Case Study 1

Suchismita Moharana and Andy Walch

October 11, 2019

Introduction

In this report, we explore `beers` and `breweries` datasets for the 51 states in the US. The steps and procedures taken in this analysis are stipulated below. We successfully merged the two datasets `Beers` dataset which contains a list of 2410 US craft beers to the `Breweries` dataset containing 558 US breweries.

```
knitr::opts_chunk$set(echo = TRUE)

library(readr)
library(plotly)

## Loading required package: ggplot2

##
## Attaching package: 'plotly'

## The following object is masked from 'package:ggplot2':
##
##     last_plot

## The following object is masked from 'package:stats':
##
##     filter

## The following object is masked from 'package:graphics':
##
##     layout

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(ggplot2)

df_beers <- read_csv("Beers.csv")

## Parsed with column specification:
## cols(
##   Name = col_character(),
##   Beer_ID = col_double(),
##   ABV = col_double(),
##   IBU = col_double(),
##   Brewery_id = col_double(),
##   Style = col_character(),
##   Ounces = col_double()
## )

df_breweries <- read.csv("Breweries.csv")
```

Analysis Questions

In this section, we address the research questions put together on this two dataset. The questions are numbered 1 to 7

1. How many breweries are present in each state?

We answer this question by using `count` function in base to count the number of breweries grouped by "state". This creates a dataframe named `df_count` with two columns `State` which is the State name and `Breweries` which is the number of breweries in a given state. Each row represents one State. Colorado has the highest number of breweries {47} where as Washington DC, South Dakota (SD), North Dakota (ND), and West Virginia (WV) tie for the least amount of breweries each with just 1.

```
#Code
df_count <- count(df_breweries, df_breweries$State, sort=FALSE)
names(df_count)[1] <- "State"
names(df_count)[2] <- "Breweries"
df_count

## # A tibble: 51 x 2
##    State Breweries
##    <fct>     <int>
##  1 " AK"         7
##  2 " AL"         3
##  3 " AR"         2
##  4 " AZ"        11
##  5 " CA"        39
##  6 " CO"        47
##  7 " CT"         8
##  8 " DC"         1
##  9 " DE"         2
```

```
## 10 " FL"        15
## # … with 41 more rows

count_wrap <- cbind(df_count[1:(nrow(df_count)/5), ],
df_count[(1+(nrow(df_count)/5)):(10+(nrow(df_count)/5)), ],
df_count[(11+(nrow(df_count)/5)):(20+(nrow(df_count)/5)),],
df_count[(21+(nrow(df_count)/5)):(30+(nrow(df_count)/5)),],
df_count[(31+(nrow(df_count)/5)):(40+(nrow(df_count)/5)),])

count_wrap

##    State Breweries State Breweries State Breweries State Breweries State
## 1     AK          7    GA          7    MD          7    NH          3    SC
## 2     AL          3    HI          4    ME          9    NJ          3    SD
## 3     AR          2    IA          5    MI         32    NM          4    TN
## 4     AZ         11    ID          5    MN         12    NV          2    TX
## 5     CA         39    IL         18    MO          9    NY         16    UT
## 6     CO         47    IN         22    MS          2    OH         15    VA
## 7     CT          8    KS          3    MT          9    OK          6    VT
## 8     DC          1    KY          4    NC         19    OR         29    WA
## 9     DE          2    LA          5    ND          1    PA         25    WI
## 10    FL         15    MA         23    NE          5    RI          5    WV
##    Breweries
## 1          4
## 2          1
## 3          3
## 4         28
## 5          4
## 6         16
## 7         10
## 8         23
## 9         20
## 10         1
```

2. Merge beer data with breweries data by brewery id. Print first 6 observations and the last six observations to check the merged file.

we merge `df_beers` and `df_breweries` dataframes by `Brewery_ID` using merge command for base R and assign the new dataframe to `df_breweries_and_beer`. We use `head()` and `tail()` to print the first and last 6 rows of the newly created `df_breweries_and_beer` dataframe respectively.

```
#Code
# merge two data frames by ID
#Code
names(df_beers)[5]<- "Brew_ID" #making the merged columns the same
df_breweries_and_beer <- merge(df_beers, df_breweries, by="Brew_ID")
names(df_breweries_and_beer)[2] <- "BeerName" #changing name.x to BeerName
names(df_breweries_and_beer)[8] <- "BreweryName" #changing name.y to
```

```
BreweryName
head(df_breweries_and_beer, 6)
```

```
##   Brew_ID      BeerName Beer_ID   ABV IBU
## 1       1  Get Together    2692 0.045  50
## 2       1 Maggie's Leap    2691 0.049  26
## 3       1    Wall's End    2690 0.048  19
## 4       1       Pumpion    2689 0.060  38
## 5       1    Stronghold    2688 0.060  25
## 6       1   Parapet ESB    2687 0.056  47
##                                 Style Ounces        BreweryName
## 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
```

```
tail(df_breweries_and_beer, 6)
```

```
##      Brew_ID                  BeerName Beer_ID   ABV IBU
## 2405     556             Pilsner Ukiah      98 0.055  NA
## 2406     557  Heinnieweisse Weissebier      52 0.049  NA
## 2407     557           Snapperhead IPA      51 0.068  NA
## 2408     557         Moo Thunder Stout      50 0.049  NA
## 2409     557         Porkslap Pale Ale      49 0.043  NA
## 2410     558 Urban Wilderness Pale Ale      30 0.049  NA
##                         Style Ounces                  BreweryName
## 2405         German Pilsener     12         Ukiah Brewing Company
## 2406              Hefeweizen     12       Butternuts Beer and Ale
## 2407            American IPA     12       Butternuts Beer and Ale
## 2408      Milk / Sweet Stout     12       Butternuts Beer and Ale
## 2409 American Pale Ale (APA)     12       Butternuts Beer and Ale
## 2410        English Pale Ale     12 Sleeping Lady Brewing Company
##               City State
## 2405         Ukiah    CA
## 2406 Garrattsville    NY
## 2407 Garrattsville    NY
## 2408 Garrattsville    NY
## 2409 Garrattsville    NY
## 2410     Anchorage    AK
```

3. Address the missing values in each column.

as shown in the code block below returns the summary of the number of NA's per column. International Bitterness Units of beer (IBU) has the highest number of NA's of all the available variables which is 1005.

```
for (i in 1:10){
  print(paste(names(df_breweries_and_beer)[i],":",
sum(is.na(df_breweries_and_beer[,i])))))}
```

```
## [1] "Brew_ID : 0"
## [1] "BeerName : 0"
## [1] "Beer_ID : 0"
## [1] "ABV : 62"
## [1] "IBU : 1005"
## [1] "Style : 5"
## [1] "Ounces : 0"
## [1] "BreweryName : 0"
## [1] "City : 0"
## [1] "State : 0"
```

```
df_breweries_and_beer_clean <- na.omit(df_breweries_and_beer)
```

4. Compute the median alcohol content and international bitterness unit for each state. Plot a bar chart to compare.

This code block then computes the median alcohol content(ABV) per state and stores the result in vector abv. It also computes median International Bitterness Units of the beer (IBU) and stores the result in ibu. Then plots a grid bar charts to comparing median ABV and median IBU in each of the 51 States.

```
#Code
abv <- tapply(df_breweries_and_beer$ABV, df_breweries_and_beer$State,
FUN=median, na.rm=TRUE)
abv
```

```
##     AK     AL     AR     AZ     CA     CO     CT     DC     DE     FL
## 0.0560 0.0600 0.0520 0.0550 0.0580 0.0605 0.0600 0.0625 0.0550 0.0570
##     GA     HI     IA     ID     IL     IN     KS     KY     LA     MA
## 0.0550 0.0540 0.0555 0.0565 0.0580 0.0580 0.0500 0.0625 0.0520 0.0540
##     MD     ME     MI     MN     MO     MS     MT     NC     ND     NE
## 0.0580 0.0510 0.0620 0.0560 0.0520 0.0580 0.0550 0.0570 0.0500 0.0560
##     NH     NJ     NM     NV     NY     OH     OK     OR     PA     RI
## 0.0550 0.0460 0.0620 0.0600 0.0550 0.0580 0.0600 0.0560 0.0570 0.0550
##     SC     SD     TN     TX     UT     VA     VT     WA     WI     WV
## 0.0550 0.0600 0.0570 0.0550 0.0400 0.0565 0.0550 0.0555 0.0520 0.0620
##     WY
## 0.0500
```

```r
ibu <- tapply(df_breweries_and_beer$IBU, df_breweries_and_beer$State,
FUN=median, na.rm=TRUE)
ibu
```

```
##   AK   AL   AR   AZ   CA   CO   CT   DC   DE   FL   GA   HI   IA   ID   IL
## 46.0 43.0 39.0 20.5 42.0 40.0 29.0 47.5 52.0 55.0 55.0 22.5 26.0 39.0 30.0
##   IN   KS   KY   LA   MA   MD   ME   MI   MN   MO   MS   MT   NC   ND   NE
## 33.0 20.0 31.5 31.5 35.0 29.0 61.0 35.0 44.5 24.0 45.0 40.0 33.5 32.0 35.0
##   NH   NJ   NM   NV   NY   OH   OK   OR   PA   RI   SC   SD   TN   TX   UT
## 48.5 34.5 51.0 41.0 47.0 40.0 35.0 40.0 30.0 24.0 30.0   NA 37.0 33.0 34.0
##   VA   VT   WA   WI   WV   WY
## 42.0 30.0 38.0 19.0 57.5 21.0
```

```r
states <- df_count[,1]
abv_percent <- abv*100 #making these values percents so that the comparisons
are easier to see on the graph
head(abv_percent)
```

```
##   AK   AL   AR   AZ   CA   CO
## 5.60 6.00 5.20 5.50 5.80 6.05
```

```r
medians <- data.frame(ibu,abv_percent)
#medians
ibu_abv <- data.frame(c(medians$ibu,medians$abv),states)
names(ibu_abv)[1] <- "Medians"
ibu_abv$Measure <- c(rep("IBU",length(ibu)),rep("ABV",length(abv)))
ibu_abv
```

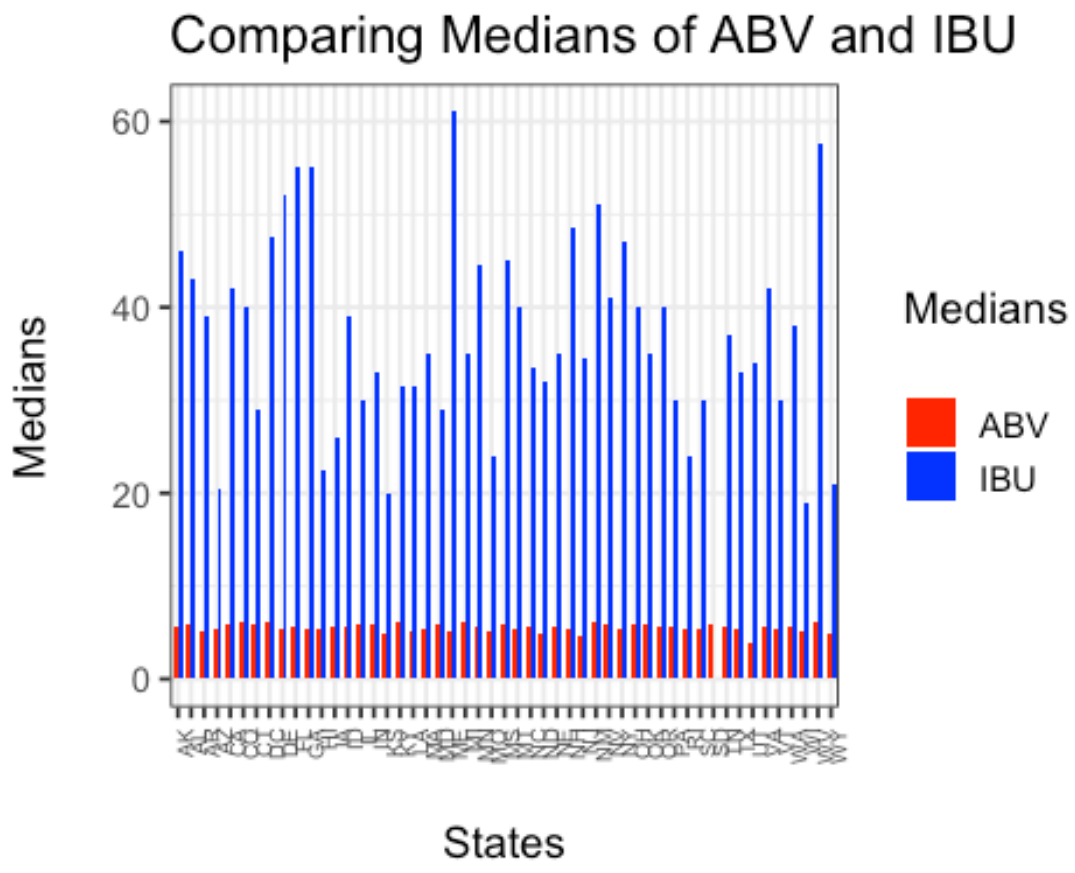
```
##     Medians State Measure
## 1     46.00    AK     IBU
## 2     43.00    AL     IBU
## 3     39.00    AR     IBU
## 4     20.50    AZ     IBU
## 5     42.00    CA     IBU
## 6     40.00    CO     IBU
## 7     29.00    CT     IBU
## 8     47.50    DC     IBU
## 9     52.00    DE     IBU
## 10    55.00    FL     IBU
## 11    55.00    GA     IBU
## 12    22.50    HI     IBU
## 13    26.00    IA     IBU
## 14    39.00    ID     IBU
## 15    30.00    IL     IBU
## 16    33.00    IN     IBU
## 17    20.00    KS     IBU
## 18    31.50    KY     IBU
## 19    31.50    LA     IBU
## 20    35.00    MA     IBU
## 21    29.00    MD     IBU
## 22    61.00    ME     IBU
```

```
## 23    35.00    MI     IBU
## 24    44.50    MN     IBU
## 25    24.00    MO     IBU
## 26    45.00    MS     IBU
## 27    40.00    MT     IBU
## 28    33.50    NC     IBU
## 29    32.00    ND     IBU
## 30    35.00    NE     IBU
## 31    48.50    NH     IBU
## 32    34.50    NJ     IBU
## 33    51.00    NM     IBU
## 34    41.00    NV     IBU
## 35    47.00    NY     IBU
## 36    40.00    OH     IBU
## 37    35.00    OK     IBU
## 38    40.00    OR     IBU
## 39    30.00    PA     IBU
## 40    24.00    RI     IBU
## 41    30.00    SC     IBU
## 42       NA    SD     IBU
## 43    37.00    TN     IBU
## 44    33.00    TX     IBU
## 45    34.00    UT     IBU
## 46    42.00    VA     IBU
## 47    30.00    VT     IBU
## 48    38.00    WA     IBU
## 49    19.00    WI     IBU
## 50    57.50    WV     IBU
## 51    21.00    WY     IBU
## 52     5.60    AK     ABV
## 53     6.00    AL     ABV
## 54     5.20    AR     ABV
## 55     5.50    AZ     ABV
## 56     5.80    CA     ABV
## 57     6.05    CO     ABV
## 58     6.00    CT     ABV
## 59     6.25    DC     ABV
## 60     5.50    DE     ABV
## 61     5.70    FL     ABV
## 62     5.50    GA     ABV
## 63     5.40    HI     ABV
## 64     5.55    IA     ABV
## 65     5.65    ID     ABV
## 66     5.80    IL     ABV
## 67     5.80    IN     ABV
## 68     5.00    KS     ABV
## 69     6.25    KY     ABV
## 70     5.20    LA     ABV
## 71     5.40    MA     ABV
## 72     5.80    MD     ABV
```

```
## 73     5.10    ME     ABV
## 74     6.20    MI     ABV
## 75     5.60    MN     ABV
## 76     5.20    MO     ABV
## 77     5.80    MS     ABV
## 78     5.50    MT     ABV
## 79     5.70    NC     ABV
## 80     5.00    ND     ABV
## 81     5.60    NE     ABV
## 82     5.50    NH     ABV
## 83     4.60    NJ     ABV
## 84     6.20    NM     ABV
## 85     6.00    NV     ABV
## 86     5.50    NY     ABV
## 87     5.80    OH     ABV
## 88     6.00    OK     ABV
## 89     5.60    OR     ABV
## 90     5.70    PA     ABV
## 91     5.50    RI     ABV
## 92     5.50    SC     ABV
## 93     6.00    SD     ABV
## 94     5.70    TN     ABV
## 95     5.50    TX     ABV
## 96     4.00    UT     ABV
## 97     5.65    VA     ABV
## 98     5.50    VT     ABV
## 99     5.55    WA     ABV
## 100    5.20    WI     ABV
## 101    6.20    WV     ABV
## 102    5.00    WY     ABV
```

```
ggplot(ibu_abv,aes(State,Medians)) + geom_bar(aes(State,Medians,
fill=Measure),stat="identity",position="dodge",width=.7)+scale_fill_manual("M
edians\n", values=c("red","blue"), labels=c("ABV","IBU")) +
labs(x="\nStates",y="Medians\n")+ theme_bw(base_size=14) + theme(axis.text.x
= element_text(angle=90,hjust=1,size=7)) + ggtitle("Comparing Medians of ABV
and IBU")
```

```
## Warning: Removed 1 rows containing missing values (geom_bar).
```



5. Which state has the maximum alcoholic (ABV) beer? Which state has the most bitter (IBU) beer?

In this code block, we identify Kentucky(KY) as the State with the maximum alcoholic beer with an `ABV` of *0.125* and Oregon (OR) as the state with the most bitter beer with an `IBU` of *138*.

```
# Code
#maximum alcoholic beer
# select the row with max ABV
df_max_abv<-
data.frame(df_breweries_and_beer_clean[which(df_breweries_and_beer_clean$ABV==max(df_breweries_and_beer_clean$ABV)),])

print(paste0("The state with the beer with maximum alcohol is ->", df_max_abv$State, " with an ABV of ", df_max_abv$ABV))

## [1] "The state with the beer with maximum alcohol is -> KY with an ABV of 0.125"

# print the state with Max alcoholic beer

#state with the most bitter beer
```

```
# select the row with max IBU
df_max_ibu<-
data.frame(df_breweries_and_beer_clean[which(df_breweries_and_beer_clean$IBU==max(df_breweries_and_beer_clean$IBU)),])

print(paste0("The state with Most bitter beer is ->", df_max_ibu$State, " with IBU of ", df_max_ibu$IBU))

## [1] "The state with Most bitter beer is -> OR with IBU of 138"
```

6. Comment on the summary statistics and distribution of the ABV variable.

To get the summary statistics of ABV by Volume variable, we are using summary function.

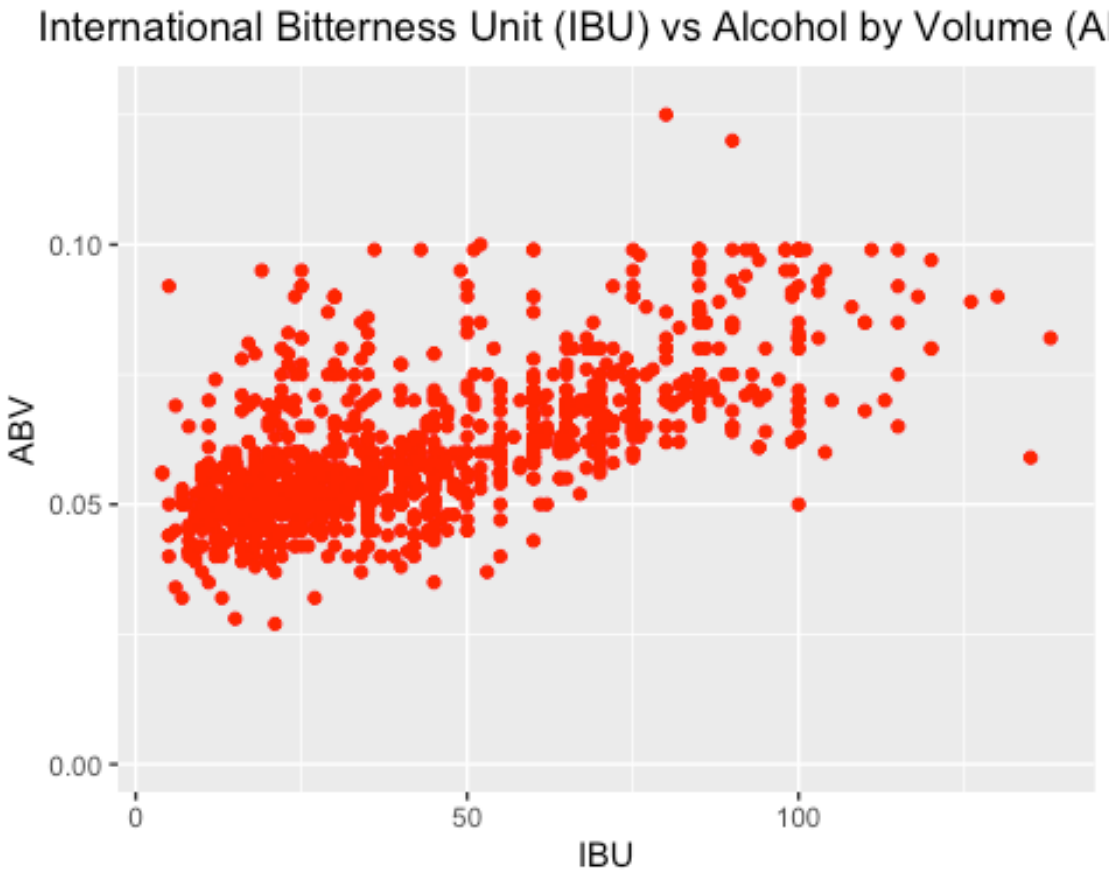
```
#Code
summary(df_breweries_and_beer_clean$ABV)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.02700 0.05000 0.05700 0.05992 0.06800 0.12500
```

7. Is there an apparent relationship between the bitterness of the beer and its alcoholic content? Draw a scatter plot. Make your best judgment of a relationship and EXPLAIN your answer.

There is a positive correlation between ABV and IBU as shown in the regression trend line in the scatter plot below. IBU increases with an increase in ABV.

```
ggplot(df_breweries_and_beer, aes(df_breweries_and_beer$IBU, df_breweries_and_beer$ABV)) + geom_point(color = ("red") , na.rm=TRUE) + labs(title = "International Bitterness Unit (IBU) vs Alcohol by Volume (ABV)", x = "IBU", y ="ABV") +theme(plot.title = element_text(hjust = 0.5))
```


International Bitterness Unit (IBU) vs Alcohol by Volume (Al
0.10
0.05
0.00
ABV
0
50
100
IBU