# Team Assignment - Our Fast Food Journey

The very best team - Group G

2020-10-24

## Introduction

We set out to collect information on 32cm Margherita pizza and 0.5l cola prices from various locations within Hungary by looking at the menus available on one of the biggest Hungarian food delivery sites called NetPincer at https://www.netpincer.hu/. In this document we first describe the data collection steps which entailed web-scraping and introduce the main characteristics of our variables. After that, we analyze how strong the connection is between our variables and we will try to see if there is any significant driver when it comes to pricing. Finally we will test whether the average price of the pizza products in our sample has a statistically significant difference when comparing prices in Budapest vs prices in the rest of Hungary. All codes and files used for this analysis are available in our github repo.

## Data collection

Our approach to collecting data was web-scraping with the goal of creating two relational data tables, one for restaurants, and another for products. We chose web-scraping to reach a the highest possible data coverage without selection bias and also to have high data-quality by reducing the chances of manual error. In our population we included pizza places only from the 6 largest Hungarian cities. As for the structure of the webpage, one has to add a location to which they want the orders to be delivered, which is something we highly leveraged. We used these options to narrow down our scope in two ways:

- We included restaurants from the chosen Hungarian cities only: Budapest, Debrecen, Szeged, Miskolc and Pecs.
  - For Budapest, we filtered for those restaurants that deliver to CEU (Central European University)
  - For the rest, we filtered for restaurants delivering to each city center
- The webpage allows the user to filter for a given type of restaurants - we chose Pizzerias.

Our code run through each webpage and collected data on every restaurant available (given the above filters) and every product of these restaurants that they advertised on the page. We created two relational data tables, namely:

- *all_restaurants_v3_w_dists.csv* which covers restaurant level information with several scraped and generated variables. We scraped the the name and address, the average ratings of users (this is an ordered qualitative variable with a scale of 1-5) and the number of given ratings. We further created geo locations (longitudes and latitudes) of the restaurants and their corresponding city center. We did this not only to calculate the km distance between restaurant vs city center, and restaurant vs CEU building, but also to be able to visualize these restaurants on a map chart.

- *all_products.csv* which has the name of the restaurant, the name of the product and the price of each product available as indicated on the website. These are saved down in our raw data folder.

We joined our relational data tables and started data cleaning where we encountered multiple challenges. All traces of our efforts can be found in the *data_cleaning.R* code, available in our repo. Our task was to find specific products in a specific size, whereas our data has multiple different products with varying features and in various sizes. And all we had was an indication of these in a string variable, which was the name of the product on the site (there wasn't any standardized naming convention for these products either - restaurants chose to update their products in any way they wanted). To sum it up, the main challenges were the following:

- **Identifying the correct product types:** We dropped all observations that were not Margherita pizzas or Colas. An extra layer of difficulty was the name difference of Margherita in English and Hungarian (Margherita vs Margareta) and the fact that discounted menu packages existed that existed as a separate observation but they were the combination of 2-3 products.
- **Identifying the exact products with exact features:** 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.
- **Looking at 32cm and 0.5l containers** Size of pizza and beverage was only available (if available at all) somewhere within the product description. Fortunately we could easily split this out into a variable since in most of the cases this was indicated at the very end of the string.

Following the above three principles and removing resulting duplicates we achieved entity-resolution and disambiguation thus relatively high data quality. Unfortunately we couldn't get 0.5l beverage price for every restaurant in our final dataframe since many restaurants either didn't serve beverages or they served in different sizes only.

# Exploratory data analysis

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 visualizes the distribution of each price variable, from which we see an interesting pattern. Both density plots look to be right-skewed, but what is even more interesting is that they seem to have two quasi-peaks, one higher peak at around 1,500 HUF for pizzas and around 350 HUF for cola beverages and a smaller one at around 1700 HUF for pizzas and 500 HUF for beverages.

| 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

We thought of two possible explanations that can explain this pattern. The first one has to do with what we assumed form the beginning; pizza and beverage prices on average might be higher in Budapest than in the other large cities. The second less likely explanation is that our population can be divided into two types of restaurants - one cheaper and probably worse quality, and one more expensive. We will explore if our first explanation makes sense.

As visible from Table 2, the mean pizza and beverage prices are much lower outside Budapest. We can not really comment yet on a general pattern, especially that the standard deviation of pizza and beverage prices are quite high in the Budapest sample. This difference might be by chance only, but it also signals that we might be on the right track. From this table we can also learn that the number of observations are quite symmetric, we have almost as many observations from Budapest as from the Rest of Hungary.

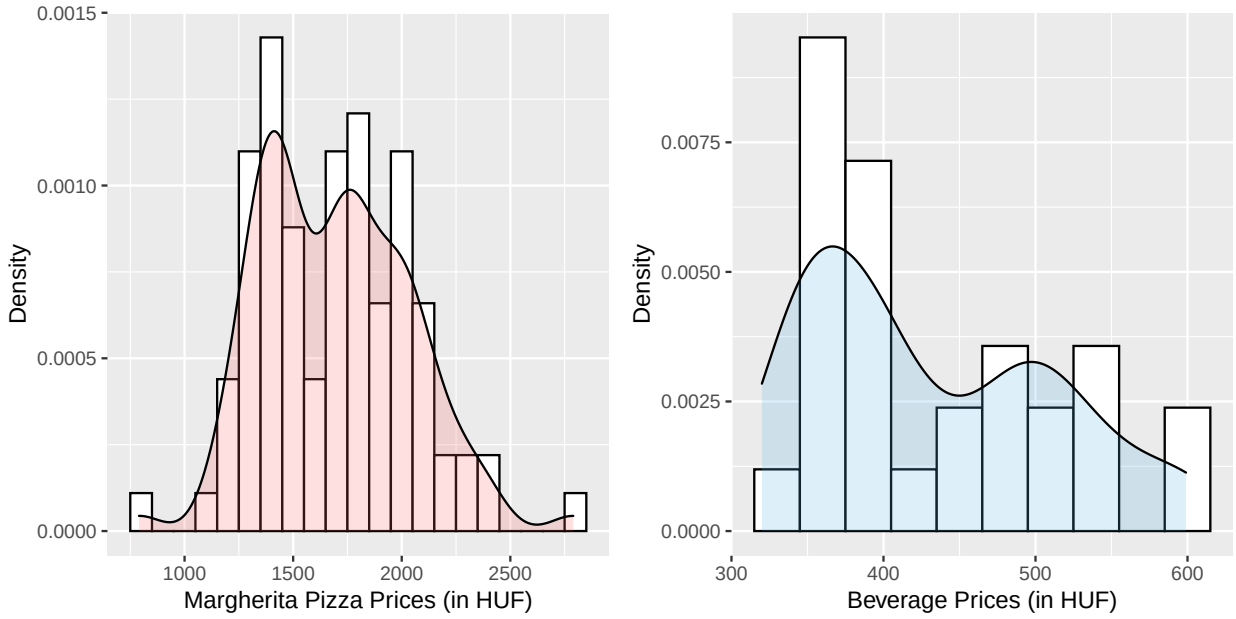


Figure 1: Price distribution for beverages and pizzas for the whole population

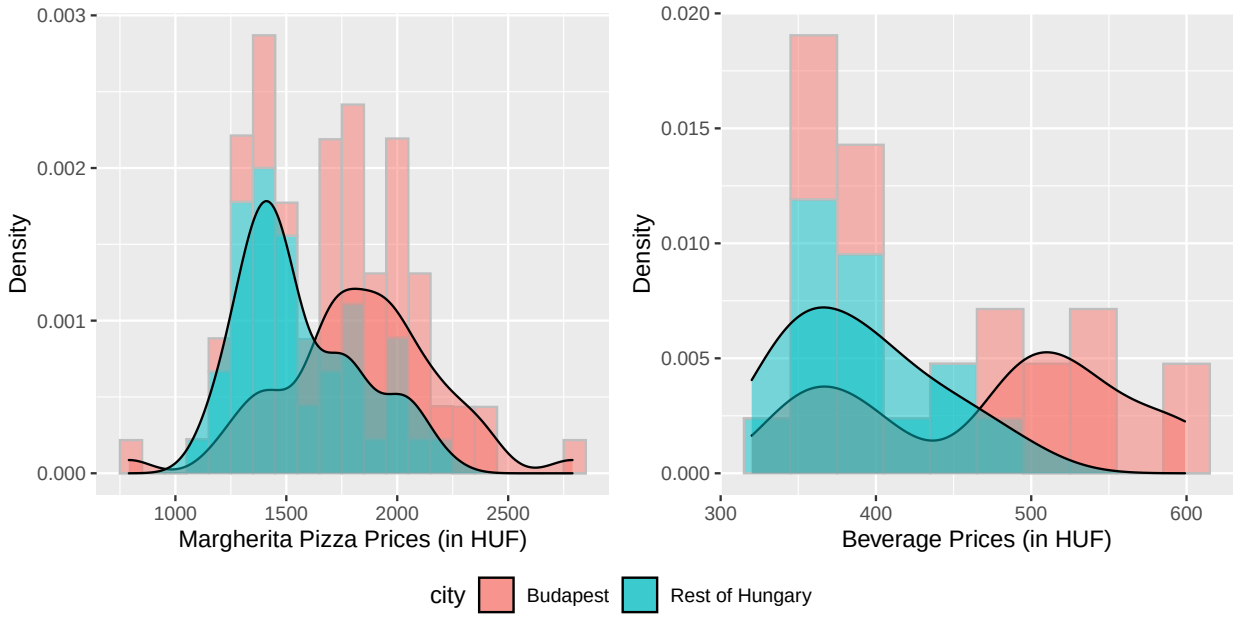


Figure 2: Price distibution of pizza and beverage prices in Budapest and Rest of Hungary

| city | variable | mean | median | std | iq_range | min | max | skew | numObs |
|---|---|---|---|---|---|---|---|---|---|
| Budapest | price_pizza | 1825.74 | 1820.50 | 364.44 | 307.50 | 790.00 | 2790.00 | -0.15 | 46 |
| Rest of Hungary | price_pizza | 1557.78 | 1490.00 | 266.87 | 420.00 | 1150.00 | 2190.00 | 0.69 | 45 |
| Budapest | price_bev | 473.43 | 494.50 | 87.36 | 137.50 | 350.00 | 599.00 | -0.21 | 14 |
| Rest of Hungary | price_bev | 389.29 | 390.00 | 48.91 | 65.00 | 320.00 | 490.00 | 0.59 | 14 |

Table 2: Summary statistics in Budapest and Rest of Hungary breakdown

We can have a closer look on the price distribution charts in Figure 2. Looks like the second peak in both BP and non-BP pizza distributions have vanished which led us to believe that the two peaks were indeed due to the average price (or rather mode) difference of prices in the distinct regions. We also added a boxplot in Figure 3 to visualize the differences.

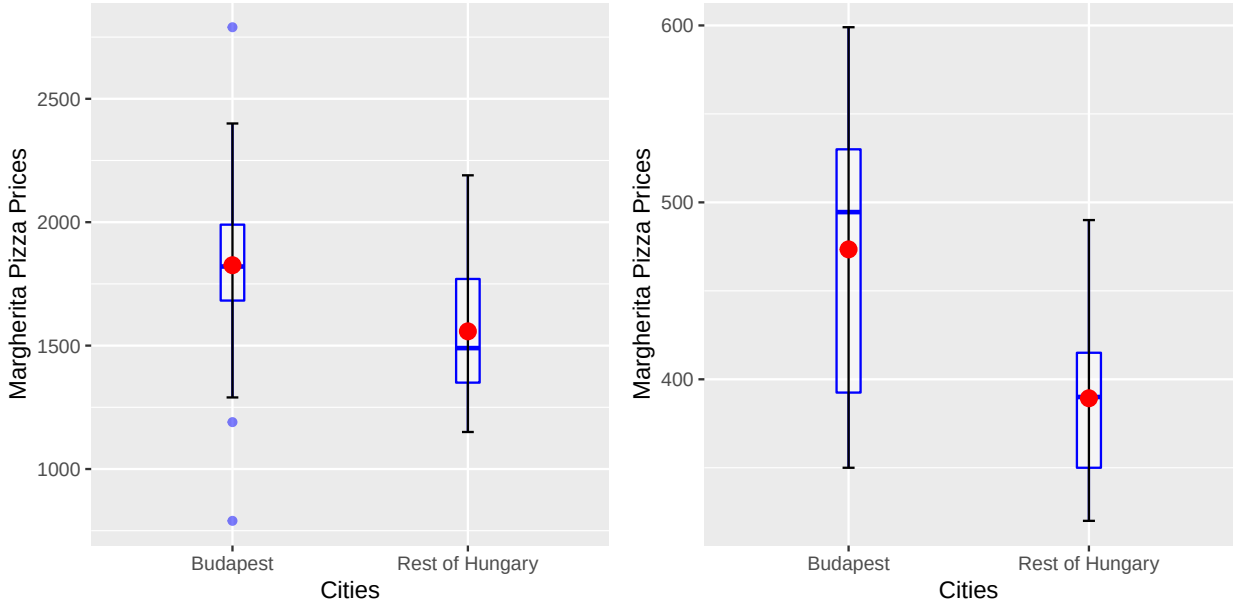


Figure 3: Box plots of pizza and beverage prices in Budapest and Rest of Hungary

## Correlation

We wanted to measure the strength and direction of the linear relationships between the various variables in our data table such as price, distance to city center,beverage price, ratings and no of ratings. We drew up a correlation matrix for the two locations Budapest and Rest of Hungary which helped us summarize the relationships of the variables against each other in both locations.

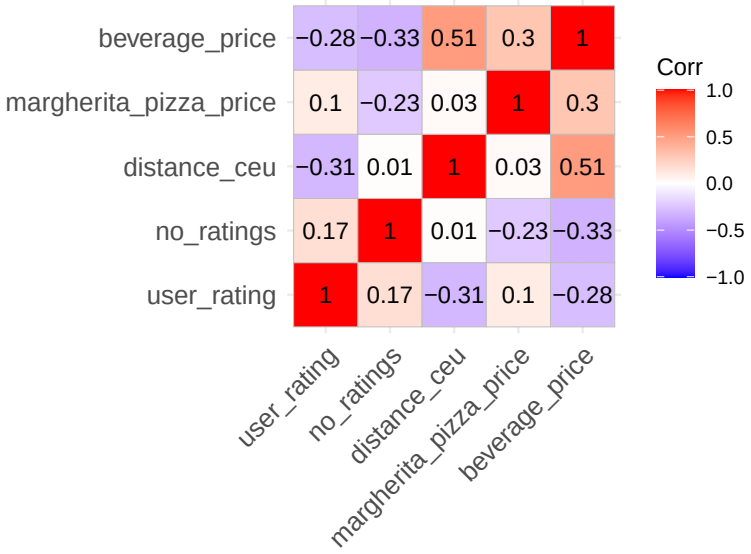


Figure 4: Correlation Matrix (Budapest)

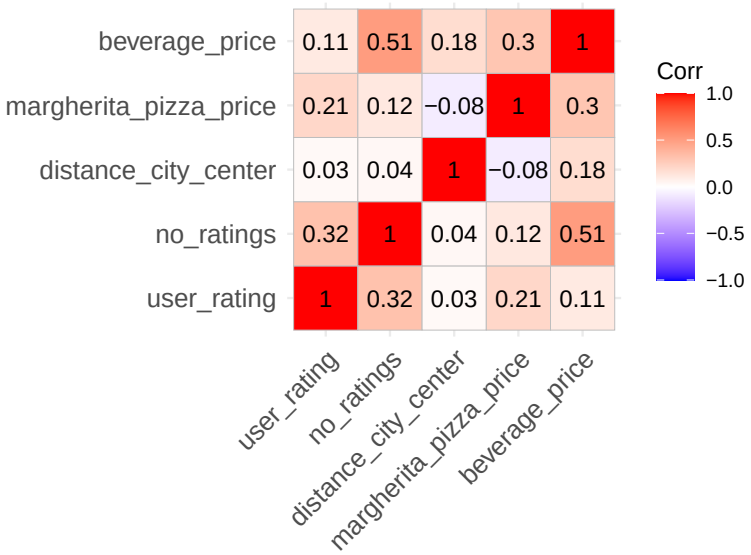


Figure 5: Correlation Matrix (Rest of Hungary)

## Margherita pizza price vs distance to city center

In Budapest,we expected the pizza price to decrease as we moved away from the city center but the correlation of 0.03 shows that there is only a minute positive linear relationship between the two variables.The price difference in regards to distance might as well be characterized as zero. In Rest of Hungary, the price of the pizza decreased a little as we moved away from city center as shown by a negative correlation of 0.08. The scatterplot of the two variables along with a trend line further shows the relationship in the two groups locations.

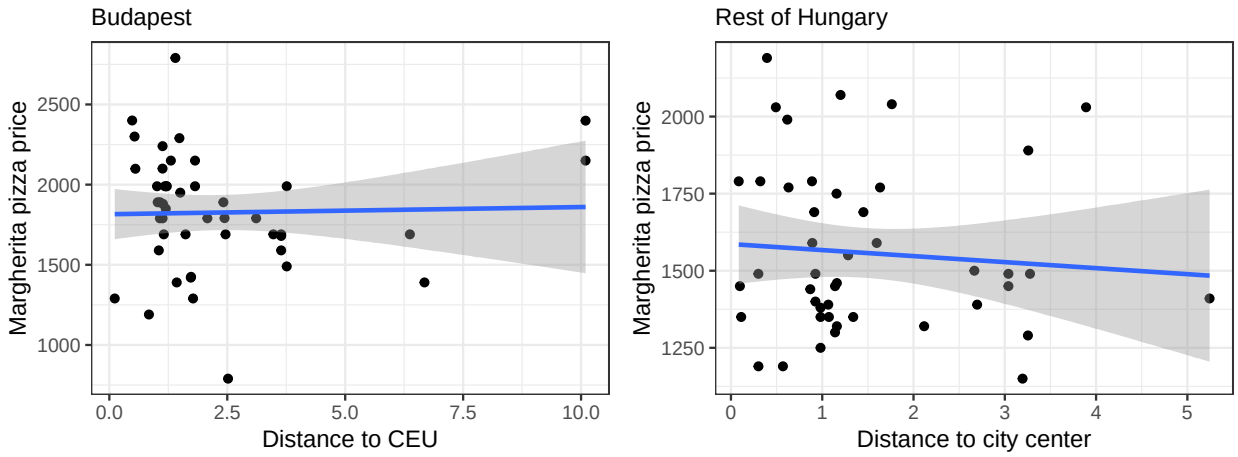


Figure 6: Scatterplot of pizza price vs distance to city center in Budapest and Rest of Hungary

## Margherita pizza price vs User rating

In Budapest, the correlation of 0.1 shows that there is a very weak positive relationship between user ratings and pizza price in a restaurant. In Rest of Hungary, the correlation is 0.21 which also signifies a weak positive relationship. The rating for restaurants that had higher prices had a better user rating. Its most likely that the quality of ingredients and taste justified the higher price and hence customers gave a slightly higher rating.

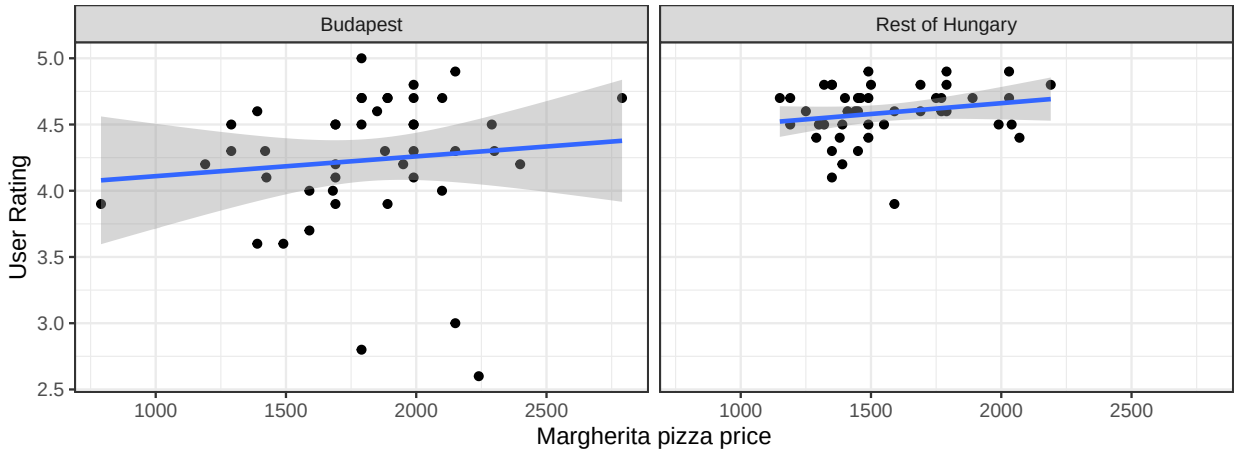


Figure 7: Scatterplot of Margherita pizza price vs User rating in Budapest and Rest of Hungary

### Distance to city center vs User rating

In Budapest, the restaurants closer to CEU (city center for Budapest) had comparatively higher ratings than the ones further away from CEU as shown by correlation of -0.31. Since many locals and tourists alike visit the city center more to eat, the restaurants aim to provide good quality food to maintain its reputation. Hence, it may be that the customers had a better experience in the city center. The number of observations related to restaurants far from city center are low so the relationship may not be entirely true and needs to be studied further.In Rest of Hungary, there is no noticeable pattern visible between distance to city center and user rating as shown by correlation of 0.03. It signifies a very weak positive correlation.

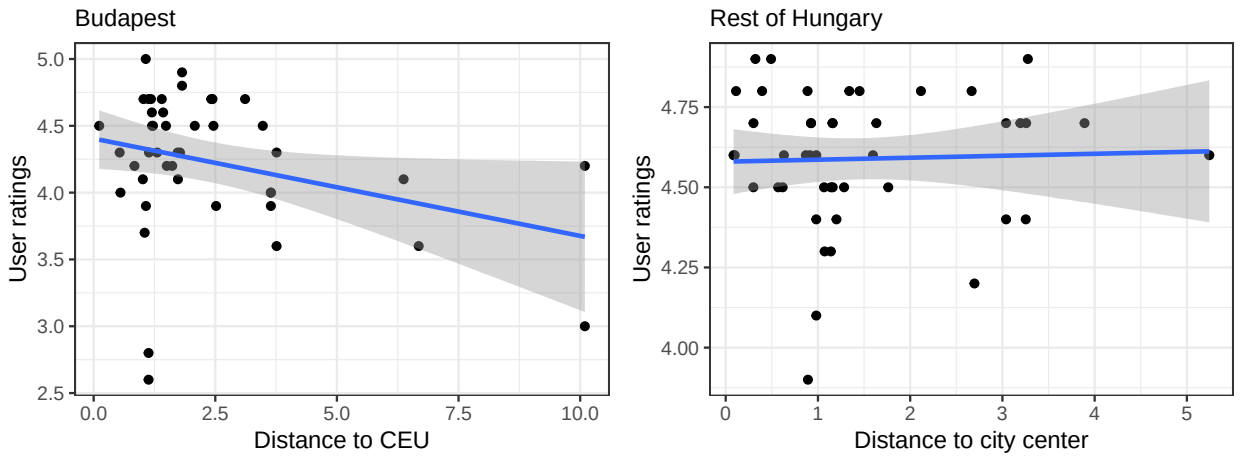


Figure 8: Scatterplot of Distance to city center vs User rating in Budapest and Rest of Hungary

## 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-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.