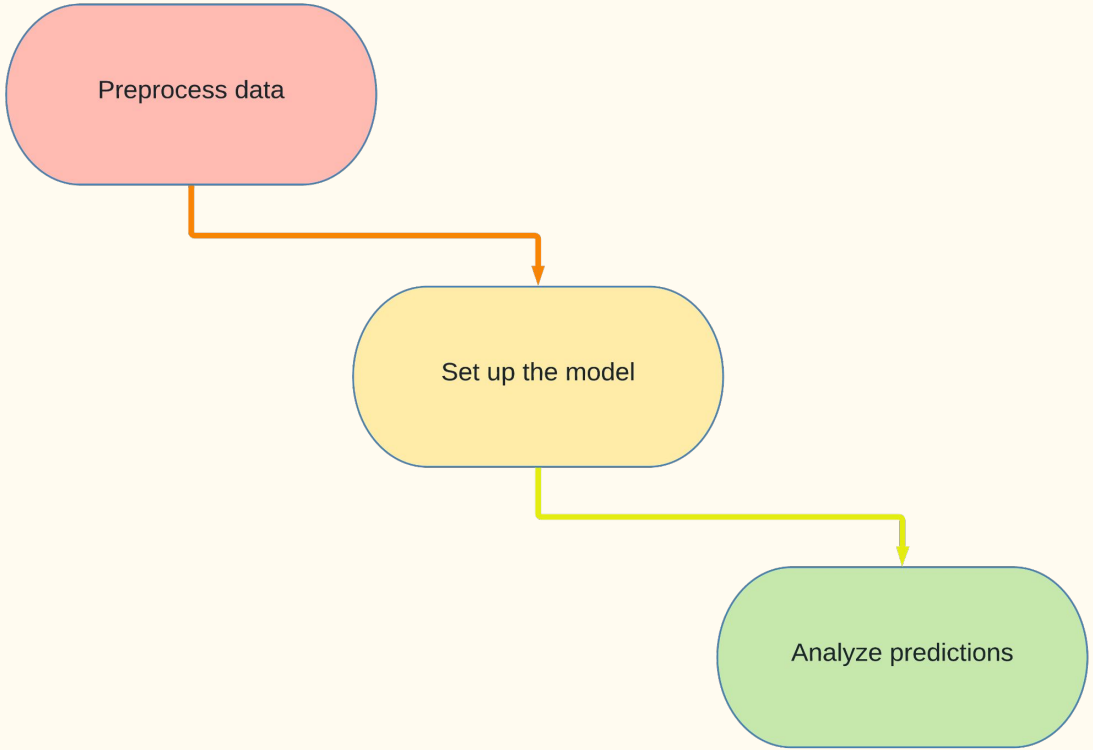
The cost of basic goods in different countries

Introduction



This report contains a statistical model that explores the relationships between grocery prices, store types, and locations worldwide. We begin with relevant assumptions about the given scenario, proceed with modeling and finish the report discussing of findings and their implications.

Data

We used the data from **26** international students who gathered information about **3** store types in **3** different countries and collected **3** different prices for **10** products. Additionally, we identified the average rental prices for **1-bedroom** apartments in the areas where the stores are located. You can find the data here.

Store types

- Cheap
- Mid-range
- Luxury

Countries

- Germany
- United Kingdom
- United States

Products

1. Apples
2. Bananas
3. Tomatoes
4. Potatoes
5. Flour
6. Rice
7. Milk
8. Butter
9. Eggs
10. Chicken breasts

Preprocessing

❗ You can find the code for preprocessing an further steps here.

This is an important first step because it leads to more manageable datasets. We conducted the following procedures:

- *Missing data*: Some entries in the original dataset contained NaN values because not all grocery stores used for data collection purposes had 3 types of products required for the model.

  Solution — we calculated the averages for 3 types of products (or less) for every entry (row) and then took the mean of all averages across the product type (column). Then we filled the NaN values with the means mentioned above to prevent reducing the small dataset to an even smaller dimension.

Preprocessing (cont.)

- *Noisy data*: Some entries in the original dataset contained very strange values (e.g., price of one egg equal to 25 GBP). These values were outliers and could cause inaccurate predictions.

  Solution — we assumed that the values resulted from errors in the data entry process and decided to remove the entries (there were 3 entries so it did not affect the dataset size much).

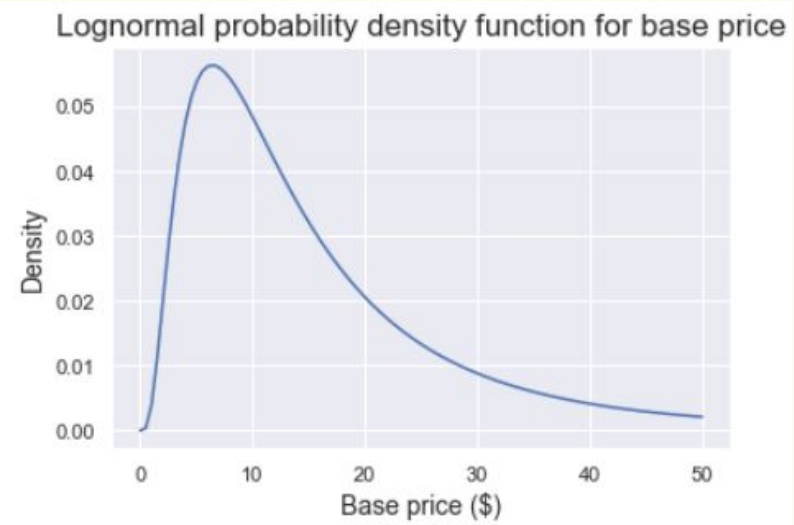
- *Inconsistent data*: Some entries in the original dataset did not correspond to the column descriptions (e.g., cities in the "country" column). This would lead to the inability to produce correct output for a specific group (e.g., location-wise).

  Solution — we designed filter functions that would either group common entries (e.g., the ones with spelling mistakes) or remove the entries that are completely inconsistent (e.g., no numerical values or zero values in the "rental price" column).

Model Setup (prior)

We need some **prior expertise** for making inference about the base average price and the influence of location and store type on it. Let's see what we know about each of the model's components:

base_price ~ lognormal(2.5, 0.8)

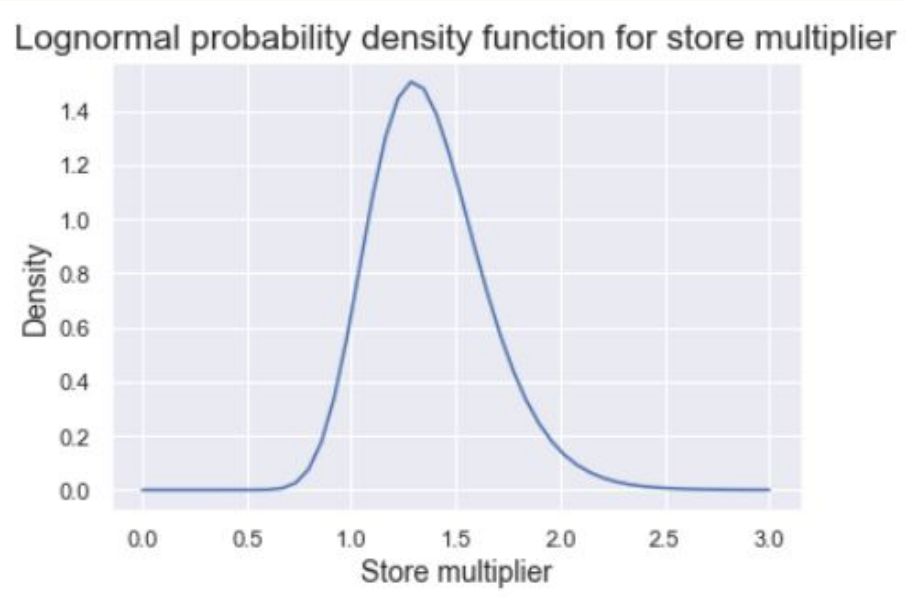


Base price

- A positive real number (support > 0).
- We assume that most values should be concentrated between \$0 and \$20 given our prior knowledge about the items (standardized to 1kg, 1L or 1 piece). However, we do not exclude the possibility of them being more expensive than \$20.
- Lognormal distribution with the mean 2.5 and standard deviation 0.8 seems like a reasonable representation of the requirements.

Model Setup (prior)

store_mult ~ lognormal(0.3, 0.2)

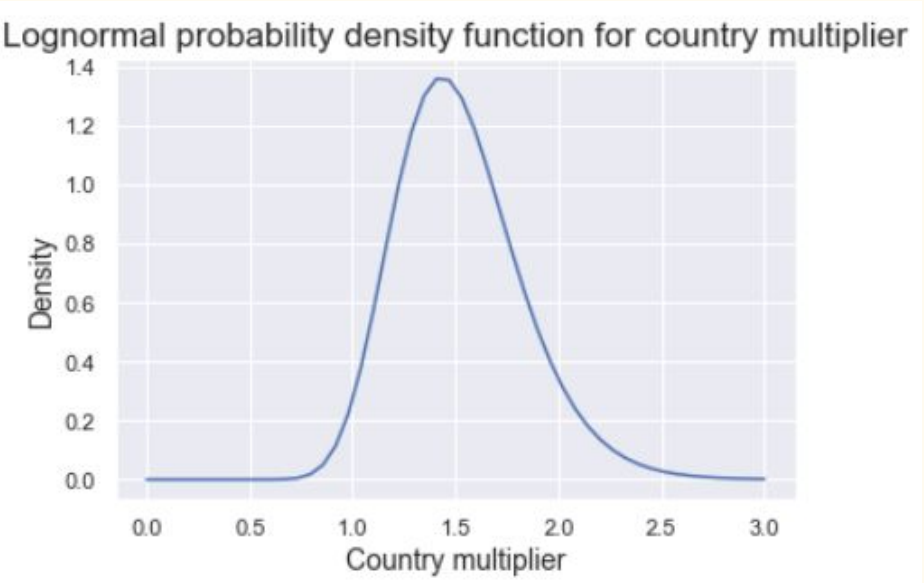


Store multiplier

- A positive real number (support > 0).
- We assume the prior to be centered approximately around 1.25 to meet the requirement of cheap stores having the multiplier less than 1 while mid-range and luxury stores having multipliers ranging between 1.1 and 1.8.
- Lognormal distribution with the mean 0.3 and standard deviation 0.2 seems like a reasonable representation of the requirements.

Model Setup (prior)

country_mult ~ lognormal(0.4, 0.2)



Country multiplier

- A positive real number (support > 0).
- We take into consideration the cost of life in the countries used in the model and conclude all of them are first-world countries, which means that the store multiplier should be higher than 1 for Germany, the UK and the USA.
- Lognormal distribution with the mean 0.4 and standard deviation 0.2 seems like a reasonable representation of the requirements.

Model Setup (likelihood)

Likelihood function

- Positive values (support > 0).
- Mean equals to the product of base price, store and country multipliers

prices ~ normal(base_prices[products_v[i]] * store_mult[perceptions_v[i]] * country_mult[countries_v[i]], sigma)

- Needs to capture price fluctuations that are caused by random factors other than the ones outlined above (standard deviation for the likelihood function).

sigma ~ gamma(2, 0.5)

Model Setup (data and parameters)

This section contains known values used by the model. Those include:

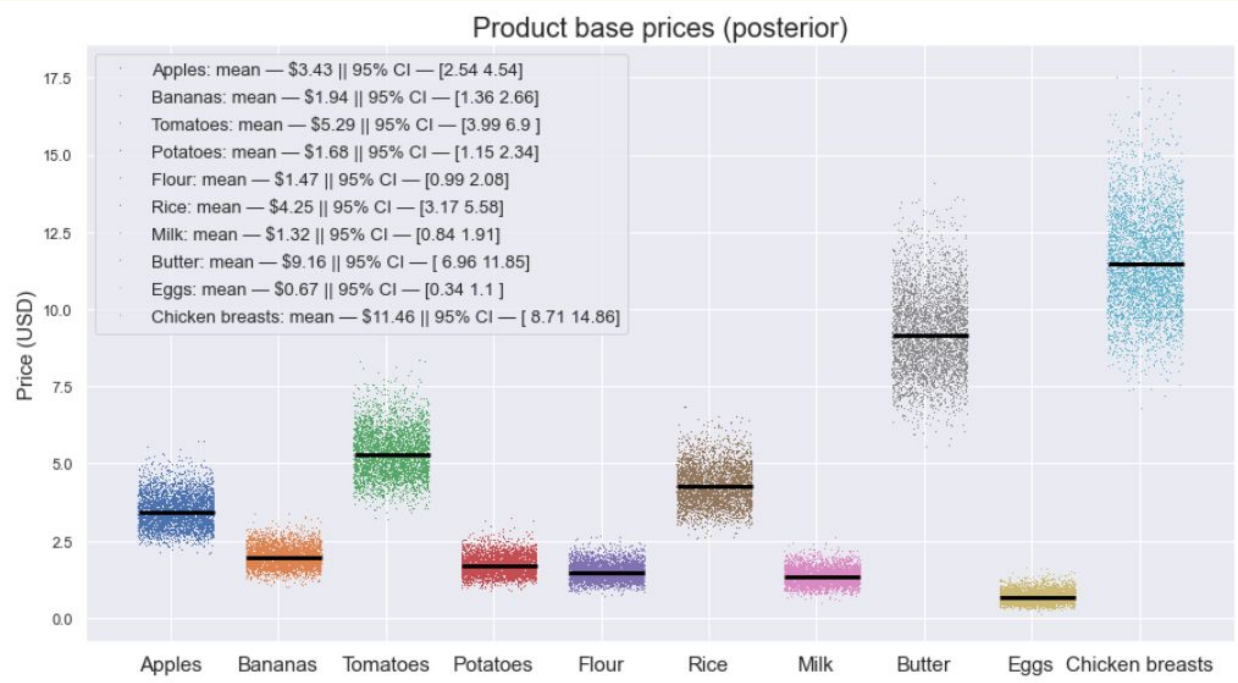
- Hyperparameters for priors and standard deviation of the likelihood function(alpha, beta)
- Number of products (10)
- Number of observations (38 after data preprocessing)
- Types of store perceptions (3)
- Number of countries (3)
- Vectors corresponding to every column mentioned above (size: number of observations)

We continue by listing all unknown quantities (parameters):

- Arrays of posterior base prices, store and country multipliers

The model calculates every unknown quantity and then combines them by taking their product and putting it as a mean of truncated normal distribution with the measure of uncertainty to produce the predictions for each product in every country and store.

Predictions (base prices)



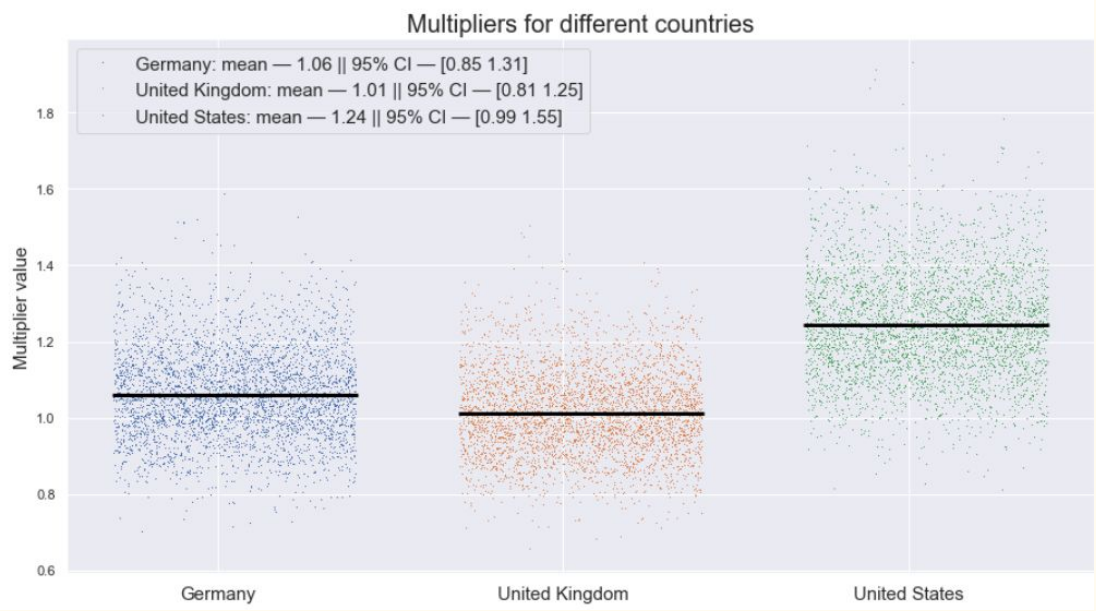
After careful analysis of the generated graph, we conclude that there is more uncertainty present (wider 95% confidence intervals) in the base prices of particular products (e.g., tomatoes, butter, chicken breasts). Thus, the prices for those items around the world might differ more than the prices for other groceries. Also, we can notice that butter and chicken are by far the most expensive products on our shopping list.

Predictions (store multipliers)

Looking at the generated plot, we can notice that the store multiplier values do not differ as much as we would expect. Interestingly, the values for mid-range and luxury stores are almost identical. There might be 2 explanations for this (where one does not exclude the other):

1. There is not enough data to create a proper distinction (prior influence).
2. The prices for the chosen products are not heavily impacted by the store type. This explanation seems sound because we have chosen a set of basic groceries that should not vary much (here, “much” means tens of dollars) between the stores.

Predictions (country multipliers)



Our analysis reveals that the final grocery prices are higher in the USA than in Germany or the UK because the multiplier for the USA is larger. However, we are surprised that all values are barely higher than 1. We think additional data with other countries will make the impact of country multipliers on the base price more explicit.

Correlation of price variation by store location and rental prices

To answer the question, we take 2 variables — country multiplier (a measure of price variation by store location) and average rental prices for each country — to calculate the correlation coefficient. Since there is a large range for rental prices (they vary by thousands of dollars) vs. average country multiplier values (they vary by ~ 0.1-0.2), we decided to normalize the average rental prices before obtaining the correlation coefficient.

The resulting value is: **r = ~0.51**

We can conclude that there is a moderate positive relationship between the above-mentioned variables, meaning that if the average rental prices increase, the average country multiplier values will also go up (with a different magnitude since the correlation is not perfect). However, since correlation is a suitable measure only for linear relationships, we need to prove that selected variables meet this requirement (e.g., create a scatter plot). Ideally, we would need more data to verify the linearity because our dataset is rather small.

Limitations and suggestions

Despite the model producing realistic results, here are some of its limitations:

1. Lack of data — it is difficult to make confident predictions with the given sample size.
2. Perception bias — different individuals perceive the store's "priciness" in different ways. For instance, it can depend on the income level, background, and other psychological/financial factors.
3. Data entry errors — despite removing the most obvious errors, we assume that there might still be some left. Even though the errors might be minor, they still have an impact on the model predictions because of the small dataset size.

The main suggestions from our side include:

1. Gathering additional data and incorporating it into the model.
2. Exploring how drastically the predictions change based on the change in prior beliefs.
3. Creating more methods that prevent the dataset from entry errors/improving the data entry form.