Mini-project report

AMIN Carine - DELOUES Damien

March 30, 2017

We have chosen to do an analysis on the alcoholic content by volume using Craft beers Dataset.

```
library(ggplot2);
library(dplyr);
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

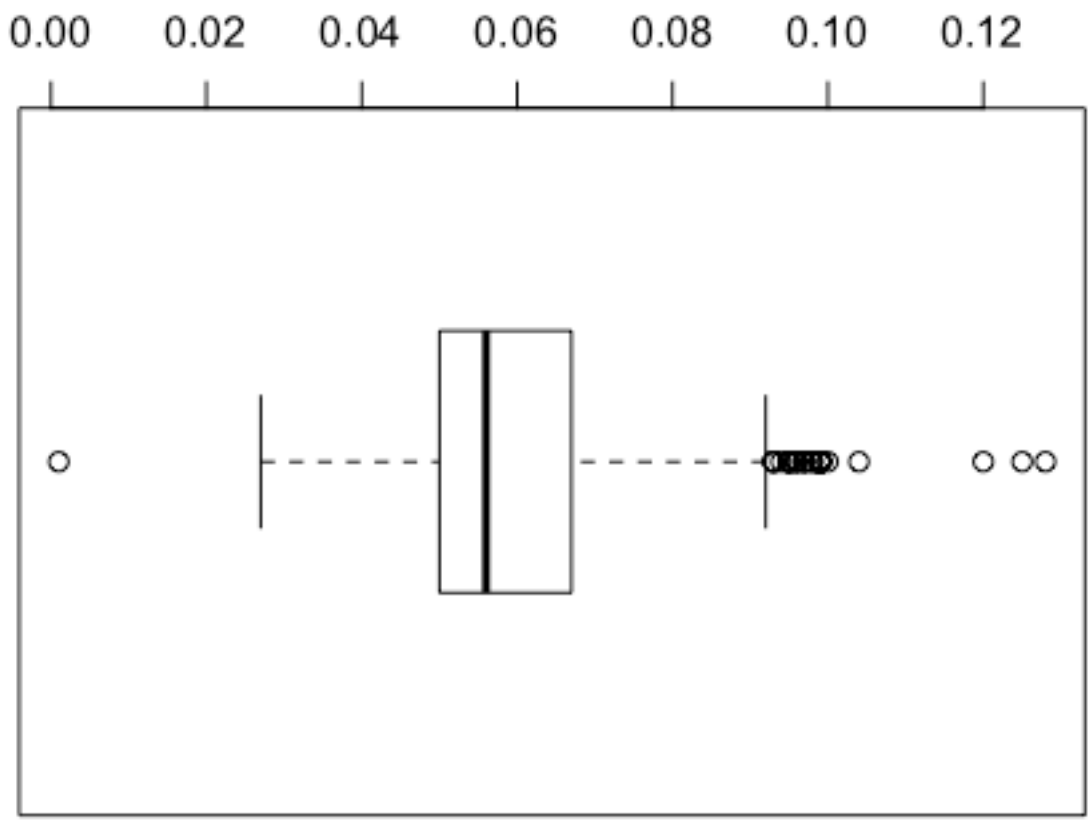
```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
beerRaw <- read.csv(file="beers.csv")
```

After loading the data on R Studio, we have selected the columns "style" and "abv", since these 2 columns are the ones we're going to analyze, and then we realized that there are NA values in the "abv" column. So we decided to omit them using the function "na.omit()".

First, we decided to do a "boxplot" to have an overview on our selection.

```
beerRaw %>% select(style, abv) %>% na.omit(beerRaw) -> beer
boxplot(x=beer$abv)
```

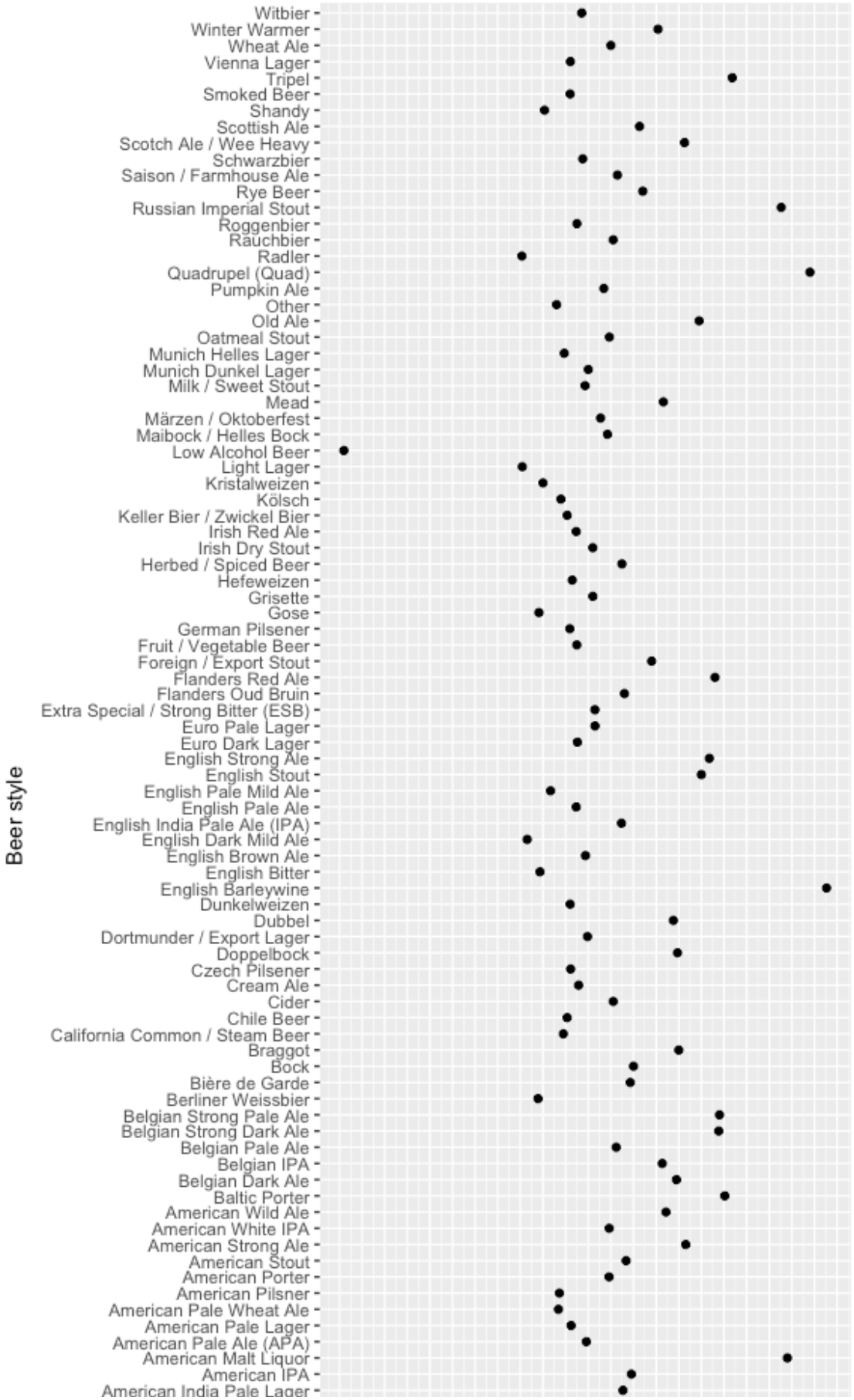

0.00
0.02
0.04
0.06
0.08
0.10
0.12

As we can see, we have a high standard deviation, which indicates that the data points are spread out over a wider range of values. We deduce that we have a wide range of beer styles, probably because there are a lot of small breweries that produce original beer styles to attract customers.

We did a graph using "ggplot".

```
 beer %>% group_by(style) %>% summarize(NbBiere=n(),MoyenneAbv=mean(abv)) ->
beer2

 beer2 %>% ggplot(aes(x=MoyenneAbv, y=style)) + geom_point()  + xlab("Average
%") +ylab("Beer style")  + theme(axis.text.x =     element_text(angle = 90 ))
+  scale_x_continuous(breaks = seq(0,0.15,by=0.005))
```


Beer style
Witbier
Winter Warmer
Wheat Ale
Vienna Lager
Tripel
Smoked Beer
Shandy
Scottish Ale
Scotch Ale / Wee Heavy
Schwarzbier
Saison / Farmhouse Ale
Rye Beer
Russian Imperial Stout
Roggenbier
Rauchbier
Radler
Quadrupel (Quad)
Pumpkin Ale
Other
Old Ale
Oatmeal Stout
Munich Helles Lager
Munich Dunkel Lager
Milk / Sweet Stout
Mead
Märzen / Oktoberfest
Maibock / Helles Bock
Low Alcohol Beer
Light Lager
Kristalweizen
Kölsch
Keller Bier / Zwickel Bier
Irish Red Ale
Irish Dry Stout
Herbed / Spiced Beer
Hefeweizen
Grisette
Gose
German Pilsener
Fruit / Vegetable Beer
Foreign / Export Stout
Flanders Red Ale
Flanders Oud Bruin
Extra Special / Strong Bitter (ESB)
Euro Pale Lager
Euro Dark Lager
English Strong Ale
English Stout
English Pale Mild Ale
English Pale Ale
English India Pale Ale (IPA)
English Dark Mild Ale
English Brown Ale
English Bitter
English Barleywine
Dunkelweizen
Dubbel
Dortmunder / Export Lager
Doppelbock
Czech Pilsener
Cream Ale
Cider
Chile Beer
California Common / Steam Beer
Braggot
Bock
Bière de Garde
Berliner Weissbier
Belgian Strong Pale Ale
Belgian Strong Dark Ale
Belgian Pale Ale
Belgian IPA
Belgian Dark Ale
Baltic Porter
American Wild Ale
American White IPA
American Strong Ale
American Stout
American Porter
American Pilsner
American Pale Wheat Ale
American Pale Lager
American Pale Ale (APA)
American Malt Liquor
American IPA
American India Pale Lager

We have put different beer styles on the vertical axis, and the rate of alcoholic content by volume on the horizontal axis.

With this graph, we can notice that most of the beer styles have an alcoholic content between 4% and 8.5%.

The beer that contains the highest alcoholic volume is the "English BarleyWine" with more than 10.5% of alcohol and the beer contains the lowest alcoholic volume is the "Low Alcohol Beer" with almost 0%.

As a conclusion, we can see that the data are quite spread out, which means that there is a very big difference between the lowest and the highest alcoholic volumes.

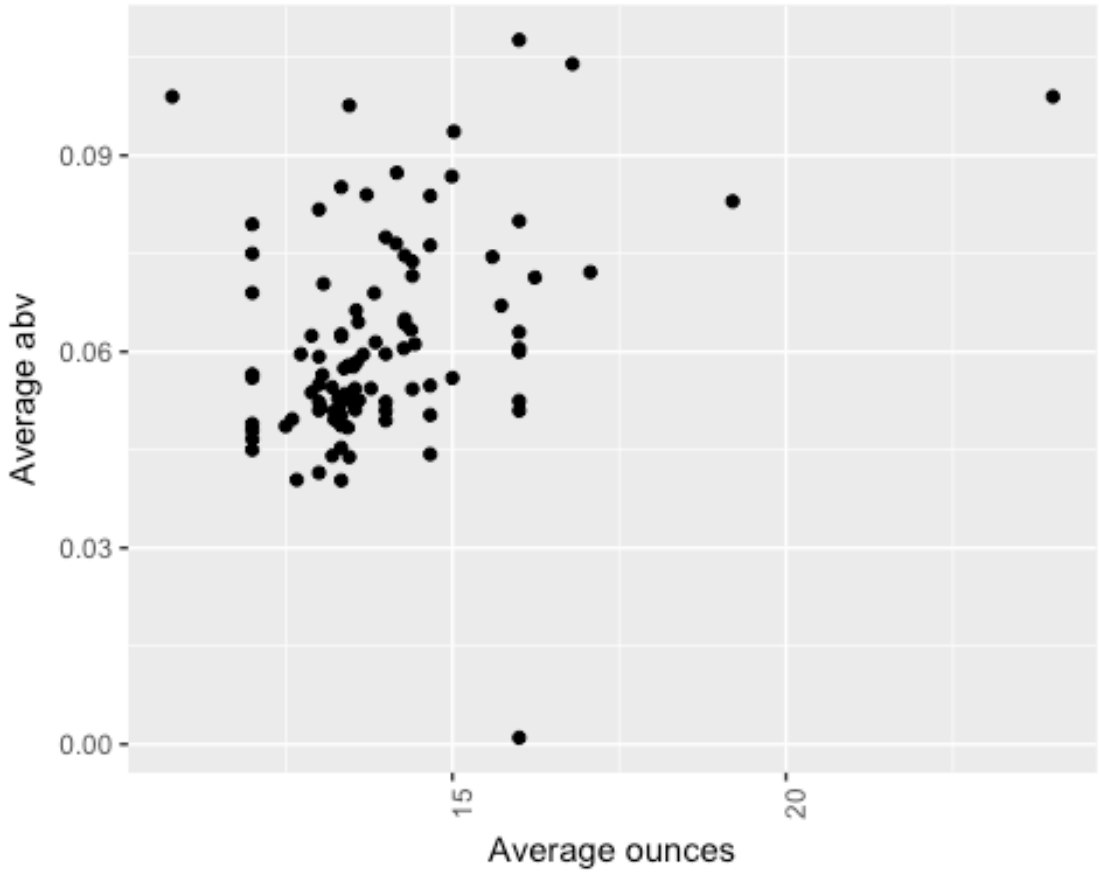
Unfortunately, we couldn't analyze the market consumption because there are some missing data such as the amount of production. It would have been interesting to have this data, and to be able to know which beer styles are the most consumed.

On the other hand, we decided to relate the average alcoholic content per beer style with the average size of a beer per style, and check if we can find any link between those two.

```
beerRaw %>% select(style, abv, ounces) %>% na.omit(beerRaw) -> beer3

beer3 %>% group_by(style) %>% summarize(NbBiere=n(), MoyenneAbv=mean(abv), Mo
yenneOunces=mean(ounces)) -> beer4

beer4 %>% ggplot(aes(x=MoyenneOunces, y=MoyenneAbv)) +
          geom_point() +
          xlab("Average ounces") +
          ylab("Average abv")  +
          theme(axis.text.x = element_text(angle = 90 ))
```



We could all think that the higher the alcoholic content, the smaller the size.

But with this graph, we can see that our hypothesis is false. Large parts of the data are focused on the same area, except few of them, which are on the extreme sides of the graph. This means that we cannot conclude any link between the average size of beer style and its alcoholic content.