Budapest Pizza Project Report - KFC (Kata, Fanni, Cosmin) Team

Data collection, difficulties & problems

The KFC team (GitHub) chose to limit the data collection to the Budapest area, with inner city on-site locations metropolitan area delivery restaurants. The Google Maps API was used in order to compile the custom Google Maps shared lists of on-site locations, gather customer reviews and compute the travel time by foot between the restaurant and CEU (as a workaround to unpredictable car travel congestion). The choice of 0.5L sparkling water was convened with the drawback that the distribution of prices, brands and mark-up would be more heterogeneous than with the margherita pizza. The district, address and restaurant variables were added to have a unique identifier for every pizza place in case the dataset were to ever be expanded. The binary indicator variable 'delivery' differentiates between restaurants' menu and online merchants' prices. This ties into the 'delivery_site' column which designates the online vendor.

Summary statistics

Table 1: Summary statistics for pizza margherita prices (first row)
and for sparkling water prices (second row) in Budapest

variable	n	Mean	Median	Std	IQ_range	Min	Max	Skew
price_pizza	40	1958.25	1920	443.8346	625	1140	2950	0.2752937
price_water	40	518.50	400	261.9312	240	200	1100	1.0688534

The average price of pizza margherita is 1958.25 HUF. As for the sparkling water, the average price is 518.5 HUF. The mean is higher than the median in case of the pizza and the water as well, indicating a somewhat skewed distribution with a longer right tail. In the next section, the distribution will be illustrated by histograms.

Histograms of product prices (kernel density showcases skewness)
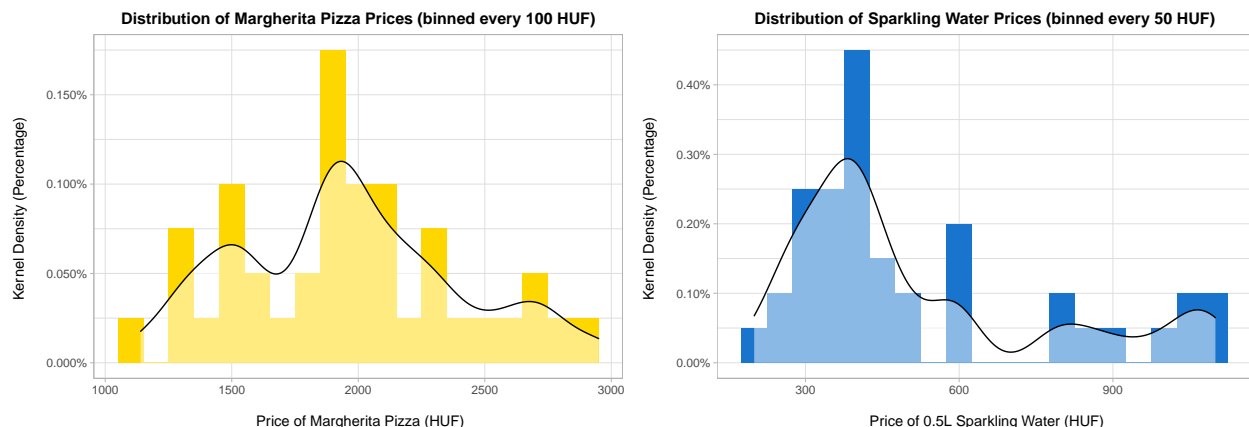
Visualizations of the price of margherita pizza conditioning on online vs offline price
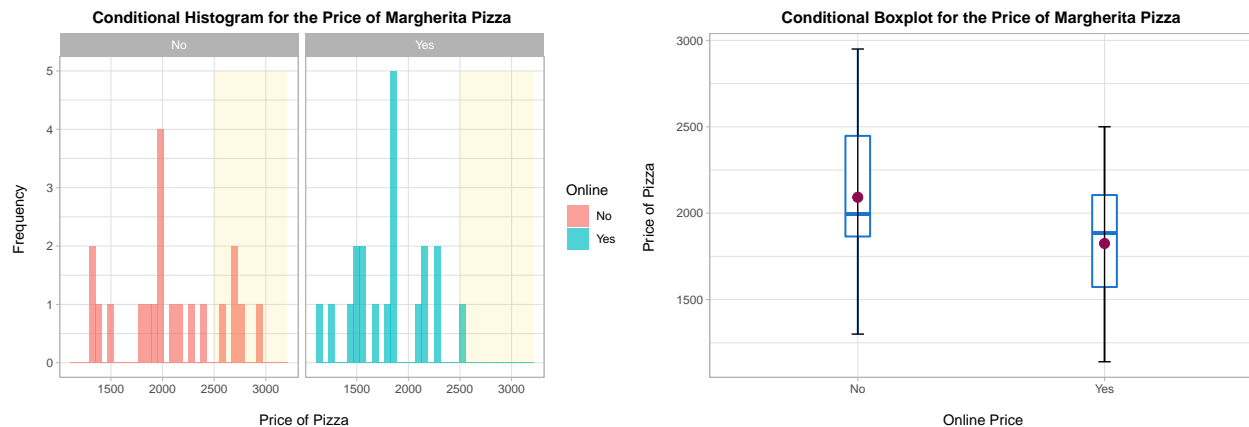
Test of whether the online and offline prices of margherita pizza are the same

The hypotheses are about whether the offline versus online price difference is zero on average in population of products represented by the data. The null hypothesis is that online and offline prices of pizza margherita are the same in Budapest. The alternative hypothesis is that online and offline prices of pizza margherita are different in Budapest. We fixed the level of significance at 5% then tested hypotheses.

```
## [1] "P-value: 0.0562"
```

```
## [1] "T-statistic: 1.9756"
```

```
## [1] "Confidence Interval: -7.5"  "Confidence Interval: 542.5"
```

Based on the results above since the p-value is larger than the level of significance we cannot reject the null hypothesis. The t-statistic is within +/- 2, also the CI is between -7 and and 542, so the difference between the online and offline prices can be zero. Thus, we can conclude that the offline and online prices of pizza margherita can be equal in Budapest.

Summary of findings regarding unconditional and conditional price distributions

Based on the histograms of product prices and the table of summary statistics it can be clearly seen that even though the price distribution for margherita pizza tends to have a few smaller modes, it closely resembles a normal distribution. With the help of binning, the distribution can be brought to look much closer like a normal distribution. The same cannot be said about the distribution of sparkling water prices, which showcases a heavy right skew, having a mean much higher than the median. This is perhaps caused by the large differences in sparkling water brands (from local Hungarian to expensive imported Italian water). As such, the range of products is much higher than with respect to the pizza. For further analysis, it would be worthwhile to inspect distributions of local sparkling water prices and imported water prices in separate graphs.

As for the conditional distribution of the price of margherita pizza it can be concluded that both the online and offline price distributions approximate a normal distribution. Furthermore, when comparing the two we can see that offline prices are higher on average than online prices since the mean of the former is higher than the latter. Also, on the histogram it is visible that the maximum online price is 2500 HUF while the maximum offline price is almost 3000 HUF. Lastly, the boxplot shows that the spread of offline prices is greater than that of the online prices.