Bogota - Coding 1 Team Project

Bogota team members: Natalia Iriarte, Péter Kaiser, Rauhan Nazir, Sára Vargha, Xibei Chen

2021/11/6

Introduction

motivating and explaining your descriptive statistics (2-3 sentence)

Data are available from: https://github.com/kanyipi/DA1/blob/main/bogota_data.csv

Data

Our main variable is the price in HUF of 2 products: regular Coca Cola 0.5l plastic bottle and Nutella 400g. The collection was carried out by anonymous visit to randomly selected stores in districts 9 and 16. (Three stores are on the edge of districts 8 and 9. But we decided to keep the data in the table as they are still valuable for our current analysis purpose.) Apart from that, we measured the busyness of the stores using number of people waiting in queue divided by the number of cashiers. In addition, the data table also includes information on the position of both products on the shelf and if they had a discount at the time of the visit, for these we used the observation method. As a complement we include qualitative information about the store (name, type, address, district and nationality).

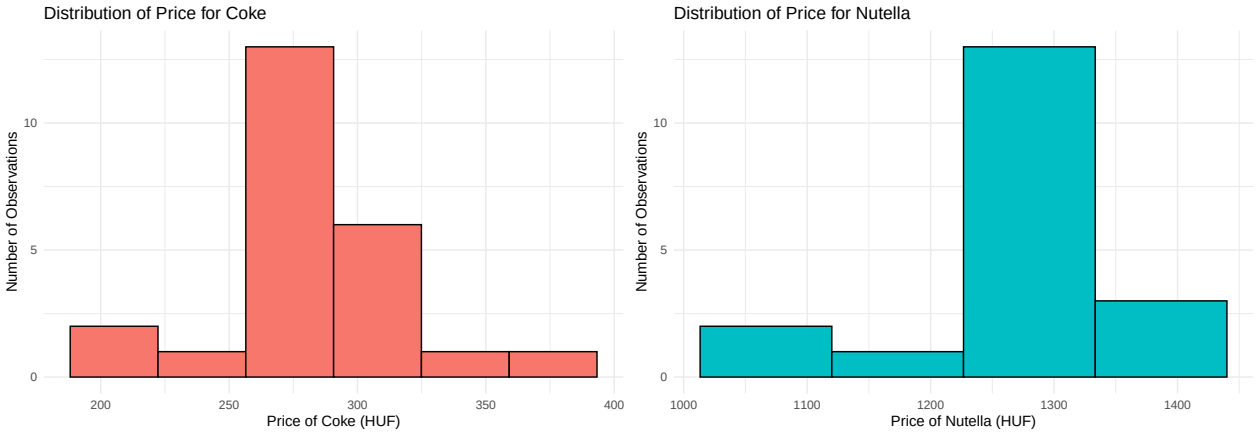
Table 1: Descriptive Statistics of Prices

Variable	Mean	Median	SD	Min	Max	Range	P05	P95	N	Missing
Price of Coke	279.33	269.5	36.54	219	390	171	223.65	334.65	26	2
Price of Nutella	1272.05	1289.0	80.23	1099	1419	320	1099.90	1374.00	26	7

We have also calculated the correlation coefficient between the price of Coke and Nutella, which is 0.36, suggesting statistically positive association between the price of both products, but rather a weak one (not as high as we assumed).

Distributions

Price of Coke and Nutella



We chose histogram over density plot, as we find histogram more of a straightforward interpretation of the distribution of the prices. We want to avoid a broken comb look by having too many bins. We think the numbers of bins we settled with portray the distribution quite well.

Busyness of Stores in Inner and Outer City

We were also curious about whether the distribution of busyness of stores are different between inner and outer city, so we created the following boxplot.

Conclusion

- The distribution of Nutella prices is way more spread out than Coke prices with over twice the standard deviation. With less missing values in Coke prices, we can tell stores are more likely to sell Coke than Nutella.
- The conditional mean of busyness is a bit lower in outer city than inner city. However, the busyness in outer city is more spread out, whereas it is more centralized in inner city.