Report Online Sales

Felipe Soares

December 15, 2017

Dataset Description

This dataset comprises data from an Online Retail store focused on wholesalers clients. As stated in the UCI website (http://archive.ics.uci.edu/ml/datasets/online+retail (http://archive.ics.uci.edu/ml/datasets/online+retail)), this online store mainly sells unique all-occasion gifts.

Attribute information

There are eight attributes in this dataset:

- InvoiceNo: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter ‘c’, it indicates a cancellation.
- StockCode: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.
- Description: Product (item) name. Nominal.
- Quantity: The quantities of each product (item) per transaction. Numeric.
- InvoiceDate: Invice Date and time. Numeric, the day and time when each transaction was generated.
- UnitPrice: Unit price. Numeric, Product price per unit in sterling.
- CustomerID: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.
- Country: Country name. Nominal, the name of the country where each customer resides.

Research question

Does the average value of the products in the orders (i.e. the average price of a product in orders) differs between November and March in the UK?

Since this online store has mainly wholesalers customers, I assume that purchases of christmas gifts are made on November, such that its customers will sell them during December. The other month, March, was chosen as the Easter in 2011 happened in April, thus, I assume that purchases were made during March.

Finally, I can then compare if there is any difference, on average, on the price of individual gifts for Christmas or Easter.

Pre-Treatment

Loading file

```
df <- read_excel("OnlineRetail.xlsx")
```

Removing entries with price and quantity <= 0 (usually related to returns)

```
df %>% filter(UnitPrice > 0 & Quantity > 0) -> df_filtered
```

Removing entries that are not products. Rationale: Products have StockCodes comprising numbers or numbers and letters. Thus, any StockCode with only letters is not a product (e.g. Amazon Fees, Postage etc).

```
df_filtered %>% filter(!grepl('^[a-zA-Z]+$', StockCode)) -> df_filtered
```

Overall data information:

```
df_filtered %>% mutate(Country = as.factor(Country)) -> df_filtered
summary(df_filtered)
```

```
##   InvoiceNo          StockCode         Description
##  Length:527918      Length:527918      Length:527918
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##     Quantity         InvoiceDate                    UnitPrice
##  Min.   :    1.00   Min.   :2010-12-01 08:26:00   Min.   :  0.001
##  1st Qu.:    1.00   1st Qu.:2011-03-28 12:23:00   1st Qu.:  1.250
##  Median :    3.00   Median :2011-07-20 13:26:00   Median :  2.080
##  Mean   :   10.56   Mean   :2011-07-04 21:31:06   Mean   :  3.279
##  3rd Qu.:   11.00   3rd Qu.:2011-10-19 13:38:00   3rd Qu.:  4.130
##  Max.   :80995.00   Max.   :2011-12-09 12:50:00   Max.   :649.500
##
##    CustomerID              Country
##  Min.   :12346    United Kingdom:484053
##  1st Qu.:13975    Germany       :  8658
##  Median :15159    France        :  8102
##  Mean   :15301    EIRE          :  7885
##  3rd Qu.:16801    Spain         :  2422
##  Max.   :18287    Netherlands   :  2322
##  NA's   :131436   (Other)       : 14476
```

From this summary one can notice that the UK has the largest amount of Plot of number of invoices per country, so we will only analyze data from that country to avoid possible noise from the other countries.

```
df_filtered %>% filter(Country == "United Kingdom") -> df_UK
```

Analysis

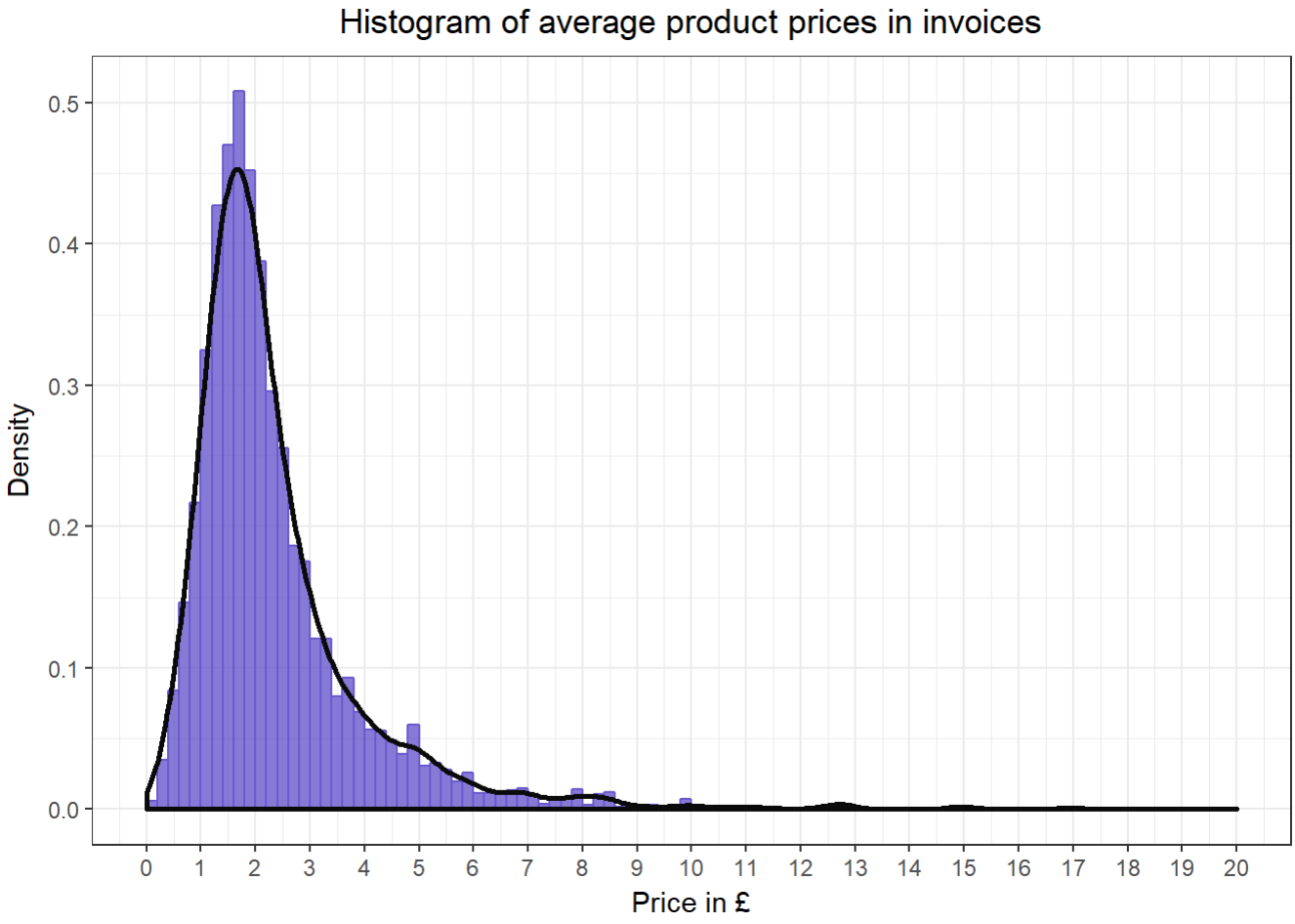
Evaluating the average price of a product in the orders. Notice that the unit price is multiplied by the purchased quantity, such that the purchasing volume is taken into account.

```
df_UK %>% mutate(InvoiceNo = as.factor(InvoiceNo)) %>%
  mutate(InvoiceMonth = as.factor(months(InvoiceDate))) %>%
  group_by(InvoiceNo,InvoiceMonth) %>% mutate(total_product = Quantity*UnitPrice) %>%
  summarize(avg_ticket_per_product = sum(total_product)/sum(Quantity)) -> df_avg
```

Ploting the histogram of the average prices generated above to give an idea about its distribution:

```
df_avg %>% ggplot(aes(avg_ticket_per_product)) +
  geom_histogram(aes(y=..density..), breaks=seq(0, 10, by = 0.2), fill = "slateblue3", col="slat
eblue3", alpha = 0.8) +
  geom_density(adjust = 2, col = "gray4", size=1) +
  xlim(0,20) + theme_bw() +
  labs(title="Histogram of average product prices in invoices") +
  labs(x="Price in £", y="Density") +
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_x_continuous(breaks=seq(0,20,1), limits=c(0,20))
```

```
## Scale for 'x' is already present. Adding another scale for 'x', which
## will replace the existing scale.
```



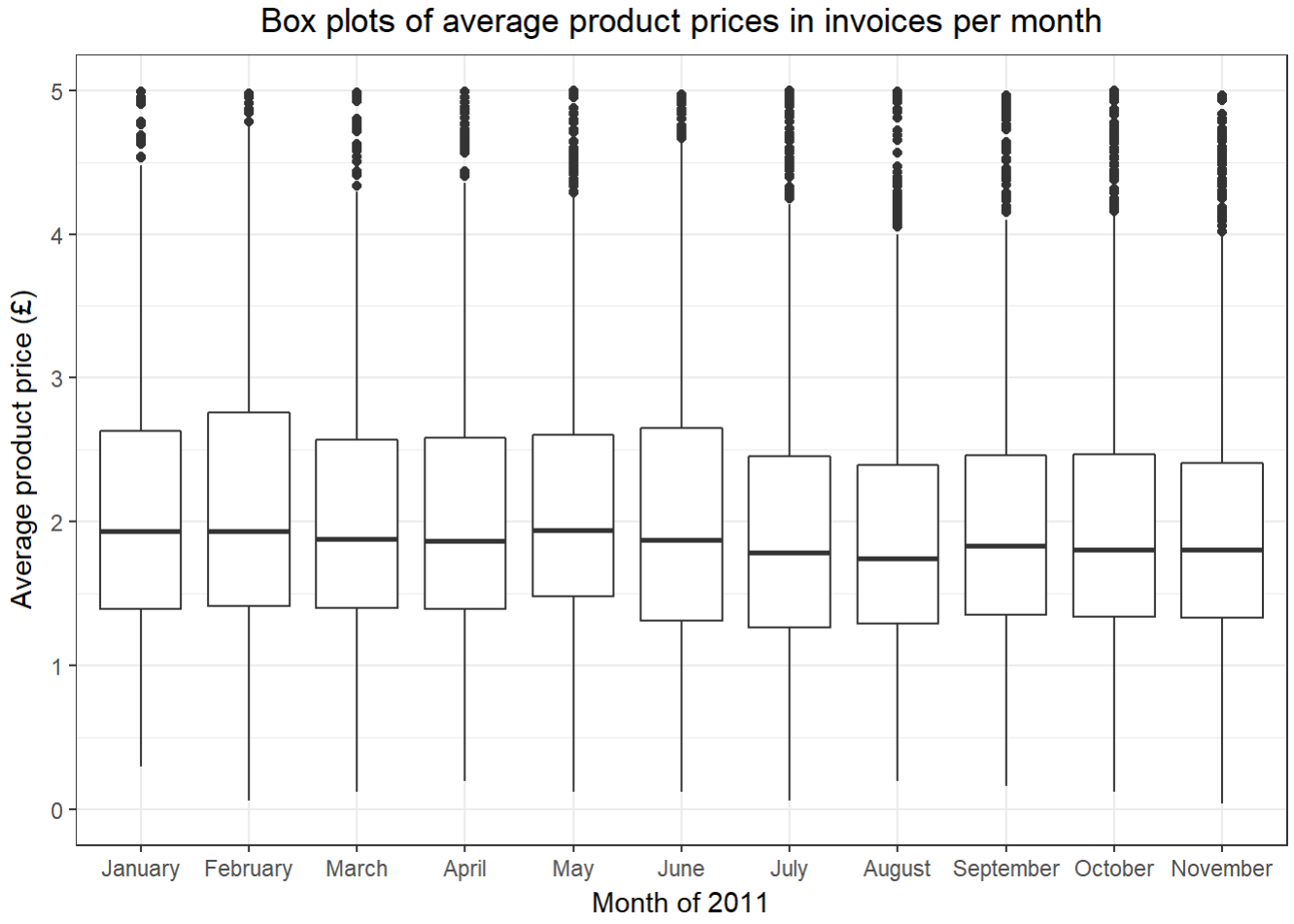
As shown in the histogram, a very few observations are greater than 6 pounds, so I will filter out those entries to avoid distortion when performing the statistical test. I will also filter out December, since that month has incomplete data (less than 30 days).

```
df_avg %>% filter(avg_ticket_per_product < 6 & InvoiceMonth != "December") -> df_avg_filtered

df_avg_filtered$InvoiceMonth <- factor(df_avg_filtered$InvoiceMonth, levels=c("January","Februar
y", "March", "April", "May", "June", "July", "August", "September", "October", "November"))
```

Verify the box plots to gain insight about the values per month:

```
df_avg_filtered %>%
  ggplot(aes(y=avg_ticket_per_product,x=InvoiceMonth)) +
  geom_boxplot() + ylim(0,5) +
  labs(title="Box plots of average product prices in invoices per month") +
  labs(x="Month of 2011", y="Average product price (£)") + theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))
```



To compare both months, November and March, I will use the t-test:

```
with(df_avg_filtered, t.test(avg_ticket_per_product[InvoiceMonth=="November"], avg_ticket_per_pr
oduct[InvoiceMonth=="March"]))
```

```
##
##  Welch Two Sample t-test
##
## data:  avg_ticket_per_product[InvoiceMonth == "November"] and avg_ticket_per_product[InvoiceM
onth == "March"]
## t = -4.6326, df = 2185.5, p-value = 3.824e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2561833 -0.1037974
## sample estimates:
## mean of x mean of y
##  2.031888  2.211879
```

Conclusion

From this test is possible to notice that average prices for Christmas gifts (month November) are statistically lower (£2.032) than for Easter (£2.212) (month March). One possible reason for this is that people usually buy gift cards and small ornaments for the house during Christmas, which may not occur on Easter.