model in STAN, and sampling from the posterior distribution to answer questions related to the base price of goods around the world. It also includes the interpretation of the results extracted from the STAN output and an analysis of the correlation between price variation by geographical location and variation in rental prices by country.

Data collection and preprocessing

Product price data was collected by students who visited different grocery stores. A total of 64 grocery stores in Germany, England, Guatemala, Vietnam, Morocco, and the US were visited to collect data. Up to 3 different prices from different brands were reported for the following products: apples, bananas, tomatoes, potatoes, flour, rice, milk, butter, eggs, and chicken breast.

Rental price data

The rental data was initially very messy. For the inputs on Germany, some inputs included the median rent, some provided the price for 70 square meters, for 1 or 2 bedrooms, etc. I decided to follow the median rent and the 70 sq. meters input. I left the median rent and 70 sq. meter values as they were and multiplied the values that has indicated price per sq. meter, times 70 since that was mentioned to be the average apartment size by other inputs in the sheet. For Vietnam, I took the mean of the ranges provided. I also converted the currency to VND for the two rows that had product prices in VND but rental price in USD. For the UK, I left everything

the way it was. For the US, some people provided a number, plus an explanation for it, some of which were around: average apartment size, or 1 bedroom. I also multiplied the inputs for 1 sq meter times 70 and the square ft prices times 747 since the inputs said that was the average apartment size.

Price conversion to GBP

I converted all prices to GBP manually in Google Sheets. The following table presents the values I used since they were the values provided by Google when I was converting the prices. The final version of the csv file with all preprocessed data which will be submitted along with this document but is also already liked to the notebook.

Currency	*Rate to USD*
Euro (EUR)	1.19
Guatemalan Quetzal (GTQ)	0.13
Moroccan Dirham (MAD)	0.11
Korean Won (KRW)	0.00089
British Pound (GBP)	1.32
Vietnamese Dong (VND)	0.000043

Table 1. Currency rates for conversion to USD

Price normalization according to units

I did the normalization in Google Sheets. I extracted only the columns for Country, Store Name, Store Perception, Rent and all 30 price columns (3 columns for each one of 10 products) to create the dataset I woud work with on the Jupyter notebook. Every step after this is carried out in the Jupyter notebook attached to this report. Just to note the price normalization was done according to the units outlined in the following table. For this report, every price mentioned for a product will follow this normalization.

Product	*Units*
Apples, Bananas, Tomatoes, Potatoes, Flour, Rice, Butter, Chicken breast	kg
Milk	liters
Eggs	count

Table 2. Units for the normalization of each product price.

Reshaping, Encoding, and Outlier Removal

In order to have a price observation for each product price recorded by the students, I used Python to reshape the dataset so that I had one row for each price observation which also specified which product was being recorded, this new array had the columns:

Country, Store (name), Perception (expensive, cheap), Rent, Price, and Product (apples, bananas). The countries, product types, and perception of the store were numerically encoded. Initial plotting of the data showed some concerning outliers for bananas, milk (Fig. 1), flour, and eggs (Fig. 2).

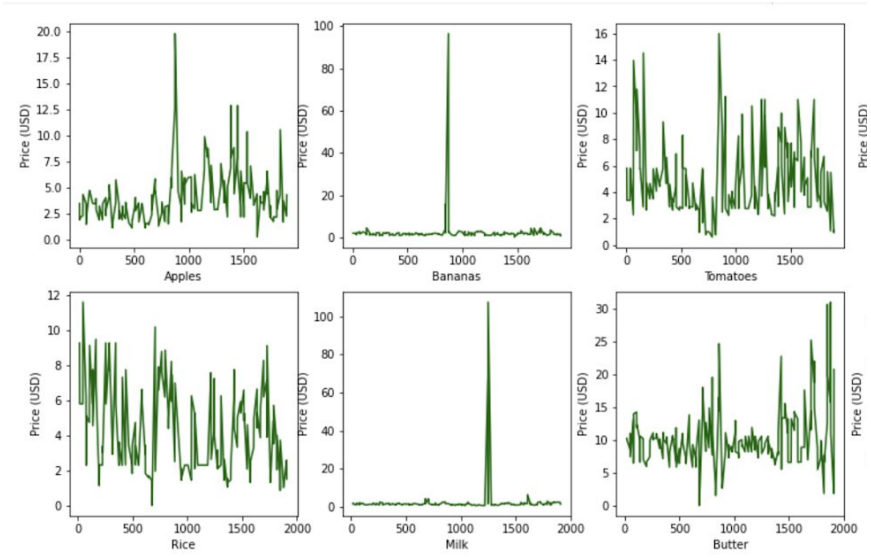

Figure 1. Initial plotting of the data: Part 1. This figure shows initial data plots for apples, bananas, tomatoes, rice, milk, and butter. We can observe a single outlier for bananas just around 100 USD/kg and another one for milk also around 100 USD/liter.

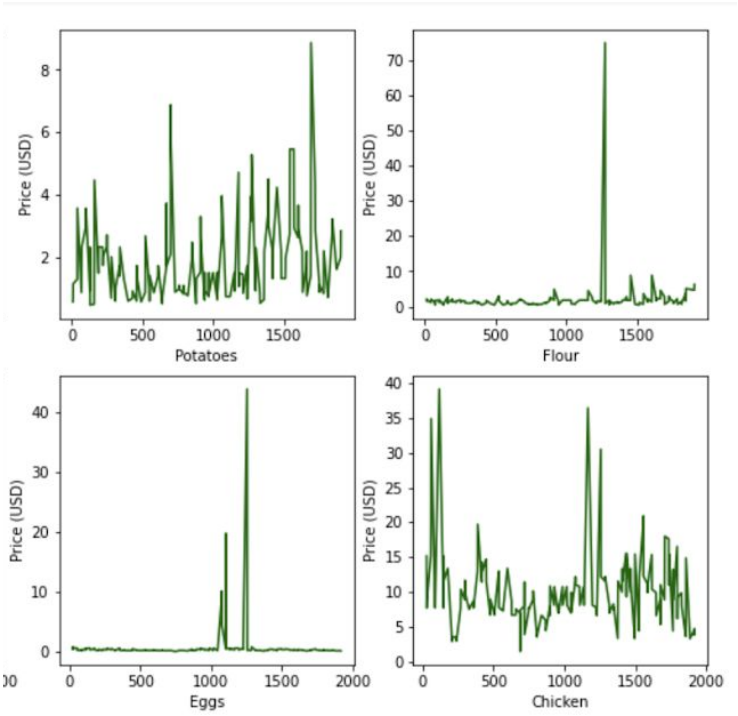

Figure 2. Initial plotting of the data: Part 2. This figure shows initial data plots for potatoes, flour, eggs, and chicken breasts. We observe an outlier for flour around 70 USD/kg and one for eggs around 40 USD/kg.

Based on the lists of maximum values, I removed a total of five outliers: one outlier from the banana prices ($96.36), one from the flour prices ($74.844), two from the milk prices ($107.448 and $66.792), and one from the egg prices ($43.956). I thought removing these outliers was a reasonable choice, since the prices come from stores in the UK that had observations for the same product, and those prices are just unrealistic considering the stores that they come from (Waitrose and Sainsbury's). Data was plotted again. (Fig. 3,4) While there were other outliers remaining, they were not as faw away as the ones removed, and I did not want to keep modifying the dataset to prevent adding more bias.

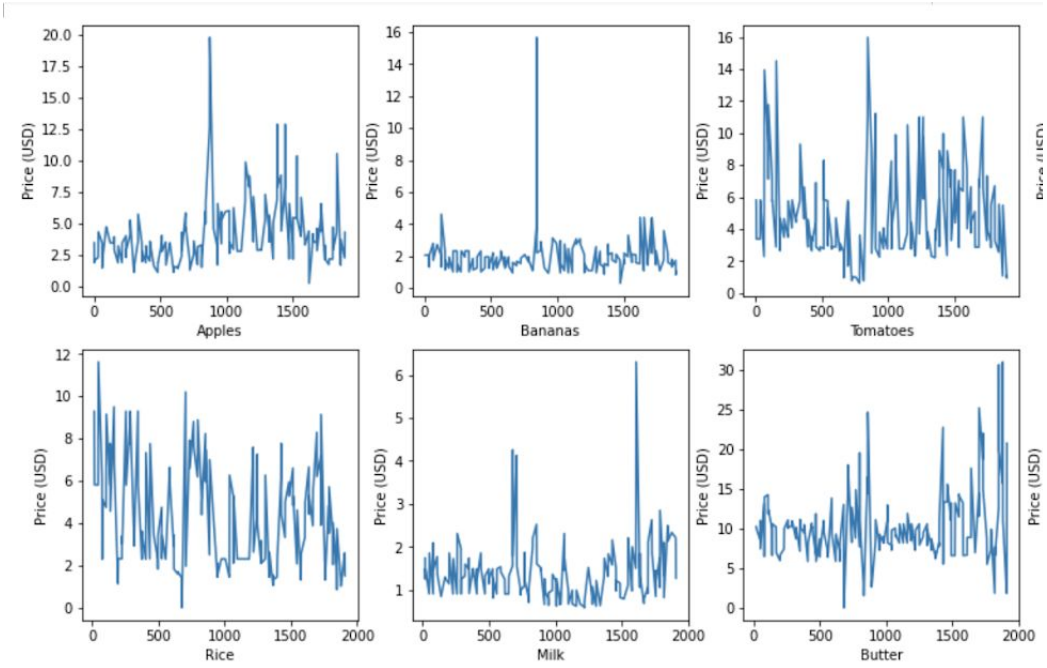

Figure 3. Data plotted after removing outliers. This figure shows initial data plots for apples, bananas, tomatoes, rice, milk, and butter.

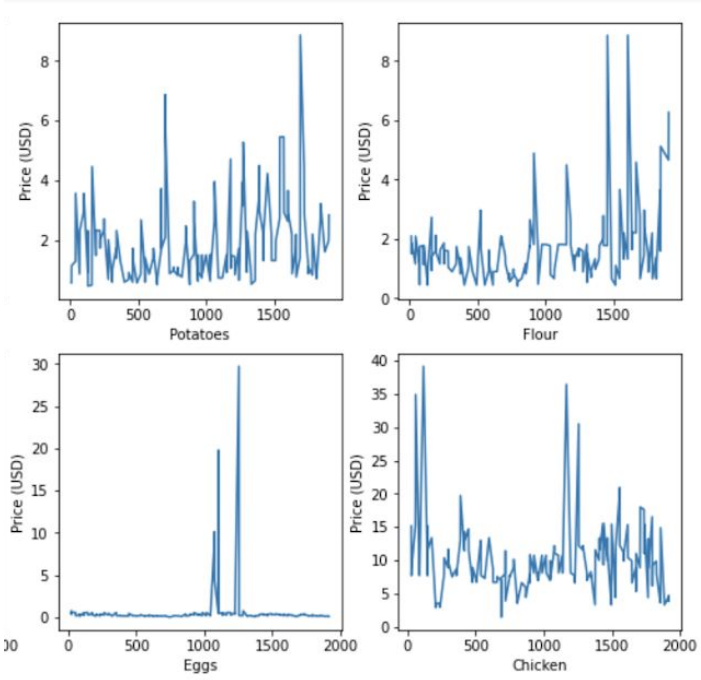

Figure 4. Data plotted after removing outliers. This figure shows initial data plots for potatoes, flour, eggs, and chicken breasts.

Building the model

Selection of distributions

I decided on a truncated Cauchy for the base price prior since the fat tails allow for higher probabilities of values far away from the mean than the normal distribution and thus more variation. I centered the Cauchy around 5, assuming that most products cost around 5 USD.

I decided to use Gamma priors for the country and store type multipliers, both centered around 1.

We will also use a Gamma(1,1) for the error, this is only to derive the standard distribution for the likelihood, since I know that some of the outliers are due to human error.

We will use a truncated normal for the likelihood, setting the price to come from a normal distribution, whose mean is given by: (*base price*) $*$ (*country multiplier*) $*$ (*store multiplier*) and whose standard deviation is given by the error. While I didn't specify the truncated distributions in the model, I set a lower limit of zero for the value of the parameter, so STAN follows that when it generates the posterior distribution.

STAN model

The STAN model was set up so it generated samples over the base price for each product, over the multipliers for each country, and the multipliers for each store. The posterior means from STAN are in Table 3. Standard errors and percentiles are shown in Figure 5. Figures 6, 7 and 8 show plots from the posterior samples extracted from the Stan output.

Product	*Posterior mean*	*Countries*	*Posterior mean*	*Store types*	*Posterior mean*
Apples	5.78	Germany	2.54	Budget	0.28
Bananas	2.8	Guatemala	2.04	Luxury	0.46
Tomatoes	7.24	Morocco	1.77	Mid-range	0.35
Potatoes	2.67	South Korea	3.45		
Flour	2.28	UK	2.45		
Rice	6.11	USA	2.75		
Milk	1.99	Vietnam	2.37		
Butter	14.02				
Eggs	1.03				
Chicken	14.14				

Table 3. Posterior means over base price, country multipliers, and store type multipliers.

```
Inference for Stan model: anon_model_48c5a5a4637074737e2ceedef0130919.
4 chains, each with iter=2000; warmup=1000; thin=1;
post-warmup draws per chain=1000, total post-warmup draws=4000.

                  mean se_mean     sd   2.5%    25%    50%    75%  97.5%  n_eff   Rhat
base_price[1]     5.78    0.07   1.97   2.75   4.38   5.49   6.83  10.39    792   1.01
base_price[2]      2.8    0.03   0.99   1.29   2.09   2.66   3.34   5.14    846    1.0
base_price[3]     7.24    0.09   2.45   3.52   5.48   6.89   8.58  13.01    758   1.01
base_price[4]     2.67    0.03   0.94   1.24   2.01   2.53   3.19   4.93    828   1.01
base_price[5]     2.28    0.03   0.83   1.03   1.68   2.15   2.74   4.22    829   1.01
base_price[6]     6.11    0.07   2.08   2.96   4.64   5.78   7.26  11.04    804   1.01
base_price[7]     1.99    0.02   0.74   0.89   1.44   1.87   2.42   3.71    922    1.0
base_price[8]    14.02    0.17   4.74   6.78  10.63  13.28  16.59  24.98    773   1.01
base_price[9]     1.03    0.01   0.46   0.35    0.7   0.96   1.28   2.13   1052   1.01
base_price[10]   14.14    0.17   4.78   6.86  10.68  13.42  16.81  25.33    771   1.01
country_mult[1]   2.54    0.03   0.87   1.16   1.89   2.43   3.06   4.48    932    1.0
country_mult[2]   2.04    0.02   0.71   0.93   1.51   1.94   2.47   3.69    957    1.0
country_mult[3]   1.77    0.02   0.61   0.79   1.33   1.69   2.13   3.16    952    1.0
country_mult[4]   3.45    0.04   1.19   1.56   2.56    3.3   4.16   6.19    972    1.0
country_mult[5]   2.45    0.03   0.84    1.1   1.83   2.35   2.96    4.4    944    1.0
country_mult[6]   2.75    0.03   0.94   1.24   2.05   2.64    3.3   4.89    937    1.0
country_mult[7]   2.37    0.03   0.82   1.08   1.76   2.28   2.85   4.22    974    1.0
store_mult[1]     0.28  5.3e-3   0.14    0.1   0.19   0.25   0.35   0.63    713    1.0
store_mult[2]     0.46  8.6e-3   0.23   0.17    0.3   0.41   0.56   1.01    709    1.0
store_mult[3]     0.35  6.7e-3   0.18   0.13   0.23   0.32   0.43   0.79    707    1.0
error              2.8  9.6e-4   0.05    2.7   2.76   2.79   2.83   2.89   2431    1.0
lp__              -2680    0.08   3.23  -2687  -2683  -2680  -2678  -2675   1518    1.0

Samples were drawn using NUTS at Sun Nov  8 15:11:31 2020.
For each parameter, n_eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor on split chains (at
convergence, Rhat=1).
```

Figure 5. Stan output. We can find the 95% confidence intervals for the posterior means here if we look under the 2.5% and 97.5% columns.

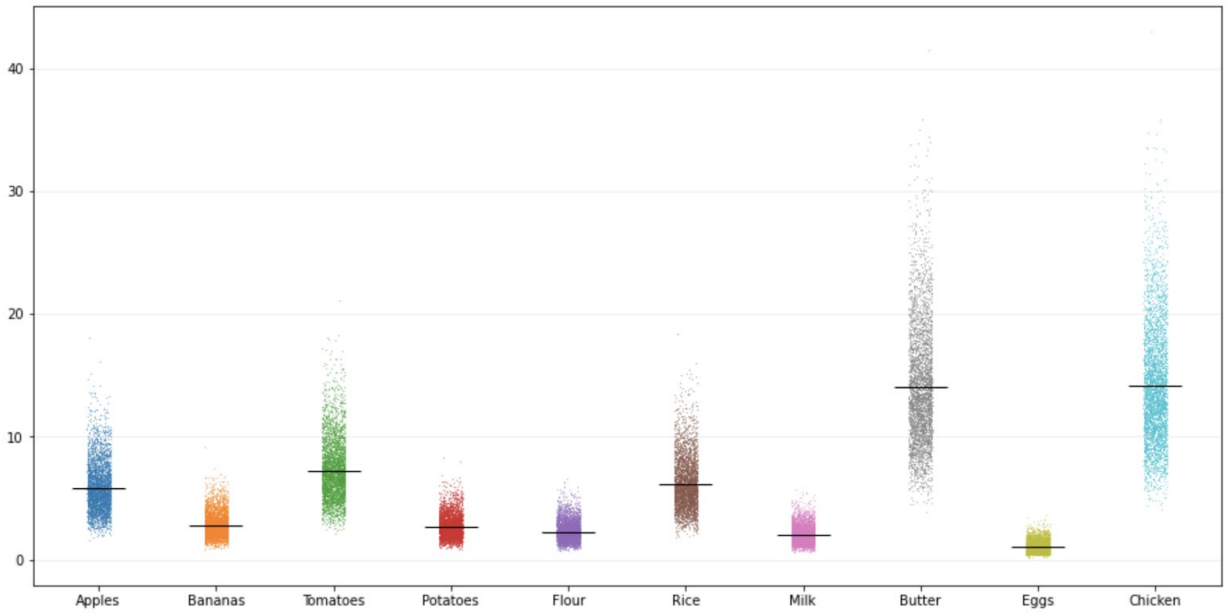

Figure 6. Posterior samples for the base price for each product.

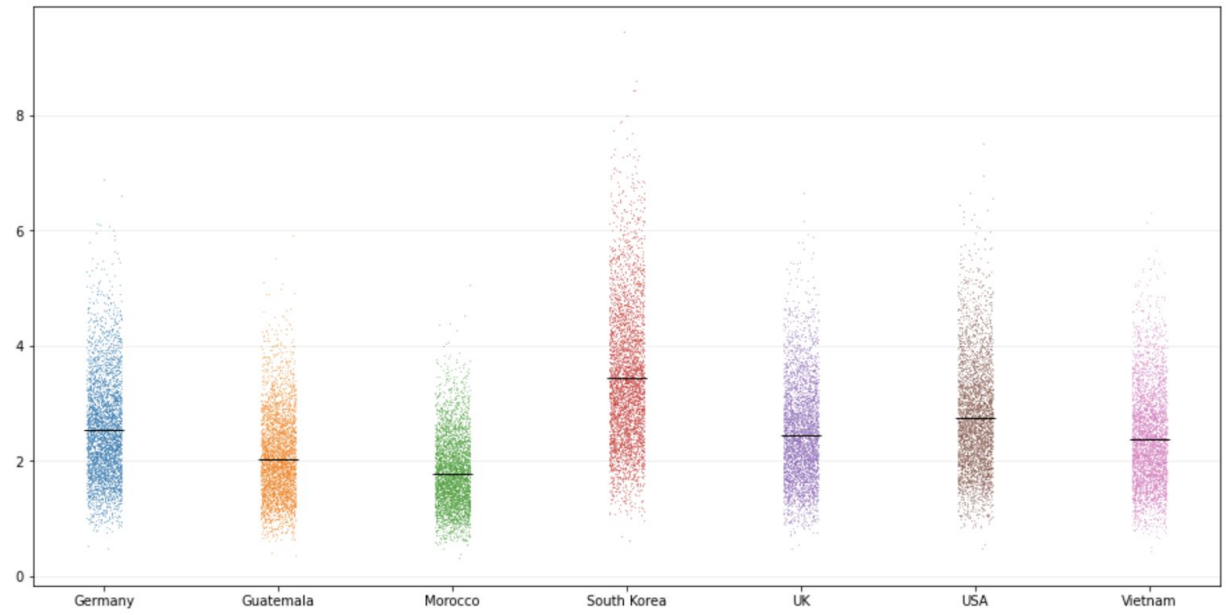

Figure 7. Posterior samples for the country multipliers

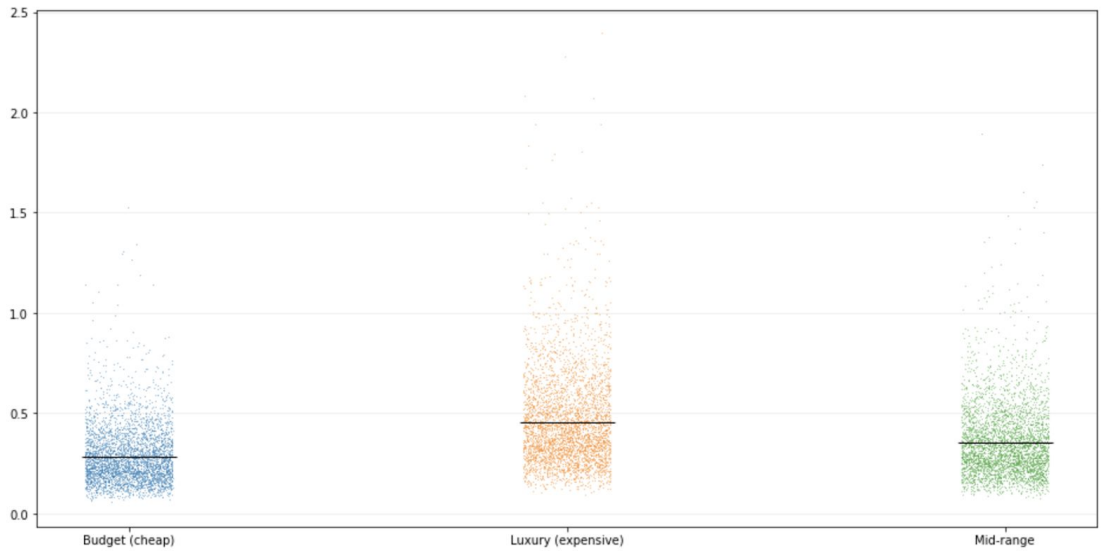

Figure 8. Posterior samples for the store type multipliers

Results

The samples from the posterior distribution over basic prices can be observed in (Fig. 5). We observe more uncertainty around the prices of butter and chicken than around the prices of potatoes, bananas, milk, and eggs since the ranges of their samples (and their 95% confidence intervals) appear to be wider. The geographical location actually has a relevant influence in the final price, with South Korea having the largest multiplier and Morocco having the smallest. The increasing size of the store type multipliers goes with increasing price perception. The geographical location has a bigger influence on the price than the price perception of the store.

Correlation with rental prices

The mean of the reported rental prices by country was computed to get the correlation coefficient between the rental price means and the country multipliers means from the STAN samples. There is a very small correlation between the variation in price by geographical location and the variation in rental prices (Fig. 8). We looked at the regression line plot and notice that the South Korea datapoint was mostly responsible for lowering the slope of the regression line.

Without the datapoint, the correlation coefficient was around 0.63 which proves how South Korea is an exception to thinking that variation in grocery prices might correlate with other living-costs measures like rental prices.

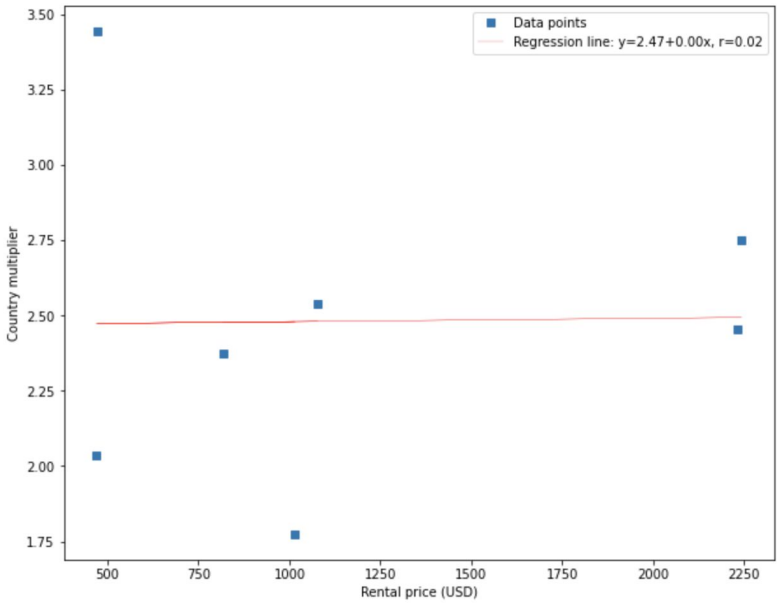

Figure 9. Regression line between mean rental prices and country multipliers.

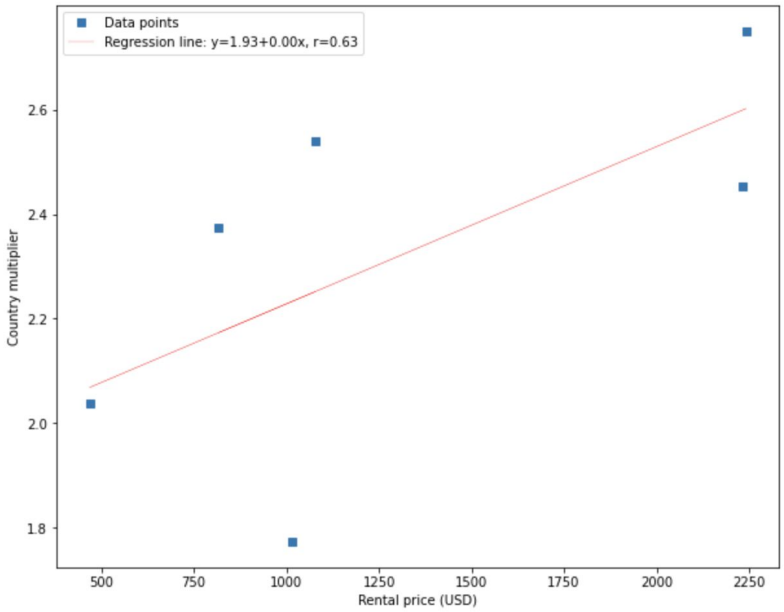

Figure 10. Regression line between mean rental prices and country multipliers without South Korea.

Link to notebook: https://github.com/marcelaradilla/cs146/blob/main/CS146_LBA_1.ipynb