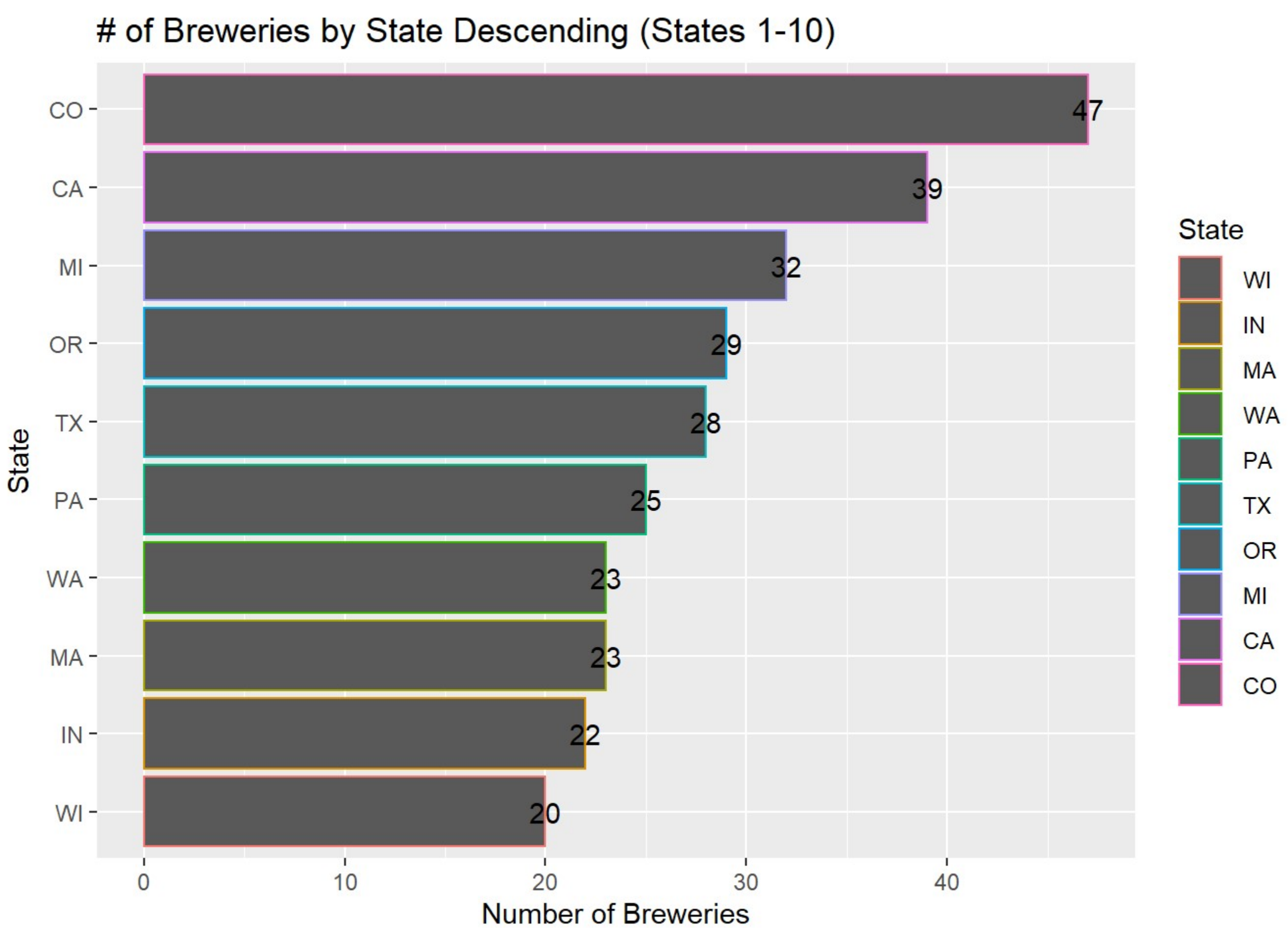
library("knitr")

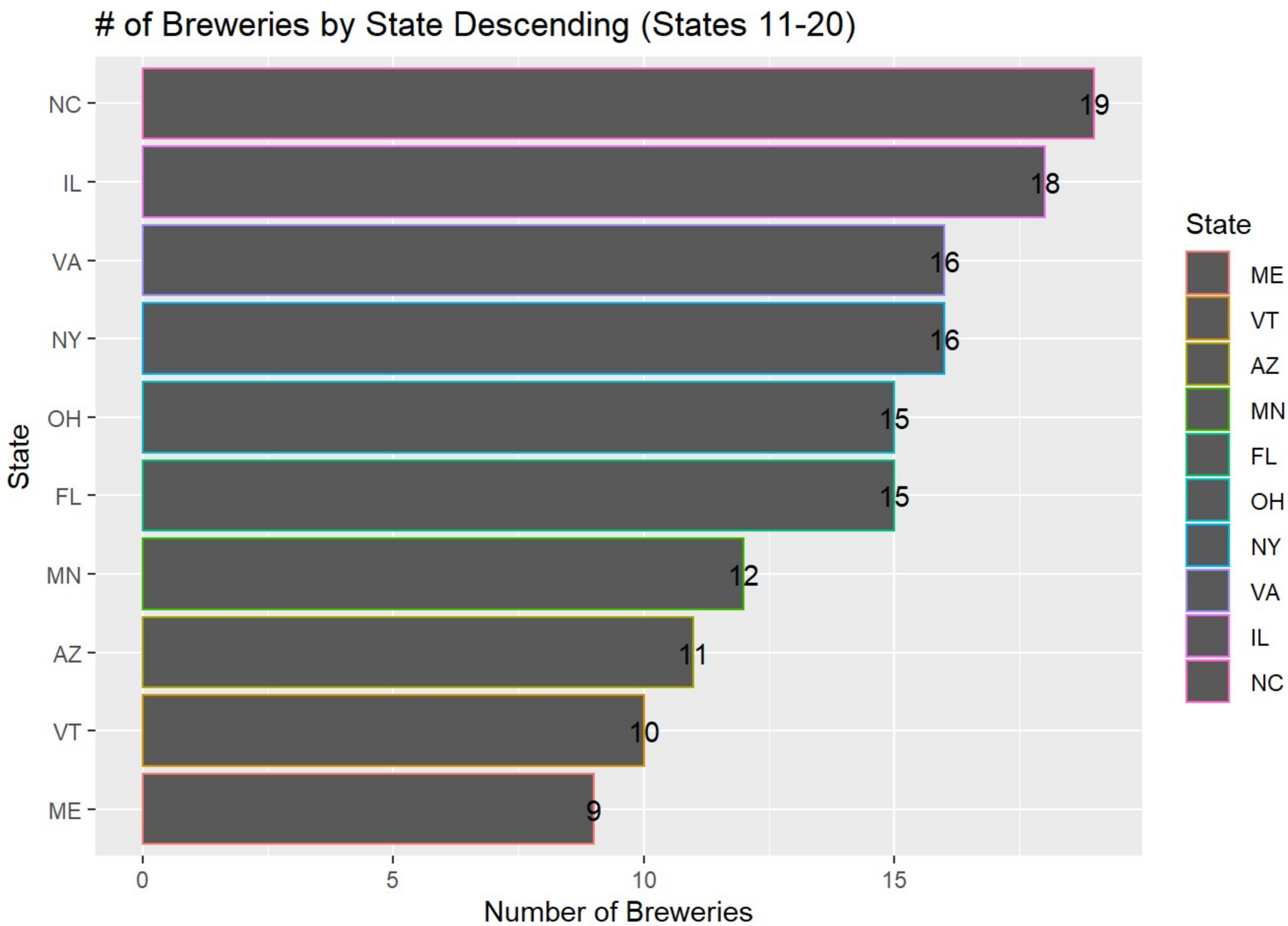
title: "DDS_Case_Study" author: "D. Dey & C. Dawson" date: "6/27/2020" output: html_document —

1. How many breweries are present in each state?

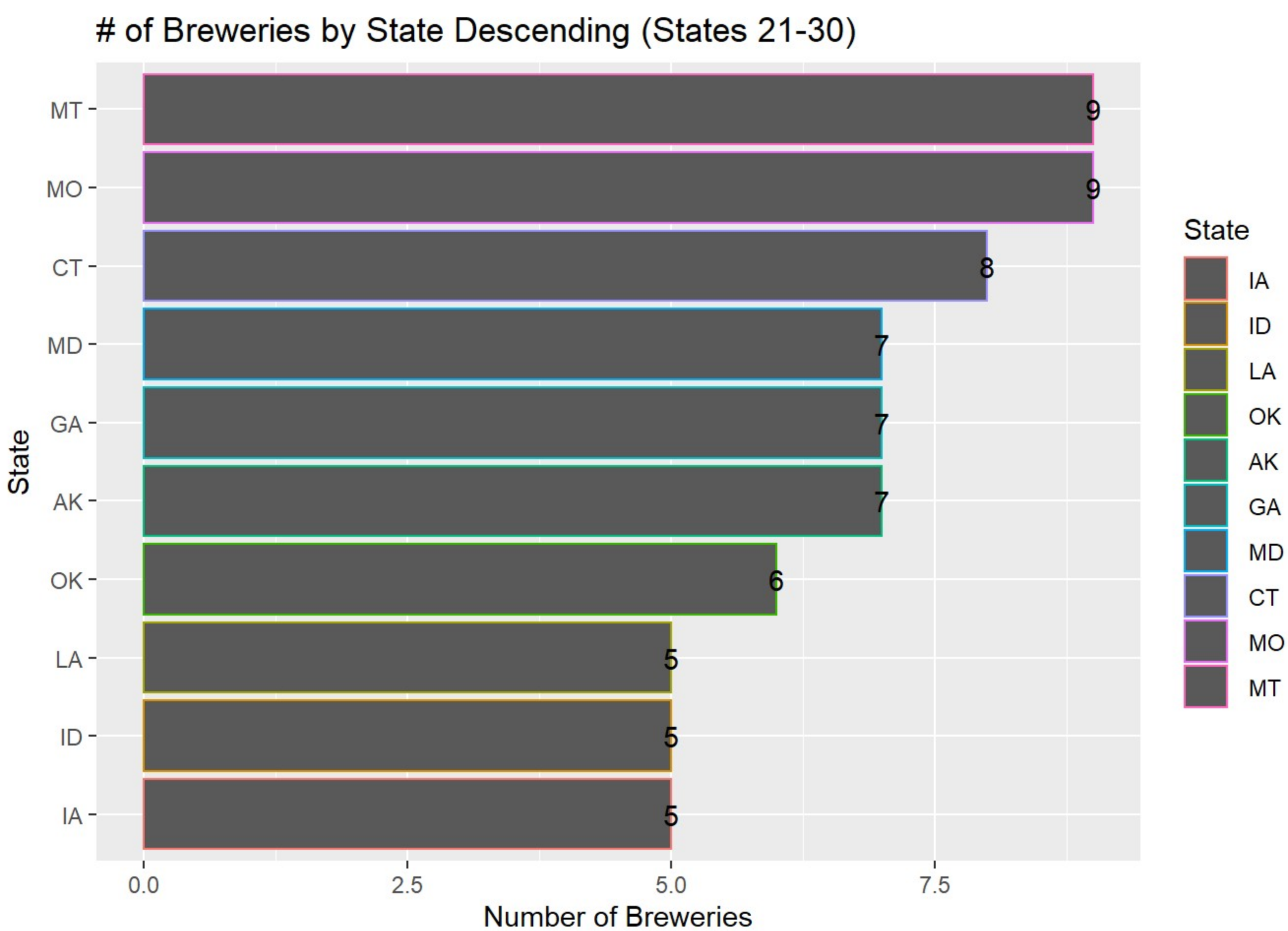
```
## # A tibble: 51 x 2
##    State     n
##    <chr> <int>
##  1 " CO"    47
##  2 " CA"    39
##  3 " MI"    32
##  4 " OR"    29
##  5 " TX"    28
##  6 " PA"    25
##  7 " MA"    23
##  8 " WA"    23
##  9 " IN"    22
## 10 " WI"    20
## # ... with 41 more rows
```

```
## There are  558  Breweries in Total within the Dataset.
```

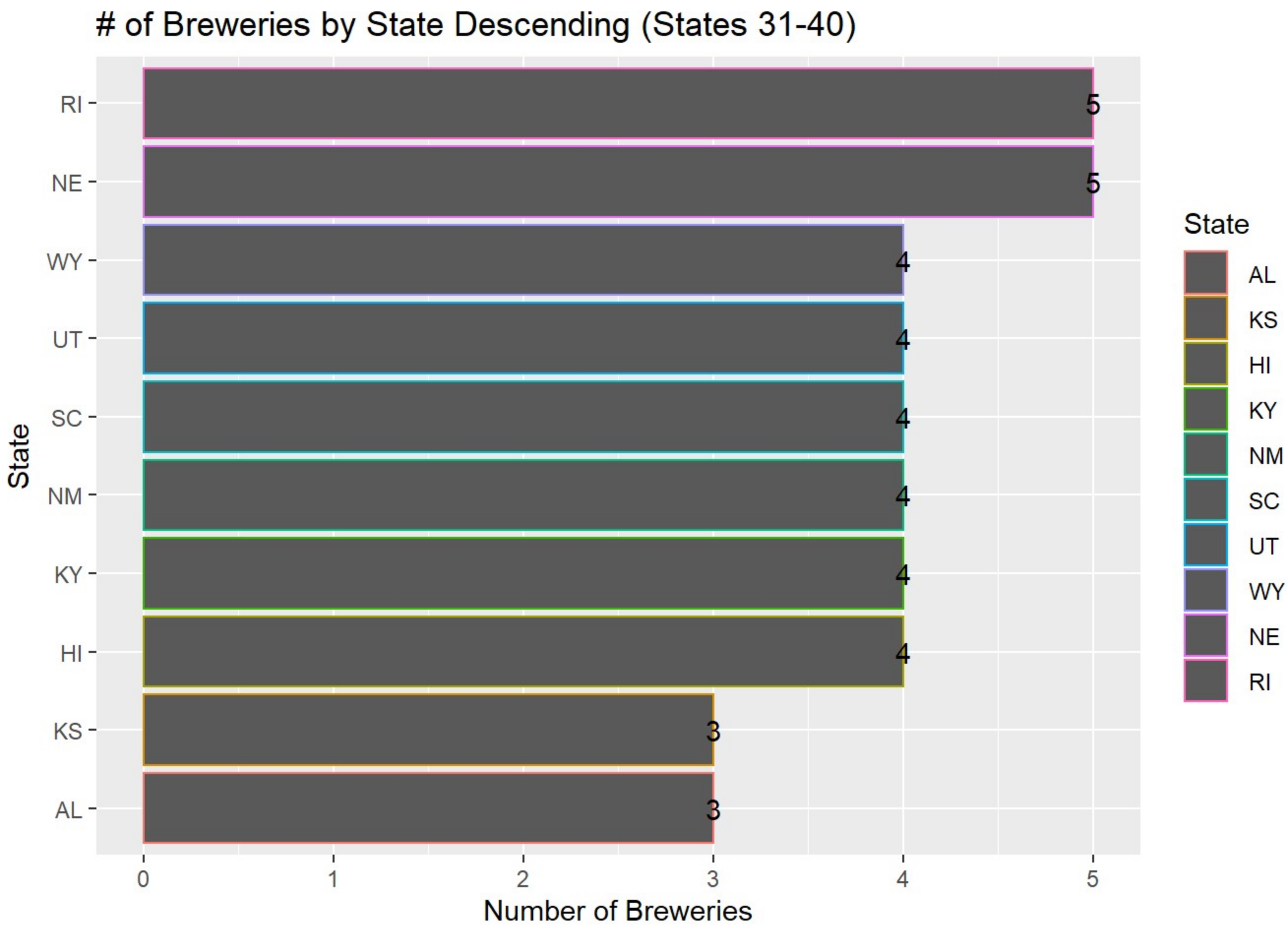




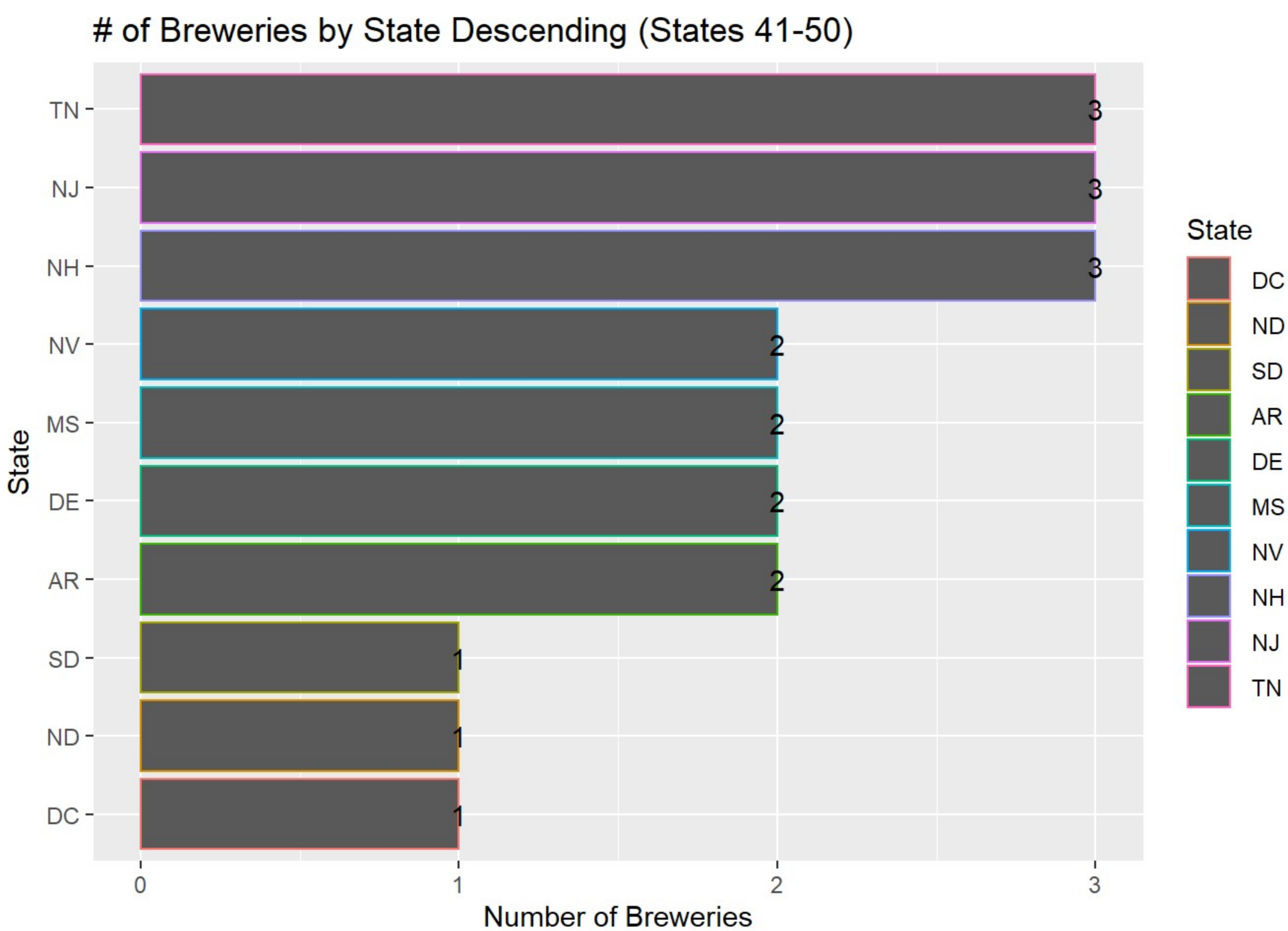
# of Breweries by State Descending (States 11-20)
NC
IL
VA
NY
OH
FL
MN
AZ
VT
ME
19
18
16
16
15
15
12
11
10
9
State
0
5
10
15
Number of Breweries
State
ME
VT
AZ
MN
FL
OH
NY
VA
IL
NC



# of Breweries by State Descending (States 21-30)
MT
MO
CT
MD
GA
AK
OK
LA
ID
IA
9
9
8
7
7
7
6
5
5
5
State
0.0
2.5
5.0
7.5
Number of Breweries
State
IA
ID
LA
OK
AK
GA
MD
CT
MO
MT



# of Breweries by State Descending (States 31-40)
RI
NE
WY
UT
SC
NM
KY
HI
KS
AL
State
5
5
4
4
4
4
4
4
3
3
0
1
2
3
4
5
Number of Breweries
State
AL
KS
HI
KY
NM
SC
UT
WY
NE
RI



# of Breweries by State Descending (States 41-50)
TN
NJ
NH
NV
MS
DE
AR
SD
ND
DC
State
3
3
3
2
2
2
2
1
1
1
0
1
2
3
Number of Breweries
State
DC
ND
SD
AR
DE
MS
NV
NH
NJ
TN

2.

Merge beer data with the breweries data. Print the first 6 observations and the last six observations to check the merged file. (RMD only, this does not need to be included in the presentation or the deck.)

```
##   Brewery_id        Name.x 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             Name.y        City
## 1                      American IPA     16 NorthGate Brewing  Minneapolis
## 2                Milk / Sweet Stout     16 NorthGate Brewing  Minneapolis
## 3                 English Brown Ale     16 NorthGate Brewing  Minneapolis
## 4                       Pumpkin Ale     16 NorthGate Brewing  Minneapolis
## 5                   American Porter     16 NorthGate Brewing  Minneapolis
## 6 Extra Special / Strong Bitter (ESB)     16 NorthGate Brewing  Minneapolis
##   State
## 1    MN
## 2    MN
## 3    MN
## 4    MN
## 5    MN
## 6    MN
```

```
##      Brewery_id                    Name.x 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                        Name.y          City
## 2405         German Pilsener     12         Ukiah Brewing Company         Ukiah
## 2406              Hefeweizen     12       Butternuts Beer and Ale Garrattsville
## 2407            American IPA     12       Butternuts Beer and Ale Garrattsville
## 2408      Milk / Sweet Stout     12       Butternuts Beer and Ale Garrattsville
## 2409 American Pale Ale (APA)     12       Butternuts Beer and Ale Garrattsville
## 2410        English Pale Ale     12 Sleeping Lady Brewing Company     Anchorage
##      State
## 2405    CA
## 2406    NY
## 2407    NY
## 2408    NY
## 2409    NY
## 2410    AK
```

3. Address the missing values in each columns.

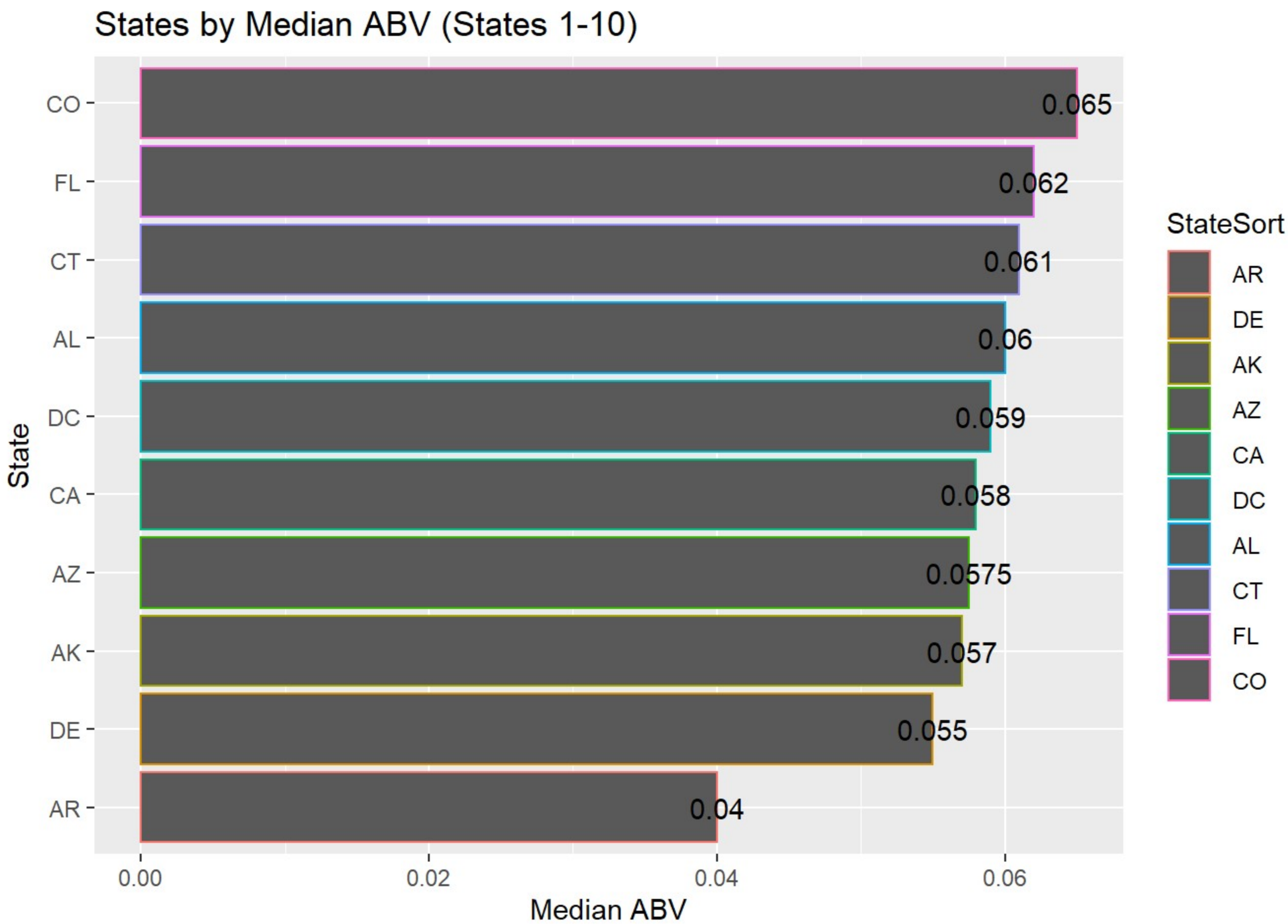
```
## There are 2410 rows before removing all rows with 'NA' from the Beer-Brewery data and 1405
thereafter.
```

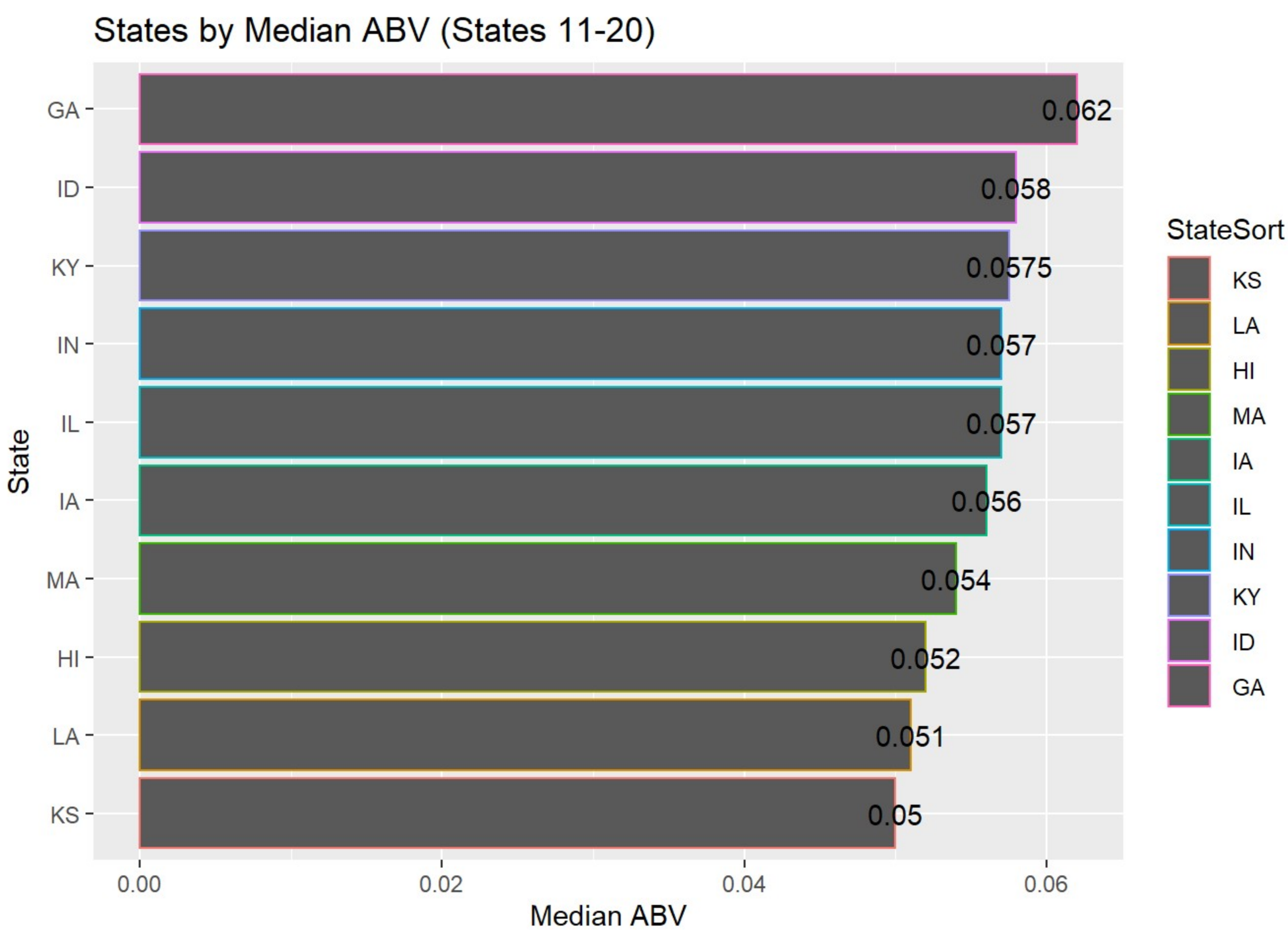
4. Compute the median alcohol content and international bitterness unit for each state. Plot a bar chart to compare.
5. Which state has the maximum alcoholic (ABV) beer? Which state has the most bitter (IBU) beer?

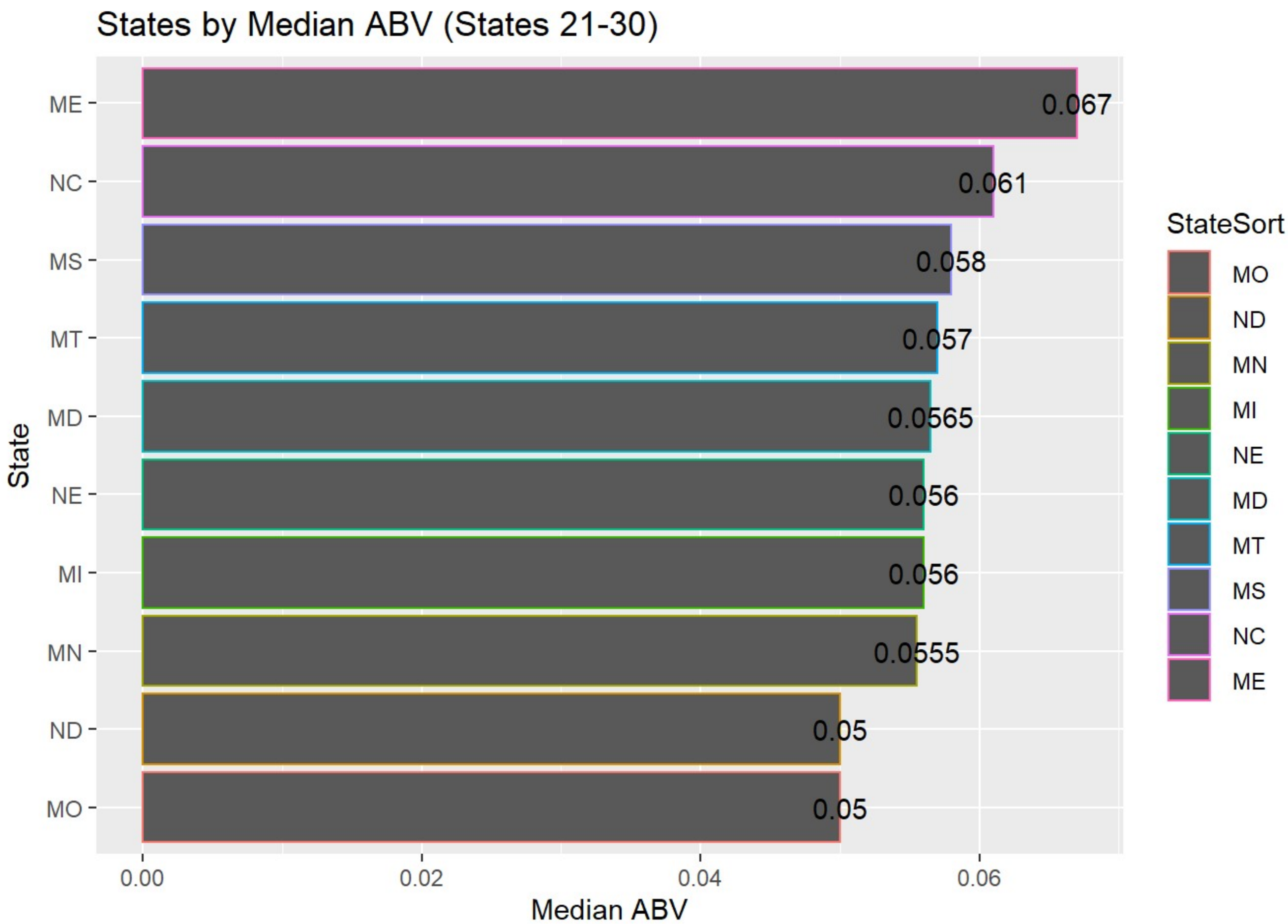
```
## # A tibble: 1,405 x 3
## # Groups:   State [50]
##    State   ABV   IBU
##    <chr> <dbl> <int>
##  1 " MN" 0.045    50
##  2 " MN" 0.049    26
##  3 " MN" 0.048    19
##  4 " MN" 0.06     38
##  5 " MN" 0.06     25
##  6 " MN" 0.056    47
##  7 " KY" 0.08     68
##  8 " KY" 0.125    80
##  9 " KY" 0.077    25
## 10 " KY" 0.042    42
## # ... with 1,395 more rows
```

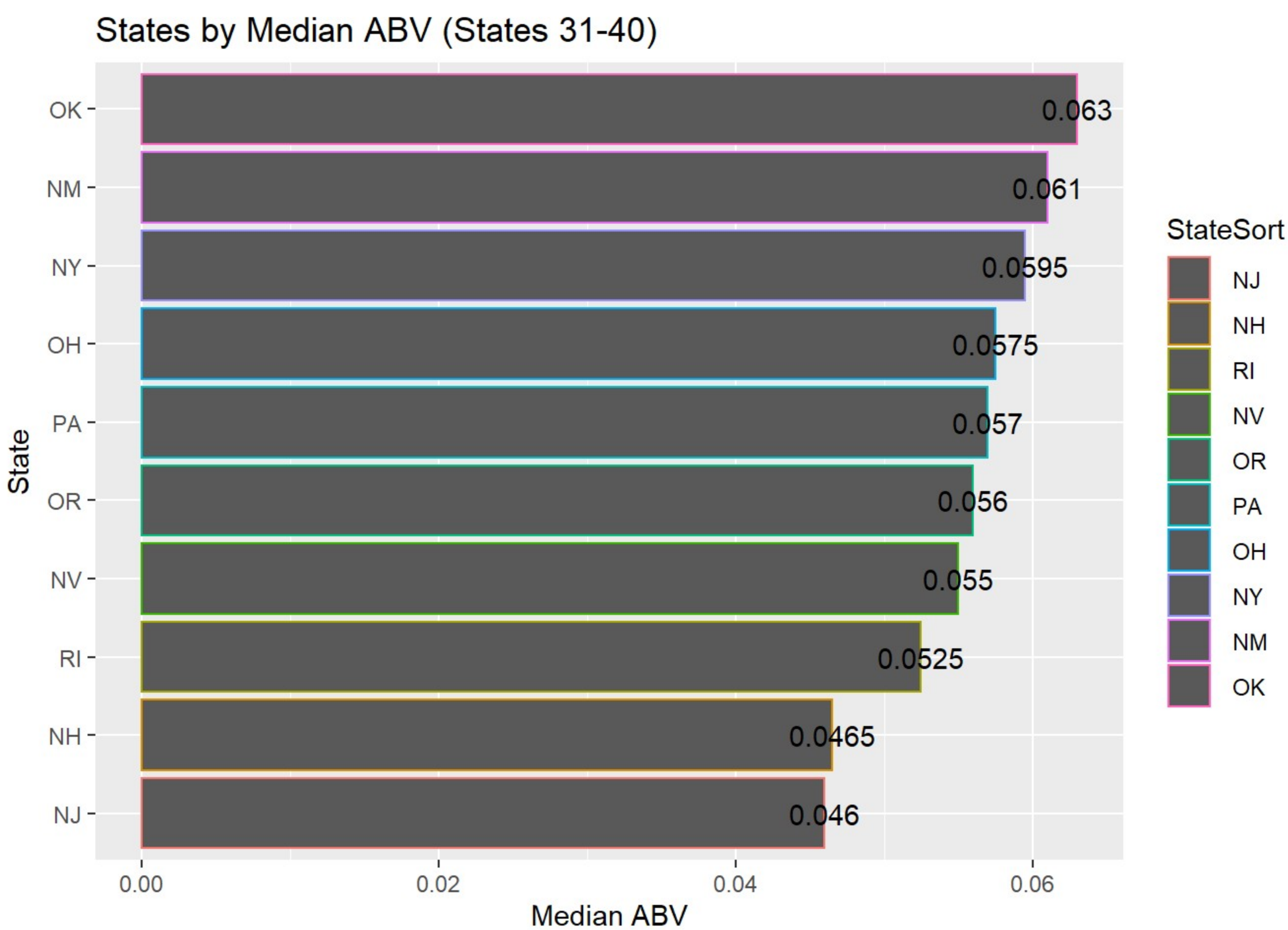
```
## `summarise()` ungrouping output (override with `.groups` argument)
```

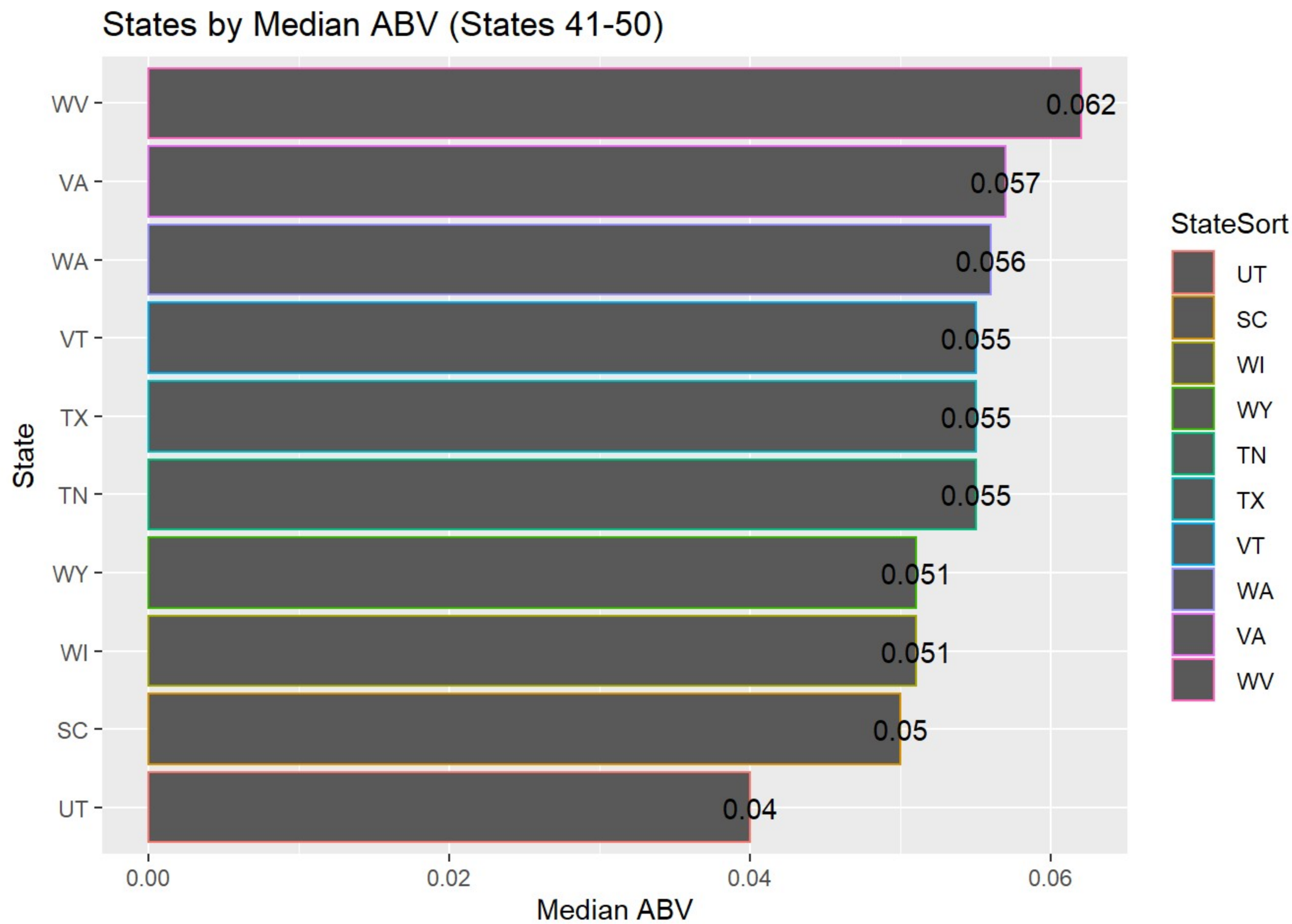
```
## # A tibble: 50 x 3
##    State ABV_Median   IBU
##    <chr>      <dbl> <dbl>
##  1 " AK"     0.057   40.9
##  2 " AL"     0.06    51.2
##  3 " AR"     0.04    39
##  4 " AZ"     0.0575  35.2
##  5 " CA"     0.058   46.3
##  6 " CO"     0.065   47.4
##  7 " CT"     0.061   40.8
##  8 " DC"     0.059   55.2
##  9 " DE"     0.055   52
## 10 " FL"     0.062   46.8
## # ... with 40 more rows
```

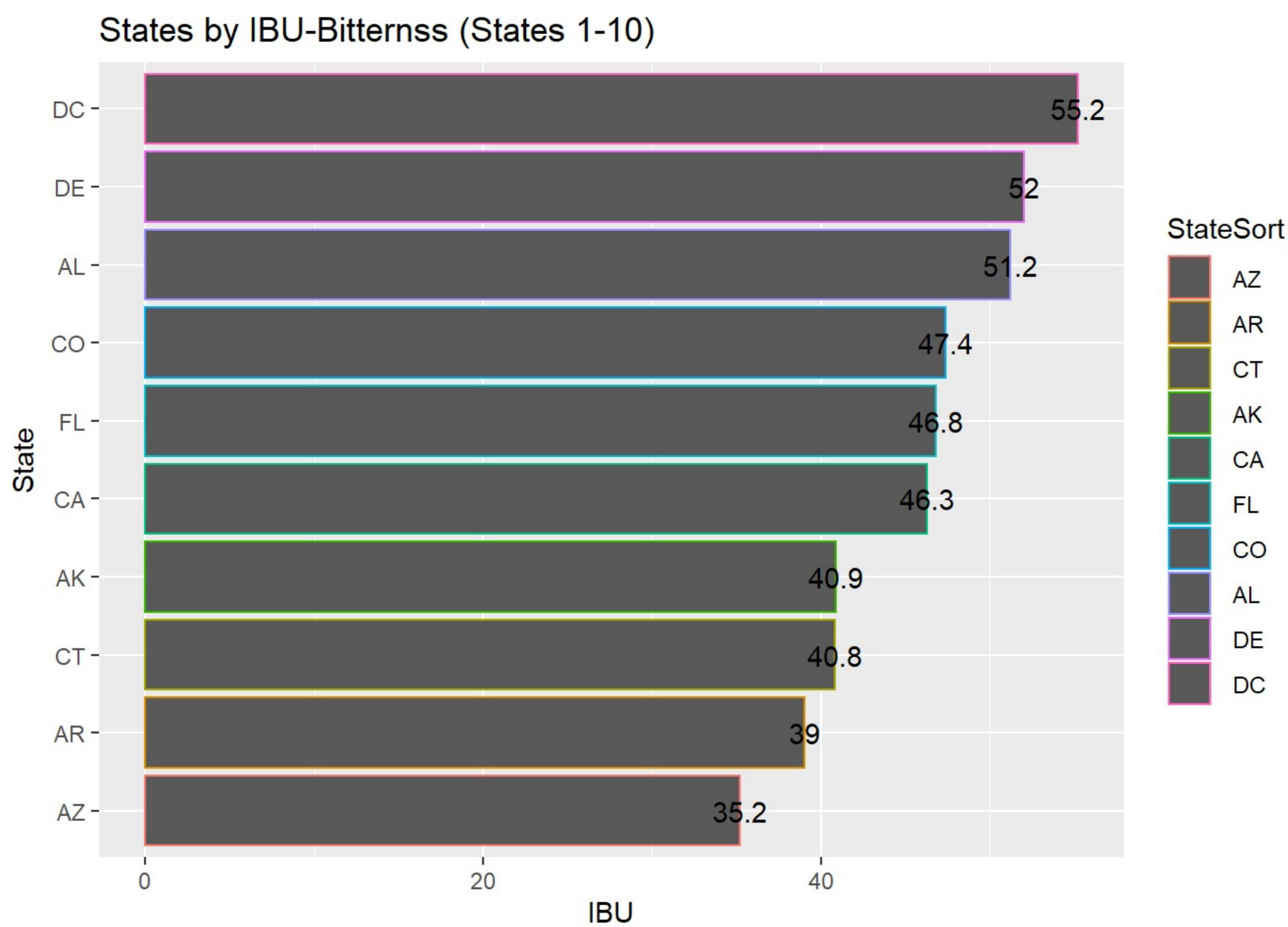

States by Median ABV (States 1-10)
CO
FL
CT
AL
DC
CA
AZ
AK
DE
AR
0.065
0.062
0.061
0.06
0.059
0.058
0.0575
0.057
0.055
0.04
State
0.00
0.02
0.04
0.06
Median ABV
StateSort
AR
DE
AK
AZ
CA
DC
AL
CT
FL
CO

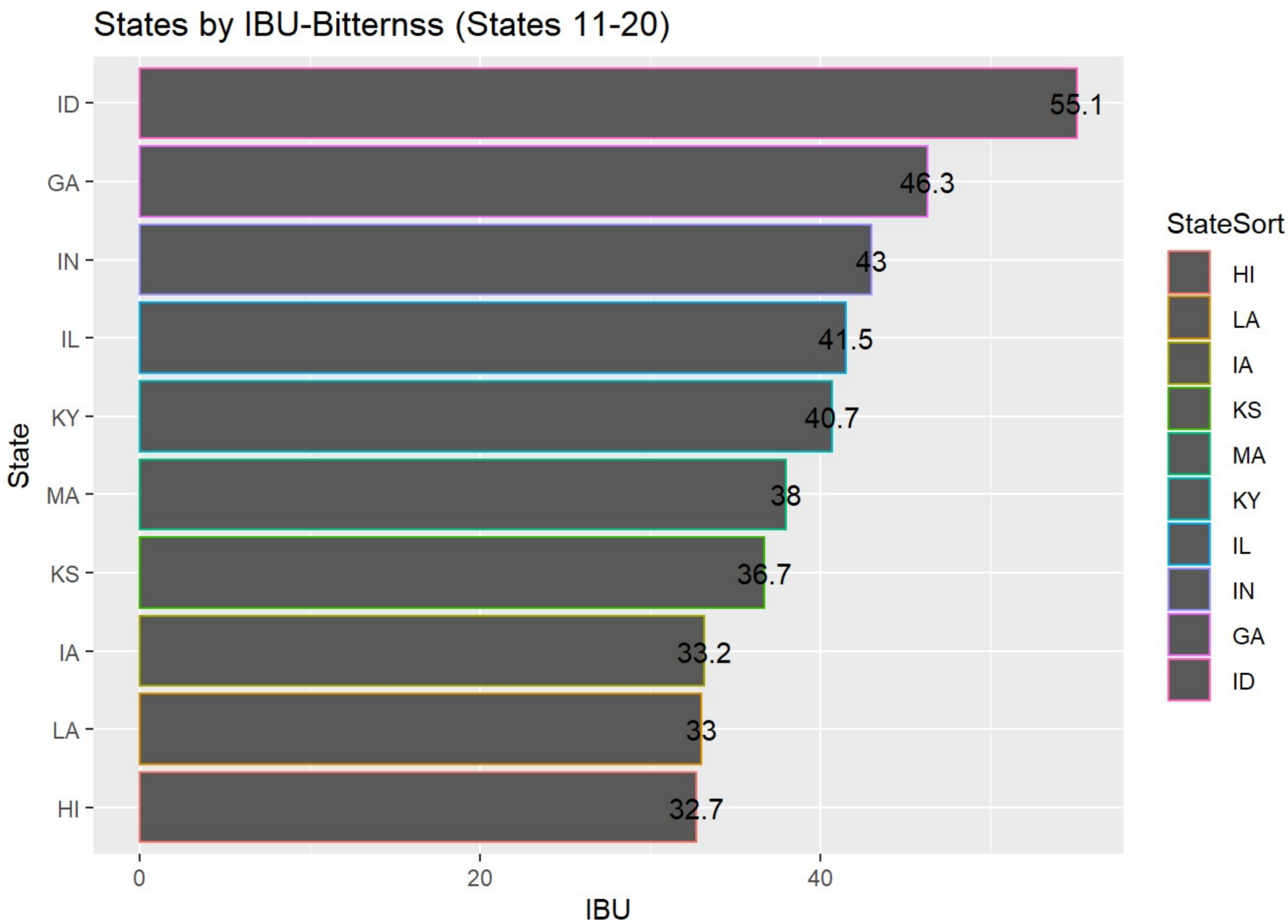

States by Median ABV (States 11-20)
GA
ID
KY
IN
IL
IA
MA
HI
LA
KS
0.062
0.058
0.0575
0.057
0.057
0.056
0.054
0.052
0.051
0.05
State
0.00
0.02
0.04
0.06
Median ABV
StateSort
KS
LA
HI
MA
IA
IL
IN
KY
ID
GA

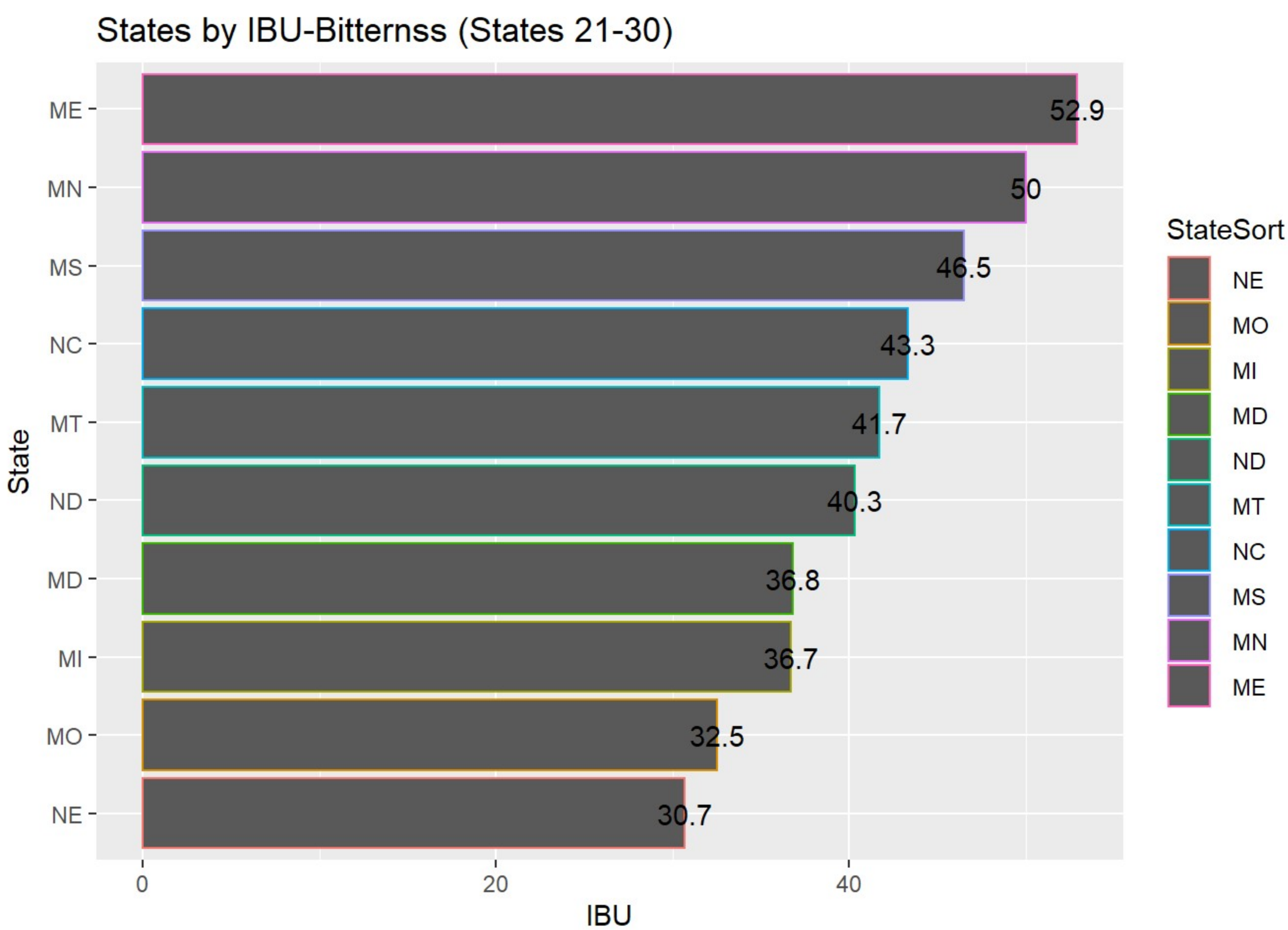

States by Median ABV (States 21-30)
ME
NC
MS
MT
MD
NE
MI
MN
ND
MO
0.067
0.061
0.058
0.057
0.0565
0.056
0.056
0.0555
0.05
0.05
State
0.00
0.02
0.04
0.06
Median ABV
StateSort
MO
ND
MN
MI
NE
MD
MT
MS
NC
ME

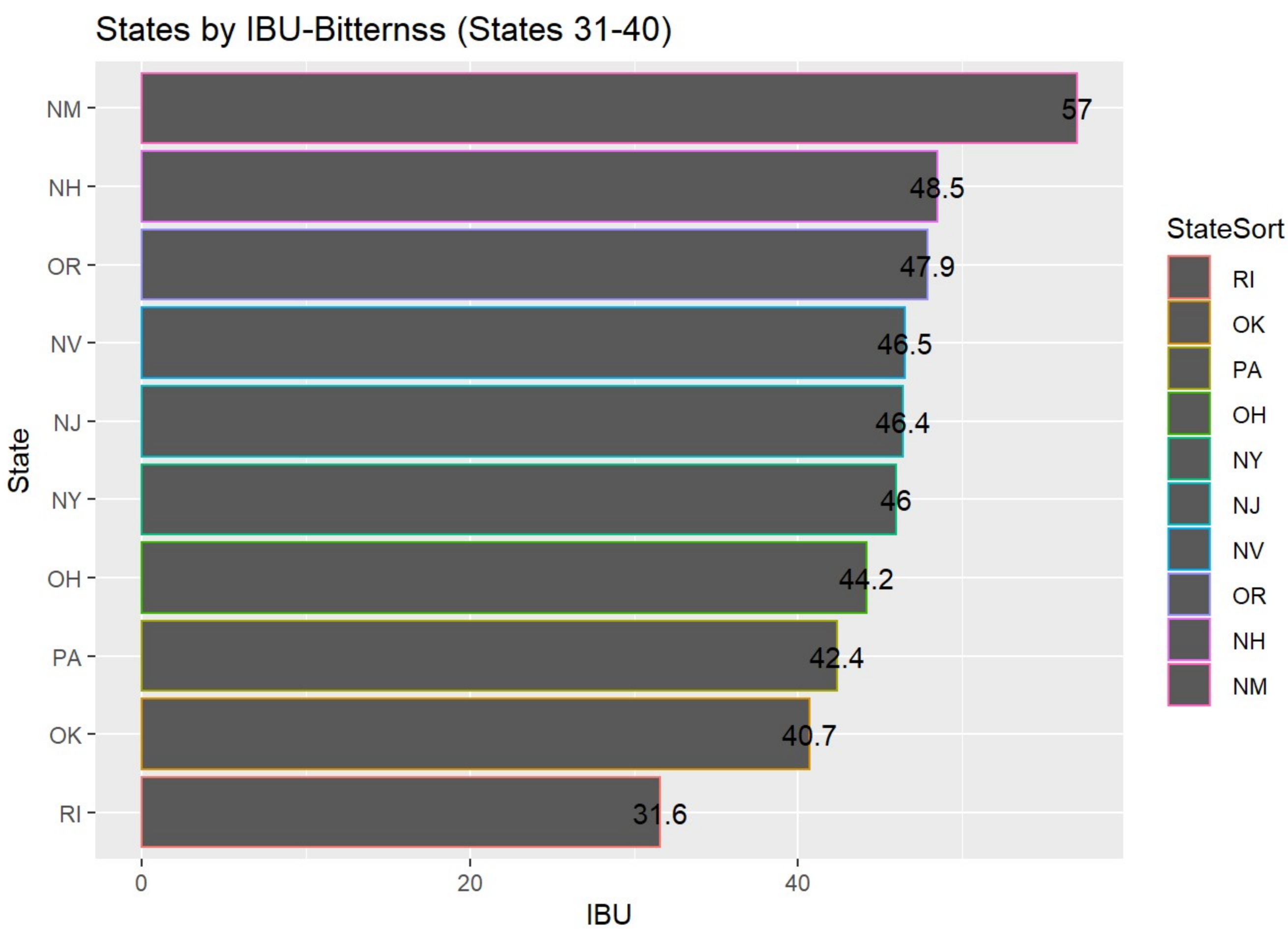

States by Median ABV (States 31-40)
OK
NM
NY
OH
PA
OR
NV
RI
NH
NJ
0.063
0.061
0.0595
0.0575
0.057
0.056
0.055
0.0525
0.0465
0.046
State
0.00
0.02
0.04
0.06
Median ABV
StateSort
NJ
NH
RI
NV
OR
PA
OH
NY
NM
OK

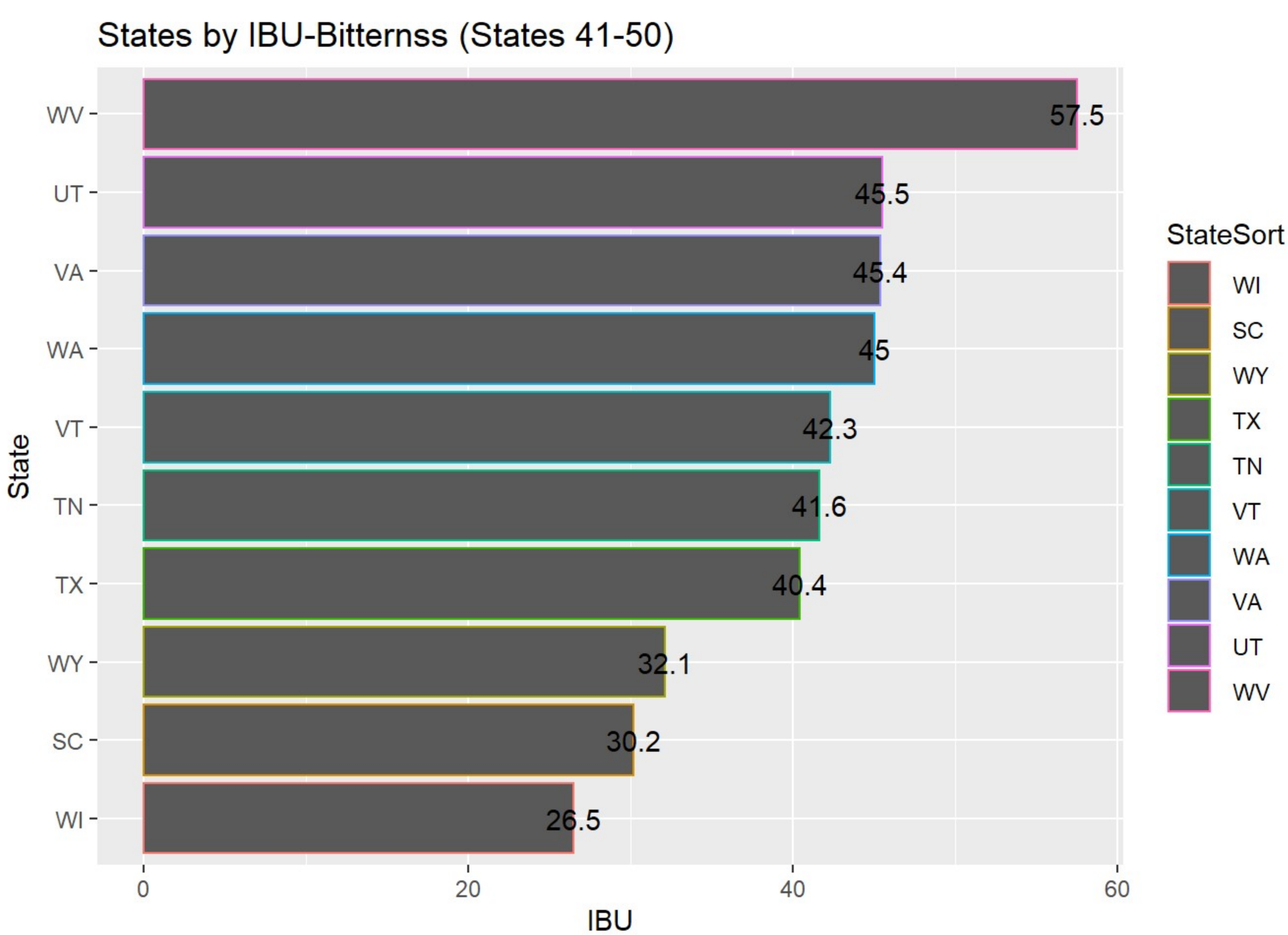

States by Median ABV (States 41-50)
State
WV
VA
WA
VT
TX
TN
WY
WI
SC
UT
0.062
0.057
0.056
0.055
0.055
0.055
0.051
0.051
0.05
0.04
0.00
0.02
0.04
0.06
Median ABV
StateSort
UT
SC
WI
WY
TN
TX
VT
WA
VA
WV


States by IBU-Bitternss (States 1-10)
State
DC
DE
AL
CO
FL
CA
AK
CT
AR
AZ
55.2
52
51.2
47.4
46.8
46.3
40.9
40.8
39
35.2
0
20
40
IBU
StateSort
AZ
AR
CT
AK
CA
FL
CO
AL
DE
DC


States by IBU-Bitternss (States 11-20)
State
IBU
ID
GA
IN
IL
KY
MA
KS
IA
LA
HI
55.1
46.3
43
41.5
40.7
38
36.7
33.2
33
32.7
0
20
40
StateSort
HI
LA
IA
KS
MA
KY
IL
IN
GA
ID


States by IBU-Bitternss (States 21-30)
State
IBU
ME
MN
MS
NC
MT
ND
MD
MI
MO
NE
52.9
50
46.5
43.3
41.7
40.3
36.8
36.7
32.5
30.7
0
20
40
StateSort
NE
MO
MI
MD
ND
MT
NC
MS
MN
ME


States by IBU-Bitternss (States 31-40)
State
NM
NH
OR
NV
NJ
NY
OH
PA
OK
RI
57
48.5
47.9
46.5
46.4
46
44.2
42.4
40.7
31.6
0
20
40
IBU
StateSort
RI
OK
PA
OH
NY
NJ
NV
OR
NH
NM

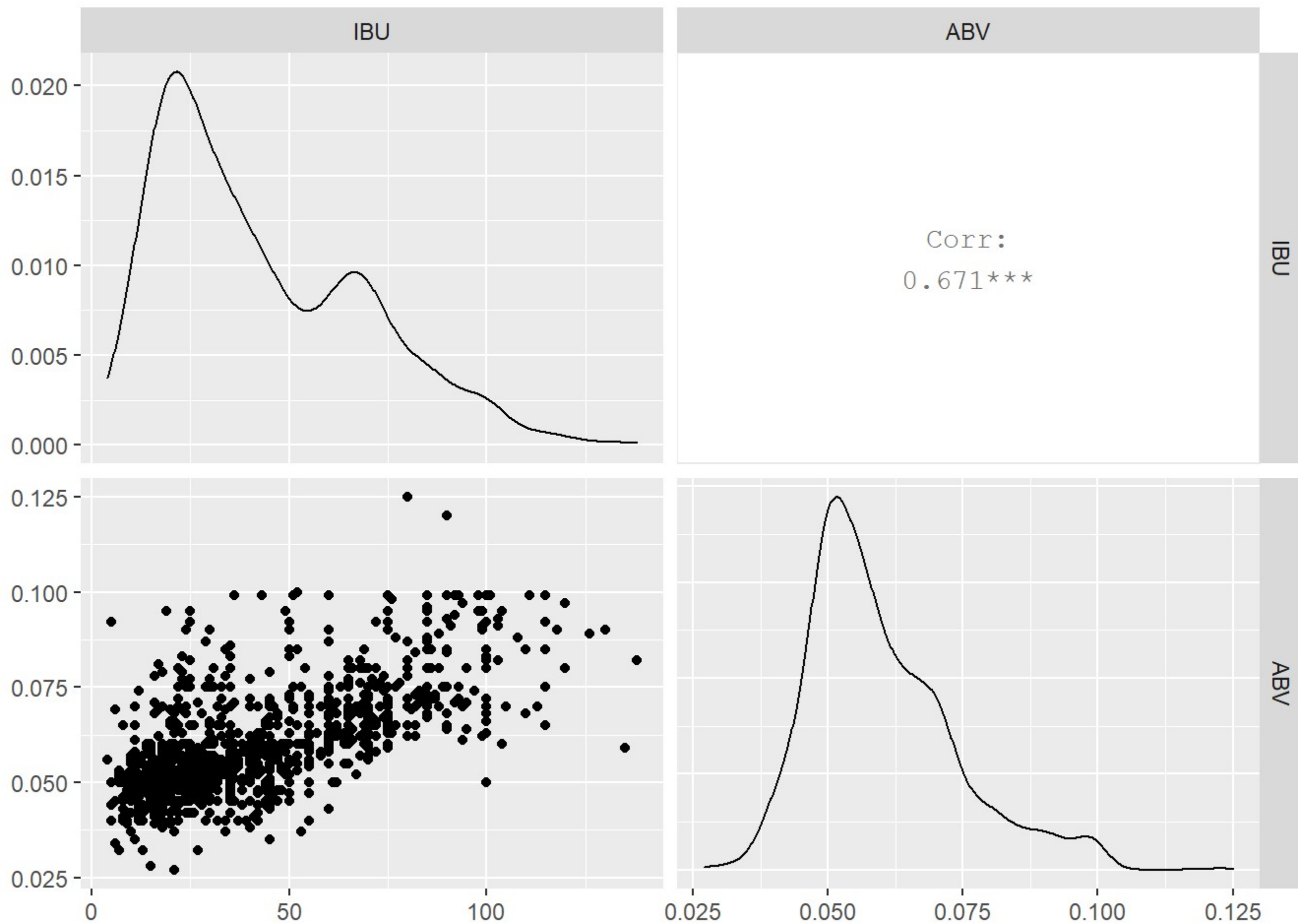

States by IBU-Bitternss (States 41-50)
State
WV
UT
VA
WA
VT
TN
TX
WY
SC
WI
57.5
45.5
45.4
45
42.3
41.6
40.4
32.1
30.2
26.5
0
20
40
60
IBU
StateSort
WI
SC
WY
TX
TN
VT
WA
VA
UT
WV

```
## The state with highest alcohol by volume (ABV) beer is  CO with a number of 146 , as per t
he given dataset
```

```
## The state with the most bitter (IBU) beer is  OR with a number of 138 , per the given data
set
```

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

```
###Summary Statistics & Distribution
Summary_Base %>% select(IBU, ABV) %>% ggpairs() + labs(main = "ABV by IBU Distribution")
```

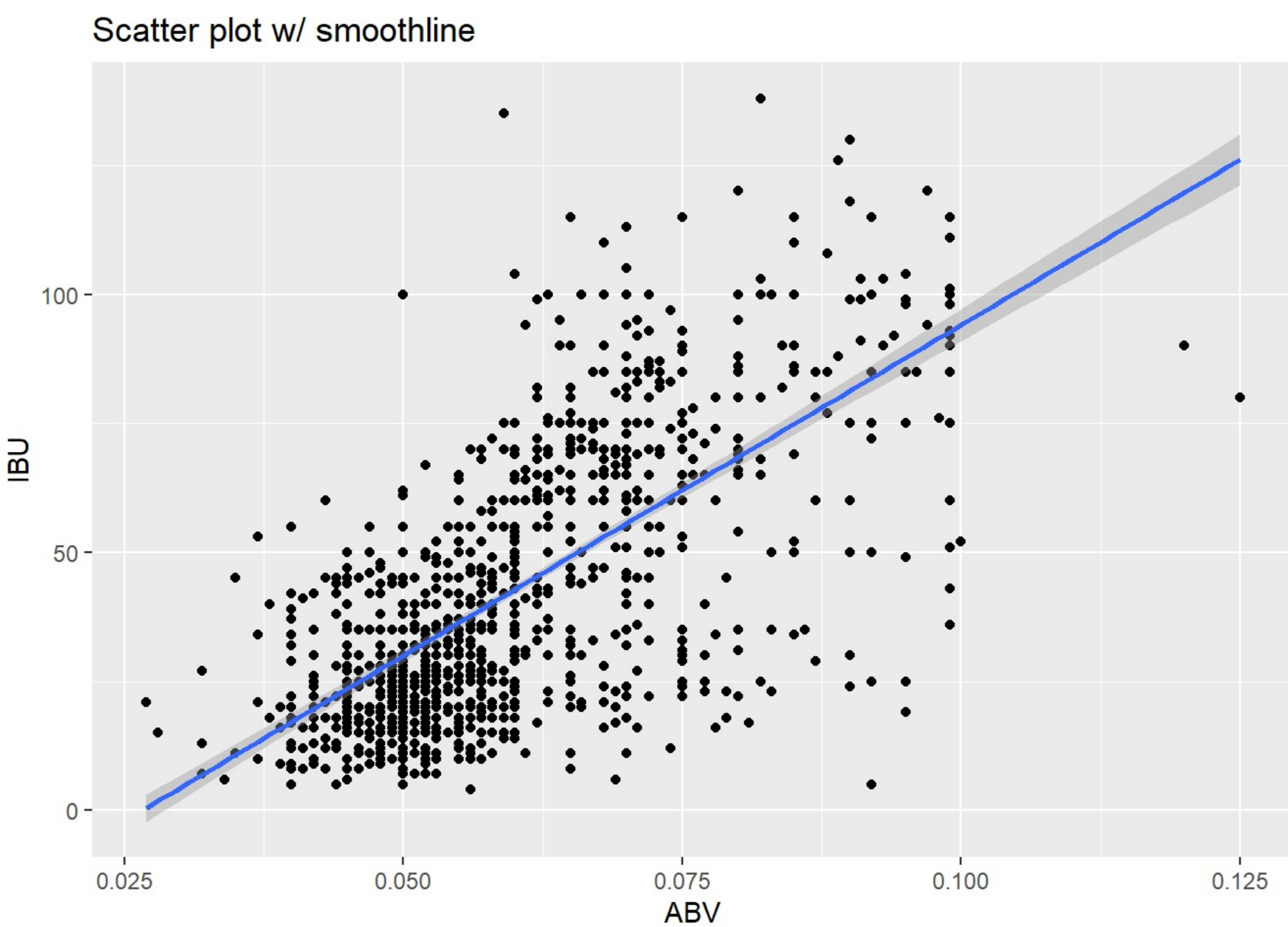



```
###Correlation test
cor.test(Summary_Base$ABV, Summary_Base$IBU) ## Pearson correlation
```

```
##
##  Pearson's product-moment correlation
##
## data:  Summary_Base$ABV and Summary_Base$IBU
## t = 33.863, df = 1403, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6407982 0.6984238
## sample estimates:
##       cor
## 0.6706215
```

```
ggplot(data= Summary_Base, aes(x=ABV, y = IBU)) +
  geom_point() +
  stat_smooth(method = lm) +
  ggtitle("Scatter plot w/ smoothline") +
  xlab("ABV") +
  ylab("IBU")
```

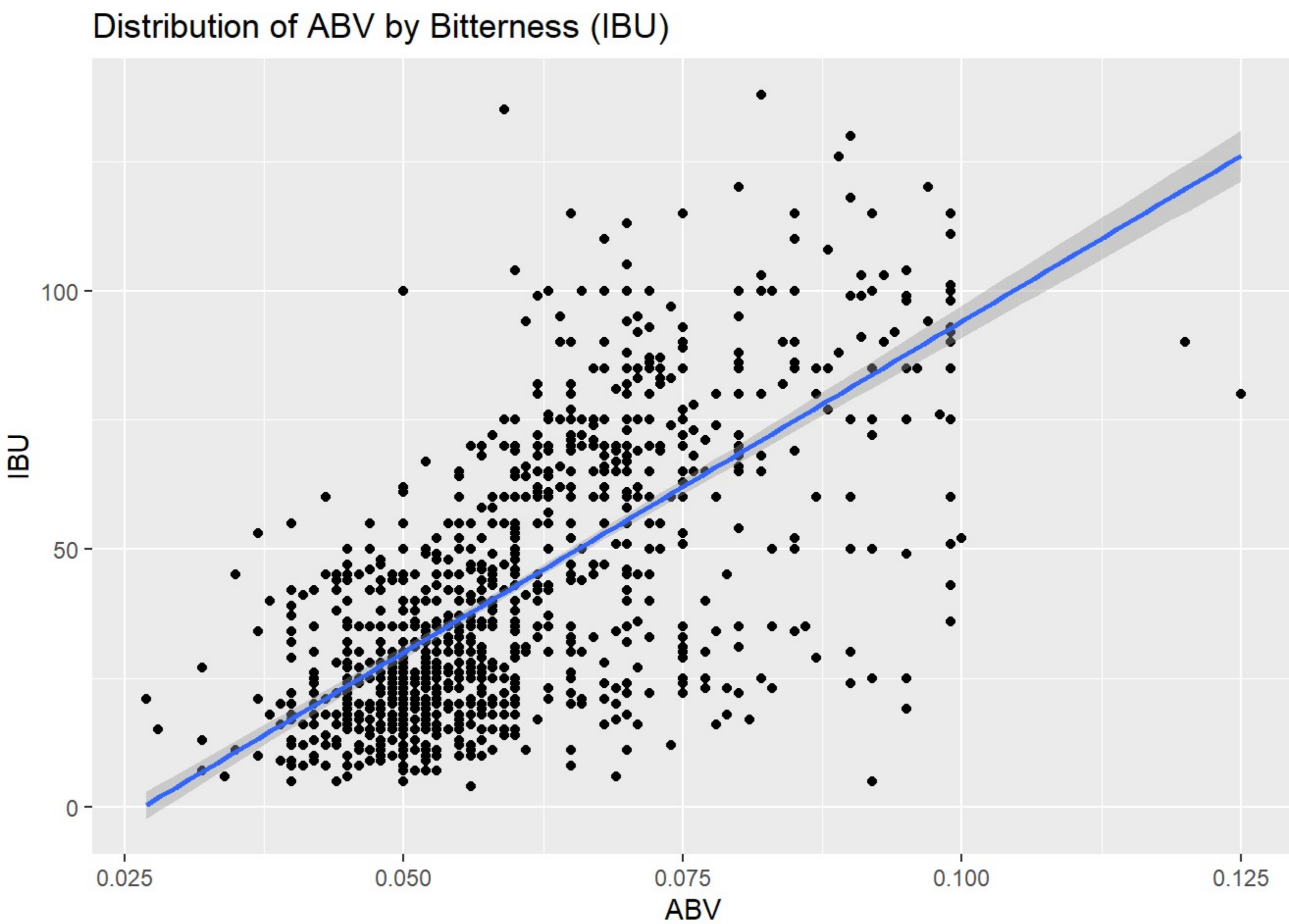
```
## `geom_smooth()` using formula 'y ~ x'
```



```
##ABV has slight left skew, while IBU is the opposite, but are close to normally distributed. Histrogram corroborates that the data normality, though it is slightly left skewed. The alcoholic content is moderately and positively correlated with the bitterness of the beer. Plot further illustrates the notion.
```

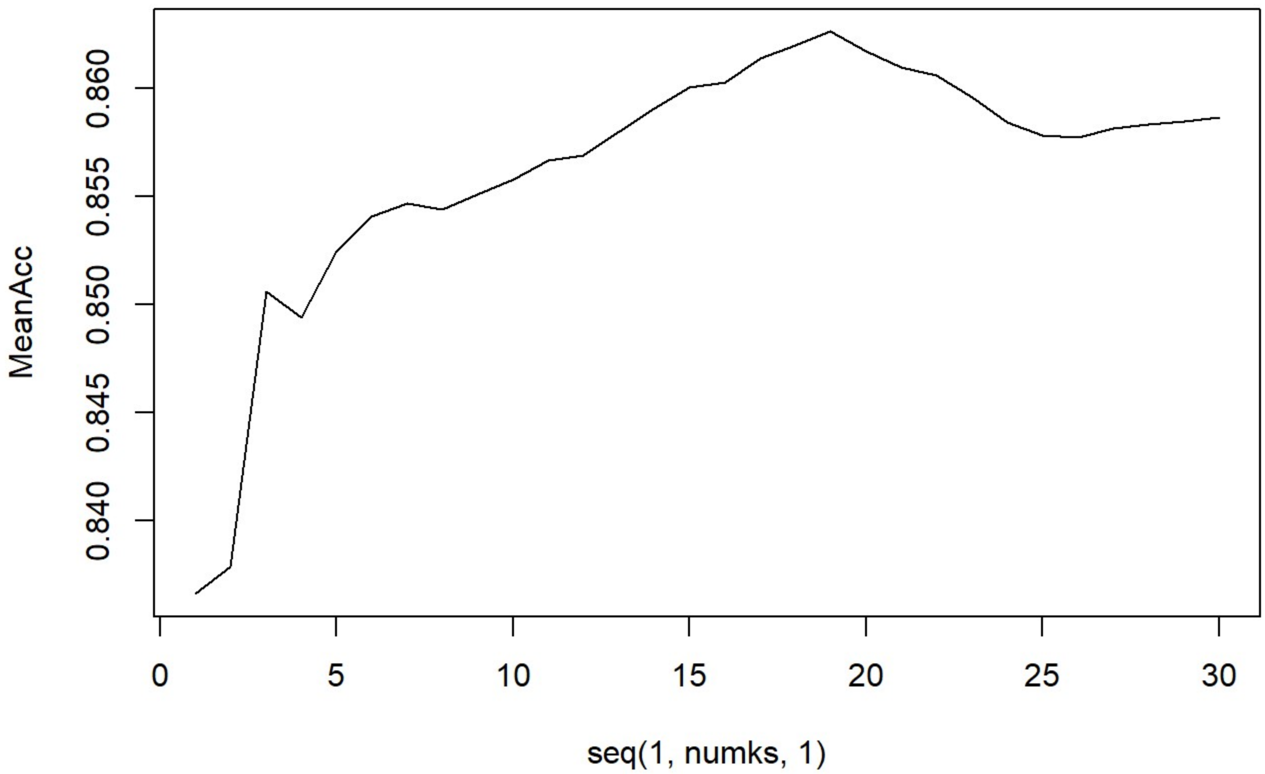
7. Is there an apparent relationship between the bitterness of the beer and its alcoholic content? Draw a scatter plot. Make your best judgment of a relationship and EXPLAIN your answer.

```
## `geom_smooth()` using formula 'y ~ x'
```



```
##
## Call:
## lm(formula = ABV ~ IBU, data = Summary_Base)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.033288 -0.005946 -0.001595  0.004022  0.052006
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.493e-02  5.177e-04   86.79   <2e-16 ***
## IBU         3.508e-04  1.036e-05   33.86   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01007 on 1403 degrees of freedom
## Multiple R-squared:  0.4497, Adjusted R-squared:  0.4493
## F-statistic:  1147 on 1 and 1403 DF,  p-value: < 2.2e-16
```

8. Budweiser would also like to investigate the difference with respect to IBU and ABV between IPAs (India Pale Ales) and other types of Ale (any beer with “Ale” in its name other than IPA). You decide to use KNN classification to investigate this relationship. Provide statistical evidence one way or the other. You can of course assume your audience is comfortable with percentages … KNN is very easy to understand conceptually.

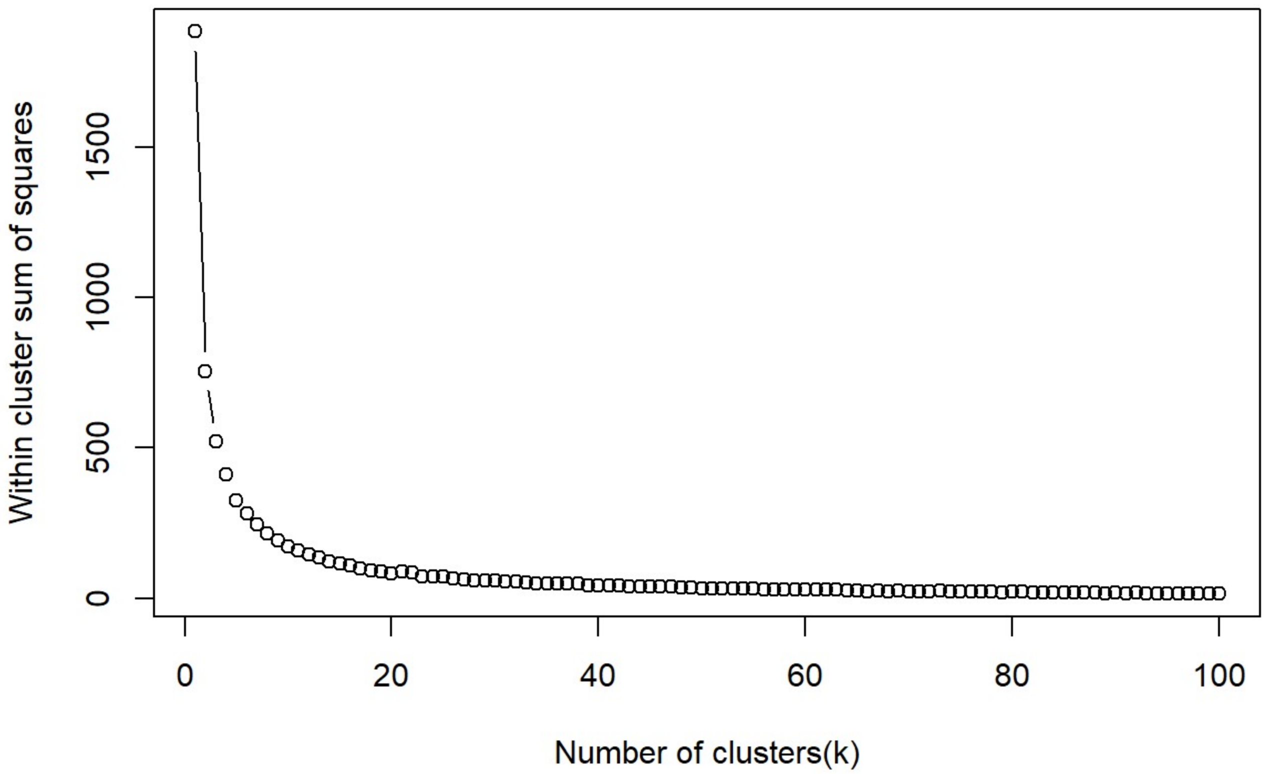



```
##
## classifications   N   Y
##               N 123  25
##               Y  16  72
```

```
## Confusion Matrix and Statistics
##
##
## classifications   N   Y
##               N 123  25
##               Y  16  72
##
##                Accuracy : 0.8263
##                  95% CI : (0.7718, 0.8724)
##     No Information Rate : 0.589
##     P-Value [Acc > NIR] : 4.489e-15
##
##                   Kappa : 0.6361
##
##  Mcnemar's Test P-Value : 0.2115
##
##             Sensitivity : 0.8849
##             Specificity : 0.7423
##          Pos Pred Value : 0.8311
##          Neg Pred Value : 0.8182
##              Prevalence : 0.5890
##          Detection Rate : 0.5212
##    Detection Prevalence : 0.6271
##       Balanced Accuracy : 0.8136
##
##        'Positive' Class : N
##
```

In addition, while you have decided to use KNN to investigate this relationship (KNN is required) you may also feel free to supplement your response to this question with any other methods or techniques you have learned. Creativity and alternative solutions are always encouraged.

```
##Knn-Means to identify the best model
set.seed(500)
k.max <- 100
wss<- sapply(1:k.max,function(k){kmeans(IPA[,1:2],k,nstart = 5,iter.max = 200)$tot.withinss})
plot(1:k.max,wss, type= "b", xlab = "Number of clusters(k)", ylab = "Within cluster sum of sq
uares")
```




```
icluster <- kmeans(IPA[,1:2],2,nstart = 20)
kmeans_matrix = table(icluster$cluster,IPA$IPA)
kmeans_matrix
```

```
##
##       N   Y
##   1  91 332
##   2 461  60
```

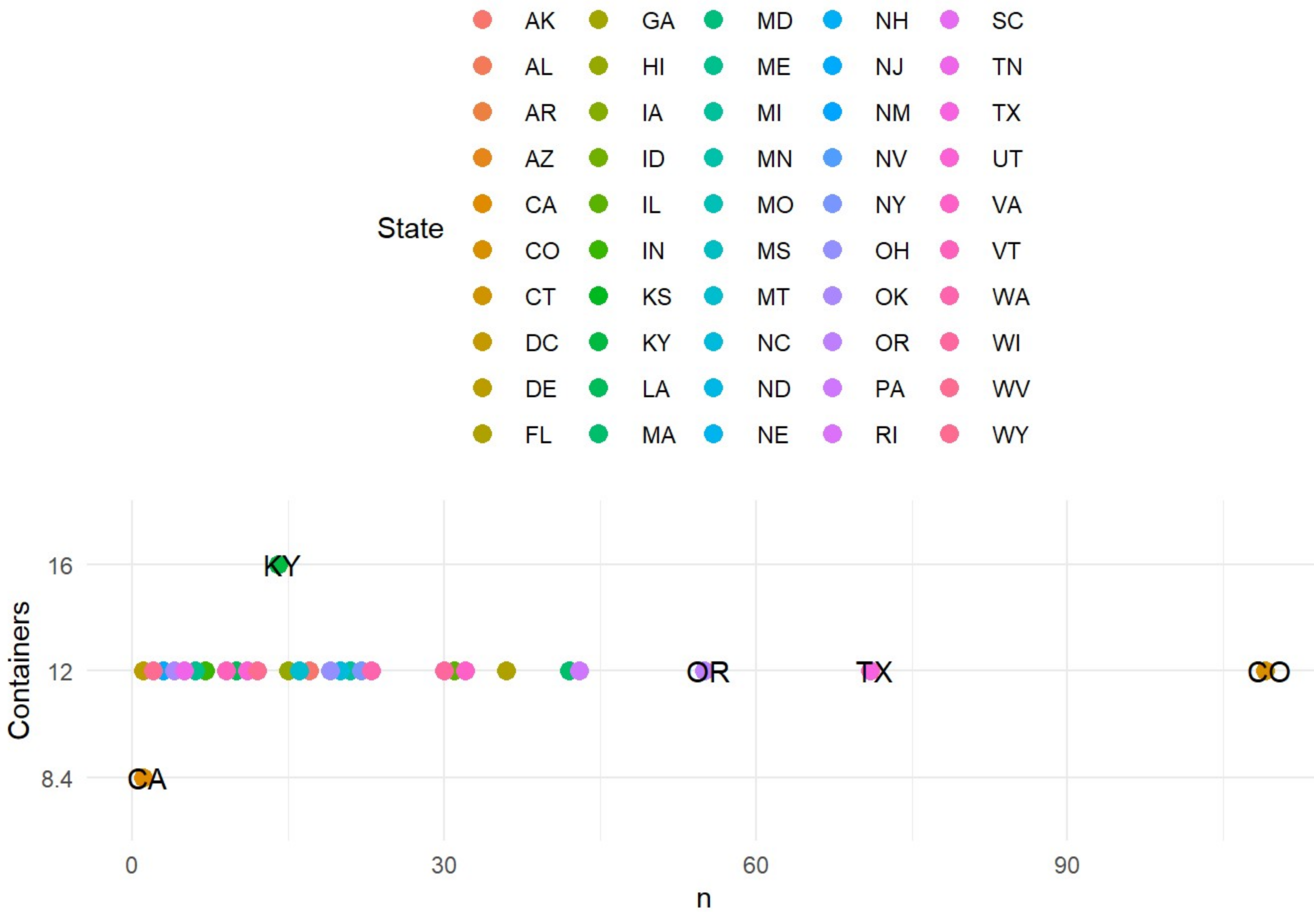
```
###Observed best k-means is 30 based on 200 iterations. Our knn-means classification with a p
recision of 88.5% and recall of 83.5%. This model better with precision, but lacks recall.
```

9. Knock their socks off! Find one other useful inference from the data that you feel Budweiser may be able to find value in. You must convince them why it is important and back up your conviction with appropriate statistical evidence.

```
## # A tibble: 1,405 x 3
##    Ounces State Containers
##     <dbl> <fct> <fct>
##  1     16 " MN" 16
##  2     16 " MN" 16
##  3     16 " MN" 16
##  4     16 " MN" 16
##  5     16 " MN" 16
##  6     16 " MN" 16
##  7     16 " KY" 16
##  8     16 " KY" 16
##  9     16 " KY" 16
## 10     16 " KY" 16
## # ... with 1,395 more rows
```

```
## # A tibble: 95 x 3
##    State Containers     n
##    <fct> <fct>      <int>
##  1 " CO" 12           109
##  2 " IN" 16            84
##  3 " CA" 12            77
##  4 " TX" 12            71
##  5 " OR" 12            55
##  6 " CA" 16            51
##  7 " PA" 12            43
##  8 " MA" 12            42
##  9 " FL" 12            36
## 10 " OR" 16            32
## # ... with 85 more rows
```

```
## # A tibble: 95 x 3
##    State Containers     n
##    <fct> <fct>      <int>
##  1 " AK" 12            17
##  2 " AL" 12             9
##  3 " AR" 12             1
##  4 " AZ" 12            21
##  5 " AZ" 16             3
##  6 " CA" 8.4            1
##  7 " CA" 12            77
##  8 " CA" 16            51
##  9 " CA" 24             3
## 10 " CA" 32             3
## # ... with 85 more rows
```


State
AK
AL
AR
AZ
CA
CO
CT
DC
DE
FL
GA
HI
IA
ID
IL
IN
KS
KY
LA
MA
MD
ME
MI
MN
MO
MS
MT
NC
ND
NE
NH
NJ
NM
NV
NY
OH
OK
OR
PA
RI
SC
TN
TX
UT
VA
VT
WA
WI
WV
WY
Containers
16
12
8.4
KY
OR
TX
CO
CA
0
30
60
90
n