craft\_brewery\_01.R

Marcelo

Thu Feb 02 12:42:47 2017

# CraftBeers dataset  
# Exploring craft Brewery data  
setwd("C:/Users/Marcelo/Desktop/DsStudy/CraftBeers")  
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

library(data.table)

## -------------------------------------------------------------------------

## data.table + dplyr code now lives in dtplyr.  
## Please library(dtplyr)!

## -------------------------------------------------------------------------

##   
## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':  
##   
## between, first, last

library(ggplot2)  
library(rmarkdown)  
  
df\_beer <- data.table(read.csv("beers.csv", header = TRUE,  
 stringsAsFactors = FALSE))  
df\_brew <- data.table(read.csv("breweries.csv", header = TRUE,  
 stringsAsFactors = FALSE))  
colnames(df\_brew) <- c("brewery\_id", "name", "city", "state")  
  
str(df\_beer)

## Classes 'data.table' and 'data.frame': 2410 obs. of 8 variables:  
## $ X : int 0 1 2 3 4 5 6 7 8 9 ...  
## $ abv : num 0.05 0.066 0.071 0.09 0.075 0.077 0.045 0.065 0.055 0.086 ...  
## $ ibu : num NA NA NA NA NA NA NA NA NA NA ...  
## $ id : int 1436 2265 2264 2263 2262 2261 2260 2259 2258 2131 ...  
## $ name : chr "Pub Beer" "Devil's Cup" "Rise of the Phoenix" "Sinister" ...  
## $ style : chr "American Pale Lager" "American Pale Ale (APA)" "American IPA" "American Double / Imperial IPA" ...  
## $ brewery\_id: int 408 177 177 177 177 177 177 177 177 177 ...  
## $ ounces : num 12 12 12 12 12 12 12 12 12 12 ...  
## - attr(\*, ".internal.selfref")=<externalptr>

str(df\_brew)

## Classes 'data.table' and 'data.frame': 558 obs. of 4 variables:  
## $ brewery\_id: int 0 1 2 3 4 5 6 7 8 9 ...  
## $ name : chr "NorthGate Brewing " "Against the Grain Brewery" "Jack's Abby Craft Lagers" "Mike Hess Brewing Company" ...  
## $ city : chr "Minneapolis" "Louisville" "Framingham" "San Diego" ...  
## $ state : chr " MN" " KY" " MA" " CA" ...  
## - attr(\*, ".internal.selfref")=<externalptr>

summary(df\_beer)

## X abv ibu id   
## Min. : 0.0 Min. :0.00100 Min. : 4.00 Min. : 1.0   
## 1st Qu.: 602.2 1st Qu.:0.05000 1st Qu.: 21.00 1st Qu.: 808.2   
## Median :1204.5 Median :0.05600 Median : 35.00 Median :1453.5   
## Mean :1204.5 Mean :0.05977 Mean : 42.71 Mean :1431.1   
## 3rd Qu.:1806.8 3rd Qu.:0.06700 3rd Qu.: 64.00 3rd Qu.:2075.8   
## Max. :2409.0 Max. :0.12800 Max. :138.00 Max. :2692.0   
## NA's :62 NA's :1005   
## name style brewery\_id ounces   
## Length:2410 Length:2410 Min. : 0.0 Min. : 8.40   
## Class :character Class :character 1st Qu.: 93.0 1st Qu.:12.00   
## Mode :character Mode :character Median :205.0 Median :12.00   
## Mean :231.7 Mean :13.59   
## 3rd Qu.:366.0 3rd Qu.:16.00   
## Max. :557.0 Max. :32.00   
##

summary(df\_brew)

## brewery\_id name city state   
## Min. : 0.0 Length:558 Length:558 Length:558   
## 1st Qu.:139.2 Class :character Class :character Class :character   
## Median :278.5 Mode :character Mode :character Mode :character   
## Mean :278.5   
## 3rd Qu.:417.8   
## Max. :557.0

head(df\_brew)

## brewery\_id name city state  
## 1: 0 NorthGate Brewing Minneapolis MN  
## 2: 1 Against the Grain Brewery Louisville KY  
## 3: 2 Jack's Abby Craft Lagers Framingham MA  
## 4: 3 Mike Hess Brewing Company San Diego CA  
## 5: 4 Fort Point Beer Company San Francisco CA  
## 6: 5 COAST Brewing Company Charleston SC

head(df\_beer)

## X abv ibu id name style  
## 1: 0 0.050 NA 1436 Pub Beer American Pale Lager  
## 2: 1 0.066 NA 2265 Devil's Cup American Pale Ale (APA)  
## 3: 2 0.071 NA 2264 Rise of the Phoenix American IPA  
## 4: 3 0.090 NA 2263 Sinister American Double / Imperial IPA  
## 5: 4 0.075 NA 2262 Sex and Candy American IPA  
## 6: 5 0.077 NA 2261 Black Exodus Oatmeal Stout  
## brewery\_id ounces  
## 1: 408 12  
## 2: 177 12  
## 3: 177 12  
## 4: 177 12  
## 5: 177 12  
## 6: 177 12

full\_beer\_data <- merge(df\_beer, df\_brew, by = "brewery\_id")  
  
head(full\_beer\_data)

## brewery\_id X abv ibu id name.x  
## 1: 0 1493 0.045 50 2692 Get Together  
## 2: 0 1494 0.049 26 2691 Maggie's Leap  
## 3: 0 1495 0.048 19 2690 Wall's End  
## 4: 0 1496 0.060 38 2689 Pumpion  
## 5: 0 1497 0.060 25 2688 Stronghold  
## 6: 0 1498 0.056 47 2687 Parapet ESB  
## style ounces name.y  
## 1: American IPA 16 NorthGate Brewing   
## 2: Milk / Sweet Stout 16 NorthGate Brewing   
## 3: English Brown Ale 16 NorthGate Brewing   
## 4: Pumpkin Ale 16 NorthGate Brewing   
## 5: American Porter 16 NorthGate Brewing   
## 6: Extra Special / Strong Bitter (ESB) 16 NorthGate Brewing   
## city state  
## 1: Minneapolis MN  
## 2: Minneapolis MN  
## 3: Minneapolis MN  
## 4: Minneapolis MN  
## 5: Minneapolis MN  
## 6: Minneapolis MN

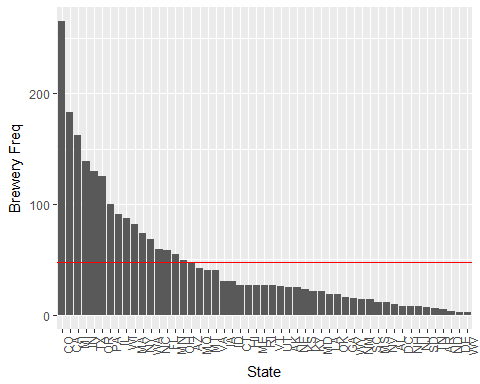
names(full\_beer\_data)

## [1] "brewery\_id" "X" "abv" "ibu" "id"   
## [6] "name.x" "style" "ounces" "name.y" "city"   
## [11] "state"

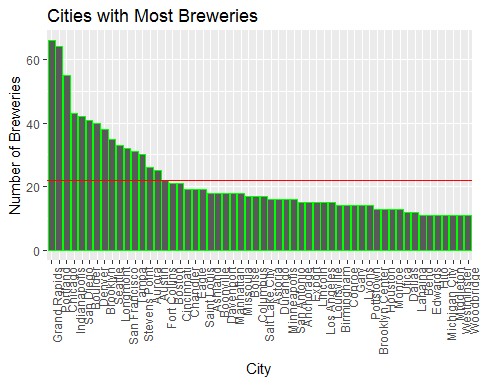
colnames(full\_beer\_data) <- c("brewery\_id", "X", "abv", "ibu", "id", "beer\_name",  
 "style", "ounces", "brewery\_name", "city", "state")  
# Dropping repeated columns  
full\_beer\_data = full\_beer\_data[,3:11]  
  
# transforming ABV into %  
full\_beer\_data$abv = full\_beer\_data$abv\*100  
  
head(full\_beer\_data)

## abv ibu id beer\_name style ounces  
## 1: 4.5 50 2692 Get Together American IPA 16  
## 2: 4.9 26 2691 Maggie's Leap Milk / Sweet Stout 16  
## 3: 4.8 19 2690 Wall's End English Brown Ale 16  
## 4: 6.0 38 2689 Pumpion Pumpkin Ale 16  
## 5: 6.0 25 2688 Stronghold American Porter 16  
## 6: 5.6 47 2687 Parapet ESB Extra Special / Strong Bitter (ESB) 16  
## brewery\_name city state  
## 1: NorthGate Brewing Minneapolis MN  
## 2: NorthGate Brewing Minneapolis MN  
## 3: NorthGate Brewing Minneapolis MN  
## 4: NorthGate Brewing Minneapolis MN  
## 5: NorthGate Brewing Minneapolis MN  
## 6: NorthGate Brewing Minneapolis MN

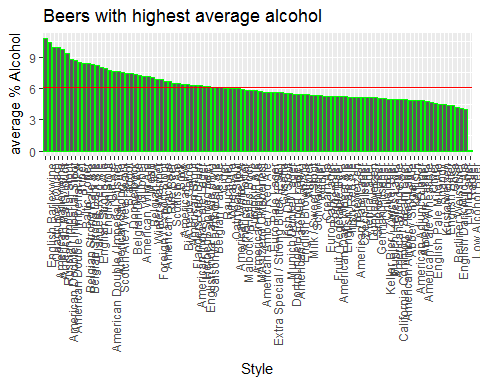
# Number of breweries x State  
breweryxstate <- full\_beer\_data %>%  
 group\_by(brewery\_name)%>%  
 group\_by(state)%>%  
 summarise(Freq = n())  
  
ggplot(data = breweryxstate, aes(x = reorder(state, -Freq), y = Freq)) +  
 geom\_bar(stat="identity") +  
 labs(x = "State", y = "Brewery Freq") +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1))+  
 geom\_hline(yintercept=mean(breweryxstate$Freq), col = "red")



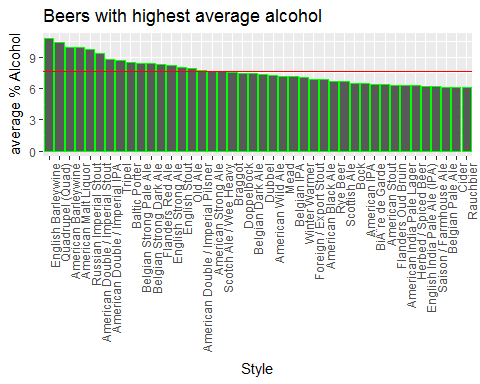
# Number of breweries x State  
breweryxcity <- full\_beer\_data %>%  
 group\_by(brewery\_name) %>%  
 group\_by(city) %>%  
 summarise(Freq = n())  
  
# subsetting for cities with more than 10 breweries  
  
breweryxcity = breweryxcity[which(breweryxcity[,2]>10),]  
  
ggplot(data = breweryxcity, aes(x = reorder(city, -Freq), y = Freq)) +  
 geom\_bar(stat="identity", col = "green") +  
 labs(x = "City", y = "Number of Breweries", title = "Cities with Most Breweries") +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1))+  
 geom\_hline(yintercept=mean(breweryxcity$Freq), col = "red")



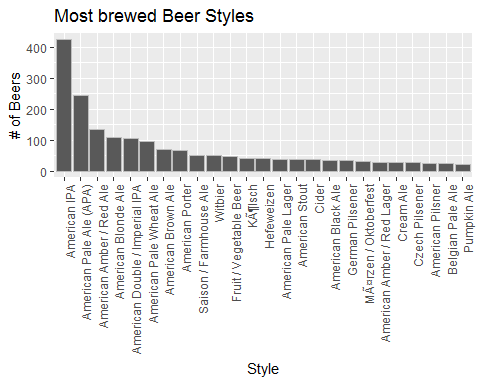
# Beer with highest average alcohol content  
  
abvxstyle <- full\_beer\_data %>%  
 filter(!is.na(abv)) %>%  
 group\_by(abv)%>%  
 group\_by(style) %>%  
 summarise(Mean = sum(abv)/n())  
  
ggplot(data = abvxstyle, aes(x = reorder(style, -Mean), y = Mean)) +  
 geom\_bar(stat="identity", col = "green") +  
 labs(x = "Style", y = "average % Alcohol",   
 title = "Beers with highest average alcohol") +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1))+  
 geom\_hline(yintercept=mean(abvxstyle$Mean), col = "red")



# Beers with more than 6% of alcohol content  
abvxstyle = abvxstyle[which(abvxstyle[,2]>6),]  
  
ggplot(data = abvxstyle, aes(x = reorder(style, -Mean), y = Mean)) +  
 geom\_bar(stat="identity", col = "green") +  
 labs(x = "Style", y = "average % Alcohol",   
 title = "Beers with highest average alcohol") +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1))+  
 geom\_hline(yintercept=mean(abvxstyle$Mean), col = "red")



# Most brewed Beer Styles  
beerxstyle <- full\_beer\_data %>%  
 group\_by(beer\_name)%>%  
 group\_by(style) %>%  
 summarise(Sum = n())  
  
beerxstyle = beerxstyle[which(beerxstyle[,2]>20),]  
  
ggplot(data = beerxstyle, aes(x = reorder(style, -Sum), y = Sum)) +  
 geom\_bar(stat="identity", col = "grey") +  
 labs(x = "Style", y = "# of Beers",   
 title = "Most brewed Beer Styles") +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1))



# Outcomes  
# - Colorado is the state with the highest number of breweries, followed by California and Minesota   
# - Grand Rapids, Portland and Chicago are the cities with the highest beer consume - NICE PLACES TO VISIT  
# - English Barleywine and Quadrupel present the highest alcohol content in % volume  
# - American IPA is by far the most brewed Beer Style