An Exploration Into Craft Beers

Lester Pi

2/1/2017

Introduction

I was browsing Kaggle one day and saw a data set about craft beers and breweries in the US. This piqued my interest, so I decided to download it and take a look at it. The main areas of interest within the data set that I wanted to take a look at is the Alcohol By Volume (ABV) and the state that each brewery is in. Note that DC is included as a state.

Some questioned I wanted to look into is if different properties of the beer has statistical significance in the ABV and whether the economic and social environment in a state factors into the ABV in beers from that state. For my analysis, I converted ABV to percentage points to make the results nicer.

I acknowledge that the data is not perfect. The data used is the most up to date I could find at the time. For the purpose of exploration of curiousity, I will not focus too much on data quality.

A Look Into Craft Beers Statistics

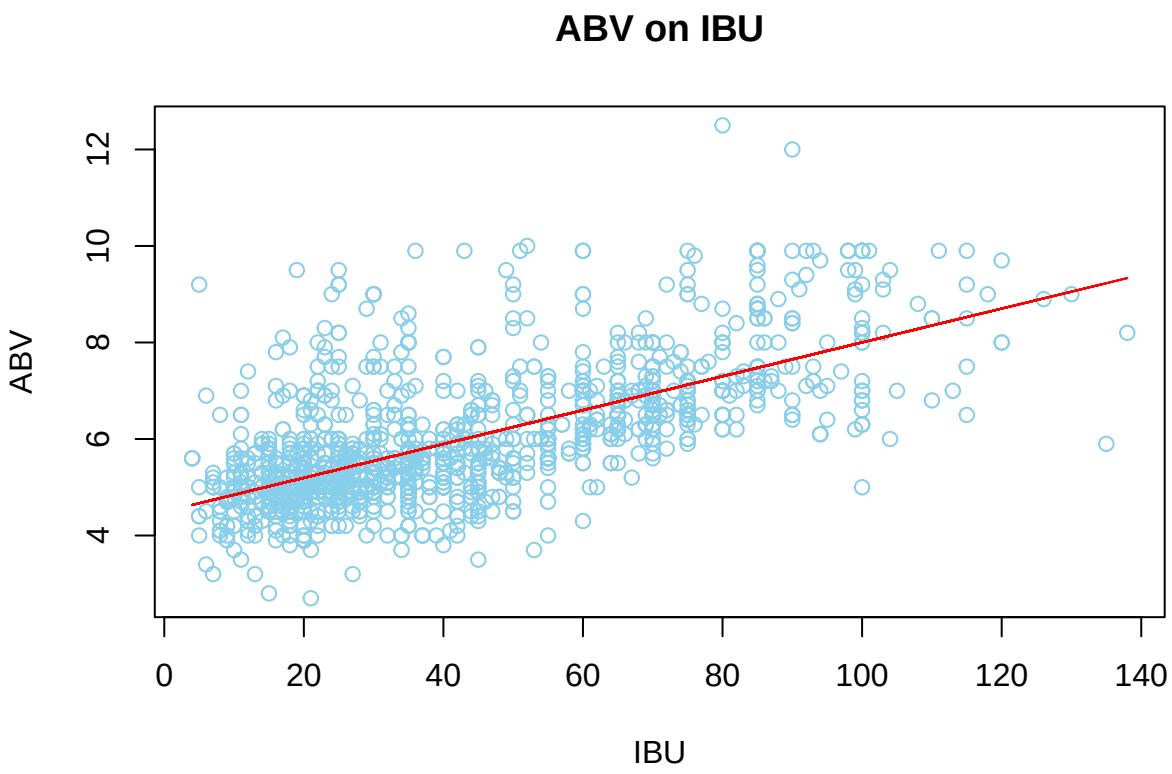
To start things off, here is a peek into the craft beer data set after merging the beers and brewery data sets.

```
##   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 abvPoints
## 1 Minneapolis    MN       4.5
## 2 Minneapolis    MN       4.9
## 3 Minneapolis    MN       4.8
## 4 Minneapolis    MN       6.0
## 5 Minneapolis    MN       6.0
## 6 Minneapolis    MN       5.6
```

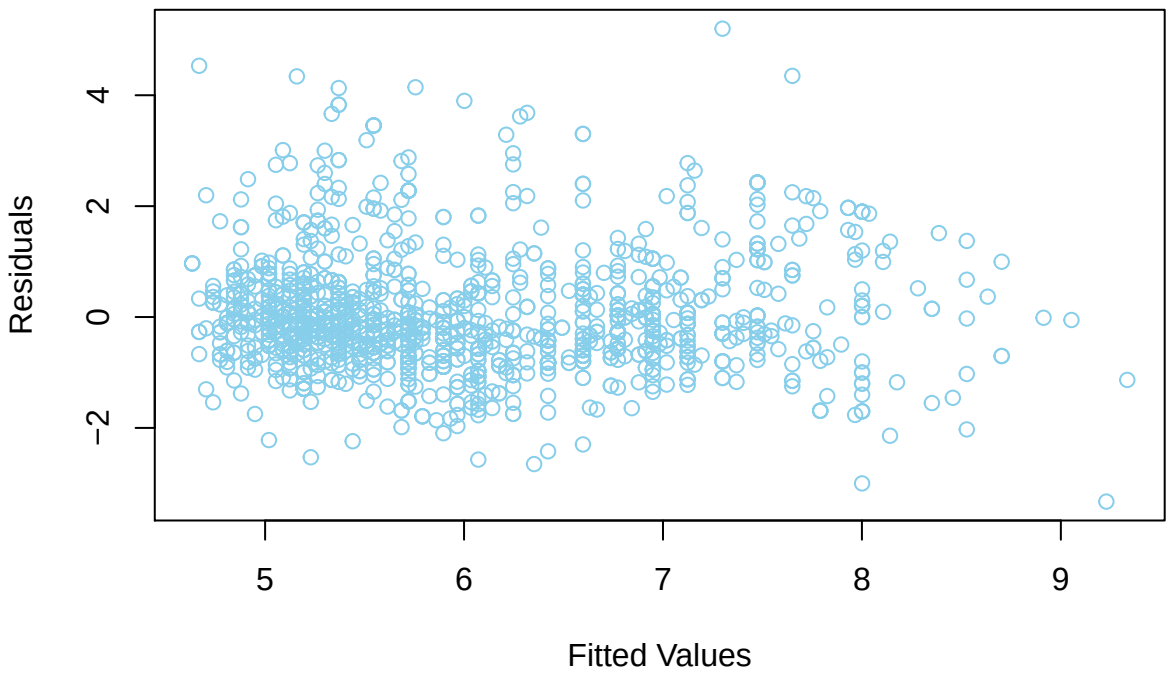
Some of the columns of interest to me, other than state and abv, are the ibu and style. The ibu column represents the International Bittering Units (IBU), which is a measurement of how bitter a beer is. The bitterness of a beer is most commonly attributed to the amount of hops used during the brewing process. This leads me to look at the beer style column, which tells us what kind of beer it is.

I ran a regression of ABV on IBU to see the effect IBU has on ABV.

```
##
## Call:
## lm(formula = abvPoints ~ ibu, data = no_ibu_na)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3288 -0.5946 -0.1595  0.4022  5.2006
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.493028   0.051772   86.79   <2e-16 ***
## ibu         0.035080   0.001036   33.86   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.007 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
```



Its easy to see there is an obvious relationship between IBU and ABV. The coefficient on ABV is strong and statistically significant and the R Squared isn't too bad either. However, the plot suggests there may be heteroskedasticity.



Looking at this plot, it makes the case for heteroskedasticity stronger. Let's run a BP test.

```
##
##  studentized Breusch-Pagan test
##
## data:  ibu_reg
## BP = 9.861, df = 1, p-value = 0.001688
```

The BP test suggests there is heteroskedasticity. However, I believe this isn't too big of a problem because there is much more variety of beer around the 5-7% ABV range than in the 8%+ range. Same is true for the higher IBU levels.

The Myth of IPAs

It is common to think that IPAs are stronger than other beers. I believe this is can be explained by looking into the bitterness relationship.

I was able to group the different types of IPAs into one IPA group through a regex match.

A side note, some good information about why IPAs have a higher ABV can be found here: http://www.titletownbrewing.com/bgk-why-are-hoppy-beers-so-strong

```
##
## Call:
## lm(formula = beers_usethis$abvPoints ~ ipas$ipa)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5949 -0.6949 -0.1949  0.4207  7.1051
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.69485    0.02972  191.60   <2e-16 ***
## ipas$ipa     1.18443    0.06086   19.46   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.257 on 2346 degrees of freedom
## Multiple R-squared:  0.139,  Adjusted R-squared:  0.1386
## F-statistic: 378.7 on 1 and 2346 DF,  p-value: < 2.2e-16
```

We can see that it looks as if being an IPA does contribute to a higher ABV. However, I would like to see if "being an IPA" actually has an effect, or is it the effect of the higher IBU associated with IPAs.

```
##
## Call:
## lm(formula = beers_usethis$abvPoints ~ ipas$ipa + ipas$ibu)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4400 -0.5879 -0.1572  0.4120  4.9960
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.426224   0.055317  80.015  < 2e-16 ***
## ipas$ipa    -0.279999   0.083722  -3.344 0.000846 ***
## ipas$ibu     0.038473   0.001447  26.582  < 2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.004 on 1402 degrees of freedom
##   (943 observations deleted due to missingness)
## Multiple R-squared:  0.4541, Adjusted R-squared:  0.4533
## F-statistic: 583.1 on 2 and 1402 DF,  p-value: < 2.2e-16
```

After controlling for IBUs in the regression, it looks as if the previous regression is a bit misleading. Being an IPA actually makes a beer have a lower ABV and an IPA's higher ABV is corelated to the IBUs, but not caused.

So, what does this all mean?

It means the other beer styles which would have the same IBU as an IPA would tend to also have a higher ABV. Since the alcohol content from beer comes from the fermentation of sugars from the malt, it makes sense that other beers with more malt would have a higher ABV after controlling for IBUs. So, "being an IPA" does not make a beer stronger. In the end, however, IPAs do tend to have a higher ABV in the market.

Breakdowning Down Beers by States

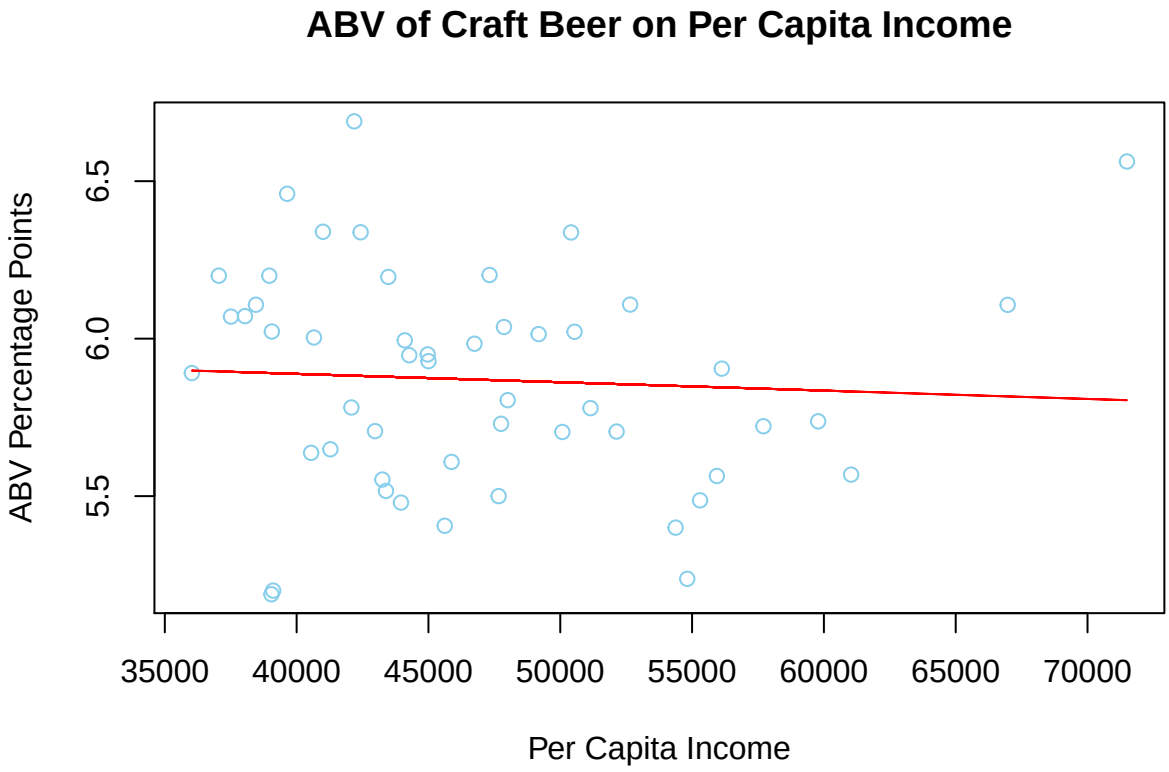
Can craft beers tell us anything about a state's economic and social environment? Probably not, but it would be interesting if it could, so lets see just for the sake of curiousity.

The data set labelled state_h contains state data on the average ABV, personal income per capita, and four different types of happiness ratings.

Note that some states have a very small sample size when it comes to craft beers and breweries, but I will continue this part of the analysis without excluding them anyways. I will average each states' ABV of craft beers so each state has an equal weight and assume this average holds since there is a lack of data and to keep things interesting.

First, I'll see if the per capita income has any effect.

```
##
## Call:
## lm(formula = abv_mean ~ state_h$per_capita_income)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.70221 -0.25652  0.05369  0.19678  0.80765
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 5.994e+00  3.007e-01   19.93   <2e-16 ***
## state_h$per_capita_income -2.652e-06  6.319e-06   -0.42    0.677
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3472 on 49 degrees of freedom
## Multiple R-squared:  0.00358,    Adjusted R-squared:  -0.01676
## F-statistic: 0.1761 on 1 and 49 DF,  p-value: 0.6766
```



Per capita income has no effect on ABV. The plot shows how the data points are all over the place and a linear fit cannot explain it.

Let's examine happiness. There is an overall happiness total score (higher = more happy), emotional & physical well-being ranking (lower = more happy), workplace environment ranking (lower = more happy), and community & environment ranking (lower = more happy).

```
##
## Call:
## lm(formula = state_h$abv ~ state_h$h + state_h$emotional_h +
##     state_h$workplace_h + state_h$env_h)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.58755 -0.24629  0.00934  0.19138  0.68193
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          5.965e+00  1.657e+00   3.599 0.000778 ***
## state_h$h           -4.743e-03  2.324e-02  -0.204 0.839219
## state_h$emotional_h -2.008e-05  1.085e-02  -0.002 0.998531
## state_h$workplace_h  7.554e-03  4.700e-03   1.607 0.114860
## state_h$env_h       -1.625e-03  5.175e-03  -0.314 0.754965
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3321 on 46 degrees of freedom
## Multiple R-squared:  0.1446, Adjusted R-squared:  0.07019
## F-statistic: 1.944 on 4 and 46 DF,  p-value: 0.1192
```

We can see there is nothing significant here, but workplace happiness has the strongest significance.

Let's break this down even more by looking at just the happiness score, since it is an aggregate score of overall happiness.

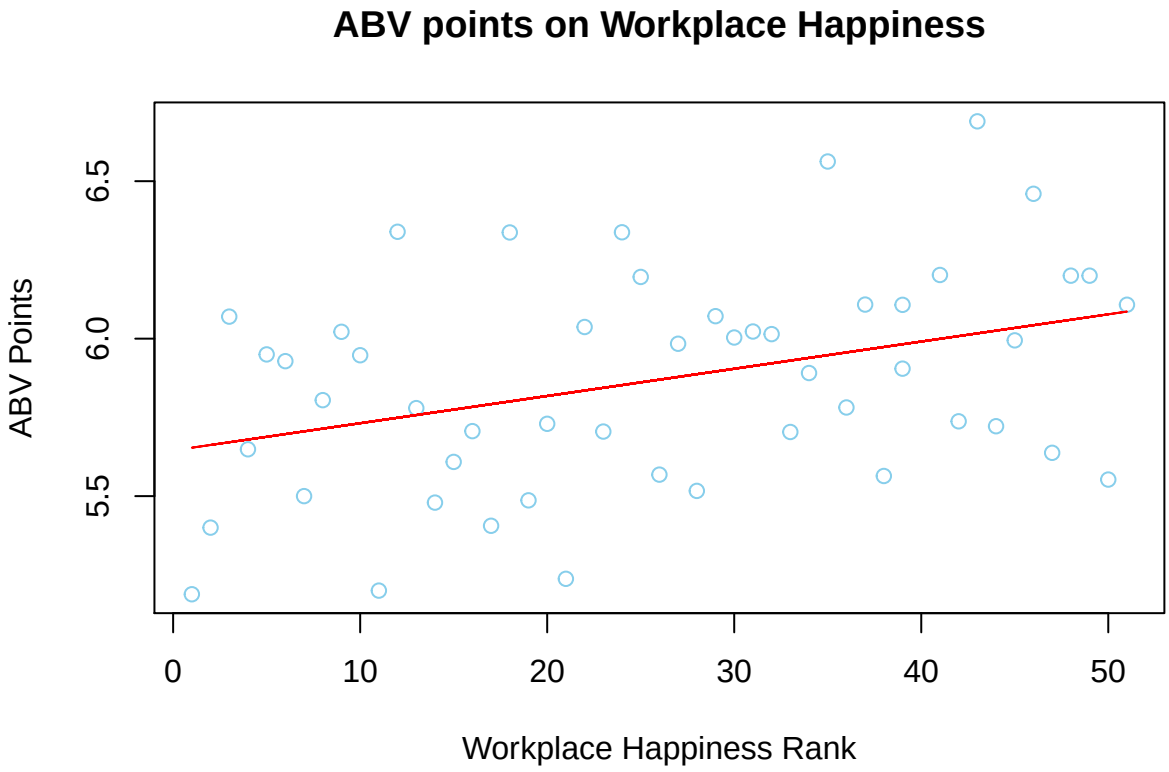
```
##
## Call:
## lm(formula = state_h$abv ~ state_h$h)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.80366 -0.24726  0.01309  0.19436  0.78095
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.364874   0.278355  22.866   <2e-16 ***
## state_h$h   -0.009451   0.005236  -1.805   0.0772 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3369 on 49 degrees of freedom
## Multiple R-squared:  0.06235,    Adjusted R-squared:  0.04322
## F-statistic: 3.258 on 1 and 49 DF,  p-value: 0.07721
```

It looks to be slightly significant, but I want to examine this further by removing irrelevant variables. I removed the intermediate regression results for easier readability.

```
##
## Call:
## lm(formula = state_h$abv ~ state_h$workplace_h)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.58913 -0.26653  0.02175  0.20223  0.67321
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         5.645120   0.091785  61.504  < 2e-16 ***
## state_h$workplace_h 0.008644   0.003075   2.811  0.00708 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3228 on 49 degrees of freedom
## Multiple R-squared:  0.1389, Adjusted R-squared:  0.1213
## F-statistic: 7.903 on 1 and 49 DF,  p-value: 0.007077
```

This is interesting. It turns out that workplace happiness rating was making the overall happiness score look more significant than it really is. It is also interesting that the more unhappy the working environment is in a state, the higher the ABV in their craft beers.

Here is a plot to visualize the results.



There is an obvious trend, even though it doesn't have a tight fit.

These results are extremely thought provoking results, perhaps more so than the relationship between alcohol consumption and workplace happiness.

Why?

Because these results can possibly tell us about consumer preferance and a consumer and producer relationship. It is possible that the breweries know their audience and cater to their local crowd. Many craft breweries only supply in their immediate areas. It is easy to see how a worse workplace environment can cause (yes, causation not correlation) a preferance for more alcohol. Now, we can say that there is a correlation between workplace happiness and ABV in craft beers that could point to the producer-consumer relationship.

It could also point to a story about who is crafting these beers. Consider this. A career paper-pusher with an incompetant manager stuck in a dead-end job decides to quit their job and pursue their dream in starting a craft brewery. Let's call this person "J". As a result of the poor work environment, J has developed a preferance for higher alochol consumption. When J sets off to create some top of the line craft beers, J brews what J likes, which turns out to be higher ABV beers that stemmed from J's preferance for higher alochol consumption.

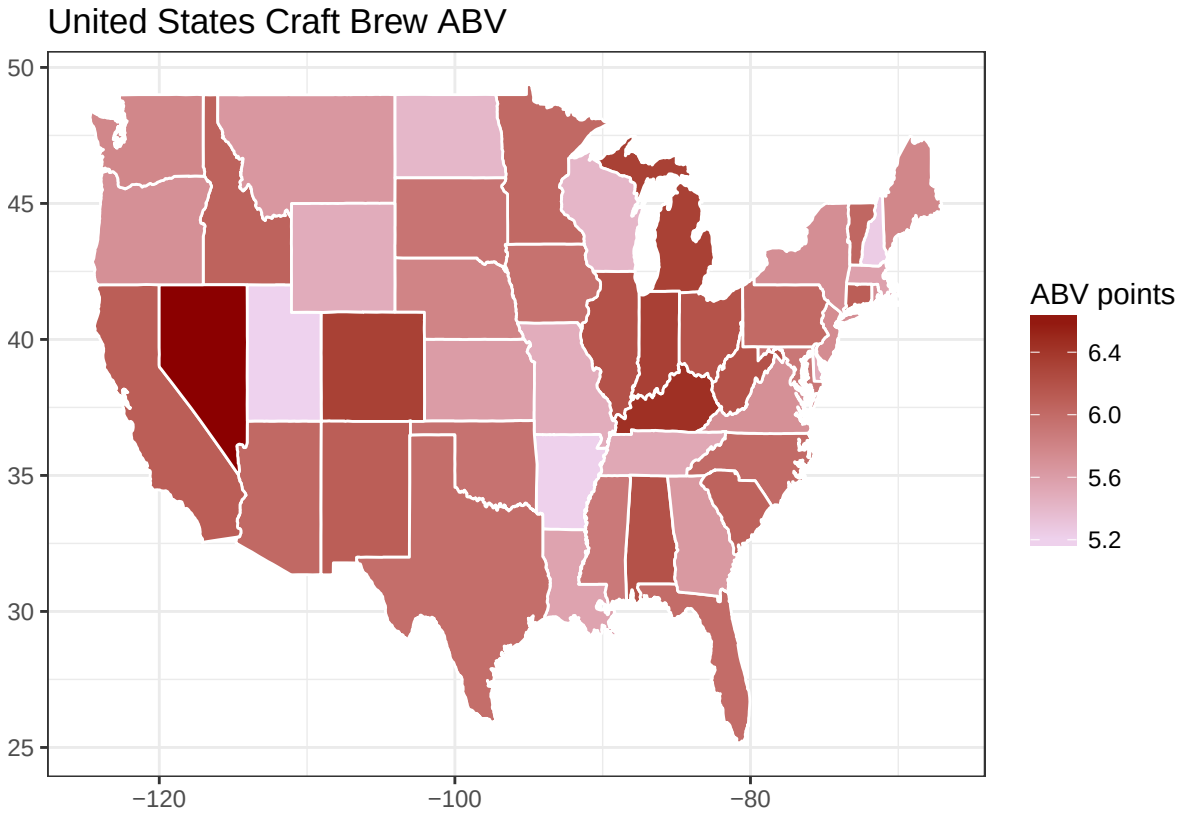
Can we find anything more out of consumer preferances? Since we found IPAs have a higher ABV, does this mean the states with a higher ABV preferance is due to a preferance for IPAs? If the assumption that the producer-consumer relationship holds, by finding the ratio of IPAs produced may give us some insight on this question.

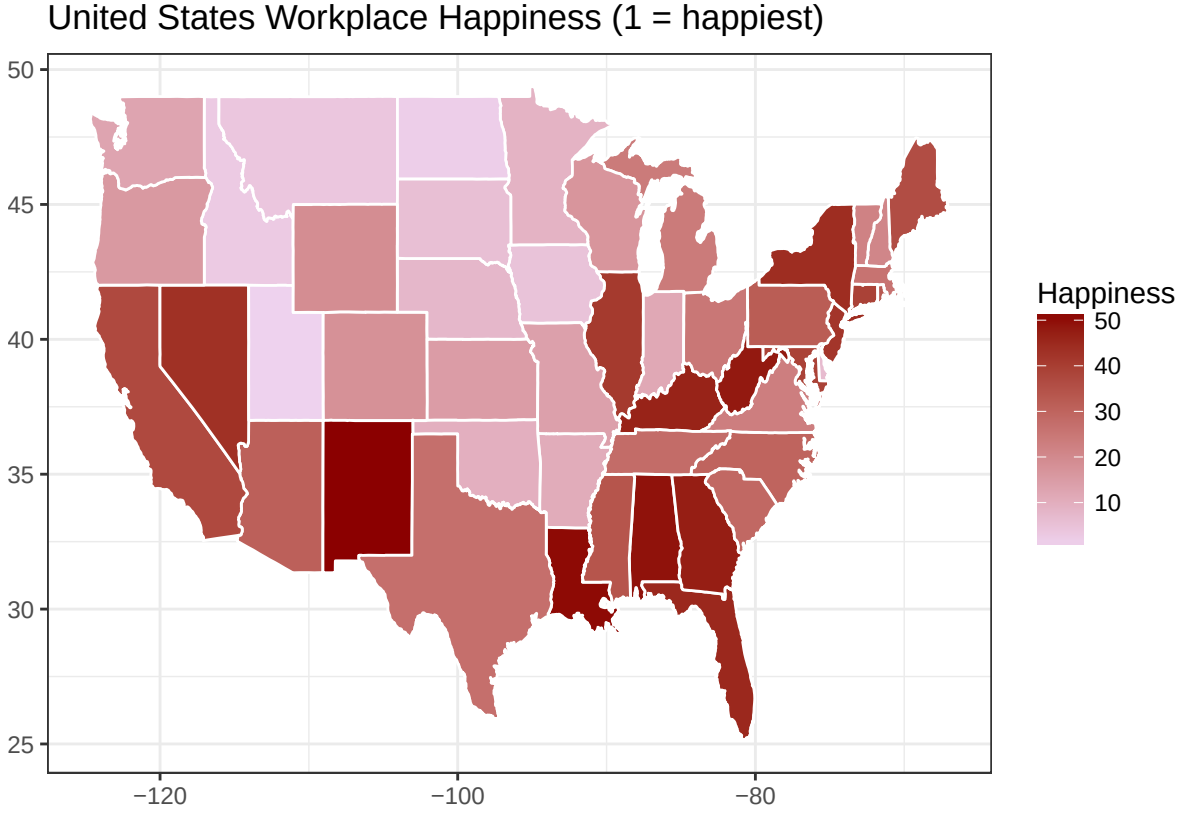
```
##
## Call:
## lm(formula = state_h$abv ~ ipa_df$ratio)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.70087 -0.24066  0.03533  0.20891  0.83020
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.8218     0.0995  58.512   <2e-16 ***
## ipa_df$ratio   0.1519     0.2756   0.551    0.584
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3468 on 49 degrees of freedom
## Multiple R-squared:  0.006158,   Adjusted R-squared:  -0.01412
## F-statistic: 0.3036 on 1 and 49 DF,  p-value: 0.5841
```

It turns out the IPA ratio does not have an effect. This only strengthens our original results on the effects of workplace happiness.

Some Cool Maps





Sources

Kaggle Craft Beer Data: https://www.kaggle.com/nickhould/craft-cans

US Happiness Data: https://wallethub.com/edu/happiest-states/6959/

US Economic Data: https://fred.stlouisfed.org/