MSDS 6306 - Case Study 1- The Beer Project

Swee Chew, Julius Mwangi, and Volodymyr Orlov

27 June, 2018

Introduction

According to a report issued by the Brewers Association in 2016, "in just four years, the economic impact of small and independent U.S. craft brewers has doubled,". The study further reported that, "the inustry contributed $68.7B to the U.S. economy and provides more than 456,000 full time jobs"[1].

Our study is geared towards providing an understanding of the beer industry and establishing possibilities for our client, on the premise that the beer industry is flourishing with an insatiable demand. We understand that for years, a few powerful breweries controlled the beer market. We also learned that craft breweries, i.e. small, independent and traditional microbreweries, have flooded the beer market and curved their market share through their fan base that comprises "drinkers who appreciate beer that is locally made, produced in small batches to high quality standards and of course, the sheer variety and multitude of flavors on offer."[2].

Our research seeks to answer the following questions:

- How many breweries are there in each state?
- Which state has the highest number of breweries?
- What are the median alcohol content and international bitterness unit for each state?
- Which state produces beer with the maximum alcohol by volume?
- Which state produces the most bitter beer?
- What is the alcoholic content and bitterness variability across the states?
- Is there a relationship between alcoholic content and bitterness of the beer?

In analyzing the beer industry's landscape we shall provide our client a bird's eye view on the overall brewing industry to ultimately support the decision of whether or not to venture into the beer industry. Due to limitation on the data provided, our report will focus on production only i.e. an understanding of how many breweries there are in the country, and the types of beer they produce.

Data

We obtained two datasets consisting beer and brewery information. The first dataset, *Beers.csv*, contains a list of 2410 US craft beers. Each beer in the beers dataset is described by *name*, *beer ID*, *alcoholic content*, *bitterness*, *brewery ID*, *style* and *ounces*. The second dataset, *Breweries.csv*, contains 558 US breweries and 2305 distinct beers. Each brewery is described by *name*, *brewery ID*, *city* and *state*. Our research involves collating and analyzing the two datasets. We shall use these two datasets to understand the breweries per state, the beer quality in terms of alcohol content and international bitterness unit, and which states have beers with the most alcohol and have the most bitter beer. As further noted in the paragraph below on "missing values", some breweries did not have all the data for their production, which might skew our report on either side. [3]

```
beers <- read.csv ('data/Beers.csv', header=T, sep=",")
breweries <- read.csv ('data/Breweries.csv', header=T, sep=",")
```

[1] https://www.brewbound.com/news/study-us-craft-beer-industry-contributes-68-billion-economy

[2] https://www.forbes.com/sites/niallmccarthy/2018/01/25/the-u-s-beer-industrys-workforce-more-than-doubled-in-a-decade-infographic/#6a92c25d1255

[3] We have no information whether the data collected was voluntary or as a result of state reporting requirement, as such we cannot vet into its accuracy hence reliability

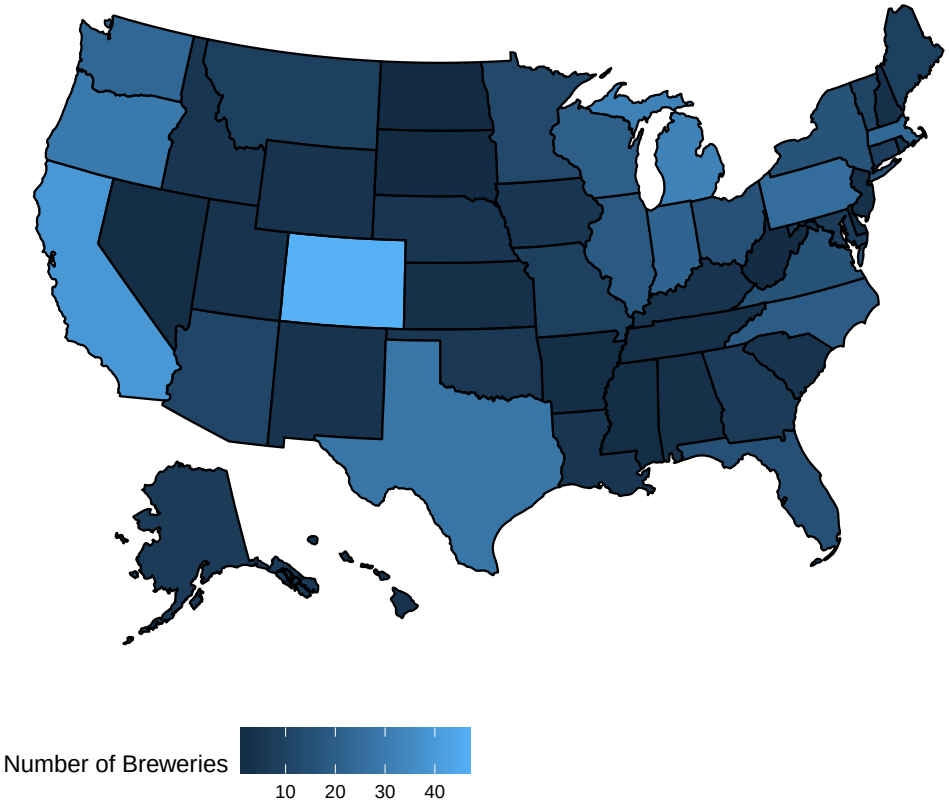
Breweries Summary

The *summary* function is used to obtain the summary statistic of *State* column within *breweries* data set. Below is a list of 51 states with the number of breweries present in each state:

```
summary(breweries$State)
```

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

```
summ <- data.frame(sapply(names(summary(breweries$State)), function(x) substring(x, 2)))
summ$Count <- as.numeric(summary(breweries$State))
colnames(summ) <- c('state', 'count')
plot_usmap(include = summ$state, data=summ, values='count') + labs(fill='Number of Breweries') + theme(
```



The top five states with the highest number of breweries are:

1. Colorado, 47 breweries
2. California, 39 breweries
3. Michigan, 32 breweries
4. Oregon, 29 breweries
5. Texas, 28 breweries

Of these, three are in the West Coast (Colorado, California, Oregon), one is in the Midwest (Michigan) and one is in the South Central (Texas). Overall, there are 23 states with five breweries and below. The states with only one brewery are: District of Columbia, North Dakota, South Dakota and West Virginia. Two of these are in the East Coast and two are in the Midwestern region of the United States. From this summary

and the map, we can see that the West Coast has a slightly larger number of breweries on average than other regions of the United States. Also, Colorado distinctly has a larger number of breweries compared to the rest of the states.

Beer Brands and Corresponding Breweries

To be able to analyze *International Bitterness Units (IBU)* and *Alcohol by Volume (ABV)* content of a beer at the state level, we've merged datasets, *breweries* and *beers*, into a single table using *Brewery_id* as a join key. The first and last 6 lines of the full table are shown below.

```
names(breweries)[names(breweries) == "Brew_ID"] <- "Brewery_id"
beers_breweries <- merge(breweries,beers,by="Brewery_id")
head(beers_breweries)
```

```
##   Brewery_id             Name.x        City State        Name.y Beer_ID
## 1          1 NorthGate Brewing  Minneapolis    MN       Pumpion    2689
## 2          1 NorthGate Brewing  Minneapolis    MN    Stronghold    2688
## 3          1 NorthGate Brewing  Minneapolis    MN   Parapet ESB    2687
## 4          1 NorthGate Brewing  Minneapolis    MN  Get Together    2692
## 5          1 NorthGate Brewing  Minneapolis    MN Maggie's Leap    2691
## 6          1 NorthGate Brewing  Minneapolis    MN    Wall's End    2690
##     ABV IBU                              Style Ounces
## 1 0.060  38                        Pumpkin Ale     16
## 2 0.060  25                     American Porter     16
## 3 0.056  47 Extra Special / Strong Bitter (ESB)     16
## 4 0.045  50                         American IPA     16
## 5 0.049  26                  Milk / Sweet Stout     16
## 6 0.048  19                   English Brown Ale     16
```

```
tail(beers_breweries)
```

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

IBUs measure parts per million of isohumulone found in a beer according to the Beer Connoisseur website[4]. It

[4]https://beerconnoisseur.com

further adds that isohumulone is the acid found in hops that gives beer its bitter bite. Its (IBU) measurement ranges from 0-100, with 100 being the highest. Bitterness, however, is relative, as often it's sweetened. On the other hand, ABV is measured as a percentage and it indicates how much of the beer is alcohol by volume. The website further stated that the IBU and ABV measurements are legally required to be imprinted on the beer.

Missing Values

Before doing analysis of IBU and ABV content of beers, we screened the data for missing values. Based on our analysis, we found out that 'ABV' and 'IBU' columns contain 62 NA's, 1005 NA's respectively. This could affect the overall reporting on the ABV and IBU measurements.

```
  colSums(is.na(beers_breweries))
```

```
## Brewery_id     Name.x       City      State     Name.y    Beer_ID
##          0          0          0          0          0          0
##        ABV        IBU      Style     Ounces
##         62       1005          0          0
```

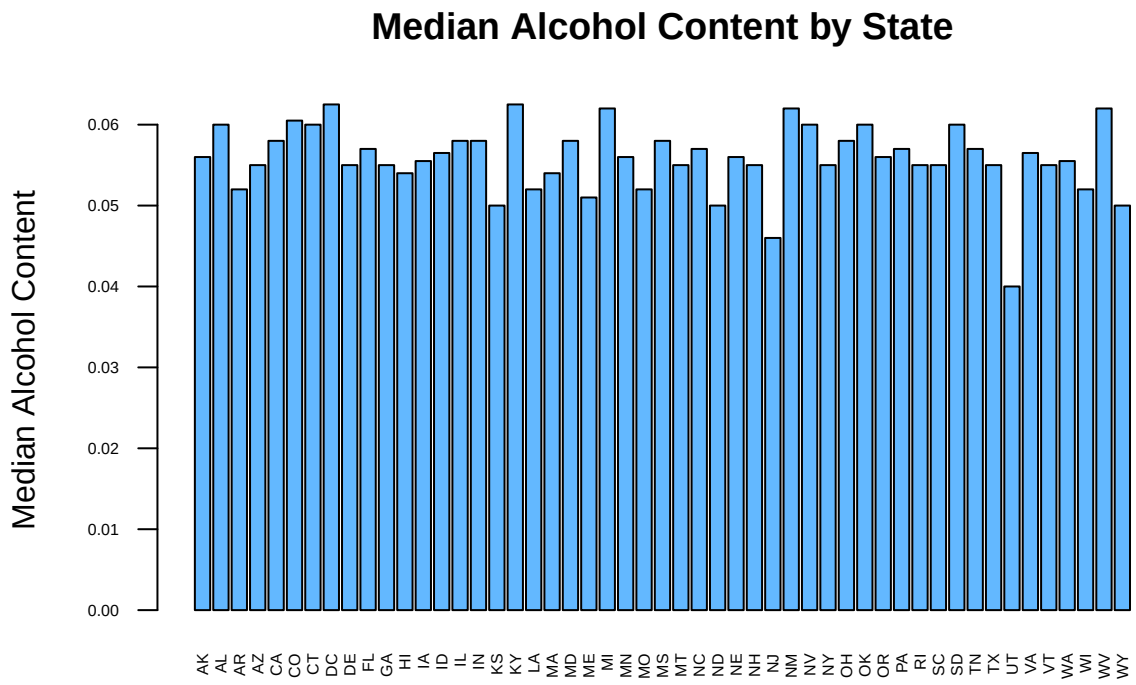
Median Alcohol Content (ABV) and International Bitterness Unit (IBU) per State

Main focus of this project is the aggregate analysis of beer production and the IBU and ABV levels. Here we look at the median IBU and ABV across the states. We found that the State of Utah has the lowest median alcohol by volume at 0.40 while Washington DC and Kentucky have the highest median alcohol by volume of the beer at 0.625.

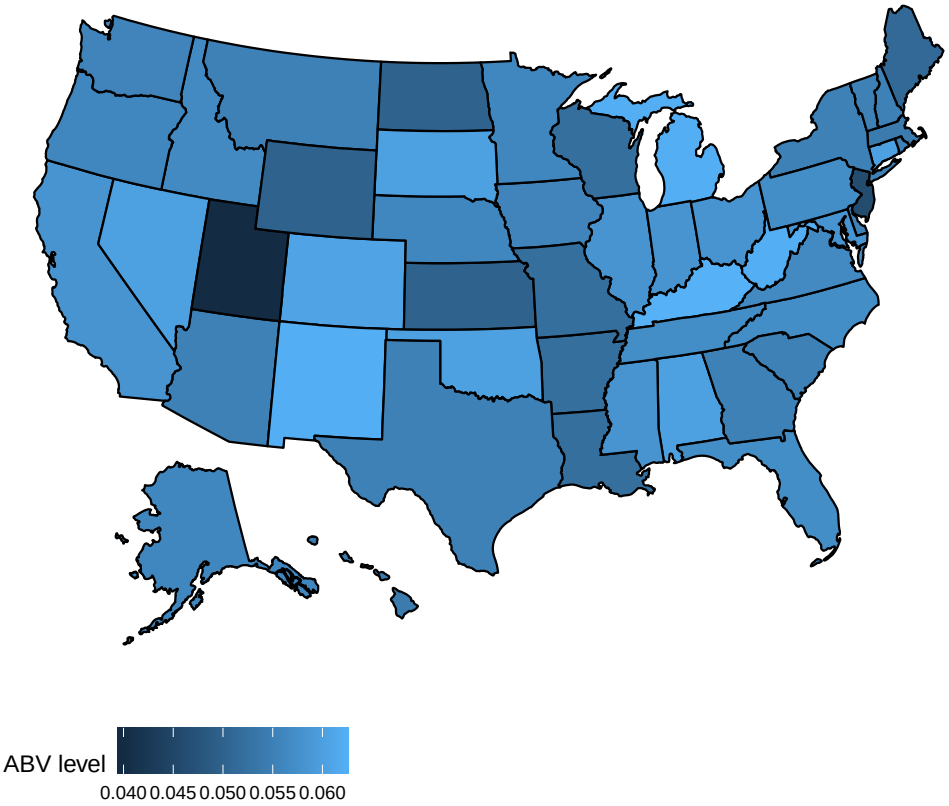
```
median_ABV <- tapply(beers_breweries$ABV, beers_breweries$State, median, na.rm = TRUE)
median_ABV
```

```
##     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
```

```
barplot(median_ABV, xlab = "State", ylab = "Median Alcohol Content", main = "Median Alcohol Content by S
```




```
abv_summ <- data.frame(sapply(names(median_ABV), function(x) substring(x, 2)))
abv_summ$abv <- as.numeric(median_ABV)
colnames(abv_summ) <- c('state', 'abv')
plot_usmap(include = abv_summ$state, data=abv_summ, values='abv') + labs(fill='ABV level') + theme(leger
```

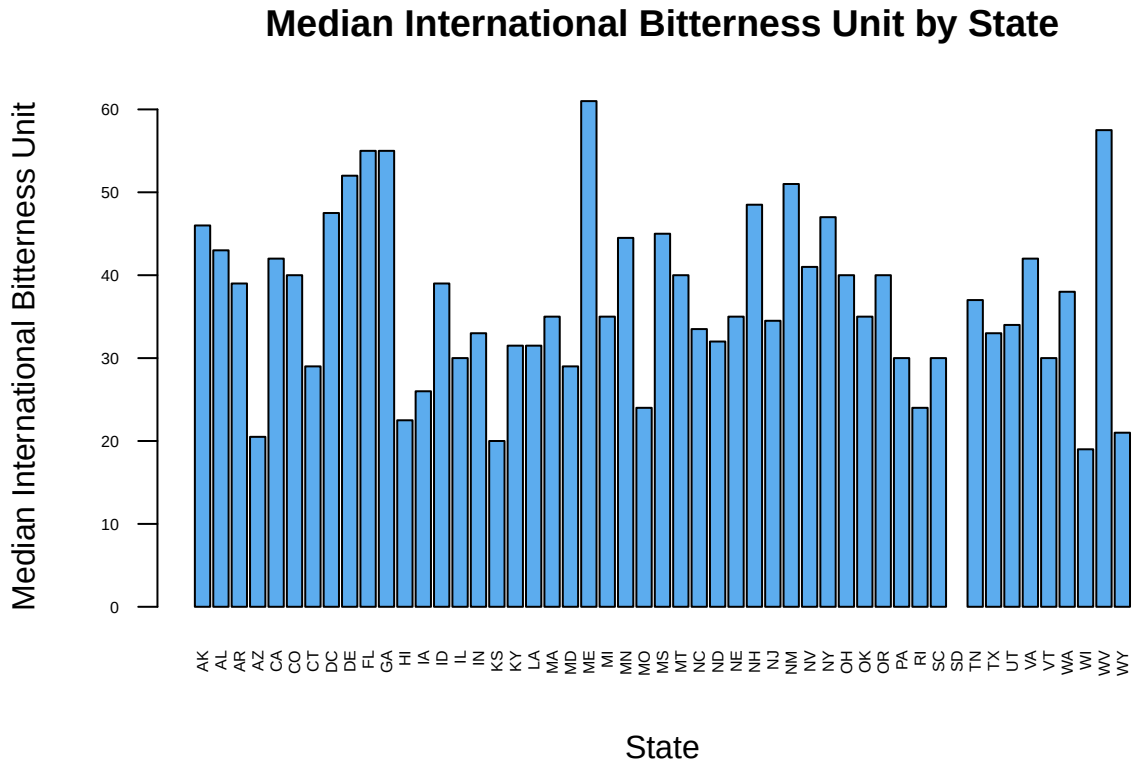


On IBU measurement, Wisconsin has the lowest IBU measurement at 19.0, while Maine has the highest at 61.0. There is no IBU data for South Dakota. This is rather odd as in the United States, IBU percentage is required to be printed on the beer. The most likely reason is that the only brewery in South Dakota did not provide the IBU information.

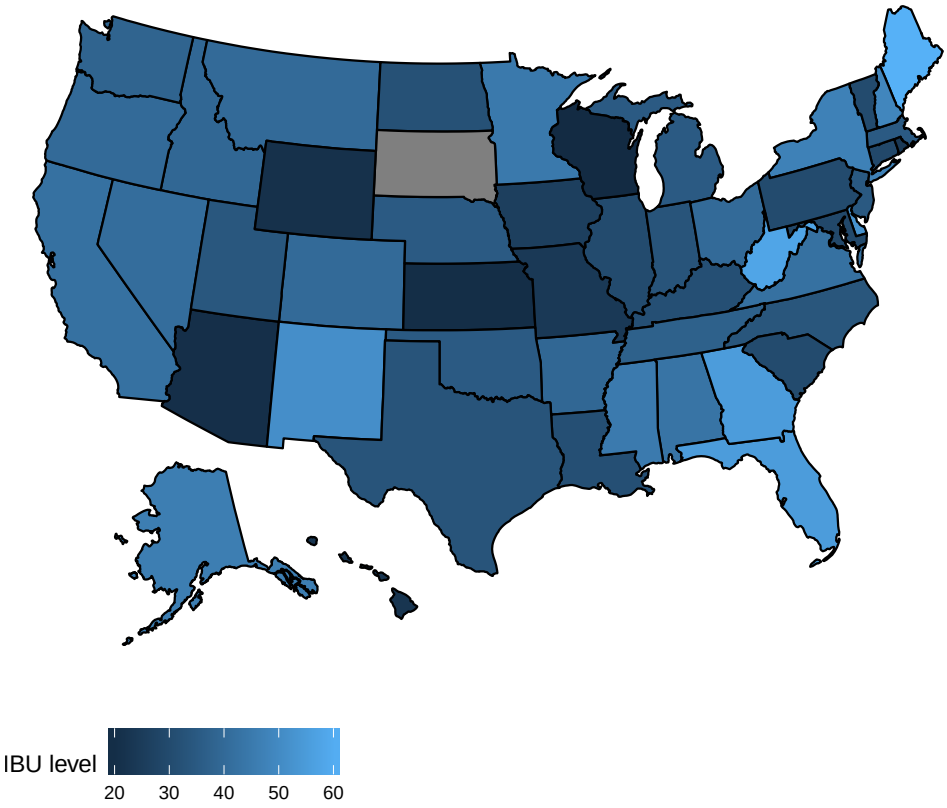
```
median_IBU <- tapply(beers_breweries$IBU, beers_breweries$State, median, na.rm = TRUE)
median_IBU
```

```
##   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
```

```
barplot(median_IBU, xlab = "State", ylab = "Median International Bitterness Unit", main = "Median Interr
```




```
ibu_summ <- data.frame(sapply(names(median_IBU), function(x) substring(x, 2)))
ibu_summ$ibu <- as.numeric(median_IBU)
colnames(ibu_summ) <- c('state', 'ibu')
plot_usmap(include = ibu_summ$state, data=ibu_summ, values='ibu') + labs(fill='IBU level') + theme(lege
```



State with the Maximum Alcoholic (ABV) Beer

As per our analysis, Colorado has the maximum alcohol by volume of the beer at 0.128, followed by Kentucky with 0.125, Indiana with 0.120, and New York with 0.100. The lowest is Delaware with 0.055.

```
max_ABV <- tapply(beers_breweries$ABV, beers_breweries$State, max, na.rm = TRUE)
max_ABV1 <- sort(max_ABV, decreasing = TRUE)
max_ABV1
```

```
##    CO    KY    IN    NY    CA    ID    MA    ME    MI    MN    NC    NJ
## 0.128 0.125 0.120 0.100 0.099 0.099 0.099 0.099 0.099 0.099 0.099 0.099
##    NV    OH    PA    TX    WI    SC    IL    NE    VT    AZ    IA    AL
## 0.099 0.099 0.099 0.099 0.099 0.097 0.096 0.096 0.096 0.095 0.095 0.093
##    DC    CT    UT    LA    OR    VA    RI    KS    MD    OK    WA    HI
## 0.092 0.090 0.090 0.088 0.088 0.088 0.086 0.085 0.085 0.085 0.084 0.083
##    FL    MO    MS    NM    MT    GA    WY    SD    AK    ND    WV    NH
## 0.082 0.080 0.080 0.080 0.075 0.072 0.072 0.069 0.068 0.067 0.067 0.065
##    TN    AR    DE
## 0.062 0.061 0.055
```

```
head(max_ABV1,1)
```

```
##    CO
## 0.128
```

State with the Most Bitter (IBU) Beer

According to a post in a beer bloggers website, humans can only detect up to about 100 IBUs in beer[5], as such any measurement above 100 is a waste as the human taste buds cannot experience the difference in bitterness.

From our analysis below, a brewery in Oregon in the West Coast produces the most bitter beer at 138 IBU. Sixteen states in total reported beer with an IBU exceeding 100.

```
max_IBU <- tapply(beers_breweries$IBU, beers_breweries$State, max, na.rm = TRUE)
```

```
## Warning in FUN(X[[i]], ...): no non-missing arguments to max; returning -
## Inf
```

```
max_IBU1 <- sort(max_IBU, decreasing = TRUE)
max_IBU1
```

```
##   OR   VA   MA   OH   MN   VT   TX   CA   DC   IN   MI   PA   NY   KS   CO
##  138  135  130  126  120  120  118  115  115  115  115  113  111  110  104
##   AL   ID   IL   NJ   NM   OK   AZ   IA   NC   MD   NV   MO   CT   UT   WA
##  103  100  100  100  100  100   99   99   98   90   90   89   85   83   83
##   FL   NH   KY   MS   MT   WI   HI   RI   WY   AK   WV   ME   ND   GA   NE
##   82   82   80   80   80   80   75   75   75   71   71   70   70   65   65
##   SC   TN   LA   DE   AR   SD
##   65   61   60   52   39 -Inf
```

```
head(max_IBU1,1)
```

```
##  OR
## 138
```

Summary Statistics for ABV Variable

Our analysis shows that the range of ABV column is from 0.001 to 0.128 percentage, the distribution is not normal and it's slightly right-skewed. Our dataset has 62 beers with missing ABV value, the breweries that chose not to provide the data could have attributed to this non-normality. We cannot tell how the distribution would have responded if all the data were reported.

```
summary(beers_breweries$ABV)
```

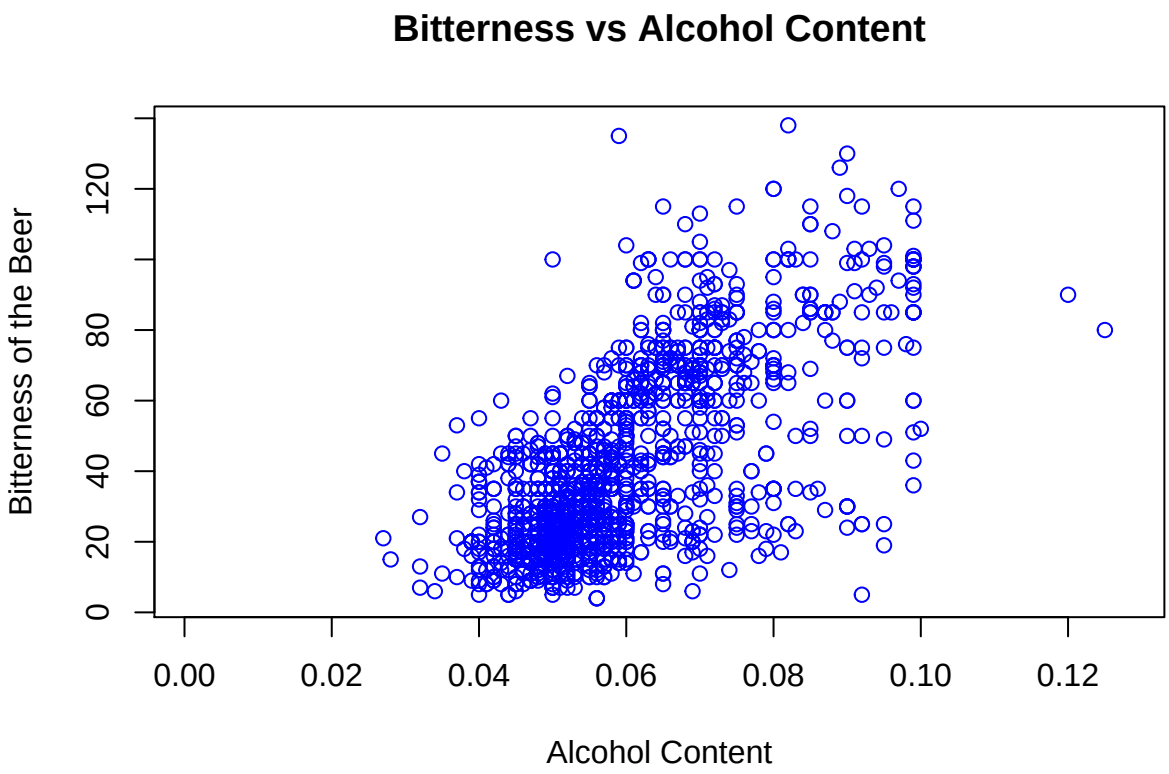
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
## 0.00100 0.05000 0.05600 0.05977 0.06700 0.12800      62
```

Relationship between Bitterness and Alcohol Content of the Beer

To better understand relationship between the bitterness of the beer and its alcohol content, we plotted the ABV data against the IBU to obtain a scatter plot below.

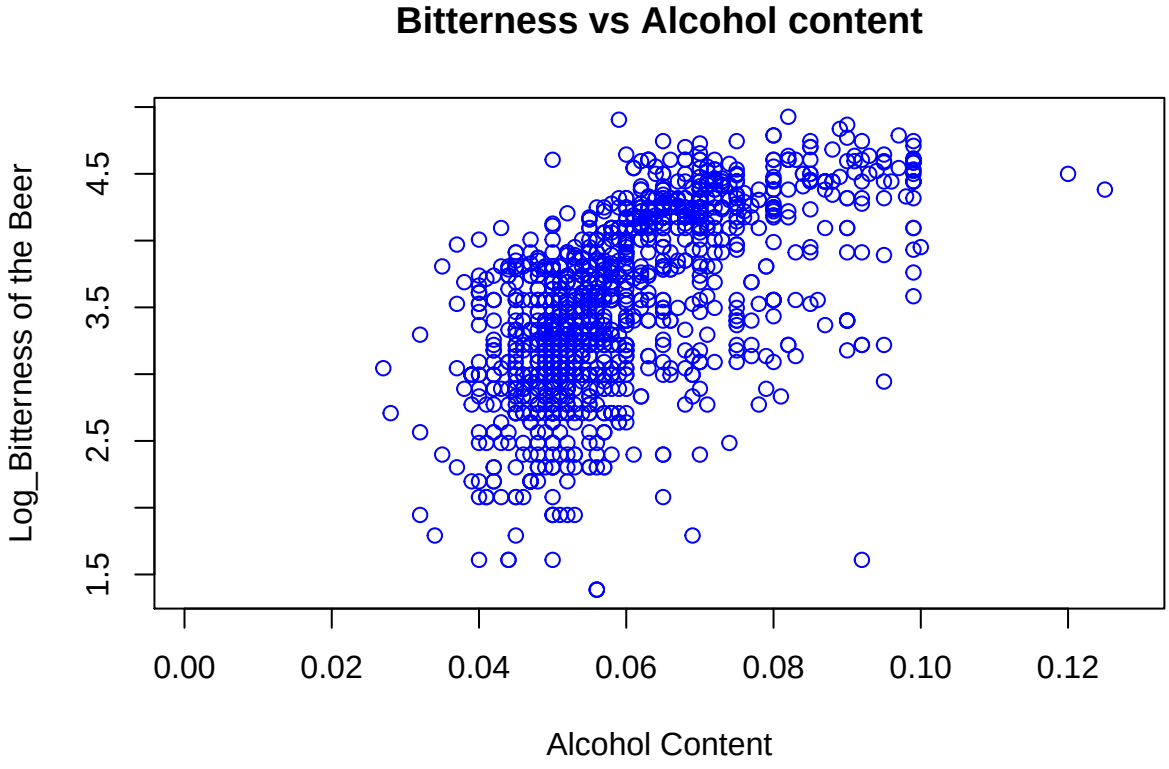
```
plot(beers_breweries$ABV,beers_breweries$IBU, xlab = "Alcohol Content", ylab = "Bitterness of the Beer"
```

[5]https://www.ratebeer.com



Based on the scatter plot, it seems that there is a relationship between the bitternees and the alcohol by volume of the beer. When ABV is plotted against log-transformed IBU we can see that this relationship might be modeled as an n^{th} degree polynomial.

```
plot(beers_breweries$ABV, log(beers_breweries$IBU), xlab = "Alcohol Content", ylab = "Log_Bitterness of
```



Conclusion

In this project we have analysed ABV and IBU content of beers from 51 states of the United States. We found that:

- Data is fairly clean but is not perfect. It has some missing values which should be investigated further.
- West Coast has a larger number of breweries per state compared to the rest of the country.
- The State of Utah has the lowest median alcohol by volume at 0.40 while Washington DC and Kentucky have the highest median alcohol by volumne of the beer at 0.625.
- Wisconsin has the lowest median IBU measurement at 19.0, while Maine has the highest at 61.0.
- Colorado has a beer with the maximum alcohol by volume at 0.128.
- Oregon has the most bitter beer at 138 IBU.
- There is a relationship between IBU and ABV variables which can be modelled as an n^{th} degree polynomial.

Code

The code with the analysis and this report is publicly available and can be found in GitHub