CS146 LBA: The Cost of Basic Goods

Huey Ning Lok

Minerva Schools @ KGI

Introduction

In this assignment, we model the cost of groceries in different neighborhoods of Berlin. Do product and store brands affect product prices, or not? Are grocery prices and the distribution of different grocery stores correlated with other cost of living measures in a city — for example, rent and real estate prices?

Importing and Pre-processing the Data

```
#import libraries
import pystan
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import scipy.stats as sts
sns.set()


#import data
data = pd.read_csv('shop.csv',encoding='latin-1')

#preprocess data for product brands
#get list of headers for product brand names
headers = [i for i in list(data.head(0))[1:] if i[-5:] == 'brand']

#search for all brand names containing the term 'bio'
bio_search = ['Bio', 'bio', 'BIO']

for i in headers:
    #replace all brand names containing the term 'no brand' with 1
    data.loc[data[i].str.contains('no brand',na=False), i] = 1
    #replace all brand names containing the term 'bio' with 2
    data.loc[data[i].str.contains('|'.join(bio_search),na=False), i] = 2
    #replace every other brand name with 3
    data.loc[data[i].str.contains('.*?',na=False), i] = 3

#take a peak at the data
data.head()
```

	Grocery store brand	Area	Apple 1 brand	Apple 1 price	Apple 2 brand	Apple 2 price	Apple 3 brand	Apple 3 price	Banana 1 brand	Banana 1 price	...	Eggs 2 brand	Eggs 2 price	Eggs 3 brand	Eggs 3 price	Chicken 1 brand	Chicken 1 price	Chicken 2 brand	Chicken 2 price	Chicken 3 brand
0	1	1	3	2.49	3	2.49	1	1.40	1	1.09	...	3	2.02	1	3.30	3	6.48	NaN	0.00	NaN
1	3	2	3	2.72	3	1.39	3	2.99	1	1.09	...	2	3.30	3	1.29	3	6.98	3	7.48	NaN
2	4	7	3	1.99	3	2.99	3	1.99	1	1.09	...	3	2.38	2	4.98	1	6.98	NaN	NaN	NaN
3	4	4	3	1.99	3	2.93	3	1.99	3	0.88	...	3	3.38	3	1.55	3	9.99	2	25.99	3
4	3	6	3	1.79	3	1.39	3	2.74	2	1.65	...	3	2.03	3	1.55	2	19.99	3	6.48	NaN

```
#define numerical code for base product types and multipliers
store_dict = { 1: 'Lidl', 2: 'Rewe', 3: 'Aldi', 4: 'Edeka' }

area_dict = {
1:"Mitte",2:"Schoneberg",3:"Neukooln",4:"Kreuzberg",5:"Friedrichshain",
             6:"Prenzlauer Berg",7:"Tiergarten",8:"Alt-Tempelhof",9:"Wedding",
             10:"Gesundbrunnen",11:"Moabit",12:"Rummelsburg",13:"Lichtenberg" }

type_dict = {1:"Apple",2:"Banana",3:"Tomatoes",4:"Potatos",5:"Flour",
             6:"Rice",7:"Milk",8:"Butter",9:"Eggs",10:"Chicken"}

brand_dict = {1:"No brand",2:"Bio brand",3:"Other"}


#preprocess data into lists of length: number of observed prices collected
prices = []
stores = []
types = []
areas = []
brands = []

for i in range(1,11):
    product = type_dict[i]
    price1 = data[[product+' 1 price',product+' 1 brand','Grocery store
brand','Area']].dropna()
    price2 = data[[product+' 2 price',product+' 2 brand','Grocery store
brand','Area']].dropna()
    price3 = data[[product+' 3 price',product+' 3 brand','Grocery store
brand','Area']].dropna()

    product_prices = price1[product+' 1 price'].values.tolist() + price2[product+'
2 price'].values.tolist() + price3[product+' 3 price'].values.tolist()
    product_stores = price1['Grocery store brand'].values.tolist() +
price2['Grocery store brand'].values.tolist() + price3['Grocery store
brand'].values.tolist()
    product_type = len(product_prices)*[i]
    product_area = price1['Area'].values.tolist() + price2['Area'].values.tolist()
+ price3['Area'].values.tolist()
    product_brand = price1[product+' 1 brand'].values.tolist() + price2[product+' 2
brand'].values.tolist() + price3[product+' 3 brand'].values.tolist()

    prices += product_prices
    stores += product_stores
    types += product_type
    areas += product_area
    brands += product_brand
```

```
#take a peak at the preprocessed data
print(prices[0:10])
print(stores[0:10])
print(types[0:10])
print(areas[0:10])
print(brands[0:10])
```

```
[2.49, 2.72, 1.99, 1.99, 1.79, 3.13, 1.99, 4.38, 1.52, 1.99]
[1, 3, 4, 4, 3, 1, 4, 3, 2, 4]
[1, 1, 1, 1, 1, 1, 1, 1, 1, 1]
[1, 2, 7, 4, 6, 1, 1, 3, 4, 3]
[3, 3, 3, 3, 3, 2, 3, 2, 3, 3]
```

Building the Stan model[1]

```
#Build the stan model
stan_code = """

// The data block contains all known quantities - typically the observed
// data and any constant hyperparameters.
data {

    int<lower=1> N;           // number of observed prices collected
    int<lower=1> T;           // number of different types of products
    int<lower=1> S;           // number of different types of stores
    int<lower=1> A;           // number of different types of areas
    int<lower=1> B;           // number of different types of brands

    // data collected
    real<lower=0> prices[N];
    int types[N];
    int stores[N];
    int areas[N];
    int brands[N];

    // fixed prior hyperparameters
    real<lower=0> alpha;
    real<lower=0> beta;

}

// The parameters block contains all unknown quantities - typically the
// parameters of the model. Stan will generate samples from the posterior
```

[1] **#probability:** I used Pystan to calculate posterior probability distributions over the base prices of different goods and store, brand and location multipliers.

```
// distributions over all parameters.
parameters {

    real<lower=0> base_price[T];       //vector containing base prices (number of
base prices == number of product types)
    real<lower=0> multiplier_store[S]; //vector containing store multipliers
    real<lower=0> multiplier_area[A];  //vector containing area multipliers
    real<lower=0> multiplier_brand[B];  //vector containing brand multipliers
    real<lower=0> sigma;              //random noise when sampling from the normal
distribution for observed prices

}


// The model block contains all probability distributions in the model.
// This of this as specifying the generative model for the scenario.
model {

    sigma ~ gamma(alpha, beta);      //generate random noise for normal
distribution

    for (i in 1:T) {
        base_price[types[i]] ~ cauchy(1.5, 2);         //generate base price
    };

    for (i in 1:S) {
        multiplier_store[stores[i]] ~ cauchy(1, 0.7);  //generate store multiplier
    };

    for (i in 1:A) {
        multiplier_area[areas[i]] ~ cauchy(1, 0.7);    //generate area multiplier
    };

    for (i in 1:B) {
        multiplier_brand[brands[i]] ~ cauchy(1, 0.7);    //generate brand
multiplier
    };

    for(i in 1:N) {
        prices[i] ~ normal(base_price[types[i]] * multiplier_store[stores[i]] *
multiplier_area[areas[i]] * multiplier_brand[brands[i]], sigma); //likelihood
function
  }

}

"""
```

```
#compile the model
stan_model = pystan.StanModel(model_code=stan_code)

#stan input data
stan_data = {
                'prices': prices,
                'types': types,
                'stores': stores,
                'areas': areas,
                'brands': brands,
                'N': len(prices),
                'T': len(set(types)),
                'S': len(set(stores)),
                'A': len(set(areas)),
                'B': len(set(brands)),
                'alpha': 1.5,
                'beta': 7
            }

# Fitting stan model to the data. This will generate samples from the posterior
over all parameters of the model.
stan_results = stan_model.sampling(data=stan_data)
print(stan_results)

# Extract the generated samples from the stan model
posterior_samples = stan_results.extract()
```

Stan Results

```
Inference for Stan model: anon_model_7c5195ec8ccd5cb3e87a60dfe62ebb75.
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]	2.17	0.02	0.47	1.35	1.84	2.12	2.45	3.17	891	1.0
base_price[2]	1.27	0.01	0.31	0.75	1.04	1.23	1.45	1.97	962	1.0
base_price[3]	3.27	0.03	0.75	1.99	2.73	3.18	3.72	4.91	882	1.0
base_price[4]	1.14	9.2e-3	0.29	0.67	0.94	1.11	1.32	1.78	980	1.0
base_price[5]	1.07	8.4e-3	0.28	0.61	0.88	1.03	1.23	1.71	1090	1.0
base_price[6]	2.89	0.02	0.67	1.74	2.41	2.82	3.31	4.35	904	1.0
base_price[7]	0.96	8.2e-3	0.25	0.54	0.78	0.93	1.12	1.55	963	1.0
base_price[8]	4.23	0.03	0.96	2.6	3.55	4.11	4.84	6.34	875	1.0
base_price[9]	2.59	0.02	0.6	1.59	2.16	2.52	2.95	3.91	869	1.0
base_price[10]	10.67	0.08	2.42	6.58	8.95	10.38	12.13	15.96	861	1.0
multiplier_store[1]	0.72	5.7e-3	0.19	0.39	0.58	0.7	0.83	1.15	1141	1.0
multiplier_store[2]	1.08	8.5e-3	0.29	0.59	0.88	1.05	1.25	1.72	1146	1.0
multiplier_store[3]	0.97	7.6e-3	0.26	0.53	0.79	0.95	1.12	1.54	1164	1.0
multiplier_store[4]	1.07	8.4e-3	0.29	0.59	0.88	1.05	1.24	1.69	1144	1.0
multiplier_area[1]	1.32	6.1e-3	0.18	1.02	1.2	1.31	1.43	1.73	843	1.0
multiplier_area[2]	0.9	4.3e-3	0.14	0.65	0.8	0.88	0.98	1.22	1101	1.0
multiplier_area[3]	0.92	4.3e-3	0.13	0.7	0.83	0.91	0.99	1.2	869	1.0
multiplier_area[4]	1.29	6.0e-3	0.17	0.99	1.16	1.27	1.39	1.68	845	1.0
multiplier_area[5]	1.3	6.2e-3	0.18	1.0	1.18	1.29	1.41	1.71	817	1.0
multiplier_area[6]	1.16	5.3e-3	0.16	0.88	1.05	1.15	1.25	1.52	897	1.0
multiplier_area[7]	1.12	5.3e-3	0.16	0.83	1.0	1.1	1.21	1.49	943	1.0
multiplier_area[8]	0.93	4.6e-3	0.15	0.67	0.82	0.91	1.02	1.25	1017	1.0
multiplier_area[9]	1.48	7.8e-3	0.27	1.0	1.3	1.46	1.64	2.07	1182	1.0
multiplier_area[10]	0.9	4.9e-3	0.17	0.6	0.79	0.89	1.01	1.29	1235	1.0
multiplier_area[11]	1.41	6.8e-3	0.2	1.06	1.27	1.39	1.52	1.88	901	1.0
multiplier_area[12]	1.13	5.4e-3	0.18	0.83	1.01	1.11	1.23	1.52	1044	1.0
multiplier_area[13]	1.0	5.2e-3	0.19	0.66	0.86	0.98	1.11	1.42	1376	1.0
multiplier_brand[1]	0.89	8.5e-3	0.27	0.43	0.69	0.86	1.05	1.51	1022	1.01
multiplier_brand[2]	1.6	0.02	0.49	0.78	1.25	1.56	1.9	2.69	1024	1.01
multiplier_brand[3]	0.86	8.1e-3	0.26	0.42	0.67	0.84	1.02	1.45	1019	1.01
sigma	1.6	6.0e-4	0.03	1.54	1.58	1.6	1.62	1.66	2553	1.0
lp__	-1366	0.1	3.99	-1375	-1368	-1366	-1363	-1359	1558	1.0

```
Samples were drawn using NUTS at Mon Nov 12 15:20:16 2018.
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).
```

Modeling Assumptions[2]

The half-Cauchy distribution was chosen as the generative prior to the base prices, store multipliers, area multipliers, and brand multipliers. This choice was motivated by the Cauchy's heavy tails, which represent a broad prior distribution where the probabilities assigned to values moving away from the mean approach 0 at a slower rate.

The store, area, and brand multipliers are centered around 1 to represent an average multiplier effect of 1. The scale parameter chosen was 0.7 to represent an averagely broad prior.

The base price prior half-Cauchy was centered around 1.5 since the data was cleaned to only contain stores from Berlin, and so a stable currency unit of Euro was safely assumed. Furthermore, the products consist of everyday goods that one would not expect to exceed 10 euros at the highest. 2 was chosen as the scale parameter to represent a broader prior, since there can be a large difference in the price range between meat and vegetables.

The normal distribution was chosen as the likelihood function, with the equation:
$f(x)$ = base_price x multiplier_store x multiplier_area x multiplier_brand as the mean and σ as the standard deviation. As the observed prices should not consist of negative numbers, the validity of using the normal distribution as the likelihood was tested by using the lowest observed price from the dataset as the mean, and multiplying it by the expected value of σ.

σ was was generated from the prior gamma distribution of α=1.5 and β=7. Gamma(α=1.5, β=7) has an expected value μ = 0.2143, and σ = 0.175 (fig 1), which is a realistic amount of noise that can be expected from the dataset.

The lowest observed price from the dataset, 0.3, was used as the mean of the normal distribution. Given a normal distribution with μ = 0.3 and σ = 0.2143 , there is only a ~0.08 probability of a random variable X being less than 0 (fig 2). Hence, the usage of the normal distribution as the likelihood function is justified since there is a low probability of sampling a negative value.

Figure 1 and figure 2 show the relevant gamma and normal distributions plots, along with the relevant statistics. The graphs were made using the University of Iowa's online probability distribution calculator.[3]

---

[2] **#distributions:** I identified appropriate distributions given the context of possible multiplier ranges and base prices of grocery goods, and justify my choices for each distribution chosen.

[3] Bognar, M., Ph.D., n.d., Probability distributions calculator. Retrieved from https://homepage.stat.uiowa.edu/~mbognar/

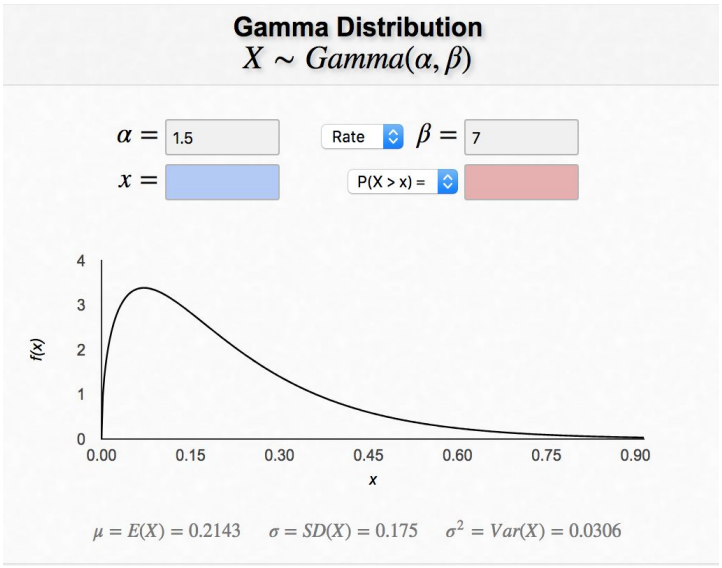


Figure 1: Gamma distribution with α=1.5 and β=7.
The statistics obtained are displayed at the bottom of the plot.

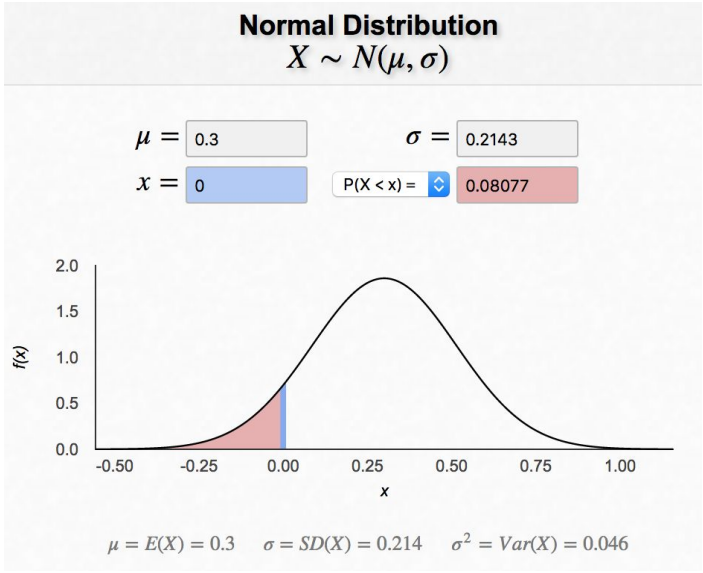


Figure 2: Normal distribution with μ = 0.3 and σ = 0.2143.
The statistics obtained are displayed at the bottom of the plot.

Other Assumptions

Human error

Some variation in the data will be purely due to human error. The data was collected by different individuals who probably went about the process with different methodologies. Furthermore, while instructions were given to record prices according to set units (1kg, 1 dozen, etc.), there is a high chance of omitting this step. Some of the data could also have been falsely constructed since there is no guarantee that the individuals actually carried out the data collection process.

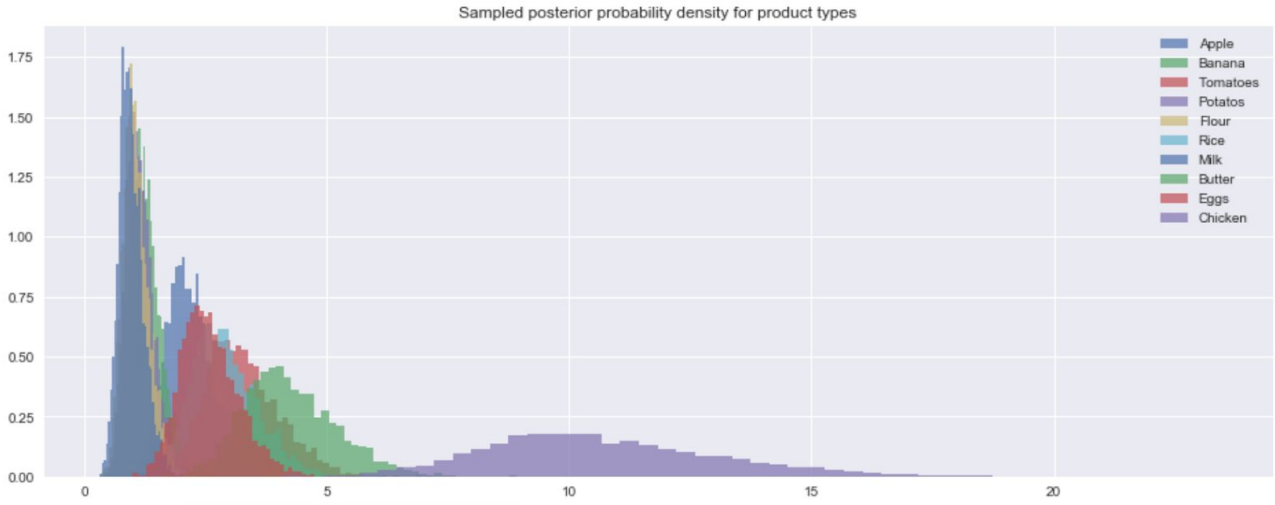
Product brand categorization

The product brands were separated into 3 distinct categories: No brand, Bio brand, and Other. It was assumed that all brands that contained the term "bio" would most likely share similar traits, and so they were placed within the same category. Due to the large variety of other brands, the "non-bio" brands were categorized together as "Other". Products with no brand recorded were categorized as "No brand".

Analysis and Discussion[4]

The basic average price for each product:[5]

	Label	Mean	Posterior 95% confidence interval	Standard Error
0	Apple	2.17	[1.3492458977100026, 3.162575776762137]	0.47
1	Banana	1.27	[0.7499940957868625, 1.9707530183460489]	0.31
2	Tomatoes	3.27	[1.9862708252062764, 4.908867958395163]	0.75
3	Potatos	1.14	[0.6711035347194378, 1.7810612135920565]	0.29
4	Flour	1.07	[0.6148647938107232, 1.7051929636281271]	0.28
5	Rice	2.89	[1.745956577960827, 4.3526133377586245]	0.67
6	Milk	0.96	[0.542958582073606, 1.5466058958416689]	0.25
7	Butter	4.23	[2.60713261102745, 6.340711619185721]	0.96
8	Eggs	2.59	[1.5906415250531818, 3.9070481150865564]	0.60
9	Chicken	10.67	[6.58679851784204, 15.958343016975801]	2.42



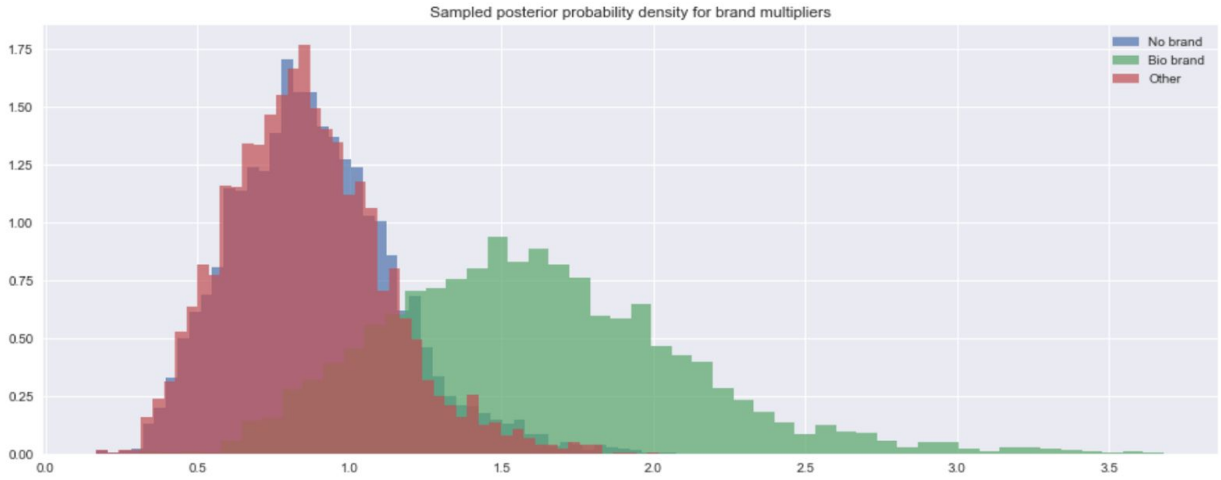
[4] **#confidenceintervals:** I incorporated 95% confidence intervals into my table results as I understood that while the mean of the distributions gives us a simple-to-reference metric of the distribution's average value, the confidence interval showed us how certain we are of our distribution by providing the range of values that we can be 95% confident that the population mean lies within. For example, we can be more certain of the values for milk and potatoes due to the small CI interval and less certain of the values for chicken due to the wide CI interval.

[5] All code for the graphs can be found in the attached Jupyter notebook or the GitHub link in the Appendix.

Effect of Various Factors on the Basic Price of the Product

Brand of the Product

	Label	Mean	Posterior 95% confidence interval	Standard Error
0	No brand	0.89	[0.4289754349616492, 1.5112206570961664]	0.27
1	Bio brand	1.60	[0.7807323112272733, 2.6923204672803607]	0.49
2	Other	0.86	[0.4170854886936737, 1.4505228437240378]	0.26



From the graph and table above, we see that Bio brand has the highest mean multiplier effect of 1.60, whereas No brand and Other brands have a virtually similar mean multiplier effect of 0.89 and 0.86 respectively. This can also be visually inferred through the graph - the posterior distributions for No brand and Other have a high area of overlap.

T-tests for the brand multipliers:[6]

- No brand is significantly cheaper than Bio brand
- No brand is significantly more expensive than Other
- Bio brand is significantly more expensive than No brand
- Bio brand is significantly more expensive than Other
- Other is significantly cheaper than No brand
- Other is significantly cheaper than Bio brand

[6] **#significance:** Although intuitive inferences can be made from observing the plot of distributions, I used t-tests to analytically calculate the statistical significance of the difference in multiplier effects, where the p-value is set at 0.05. The t-test is arguably a more accurate way of determining the difference in multiplier effects since it considers the sample size and variance of the datasets. The t-test is used here since there are a small number of comparisons to be made, and so the significance statements can be easily interpreted.

Brand of the Grocery Store

	Label	Mean	Posterior 95% confidence interval	Standard Error
0	Lidl	0.72	[0.39135053828963273, 1.145289972071517]	0.19
1	Rewe	1.08	[0.5930721067094563, 1.7219251209448216]	0.29
2	Aldi	0.97	[0.5351883347630948, 1.5357320718311804]	0.26
3	Edeka	1.07	[0.5879480203474513, 1.6875388867760923]	0.29

From the graph and table above, we see that Rewe and Edeka have the highest mean multipliers of 1.08 and 1.07 respectively, whereas Lidl has the lowest mean multiplier of 0.72.

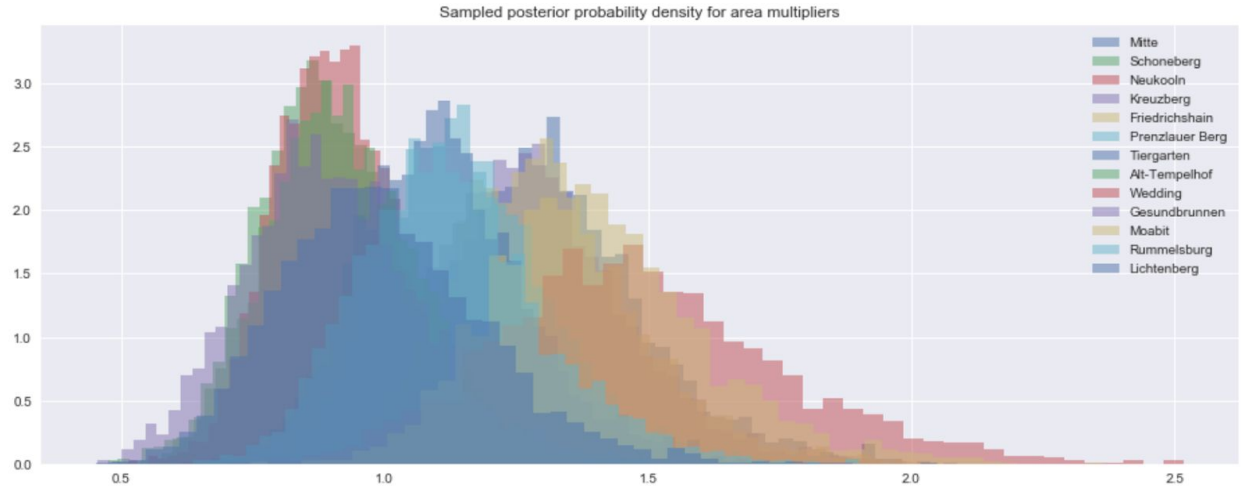
There is a high overlap area between the posterior distributions for Rewe, Edeka, and Aldi, whereas Lidl's distribution mean can be observed to peak to the left of the other 3 distributions.

T-tests for the store multipliers:

- Lidl is significantly cheaper than Rewe
- Lidl is significantly cheaper than Aldi
- Lidl is significantly cheaper than Edeka
- Rewe is significantly more expensive than Lidl
- Rewe is significantly more expensive than Aldi
- Rewe is not significantly more expensive than Edeka
- Aldi is significantly more expensive than Lidl
- Aldi is significantly cheaper than Rewe
- Aldi is significantly cheaper than Edeka
- Edeka is significantly more expensive than Lidl
- Edeka is not significantly more expensive than Rewe
- Edeka is significantly more expensive than Aldi

Geographical Location of the Grocery Store

	Label	Mean	Posterior 95% confidence interval	Standard Error
0	Mitte	1.32	[1.0190909835186404, 1.7286391492890913]	0.18
1	Schoneberg	0.90	[0.6521564402537261, 1.2209006288385988]	0.14
2	Neukooln	0.92	[0.69765425852583, 1.1973346593351333]	0.13
3	Kreuzberg	1.29	[0.985661704684716, 1.684046122082647]	0.17
4	Friedrichshain	1.30	[0.9970416973838953, 1.7060082486675912]	0.18
5	Prenzlauer Berg	1.16	[0.8817167128311205, 1.5189216447033138]	0.16
6	Tiergarten	1.12	[0.8341644230801579, 1.4881559553269486]	0.16
7	Alt-Tempelhof	0.93	[0.6673330895318058, 1.2536873646793893]	0.15
8	Wedding	1.48	[1.0056344988557337, 2.0689253888704733]	0.27
9	Gesundbrunnen	0.90	[0.6056233438688825, 1.284809136665662]	0.17
10	Moabit	1.41	[1.0583677190190437, 1.8807586344419254]	0.20
11	Rummelsburg	1.13	[0.8254087357041596, 1.5153977056813555]	0.18
12	Lichtenberg	1.00	[0.6623953893926257, 1.4170242334889591]	0.19



From the graph and the table, we see that most of the neighborhood multiplier posterior distributions overlap, and the multiplier means lie within the range of 0.90 (Schoneberg & Gesundbrunnen) on the low end, and 1.48 (Wedding) on the high end.[7]

[7] **#dataviz:** I plotted the distributions and used their graphical positions, spread and overlap in order to gain intuitive insights and inferences into the significance of various multiplier effects. While the table provides relevant descriptive statistics of the distributions (mean, 95% CI and standard error), by plotting my results graphically, I am able to easily convey the message of my findings to an audience, and better understand it myself.

The Strength of Brand and Location Multiplier Effects and their Influence on Price Variation between Shops

The strength of the brand and location effects have been illustrated above, with a higher mean multiplier representing a stronger multiplier effect, and a lower mean multiplier representing a weaker multiplier effect. The mean is used to represent the multiplier effect with an understanding that the certainty of the mean value varies with the distribution variance, and so the multiplier effect differences have to be interpreted within the appropriate context of variance and 95% CIs as shown in the tables above. The summary is as follows:

Brand multiplier

- Highest multiplier mean for brand: Bio brands (1.60)
- Lowest multiplier mean for brand: Other (0.86)
- The range of the means is 1.60 - 0.86 = 0.74. This means there is an average of 74% difference in mean price depending on whether you buy no brand, bio brand, or other.

Store multiplier

- Highest multiplier mean for store: Rewe (1.08)
- Lowest multiplier mean for store: Lidl (0.72)
- The range of the means is 1.08 - 0.72 = 0.36. This means there is an average of 36% difference in mean price depending on whether you buy from Lidl, Rewe, Aldi or Edeka.

Area multiplier

- Highest multiplier mean for area: Wedding (1.48)
- Lowest multiplier mean for area: Schoneberg & Gesundbrunnen (0.90)
- The range of the means is 1.48 - 0.90 = 0.58. This means there is an average of 58% difference in mean price depending on which neighborhood the product is located within.

Assuming that the means of the various distributions can be safely generalized to represent the multiplier effect (despite differing variances), the brand multiplier has the greatest influence on price variation between shops (range of means = 0.74), whereas the store multiplier has the lowest influence on price variation between shops (range of means = 0.36).[8]

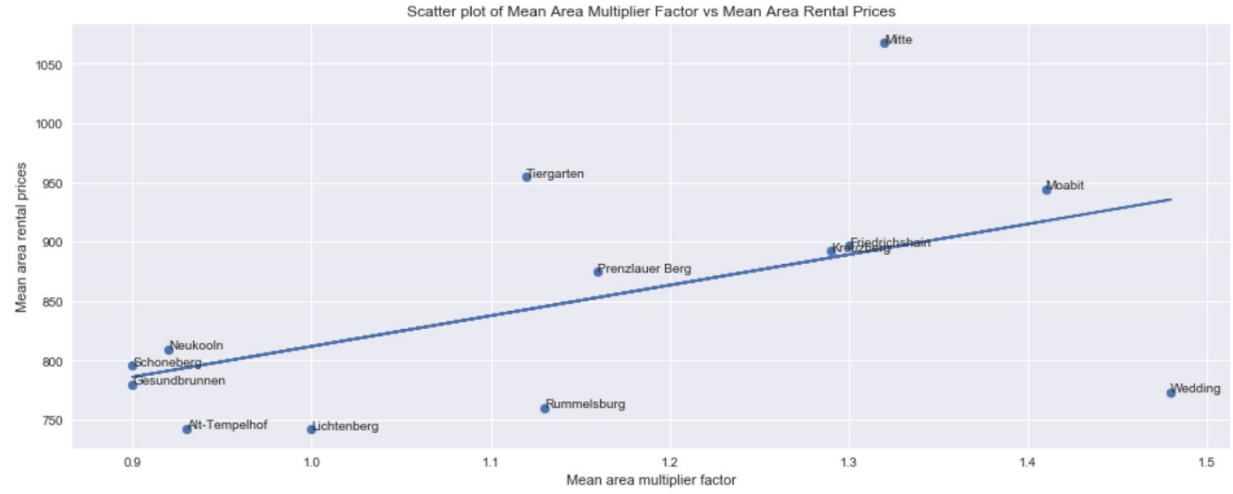
[8] **#descriptivestats:** I considered the relevant descriptive statistics calculated in the earlier section when justifying the method that I chose to measure the difference between multiplier effects. I acknowledged that using the distribution mean alone to represent the multiplier effect could only provide a rough estimation of the difference in multiplier effects, and I encourage the audience to refer to the variance and confidence intervals in the tables above in order to determine the certainty of the estimation. I intentionally avoided using a more analytical method of measuring difference significance, e.g. t-tests, due to the high number of comparisons that would need to be made, which could potentially decrease the interpretability of my results.

Correlation of Price Variation by Geographical Location vs Rental Price Variation in Berlin

Rental price data was obtained from the 2017 map of rental prices by UBahn and SBahn station.[9]

	Label	Mean	Rental price means
0	Mitte	1.32	1067.60
1	Schoneberg	0.90	795.80
2	Neukooln	0.92	809.60
3	Kreuzberg	1.29	892.60
4	Friedrichshain	1.30	896.60
5	Prenzlauer Berg	1.16	874.50
6	Tiergarten	1.12	954.67
7	Alt-Tempelhof	0.93	742.67
8	Wedding	1.48	773.33
9	Gesundbrunnen	0.90	779.80
10	Moabit	1.41	944.50
11	Rummelsburg	1.13	760.00
12	Lichtenberg	1.00	742.33

```
Pearson Correlation Coefficient: 0.5283170763686273
```



The mean area rental prices and mean area multipliers show a positive correlation, with a Pearson Correlation Coefficient of 0.53.

Some of the more prominent outliers include Wedding and Mitte, whose x,y points are (1.48,773.33) and (1.32, 1067.60) respectively. These two neighborhoods lie on opposite ends of

[9] https://www.immobilienscout24.de/content/dam/is24/ibw/dokumente/mietmap-berlin-2017.jpg

the spectrum - with Wedding having an unusually low mean rental price given its corresponding multiplier effect, and Mitte having an unusually high rental price given its corresponding multiplier effect.

While the positive correlation between area multiplier factor and area rental price seems intuitive, it has to be noted that the data collection method for the rental prices were rather crude. I hand-picked prices from the map by googling train stations that were closest to a given neighborhood, and selected the 5 closest neighboring stations to get the neighborhood's mean rental price. In cases where the train stations were rather isolated from other stations (low train station density), I would select fewer stations for the calculation of the average rental price.[10]

[10] **#correlation:** I identified the positive correlation between mean area multiplier factor and mean area rental prices by calculating the Pearson correlation coefficient. I noted that the positive correlation could be weaker than observed due to the shallow data collection method. The small number of data points means that the outliers might not be irregularities, but simply neighborhoods that are underrepresented in the data. As such, it might be appropriate to assign the outliers a higher weight, which would alter the correlation direction. Furthermore, the dataset only consists of a small number of neighborhoods. An extension of the linear regression line might reveal a different trend.

Appendix

Code:
https://github.com/hueyning/CS146-repo/blob/master/Lba/CS146%20Location%C2%AD-based%20Assignment.ipynb

Raw data:
https://github.com/hueyning/CS146-repo/blob/master/Lba/shop.csv

Store: ALDI, Ostseestraße 25
Visited: 10/27/2018 13:24:06

Store: EDEKA, Annenstraße
Visited: 10/29/2018 13:55:00