DA1 Assignment 2021 Fall, Lima Group

During the **preparation phase**, our group held a meeting to clarify the methodology of data collection. We decided to validate the pricing of a .5L plastic Coca Cola bottle. We then decided on a secondary product that is widely available and easily recognizable. After considering different types of chocolate brands and the podtential for confusion on variety, we settled with an easily recognizable Snickers Bar (50 grams) since unlike other brands, this item does not have many variations. It also was easy to identify since only ¼ of our group speaks Hungarian and this may make it difficult to decipher local candy bars. Secondly, we defined variables for the shops that are easy to measure and as objective as possible. We decided to measure 1) the size of shops (text), 2) number of cashiers (numeric), and 3) chain (binary). For the size of the shops, since we cannot easily measure shopping area or building size, we used features that we think indicate size. For example, a shop was considered small if there were only hand baskets, a shop was medium if there were push carts, and finally a shop was large if there was an attached parking lot. Additionally, we recorded the number of cash registers since this could be an important factor of popularity, and consequently of price. Our last variable, chain, tells whether the shop is part of any national or global brands. We wanted to see if this aspect correlates with price. We deliberately included variables with different data types (text, numeric, binary) as we think this makes for a better practice table. Lastly, we chose target districts VIII. and XX. Both being on Pest side may make for better comparison of inner-outer city which would reduce any complexities with a Pest-Buda comparisons.

Even with well-defined goals, we encountered several questions during the **collection phase**. Regarding our **variables**, initially we didn't clarify whether to record the number of cashier employees or the number of cash register machines. We had to clarify during our data collection phase to record the machines to avoid introducing bias for rush-hour increase in employee numbers. Cash register numbers are very unlikely to change over the few hours of our data collection. Additionally, at one store, we encountered self-service cash registers, we decided to include these as a cash register since they were open and used during our stay. Removing them from our totals would likely distort the data more. As mentioned, our size variable worked as intended since only the biggest shopping center had an attached parking lot. Unfortunately, this was the only large store we surveyed in our districts so we do not have additional data for comparison. This may affect the results we see when comparing prices against store sizes.

The largest issue we see from our data collection experience is the actual sampling and **selection of store locations**. We split into two separate groups of two individuals to "randomly" walk around the districts to gather data from stores. However, this approach led to selection bias. For example, stores in the VIII district were clustered together in a few high-density blocks which made it easy to gather data from 2-3 stores at once. We did not explore stores in the outer limits of the VIII district which may have different pricing than the stores we surveyed which were close together. A truly random way of sampling stores would be to have an unbiased approach to choosing the stores before hand perhaps via an algorithm ingesting Google Maps API data so we did not choose the stores ourselves based on proximity to the Metro, other stores, and more.