Hop Cultivar Analysis

Art Steinmetz

2020-12-02

Introduction

Typically, I will think about hopping with a style archetype in mind, I will survey a list of hops, read about the attributes, chemical compositions and then choose the hop bill. I seldom think about the universe of hop cultivars as a whole and what meta-profile it might have. For this note, I will look at aggregate features of multiple cultivars with a focus on the top of the hop taxonomy, purpose, which can be either bittering, aroma or dual-purpose. Can we analytically determine the purpose of a hop using it's chemical composition? The simple answer is, of course, "yes." High-alpha cultivars are usually bittering hops. **The purpose of this exercise is not to discover the obvious but to explore the techniques we can bring to bear to do meta-analysis and visualization on large data sets of brewing information.**

Getting the data and putting it into analyzable form is the biggest part of these projects. We need nice tables of data but often we have unstructured or semi-structured text. We are in luck today because the good folks at brewcabin.com have put the data from 147 hop cultivars into a publicly accessible Google spreadsheet that we can easily import.

Less structured data is more of a pain but do-able. The source for much of the hops spreadsheet was information from Yakima Chief Hops. I assume they hand scraped this PDF file but we could build a robot to do it (with permission, of course). You can look at a project I did here for an example of data scraping from within PDF files.

The format of this note is a data science notebook. The code to replicate the work and present all the data is presented along with the descriptive text. In the web page version there is a button at at the top right labeled "code" which you can use to toggle code visibility on or off. **I recommend you turn code visibility "off" to begin with. This is more readable that way.**

Load the data from the spreadsheet

```
## # A tibble: 147 x 24
##    Variety Origin Type  `Alpha Acid LOW~ `Alpha Acid HIG~ `Beta Acid LOW ~
##    <chr>   <fct>  <fct>            <int>            <int>            <int>
##  1 Ahtanu~ Unite~ Aroma                5                6                5
##  2 Amaril~ Unite~ Aroma                8               11                6
##  3 Aramis  France Aroma                7                8                3
##  4 Aurora  Slove~ Aroma                7                8                2
##  5 Blanc   Germa~ Aroma                9               12                4
##  6 Bobek   Slove~ Aroma                3                7                4
##  7 Brewer~ Germa~ Aroma                5                8                2
##  8 Cascade Unite~ Aroma                5                9                6
##  9 Citra®  Unite~ Aroma               11               15                3
## 10 Crystal Unite~ Aroma                2                4                4
## # ... with 137 more rows, and 18 more variables: `Beta Acid HIGH (%)` <int>,
## #   `Total Oil LOW (mL/100g)` <int>, `Total Oil HIGH (mL/100g)` <int>,
## #   `Co-Humulone LOW (%)` <int>, `Co-Humulone HIGH (%)` <int>, `Myrcene LOW (%
## #   of total oil)` <int>, `Myrcene HIGH (% of total oil)` <int>, `Caryophyllene
## #   LOW (% of total oil)` <int>, `Caryophyllene HIGH (% of total oil)` <int>,
## #   `Humulene LOW (% of total oil)` <int>, `Humulene HIGH (% of total
## #   oil)` <int>, `Farnesene (% of total oil)` <chr>, `Description
## #   (ychhops.com)` <chr>, `Aroma (ychhops.com)` <chr>, `Beer Styles` <chr>,
## #   Substitutions <chr>, `Sources (DO NOT IMPORT)...23` <chr>, `Sources (DO NOT
## #   IMPORT)...24` <chr>
```

Clean up the Data

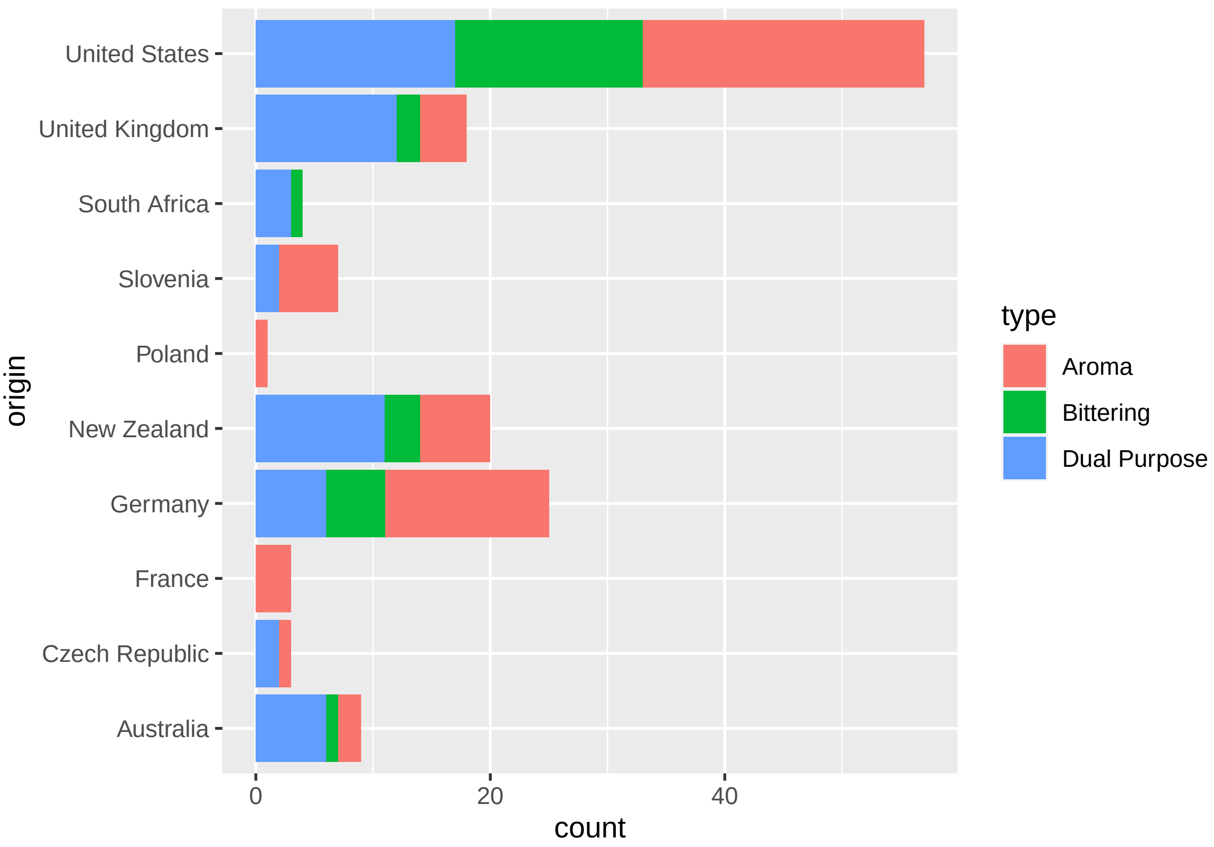
While importing the spreadsheet is easy the data still needs a lot of cleaning up. The chemical constituents are expressed as high and low ranges. I choose to condense them into average values. Further the words used to describe qualitative attributes like aroma and beer style are inconsistent. Are “citrusy” and “citrus” the same thing? Misspellings are common, German words, especially. We also strip out “stop words” like “and”,“the”,“lots”, etc, that don’t really say anything. Fixing these is a tedious art but critical to success. This part is actually the bulk of the code and uses the most run time in the notebook. One thing that’s confusing about our cleaned data is that I’ve embeded all the descriptive terms nested as lists in single columns. That will be make running our models easier.

```
## # A tibble: 147 x 14
##    variety type  origin alpha  beta cohumulone myrcene caryophyllene humulene
##    <chr>   <fct> <fct>  <dbl> <dbl>      <dbl>   <dbl>         <dbl>    <dbl>
##  1 Ahtanu~ Aroma Unite~   5.5  5.5        32.5    52.5         10.5      18
##  2 Amaril~ Aroma Unite~   9.5  6.5        22.5    69            3        10
##  3 Aramis  Aroma France   7.5  3.5        20.5    40            7        20.5
##  4 Aurora  Aroma Slove~   7.5  3          23      22.5          7.5      22.5
##  5 Blanc   Aroma Germa~  10.5  4.5        24      62.5          1         1.5
##  6 Bobek   Aroma Slove~   5    5          29      37.5          5        16
##  7 Brewer~ Aroma Germa~   8    3.75       44      45.5          8.75     22.5
##  8 Cascade Aroma Unite~   7    6.5        32.5    52.5          7        17
##  9 Citra®  Aroma Unite~  13    3.5        22      65            6.5       9.5
## 10 Crystal Aroma Unite~   3    5          23      52.5          6        21
## # ... with 137 more rows, and 5 more variables: farnesene <dbl>,
## #   total_oil <dbl>, aroma <list>, beer_styles <list>, substitutions <list>
```

Exploratory Data Analysis (EDA)

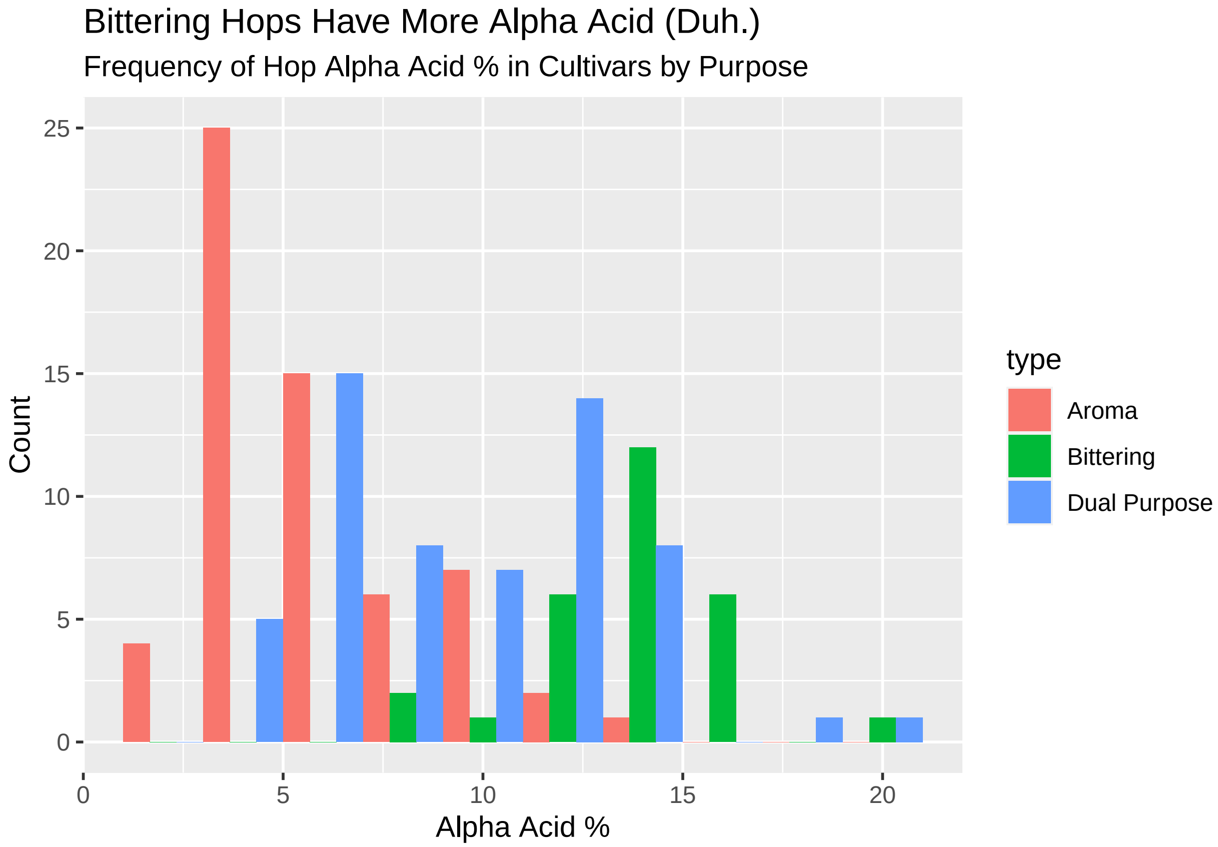
Origin

We’ll start by looking at some basic features of the data set. In this spreadsheet the U.S. represents the majority of the cultivars. While the folks at Brew Cabin claim this is “All Hop Varieties on Earth,” you can go to the growers association web pages for any of these countries and find many more. Caveat Emptor.

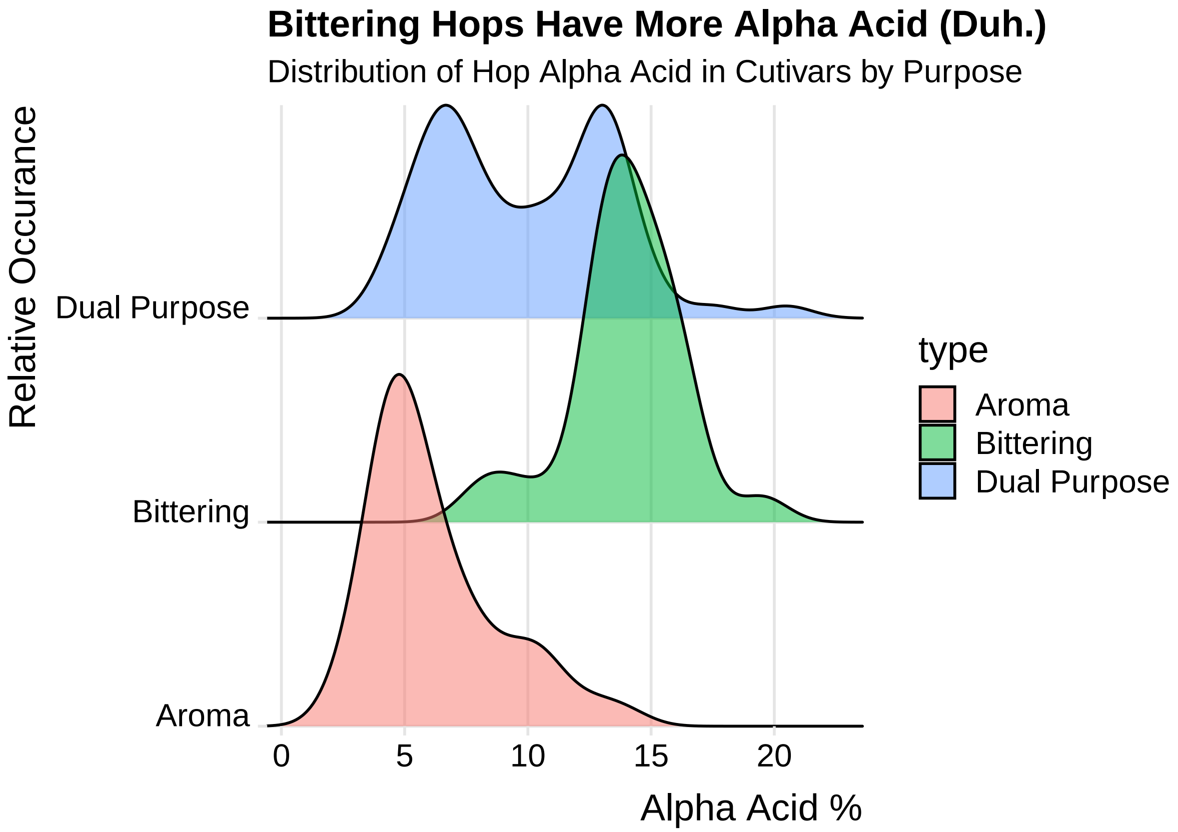

United States
United Kingdom
South Africa
Slovenia
Poland
New Zealand
Germany
France
Czech Republic
Australia
origin
0
20
40
count
type
Aroma
Bittering
Dual Purpose

Alpha Acids

Every brewer knows that high alpha hops are for bittering. Let's demonstrate this.



Most of the green bars (bittering) have more than 10% alpha acids and most of the red bars (aroma) have less. This chart could be clearer. Let's do a "density" plot instead.



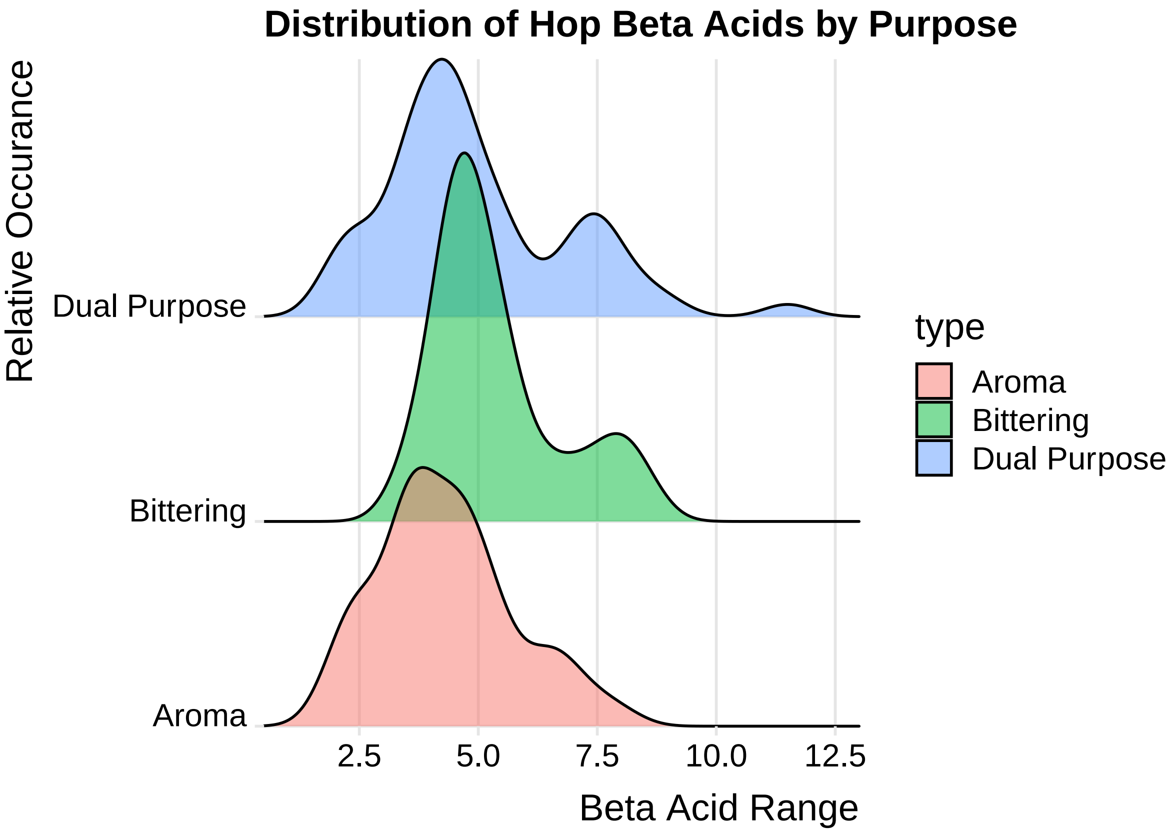
This is a much more pleasing plot and actually reveals a surprise. Dual purpose hops are not spread evenly across the alpha acid spectrum but are bi-modally distributed. Is this because growers select for bittering or aroma and brewers put the dual purpose label on later?

Now let's look at two other major components, beta Acid and total oil concentration. Unlike alpha acid, there is not clear story.

Beta Acids

These don't seem related to purpose at all.

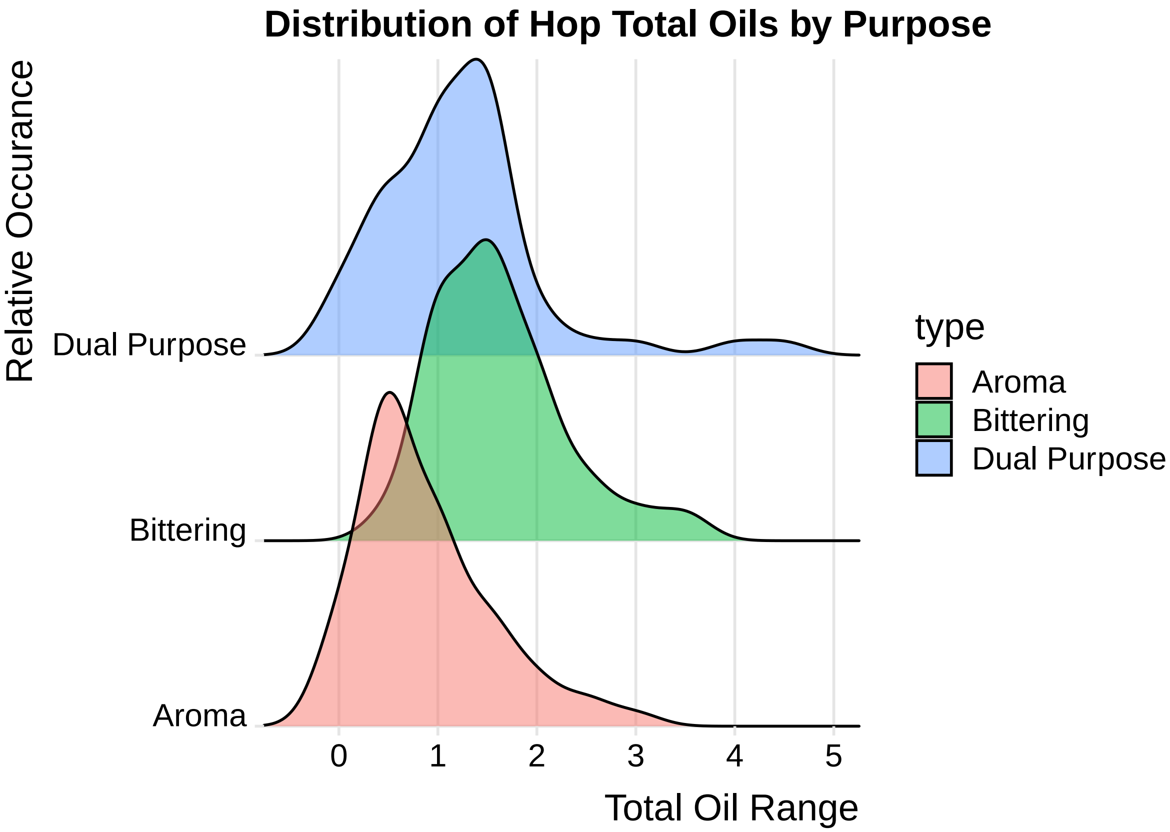
```
## Picking joint bandwidth of 0.5
```



Total Oil

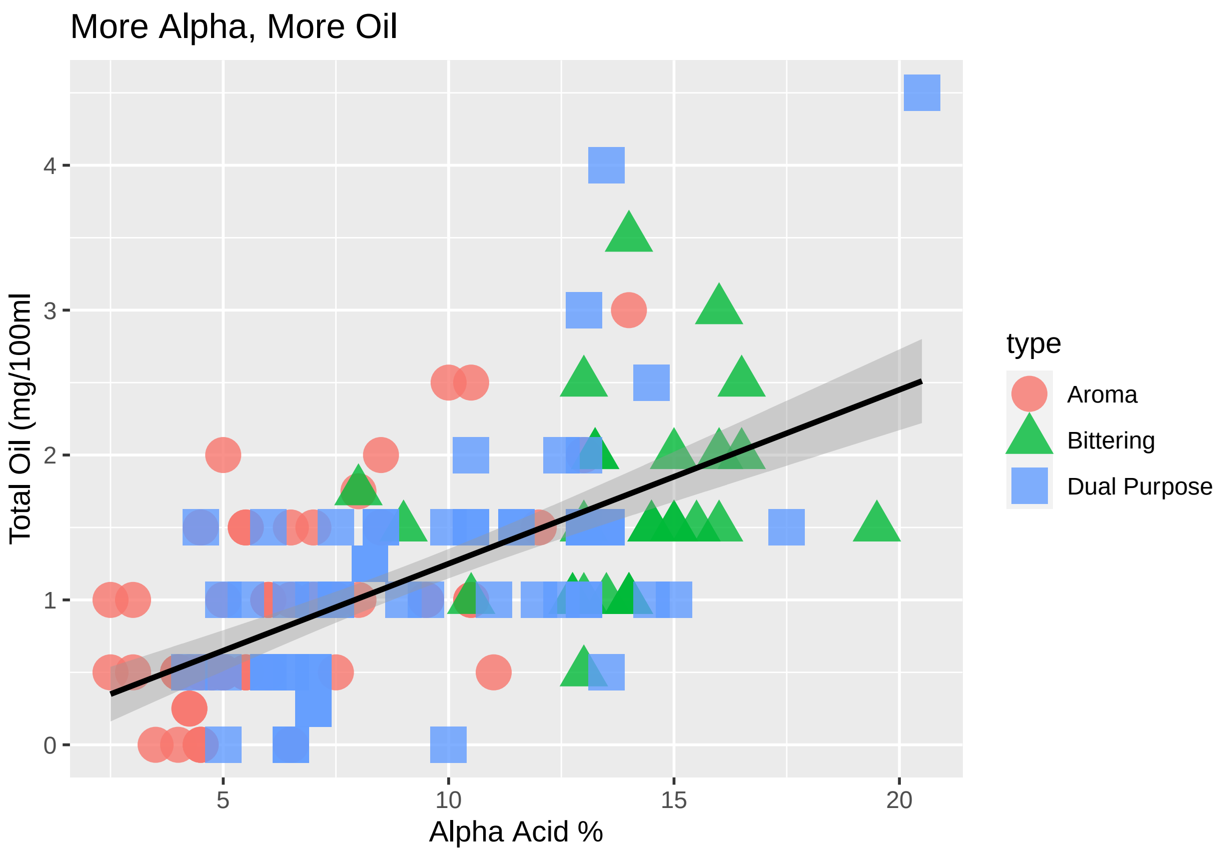
Oils are where the aroma components reside. We might suspect that more oil is associated with aroma hops but here we can see that bittering hops also have relatively more more oils.

```
## Picking joint bandwidth of 0.252
```



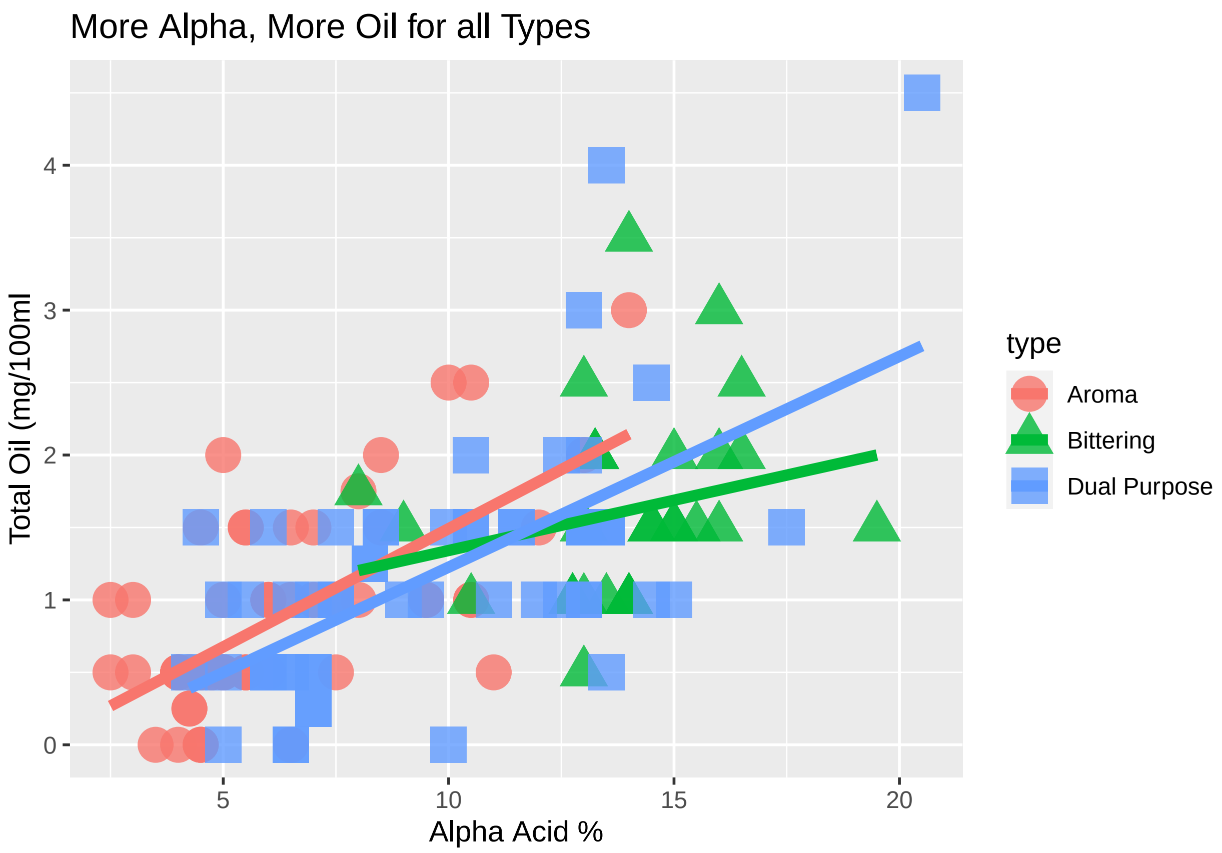
We can also see how hop constituents relate to each other. We've seen how higher alpha hops are generally higher oil hops,also. A simple scatter plot shows the relationship. It's not tight but more alpha means more everything.

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



The relationship of “more alpha is more everything” is true for all three types.

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

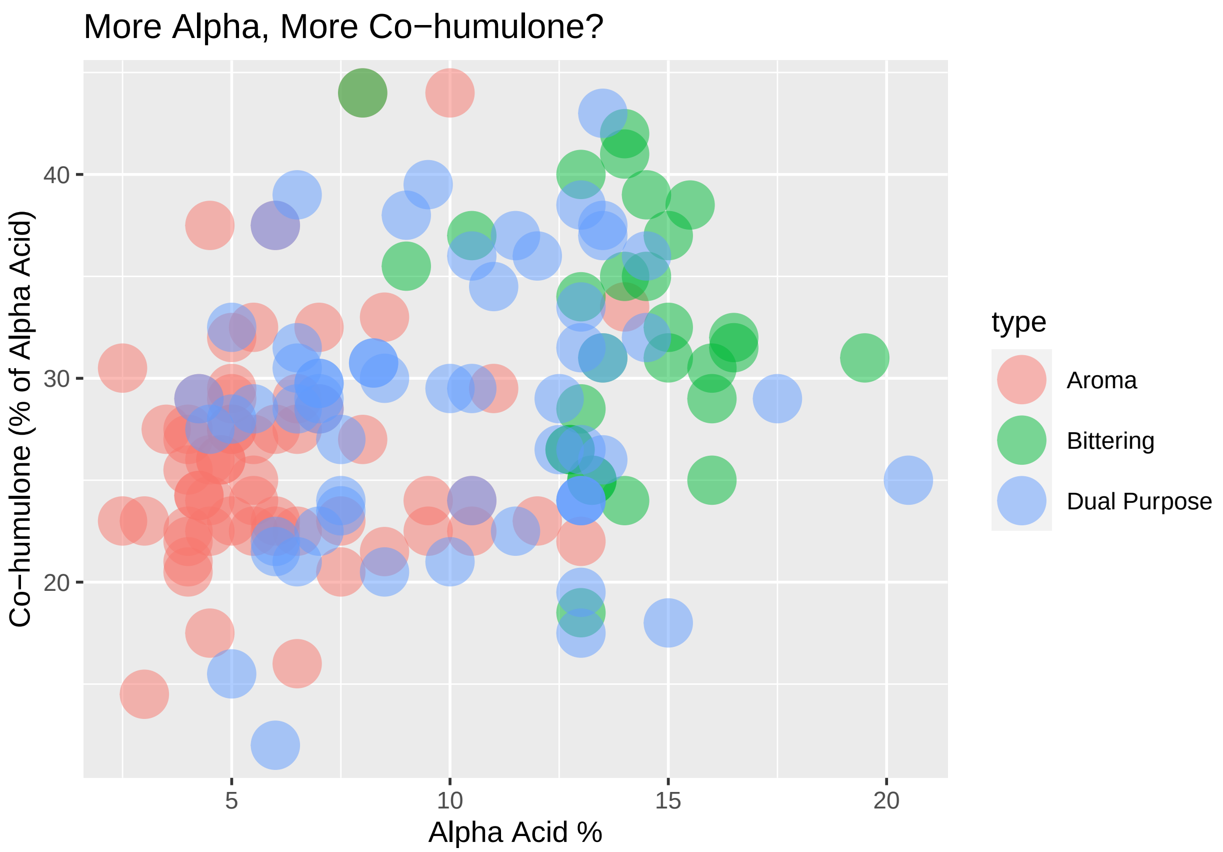


## Other Components

So far, only alpha acid, among alpha, beta and total oil, seems to distinguish hop types. Let's dig further and look at other constituents.

Co-Humulone

Co-humulone has a rap for contributing harshness but that has been somewhat debunked of late. Below we can see that bittering hops have a higher average fraction of co-humulone making up the alpha acid percentage than aroma hops, but the relationship is very loose.

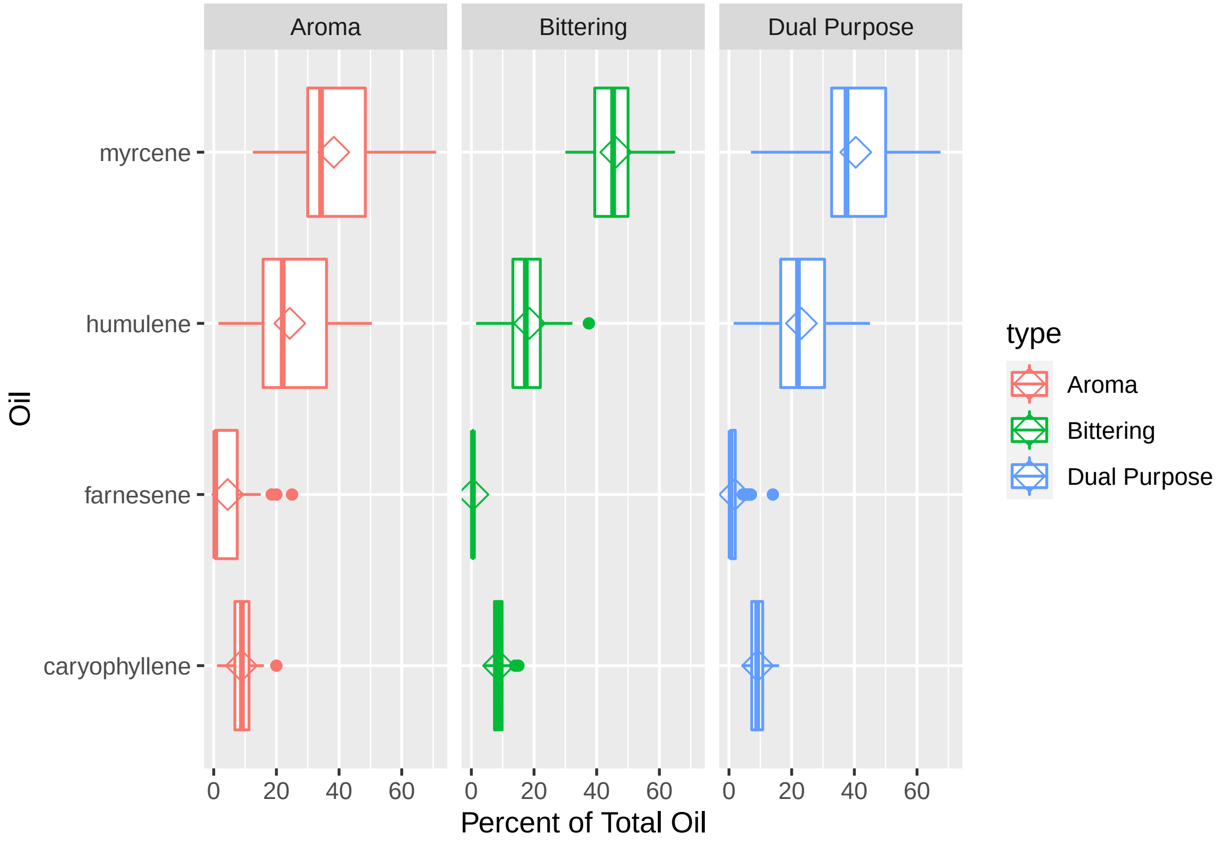


Individual Oils

We've looked at total oils. Does the makeup of the oils have a bearing on classification? We'll switch up the visualization and use boxplots to visulize the range of oils in all varieties by type. These plots show the median, the mean, the interquartile range and big outliers. As we can see the range of oil fractions doesn't change much among varieties, though the extremes are pretty wide. Nor do we see big differences in average fractions by type. We do notice wider variability among aroma hop fractions than for bittering hops.

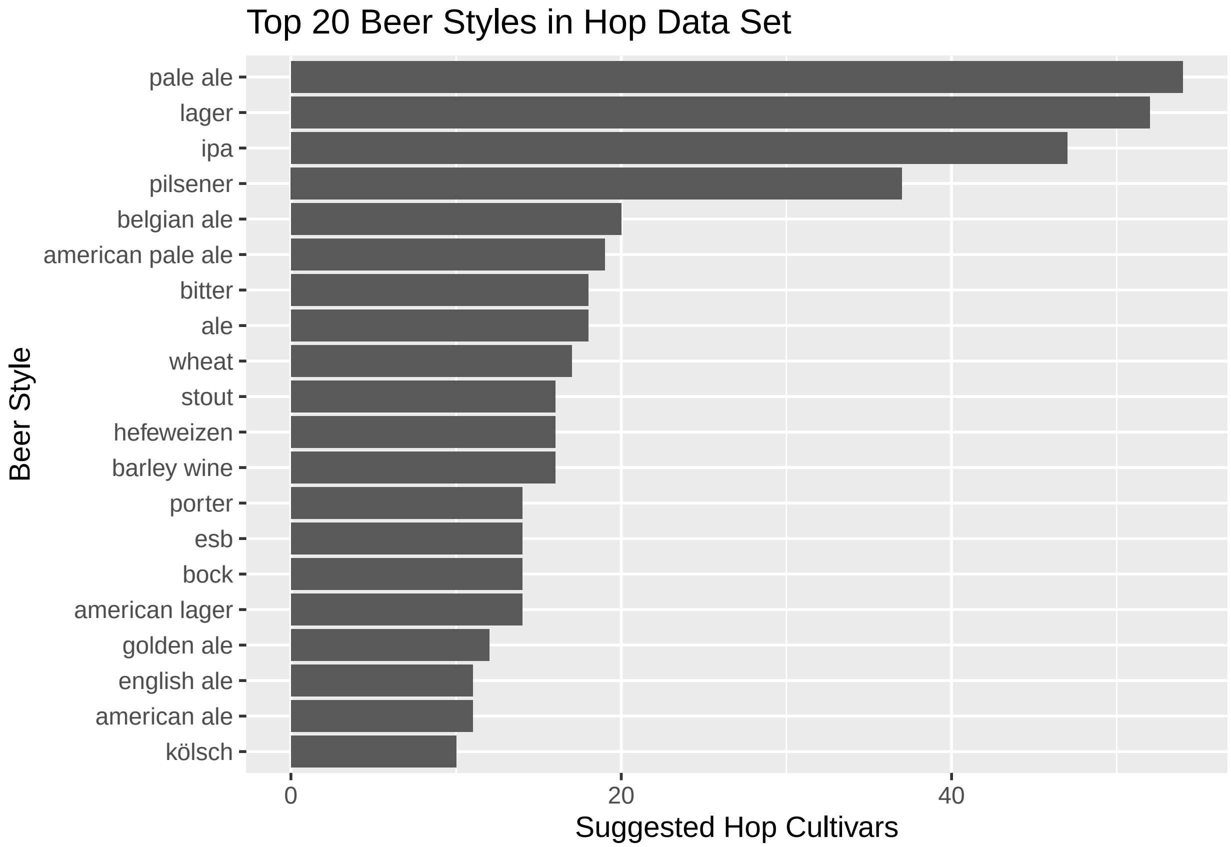
```
## Warning: Removed 33 rows containing non-finite values (stat_boxplot).
```

```
## Warning: Removed 33 rows containing non-finite values (stat_summary).
```

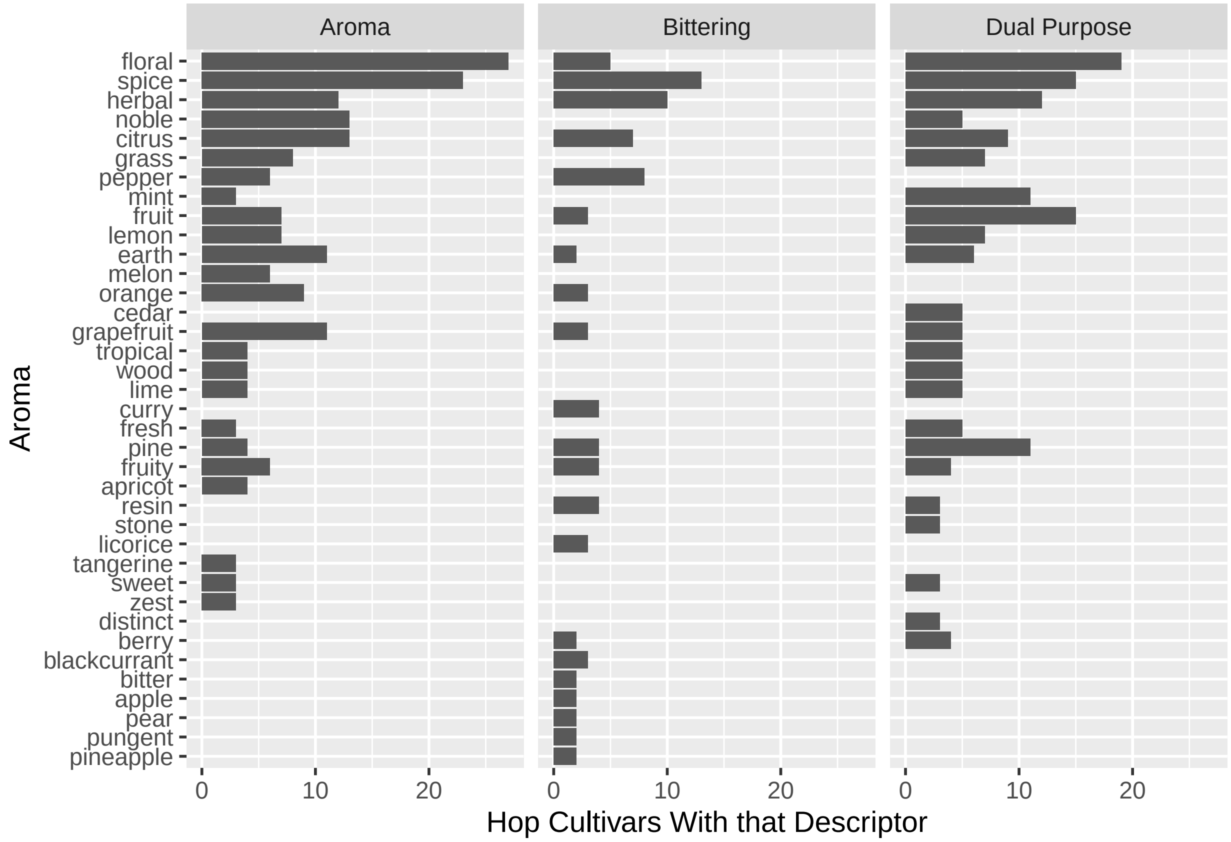


Qualitative Descriptors

Thus far we've looked at the quantitave featues of the hop data set. Since we turned the prose-style qualititative descriptors into word lists, we can examine those as well. Which beer styles are the most flexible with respect to hop choice? Maybe this chart answers the question or maybe it just answers the question "what styles of beer are most popular among brewers?"



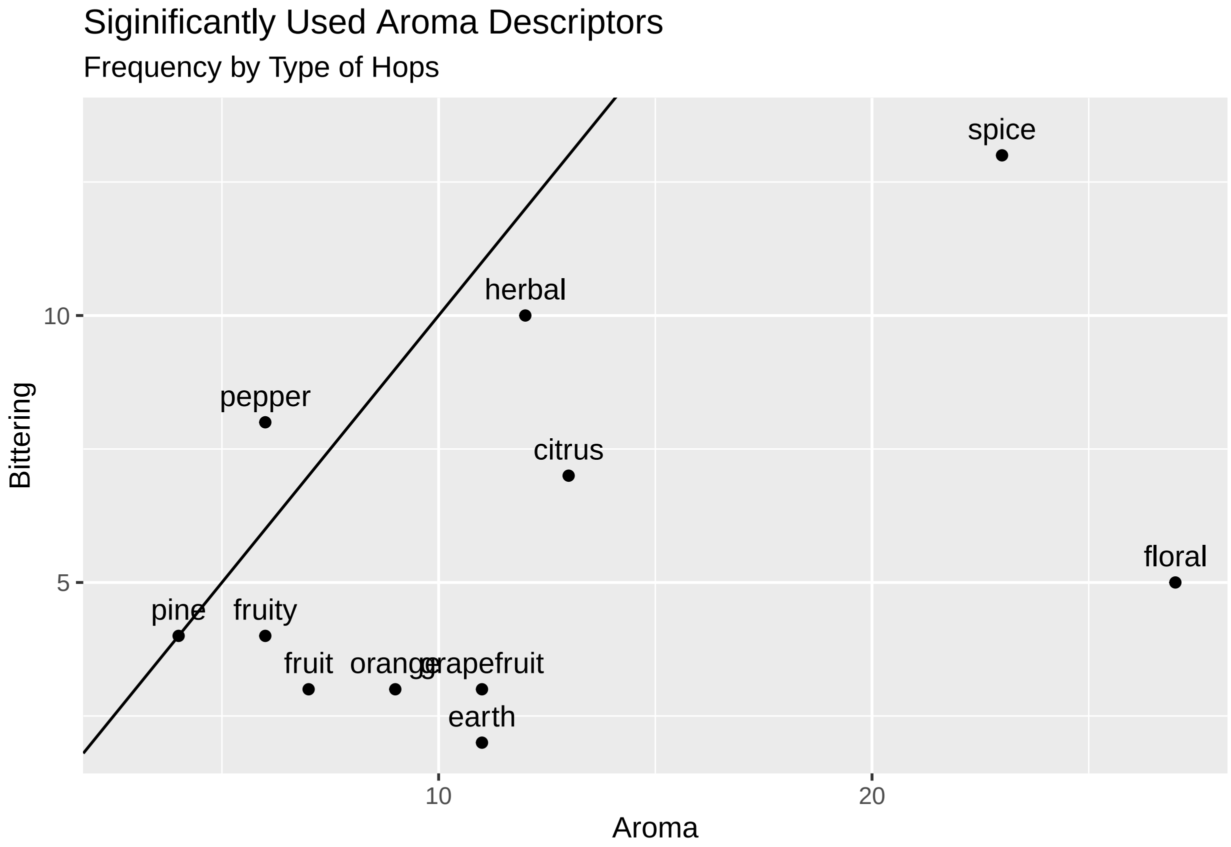
More interesting, perhaps, are the choices of aroma descriptors by hop type. “Noble” is never used to describe bittering hops, as we would expect. “Floral” is very popular for aroma hops.



The only significantly used aroma word that is applied more to bittering hops is “pepper.”

```
## Warning: Removed 26 rows containing missing values (geom_point).
```

```
## Warning: Removed 26 rows containing missing values (geom_text).
```

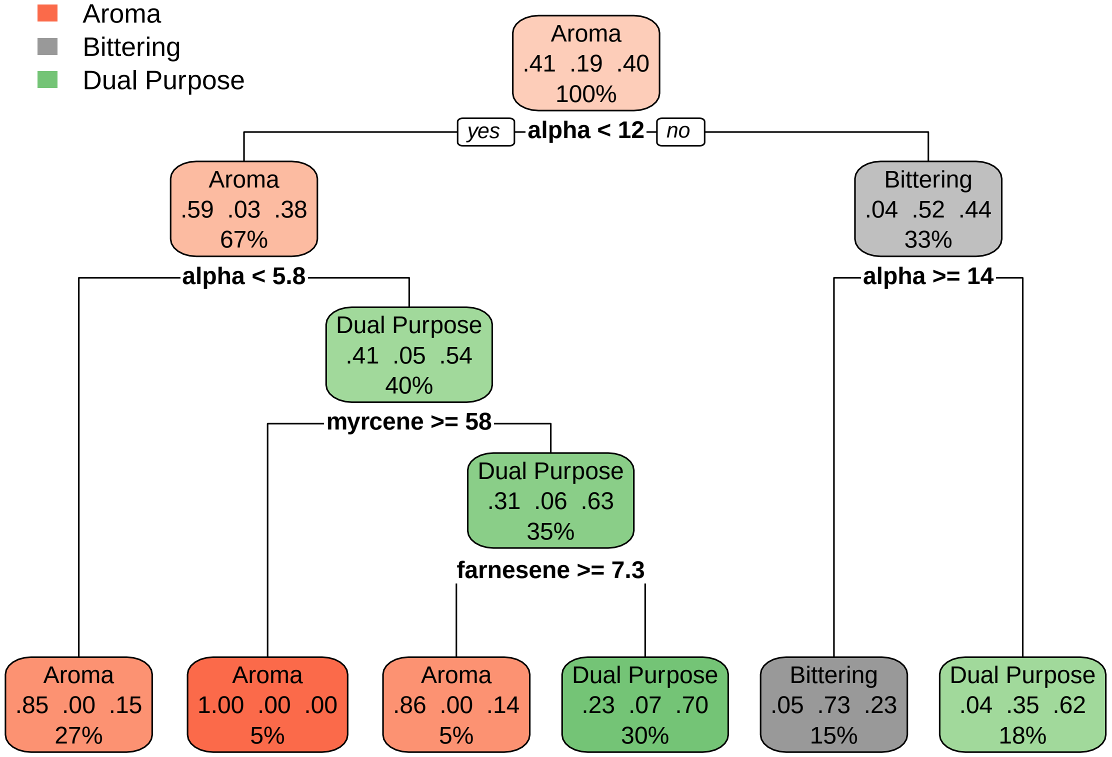


Machine Learning

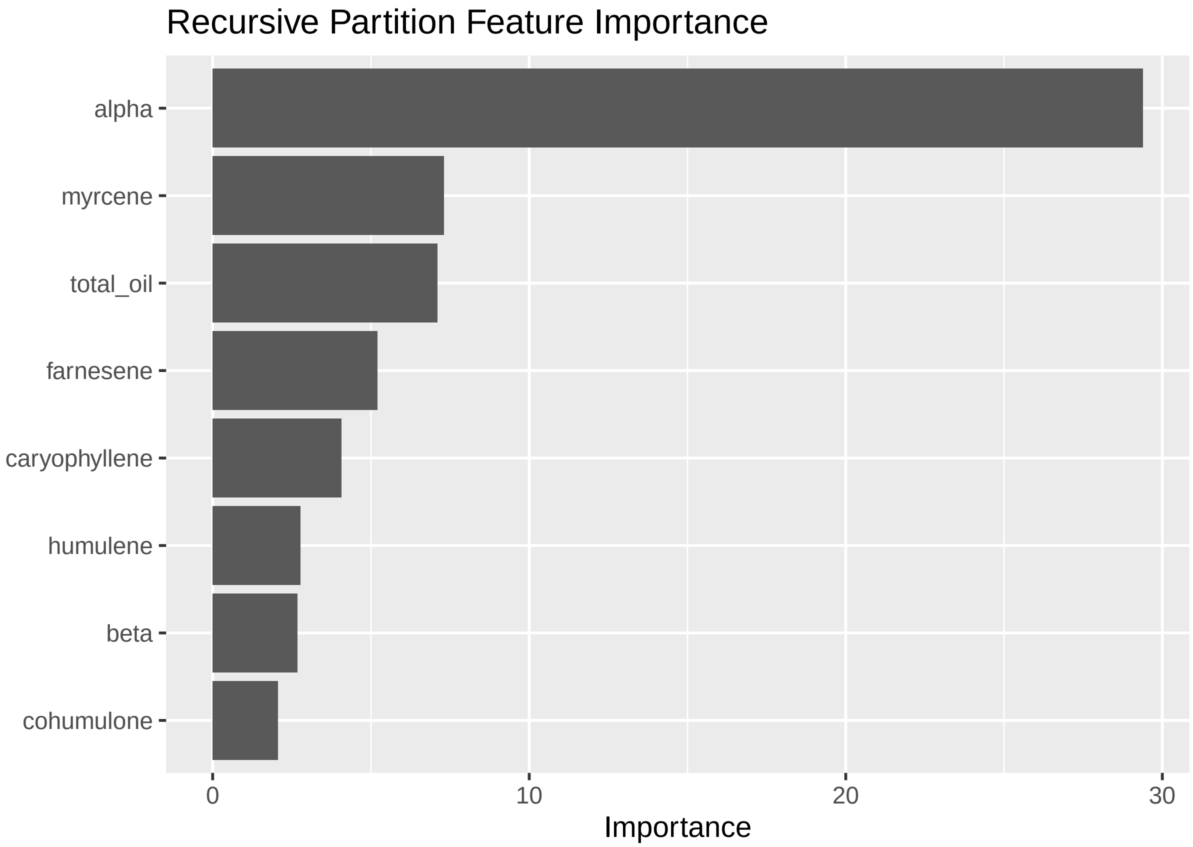
Finally, let's apply the intuition we've gained by kneading the data set with our hands to bake up some predictive models. Can we use the data to predict how a hop will be classified? Again, this is not super interesting becuse we know what makes a bittering hop, but lets get a little more precise and also see if we can take a crack at predicting what makes a "dual purpose" hop.

Recursive Partitioning

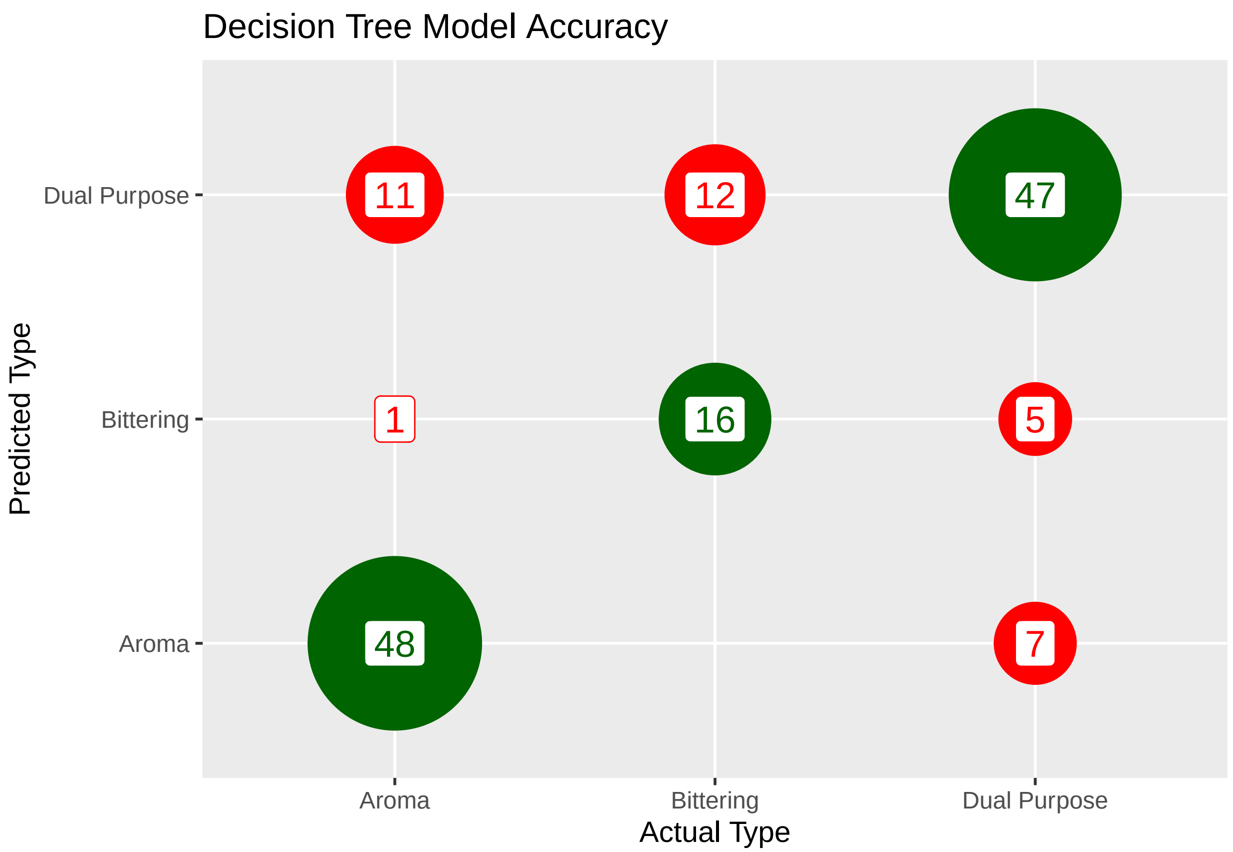
The first tool we'll try is "recursive partitioning." This is an attractive choice because the model gives us an easily explainable descision tree. For this we'll use only our quantitative variables.



What I find interesting here is that farnesene, which is typically the smallest fraction of oil, is relevant for selecting between aroma and dual purpose hops. We have to be cautious because, given the small sample size, it could be coincidence. How important are each of these features? Again, farnesene is the only surprise. Co-humulone is least important.



Finally, how accurate is our model? It got 25% wrong.

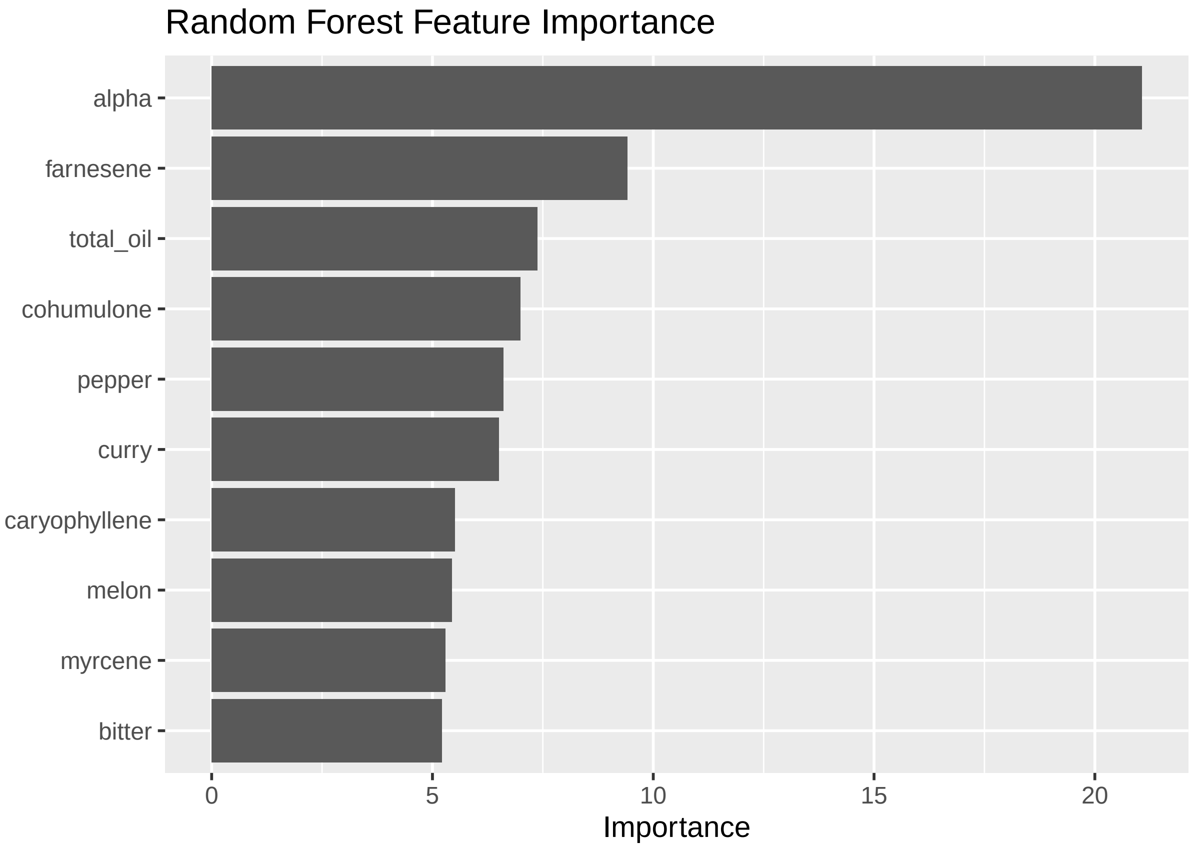


Random Forest

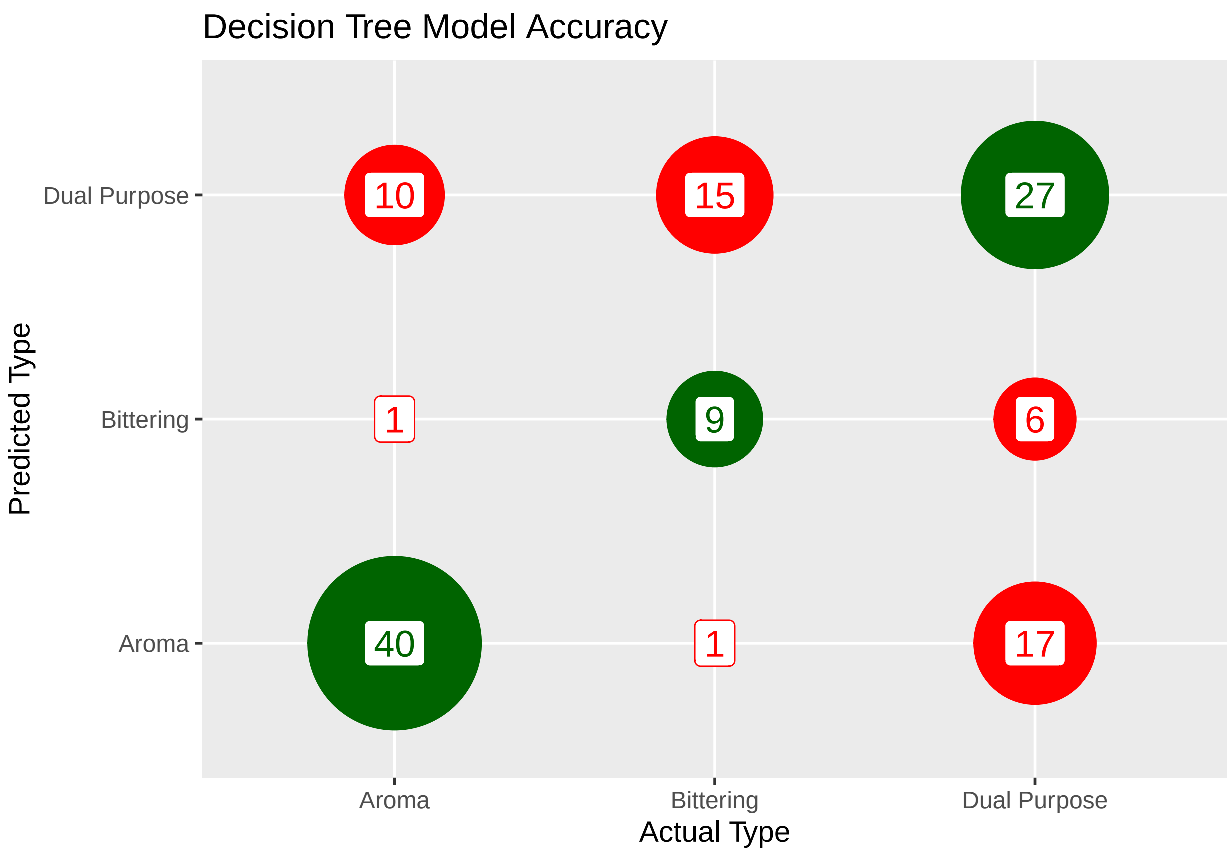
There are many sophisticated machine learning models suitable for use on massive data sets. Most are overkill for this small set but let's play with one, "random forest." It will allow us to add qualitative descriptor words as features. In this case we'll use aroma. One caution right off the bat. Do the people who ascribed these aromas know the variety and purpose of the cultivar beforehand? Almost certainly. That introduces a bias in descriptors, I assume. We have to manipulate the data set again by turning each aroma word into a unique "dummy" variable in the data set. Each word gets its own column with "1" if it is used to describe the hop and "0" if it not. This expands the size of the set to 141 columns but still pretty small in the machine learning world.

As our models get more sophisticated we lose something in explainability. We could show the decision tree but it would be a tangled spider web without much intuition. This model runs 500 diffent decision trees to find the best one.

Once again, alpha is the most important variable but several aroma words make the list as well.



How accurate is the model? The error rate of this model is actually higher at 36%. Adding features and complexity didn't improve results. This may be beacuse several hop varietys have no aroma information at all, so our sample size is smaller and the results are not completely comparable.



We are cheating in both of these cases for the sake of illustration. Technically, if we want to do predictions, we should split our data into training and test sets so we do our predictions "out of sample." This data set is too small to do this effectively.

Conclusion

This has been a practice exercise just to demonstrate some of the tools we can use to explore relationships in brewing data. We've tortured a small data set to wring every last insight out of it. I look forward to delving in to larger brewing data sets that have more mysteries to uncover.