Budapest Pizza Report

KFC (Kata, Fanni, Cosmin) Team

October 25th, 2020

The purpose of the report is to explore data about pizza prices in Budapest. The following repository contains the dataset, walkthrough explanation of the variables and the R codes used to create this descriptive analysis: GitHub

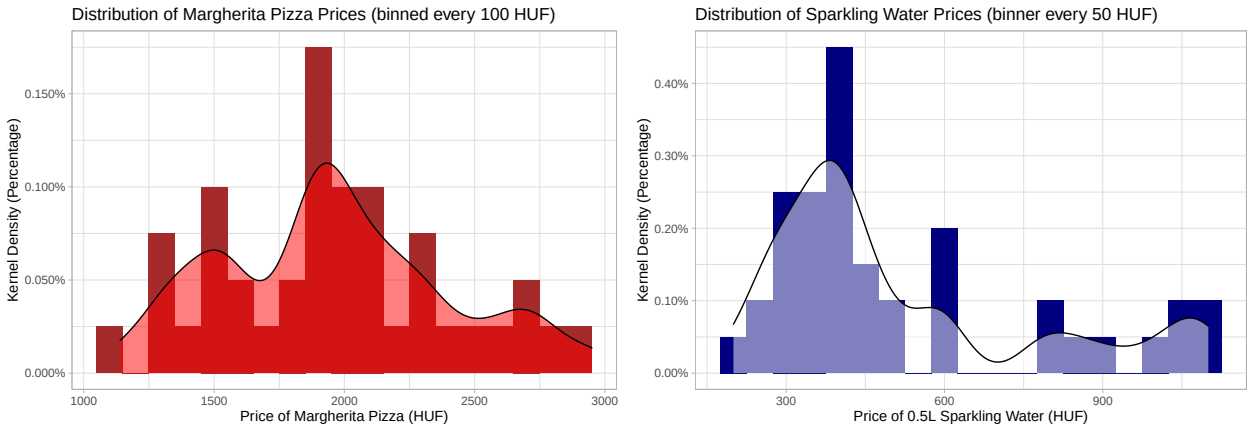
Data collection, difficulties & problems

The KFC team chose to limit the data collection to the Budapest area, with on-site prices being gathered only within the inner city districts (i.e. 2, 5, 6, 7, 8, 9, 11, 13), while delivery prices were documented irrespective of inner city district limits. The Google Maps API was used in order to compile the list of on-site locations, gather customer reviews and compute the travel time between the restaurant and CEU. The choice of a margherita pizza made the collection of prices straight-forward. However, the task of gathering optional 0.5 liter beverage prices proved to be a challenge due to inconsistencies across lower-tier and more upscale restaurants. The choice of sparkling water was convened with the drawback that the distribution of prices, brands and mark-up would be much more heterogeneous than with the margherita pizza.

Variables like district, address and restaurant were added for reference purposes and to be able to have a unique identifier for every pizza place in case the dataset were to ever be expanded (for example by adding multiple 'Pizza Forte' franchises throughout Budapest). The binary indicator variable 'delivery' was conceptualized in order to differentiate between prices as quoted on the restaurants' menus (thus considered 'on-site') and prices as quoted by various online merchants. The creation of this variable leads into the 'delivery_site' column which designates the online vendor from which the specific prices were gathered. This alleviates potential issues with regards to price fluctuations because the intermediary merchant is logged and can be researched in the case of unwanted mark-ups. Lastly, since distance from the CEU Nador 15 building is hard to quantify with regards to driving a car (as there are numerous one-way streets and other detours which might add unnecessary distance units to the computation), the shortest walking distance was chosen and the metric of choice was minutes.

Histograms of product prices (kernel density showcases skewness)

The following plots describe the distribution of margherita pizza prices and sparkling water prices.



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 showcase 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.

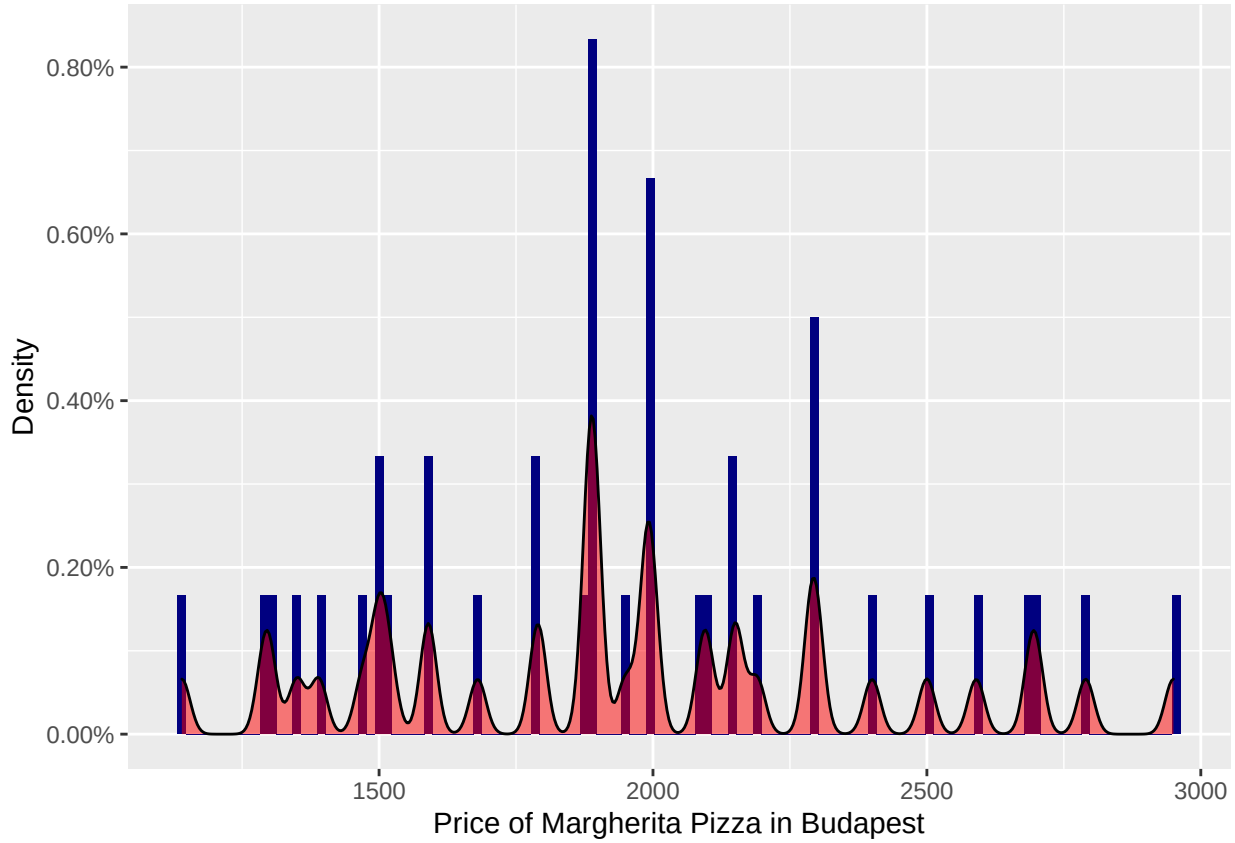
Summary

```
summary(pizza_data$price_pizza)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1140    1590    1920    1958    2215    2950
```

Plots

```
ggplot( data = pizza_data , aes( x = price_pizza ) ) +
  geom_histogram( aes(y = ..density..) , fill = 'navyblue' , binwidth = 15 ) +
  geom_density( aes(y = ..density..) , fill = 'red' , bw = 15, alpha = 0.5 ) +
  labs(x='Price of Margherita Pizza in Budapest',y='Density')+
  scale_y_continuous(labels = scales::percent)
```



Tables

```
library(moments)
pizza_stat <- pizza_data %>% summarise(
  mean     = mean(price_pizza),
  median   = median(price_pizza),
  std      = sd(price_pizza),
  iq_range = IQR(price_pizza),
  min      = min(price_pizza),
  max      = max(price_pizza),
  skew     = skewness(price_pizza),
  numObs   = n() )
knitr::kable(pizza_stat,caption="Summary stat for hotel prices")
```

Table 1: Summary stat for hotel prices

mean	median	std	iq_range	min	max	skew	numObs
1958.25	1920	443.8346	625	1140	2950	0.2752937	40

Better table

```r
library(xtable)

xtb <- xtable(pizza_stat, type = "latex", caption = "Summary statistics - elegant package")
print(xtb, comment=FALSE, include.rownames=FALSE)
```

mean	median	std	iq_range	min	max	skew	numObs
1958.25	1920.00	443.83	625.00	1140.00	2950.00	0.28	40

Table 2: Summary statistics - elegant package

Summary of results