

# Team Assignment - Our Fast Food Journey

The very best team - Group G

2020-10-24

## Introduction and data collection

We set out to collect information on 32cm Margherita pizza and 0.5l cola prices from various locations within Hungary by web-scraping the menus available on one of the biggest Hungarian food delivery sites called NetPincer. All codes and files used for this analysis are available in our github repo.

We chose web-scraping in order to reach the highest possible data coverage without selection bias and also to have high data-quality by reducing the chances of manual error. We scraped data from Pizzerias from the 5 largest Hungarian cities (Budapest, Debrecen, Szeged, Miskolc and Pecs). We collected every information that was available on the site (e.g. rating, number of ratings, address) on the restaurants, their advertised products and created geo locations from their addresses so that we can calculate their distance from the city center in an automated manner.

We had 3 main challenges when cleaning the scraped data. We needed to **identify the correct product types** and We dropped all observations that were not Margherita pizzas or Colas. We also had to **identifying the exact products with exact features** which was not easy since Margherita pizzas existed in the data with different features (e.g. low carb, vegan, thin crusted, with olives etc...). We eliminated all observations with such extra features. After that we **looked at 32cm and 0.5l containers only** and dropped observations with different sizes.

We tackled the above only by employing a number of transformations on a string variable which was the product description on the site. We also dropped any resulting duplicates to arrive to a tidy table. By that we entity-resolution and misambiguation thus had relatively high data quality. Unfortunately many restaurants didn't serve beverages or they served it in different sizes - in these cases we marked the beverage price with NA.

## Descriptive statistics

Our dataframe holds 91 restaurants with Margherita pizza prices out of which only 28 served cola beverages in 0.5l containers as visible in Table 1. Figure 1 visualises the distribution of each price variable, from which we can see that the density plots are right-skewed with a local and a global maximum. We can see in Figure 2 that the two maximum exist due to average prices being generally higher in our Budapest sample than in the other large cities.

variable	mean	median	std	iq_range	min	max	skew	numObs
Pizza prices	1693.23	1690.00	345.51	555.00	790.00	2790.00	0.36	91
Beverage prices	431.36	400.00	81.62	142.25	320.00	599.00	0.53	28

Table 1: Summary statistics of all observations

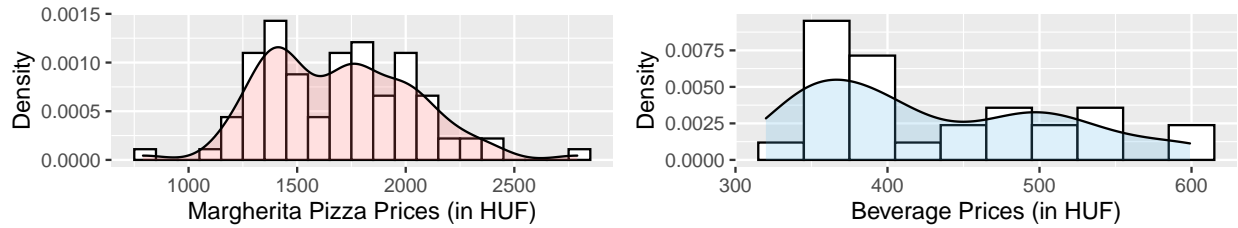


Figure 1: Price distribution for beverages and pizzas for the whole population

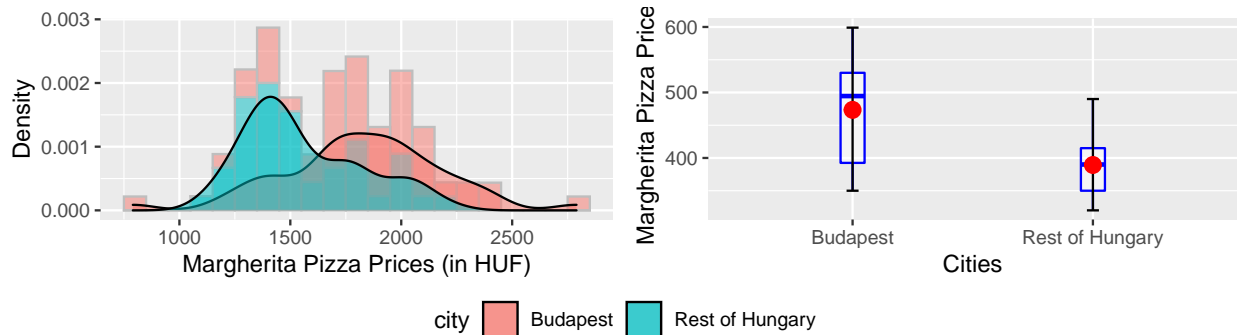


Figure 2: Price distribution of pizza and beverage prices in Budapest and Rest of Hungary

## Correlation

In addition we made some experiments on correlations, which is available in the file correlation.pdf.

## Hypothesis testing

### Two sample t-test - Analyzing 32 cm Margherita pizza prices

**Question:** Is the average price of marheritha is the same in Budapest vs. Rest of Hungary (other big cities: Debrecen, Szeged, Pécs, Miskolc)

**H0:** *avg. price of pizza in Budapest - avg. price of pizza outside of Budapest == 0*

**H-Alternative:** *avg. price of in Budapest - avg. price outside of Budapest != 0*

We test for equality of average prices. We can reject N0, if  $p < 0.05$ .

**Result:**  $t = 4.0079$ ,  $df = 82.502$ ,  $p\text{-value} = 0.0001337$ , 95 percent confidence interval:  $[134.9711; 400.9516]$ . The p-value is way smaller than the defined 5%, hence, we CAN reject the Null hypothesis. The probability of making a false positive error is about 0.013%. True difference in means is not equal to 0, hence avg. price of pizza in Budapest vs. avg. price of pizza outside of Budapest differs significantly.