DA1 Assignment - Task 2

Halmschlager, Kovacs, Szokolics

9/25/2019

Introduction

During a previous task (task 1) we collected the prices of two products ((1) Coca Cola, 0.5l plastic bottle and (2) Orbit, chewing gum, peppermint, single pack) and data on the shops where the prices were collected (address, google rating, shop type, number of cashiers). This data collections forms the basis for the following task:

Task 2 - Description: Create a report describing your data. (6p)

- Submit a single pdf, code and data in a zipped file
- Discuss the data collection, difficulties, problems. How you picked product 2. How did you decide on store features to record and how did you code it? (2-3 paragraphs) [1p]
- Present descriptive statistics of prices. You may do it with one or two tables as you see fit [1p]
- Show two descriptive graphs of price distributions of your products in the whole data (ie two districts merged). [0.5p]
- Show two descriptive graphs of price distributions of Coca Cola in the two districts merged [0.5p]
- Test if the price of Coca Cola is the same in the two districts. [1p]
- Summarize your findings regarding price distributions [1p]
- Now pick any feature of stores and create a binary variable. Pool your data across districts. Compare prices by this new variable and discuss. [1p]"

Data Collection

Difficulties encountered:

- To find a product that is widely available in all kinds of shops.
- Decide on how to measure quality and size of a store. For example it is not possible to find out the exact square meter of a store. What else could reflect if a store was large or small?
- Some small shops did not have price tags for all products so it was necessary to purchase the product on order to assess the price.
- Some shop assistants asked not to take any pictures.

Product 2: The second product (Orbit, chewing gum, peppermint, single pack) was chosen mainly because it is widely available in shops, gas stations, local stores etc. It is also easy to define in terms of brand, size and flavour.

Store Features: The store features (Google rating, Shop type, Number_of_cashiers) were chosen to reflect the quality, the type and the size of the shop. For quality we decided to use an official rating rather than an individual assessment, and included the google rating for the individual shop. The type of shop was chosen among a list of types we had agreed on before collecting the data. To express the size of a shop we counted the number of cash desks in the shop.

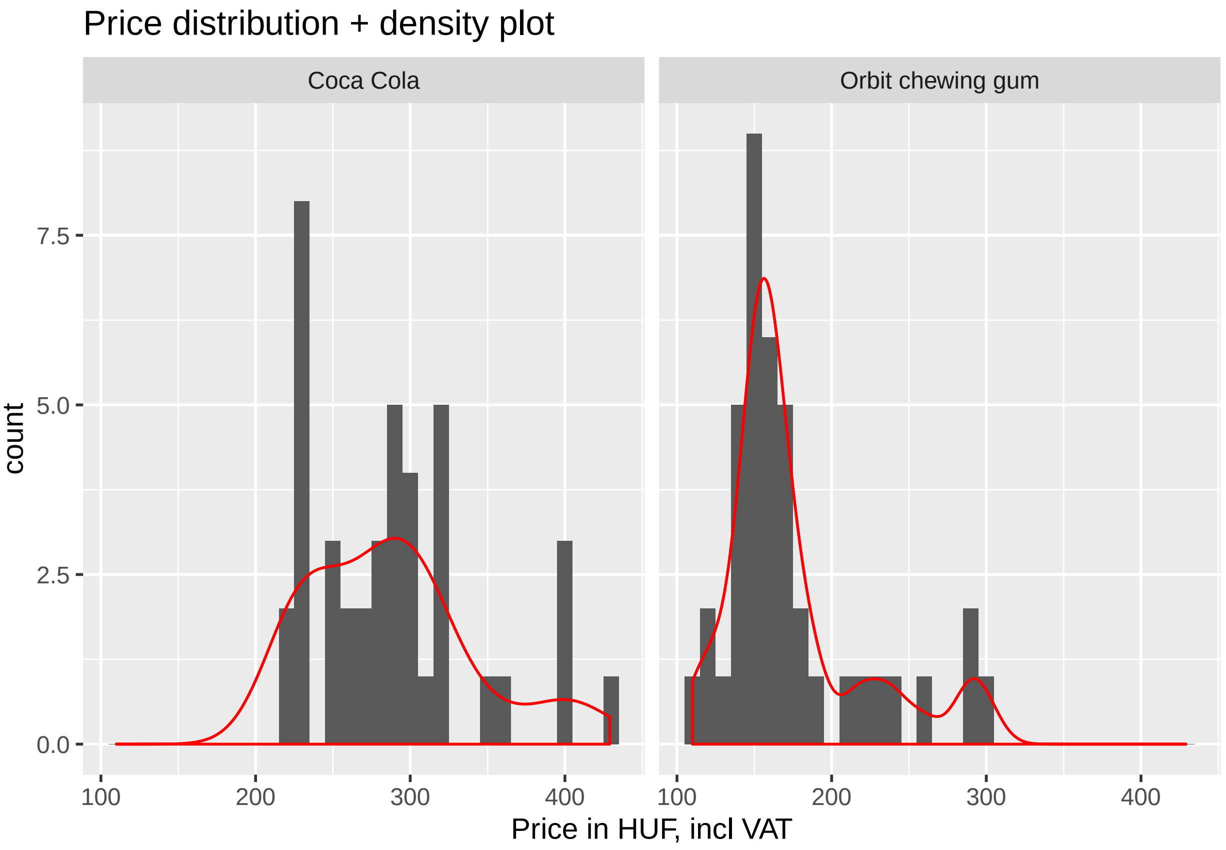
Descriptive statistics of prices

```r
# import data
products <- read_csv("da1-asgn-Halmschlager_Kovacs_Szokolics-data.csv")

# min,max,avg price of product 1 and 2
products %>% group_by(Product) %>% summarize(Mean=mean(Price), Min=min(Price), Max=max(Price))
```

```
## # A tibble: 2 x 4
##   Product             Mean   Min   Max
##   <chr>              <dbl> <dbl> <dbl>
## 1 Coca Cola           288    219   429
## 2 Orbit chewing gum  175.    110   300
```

Descriptive graphs of price distribution

```
# The following chart shows the price distribution per product (grey bars) and the respective density p

ggplot(products, aes(Price)) +
  geom_histogram(aes(y=..count..), binwidth=10) +
  geom_density(aes(y=10*..count..),colour="red") +
  facet_grid(. ~ Product) +
  labs(x = "Price in HUF, incl VAT",
      title ="Price distribution + density plot")
```

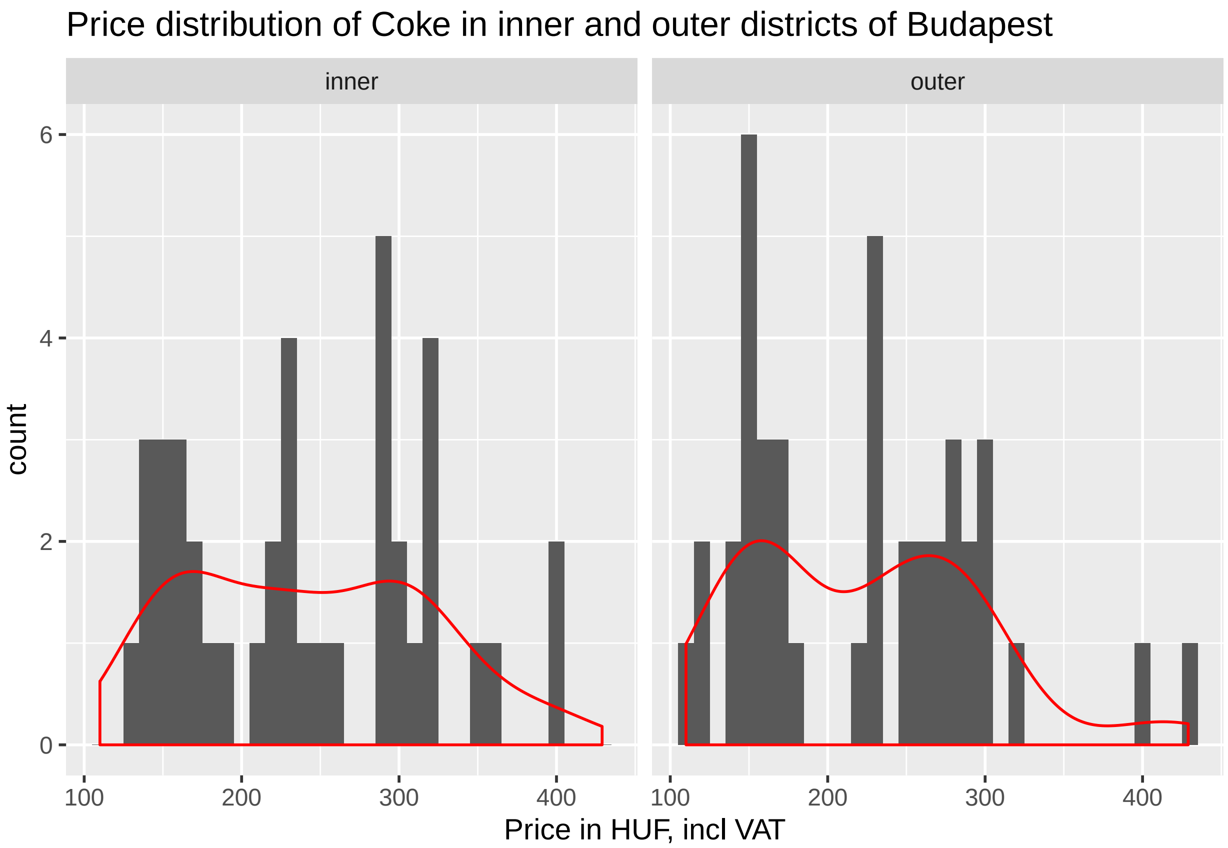


Descriptive graphs of prices for Coca Cola

```
# Show two descriptive graphs of price distributions of Coca Cola in the two districts merged [0.5p]

# add inner/outer district variable to dataset
products %>% group_by(Address_ZIP) %>% summarise(count = n()) # shows distinct ZIP codes in dataset
```

```
## # A tibble: 5 x 2
##   Address_ZIP count
##           <int> <int>
## 1          1111    12
## 2          1114     8
## 3          1117    20
## 4          1171     6
## 5          1173    35
```

```
Districts <- data.frame(Zip = c(1111,1114,1117,1171,1173), District = c("inner","inner","inner","outer"

products2 <- products %>% left_join(Districts, by =c("Address_ZIP" = "Zip")) # join district to product

ggplot(products2, aes(Price)) +
  geom_histogram(aes(y=..count..), binwidth=10) +
  geom_density(aes(y=10*..count..),colour="red") +
  facet_grid(. ~ District) +
  labs(x = "Price in HUF, incl VAT",
      title ="Price distribution of Coke in inner and outer districts of Budapest")
```



Price comparison for Coca Cola in different districts

To test if the price of Coca Cola is the same in the two districts, we are using three different methods. First we look at the min,max and avg prices of coke in both districts, then we run a t.test, followed by a basic regression model to test whether there is a significant difference between prices in inner and outer districts.

Min, max, avg prices

```
# min,max,avg price of product 1 and 2
products2 %>% filter(Product_ID == 1) %>% group_by(District) %>% summarize(Mean=mean(Price), Min=min(Pr
## # A tibble: 2 x 4
##   District  Mean   Min   Max
##   <chr>    <dbl> <dbl> <dbl>
## 1 inner     299.   219   399
## 2 outer     278.   219   429
```

From this table we can assume that the price span is higher in the outer districts than in the inner districts.

T-test:

```
# The null hypothesis: the values (price of coca cola) in the inner and outer district are the same dis
t.test(Price ~ factor(District), data=products2[products2$Product_ID==1,])
```

```
##
##  Welch Two Sample t-test
##
## data:  Price by factor(District)
## t = 1.2796, df = 38.956, p-value = 0.2083
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -12.41626  55.17341
## sample estimates:
## mean in group inner mean in group outer
##            298.9500            277.5714
```

The t-test shows p > 0.05 which means there is no significant difference in the distribution of values between the two groups (inner vs outer district).

Regression:

```
# + + + + should this part be included? The result is the same as in the t-test, p = 0.208

# basic regression model, checking for a relationship between Price of Product 1 and District
mod1 <- lm(Price ~ District, data=products2[products2$Product_ID==1,])

# Compare estimates and level of significance from model1:
summary(mod1)
```

```
##
## Call:
## lm(formula = Price ~ District, data = products2[products2$Product_ID ==
##     1, ])
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -79.95 -48.57   0.05  20.05 151.43
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     298.95      11.96  24.992   <2e-16 ***
## Districtouter   -21.38      16.71  -1.279    0.208
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 53.5 on 39 degrees of freedom
## Multiple R-squared:  0.04026,    Adjusted R-squared:  0.01565
## F-statistic: 1.636 on 1 and 39 DF,  p-value: 0.2084
```

The regression model also shows $p > 0.05$ which means there is no significant difference in the distribution of values between the two groups (inner vs outer district).

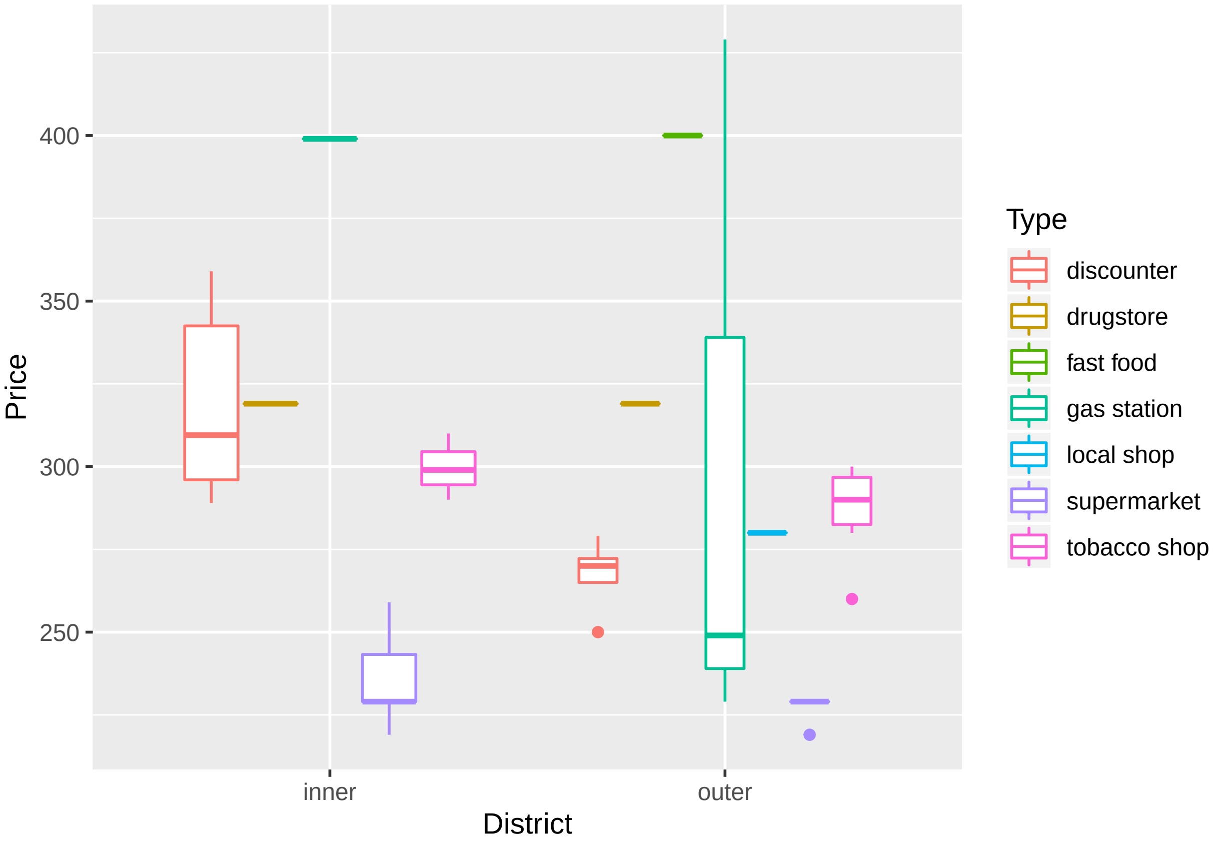
Regression with control variable:

```r
# simple model with indicators for inner vs outer district, controlling for shop type.
mod2 <- lm(Price~District + Type, data=products2[products2$Product_ID==1,])

# Compare estimates and level of significance from model2:
summary(mod2)
```

```
##
## Call:
## lm(formula = Price ~ District + Type, data = products2[products2$Product_ID ==
##     1, ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -99.269 -15.831  -0.003  14.725 100.731
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        310.831     12.023  25.852  < 2e-16 ***
## Districtouter      -31.828     11.636  -2.735 0.009948 **
## Typedrugstore       16.126     20.813   0.775 0.443985
## Typefast food      120.997     37.425   3.233 0.002779 **
## Typegas station     49.266     19.342   2.547 0.015709 *
## Typelocal shop       0.997     37.425   0.027 0.978906
## Typesupermarket    -64.728     15.331  -4.222 0.000178 ***
## Typetobacco shop     1.276     16.404   0.078 0.938445
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35.06 on 33 degrees of freedom
## Multiple R-squared:  0.6512, Adjusted R-squared:  0.5773
## F-statistic: 8.803 on 7 and 33 DF,  p-value: 4.469e-06
```

With the previous t-test without controlling for shop type there is very little difference between inner and outer district, but in the regression model which includes the control variable the effect on the dependent variable becomes more apparent (Districtouter p-value < 0.05).

```r
ggplot(products2[products2$Product_ID ==
    1, ], aes(x=District, y=Price, colour=Type)) +
  geom_boxplot()
```



Price comparison for both products in both districts by shop size

The size of the shop should be visible through the number of cashiers available in the particular shop. We are going to distinguish between small shops (1 cashier) and medium/large shops (>1 cashier). The dummy variable is going to be 1 = small, 0 = not small

```
#
products3 <- products %>% mutate(Size = ifelse(Number_of_cashiers==1,1,0))
products3 %>% group_by(Size) %>% summarize(count = n())
```

```
## # A tibble: 2 x 2
##    Size count
##   <dbl> <int>
## 1     0    43
## 2     1    38
```

Now we compare the prices among all districts with regards to the size of the shop.

Regression:

```
# + + + + should this part be included? The result is the same as in the t-test, p = 0.208

# basic regression model, checking for a relationship between Price of Product 1 and District
mod3_1 <- lm(Price ~ Size, data=products3[products$Product_ID==1,])

# Compare estimates and level of significance from model1:
summary(mod3_1)
```

```
##
```

```
## Call:
## lm(formula = Price ~ Size, data = products3[products$Product_ID ==
##     1, ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.682 -48.682  -9.947  20.053 151.318
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   277.68      11.39  24.388   <2e-16 ***
## Size           22.27      16.73   1.331    0.191
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 53.41 on 39 degrees of freedom
## Multiple R-squared:  0.04346,    Adjusted R-squared:  0.01894
## F-statistic: 1.772 on 1 and 39 DF,  p-value: 0.1909
```

The regression model shows $p > 0.05$ which means there is no significant difference in the distribution of values between small size shops and medium/large size shops.