Analysis of Beer Hops

COMP4447 – Data Science Tools 1

Fall 2021 – Final Project

Tomer Danon & Romith Challa

https://github.com/rc-9/tools1_project

Abstract

Beer is one of the oldest and most preferred beverages by humans on earth. Although it has been brewed and consumed for centuries, one of its main ingredients has changed throughout the years. It wasn't until the 16th century when hops became more widely used and replaced the use of gruit, an herb mixture that was used to bitter and flavor beer. The hop plant Humulus Lupulus is a member of the Cannabaceae family of flowering plants. The flower of this plant is called a hop and may also be referred to as a seed cone or strobile. Hops are used as a bittering, flavoring, and stabilizing agent in beer. They consist of oils and other compounds that are largely responsible for the unique flavors and aromas present in each hop variety, as well as the bitterness it provides.

Having worked in the brewing industry for several years, one finds themselves in conversations about the aromas, flavors, and bitterness of different varieties of hops from around the world. Some varieties are better used for bittering, while other varieties are better used for aroma and flavor. Beers with hops from New Zealand have a unique profile that seems very different than hops from Europe. There is something about South African hops profiles that are not found in hops from the United States. This leads to the questions:

- *What is the relationship between the oil concentrations and aroma tags of hops from the same continent?*
- *What is the relationship between the brewing values of hops and the brewing purpose?*
- *Could a model be developed to accurately predict the origin and purpose of a hop based on the brewing values and oil concentrations?*

Background

Over the years, brewers and scientists have studied which compounds in the hop affect the aroma, flavor, and bitterness in beer. Hops high in aroma and low in bitterness would be used for their aroma properties, while hops low in aromatic oils and high in bitterness would be labeled as bittering hops. Hops that have the best of both worlds are considered dual purpose. In the dataset, this is represented by the **Purpose** feature. Two of the most prominent brewing values that relate to this are alpha acids and beta acids.

Alpha acids (α acids) are a class of chemical compounds found in the resin glands of the hop plant flowers. They are the source of bitterness in beer, and they possess anti-bacterial properties. The bitterness level is a result of a process called isomerization which happens in the boiling stage of the brewing process. The degree of isomerization and hence the bitterness are highly dependent on the length of time the hops are boiled. Longer boil times result in isomerization of more alpha acids and therefore increase the bitterness. Beta acids are compounds in hop resins that contribute volatile aromatic and flavor properties. They contribute no bitterness. The ratio of alpha to beta acids represents the degree to which bitterness fades during storage. 1:1 ratio is common in aroma varieties, while 3:1 ratio is found in more alpha-dominant varieties. These relate to the following features of the dataset: **Alpha Acid % - Min/Max/Avg, Beta Acid % - Min/Max/Avg, Alpha-Beta Ratio – Min/Max/Avg.**

Co-Humulone % of Alpha relates to the bitterness type and feel. When added to the boil, hops with low co-humulone are generally associated with smoother bitterness, while hops with higher co-humulone are associated with sharper bitterness. The final beer greatly depends on the balance of its' ingredients. Depending on the other hops or the style of beer, one may choose a hop with higher or lower concentration of co-humulone or other compounds. In the dataset, this is represented by the features: **Co-Humulone as % of Alpha – Min/Max/Avg.**

In addition to the acids described above, the hop contains oils that relate to the flavor and aroma of each unique hop variety. These oils are highly volatile, and not sufficiently soluble. They are easily boiled off but add flavor and aroma to the finished beer when added very late in the boil, during fermentation, or in a late hopping process called dry-hopping. In the dataset, he total oil measurement is represented by the feature: **Total Oils (mL/100g).**

The two most notable oils that make up majority of the total oils are myrcene and humulene. Myrcene (β-myrcene) is mostly associated with citrusy and resinous-pine aromas. It is the most abundant of all the oils in hops and is the most potent since it has the lowest odor threshold of 13 ppb. Hops with high myrcene include Amarillo, Citra, Simcoe, and Cascade. Humulene (α-humulene and α-caryophyllene) is mostly associated with woody and pine aromas; however, hops with higher amounts tend to be more floral, herbal, and black pepper in character. Hops with high α-caryophyllene percentage include Vanguard, Perle (GR) and East Kent Goldings. In the dataset, these are represented by the features: **Myrcene – Min/Max/Avg** and **Humulene – Min/Max/Avg.**

Two other oils in the hop are Caryophyllene and Farnesene. Caryophyllene (β-caryophyllene) is mostly associated with black pepper, spiciness, and herbal aromas. Farnesene (α-farnesene and β-farnesene) is mostly associated with floral aromas, and slightly associate with notes of woods and citrus. Hops high in farnesene include Tettnanger, Sterling and Saaz. In the dataset, these are

represented by the features: **Caryophyllene – Min/Max/Avg** and **Farnesene - Min/Max/Avg.**

The standards for aroma identification were initially formed in 2011 by the BarthHass Group (BHG) with the help of Frank Rittler, a world-renowned and experienced perfumer based in Düsseldorf, Germany. In 2018, the BarthHass Group (BHG) finally introduced this set of standards with the name HOPSESSED®, based on the initial twelve aroma categories noted by Rittler. The twelve categories of HOPSESSED® are described in Appendix 1. The aroma tags in the data were partly based off the twelve categories and their subcategories. For this research they were eventually made into Boolean features in the dataset.

Literature Review

While there is some research on hops, much of it is focused on breeding, genetics, and chemistry. In the search to find similar work, this Github repository was discovered. The hop data were scraped from sources different than the one used for this project, and a couple extensive visuals are presented.

- https://github.com/vieuxsinge/hops-datasets

In these two references, researchers used classification techniques on hop photos:

- https://www.kaggle.com/scruggzilla/hops-classification
- https://www.researchgate.net/publication/354093321_Dataset_for_Hop_Varieties_Classification

Here are three notable institutes that carry out research on beer hops:

- https://en.wikipedia.org/wiki/Hop_Research_Center_H%C3%BCll
- https://www.hopresearchcouncil.org/
- https://www.barthhaas.com/en/world-of-flavor/hopsessed

Dataset

The source of the data: https://beermaverick.com/hops/

Beer Maverick has compiled one of the largest and most extensive databases of current beer hop varieties. Much of this data were sourced from hop farms, hop breeders, and hop sellers. It is mentioned by the company that for some hops, contradictory data was found. This explains why the data is provided as minimum, maximum, and average. Without measuring again, it would be hard to know which of these values are true and which may be a mistake. This may lead to some outliers in the data. Throughout this research and analysis, outliers were not obvious, nor did they appear to distort the results. It is also important to note that each yearly crop can yield hops that have slightly different qualities, so ranges provided are based on history.

Data Scraping

Please refer to file: step1_scraper.ipynb.

The data were web-scraped directly from Beer Maverick's website using the library **BeautifulSoup**. Once it is confirmed with **urllib.robotparser** that the website would allow the scrape, a connection is establish using the **requests** library.

1. The hop names are scraped into a list from: https://beermaverick.com/hops/
2. This is used to build a link for each individual hop. For example: https://beermaverick.com/hop/citra/
3. From there, there are three points of interest:
    a. Hop profile table – first table on the page
    b. Flavor & aroma tags – center of the page
    c. Brewing values – table at the bottom of the page

A scraper generator function is created to handle the scrape that will yield a dictionary with data for each hop. Each dictionary represents a row in the dataframe. The dataframe is created from all the dictionaries using **pandas** library and saved to a csv file (raw_hops_main.csv). Some reference materials containing meta-data were scraped and saved into csv files. They are described below:

- The Hop Substitution table from this page: https://beermaverick.com/hops/hop-substitutions-chart/, and saved as csv (raw_ref_hops_substitutions.csv)
- Flavor & aroma meta-data scraped from this page: https://beermaverick.com/the-science-behind-identifying-hop-aromas/, and saved as csv (raw_ref_aroma_types.csv)
- Brewing values meta-data scrape from this page: https://beermaverick.com/hop/newport/, and saved as csv (raw_ref_brew_values.csv)

Data Cleaning

Please refer to file: step2_cleaner.ipynb

Use of **pandas** dataframe methods and processes to clean the raw data that were scraped.

raw_hops_main.csv

- Removed columns: 'Unnamed: 0', 'Scraping Status', 'Cultivar/Brand ID:', 'International Code:', 'Ownership:', 'Hop Storage Index (HSI)'
- Cleaned characters of no use.

- ‘Hope Name’ as the index
- Make ‘Country’ and ‘Purpose’ categorical data.
- The brewing value columns have min/max/avg in them. Created new columns for each of those values.
- Fill empty values with np.nan from **Numpy** library
    - Cleaned file: cln_hops_brewvalues.csv
- Convert the column ‘Flavor & Aroma Profile’, which consists of lists of tags, to Boolean columns for each tag.
    - Cleaned file: cln_hops_profile.csv

raw_ref_hops_substitutions.csv

- ‘Hop Name’ made into index.
- ‘Substitutions’ converted from comma separated strings, to list of strings.
- Cleaned file: cln_ref_hops_substitutions.csv

raw_ref_aroma_types.csv

- ‘Aroma Type’ made into index.
- ‘Aromas’ and ‘Compounds’ converted from comma separated strings, to list of strings.
- Cleaned file: cln_ref_aroma_types.csv

raw_ref_brew_values.csv

- Cleaned characters of no use.
- Cleaned file: cln_ref_brew_values.csv

Missing values in the data were changed to np.nan values. These rows will be dropped for cleaner visualizations and to satisfy data requirements for the model building.

Exploratory Data Analysis

It is important to explore and understand the data prior to the implementation of analytics. A few functions that can help with this are df.info() and df.describe().

- df.info() returns general information about the dataframe size, column datatype, etc.
- df.describe() performs basic analysis and returns the count, mean, std, min, max, and quartiles of each column.

The rest of the exploratory analysis consists of visualizations and conclusions. The most important visuals are presented in the last section of this paper. Please refer to eda_and_summary_visuals.ipynb for complete EDA.

Feature Engineering

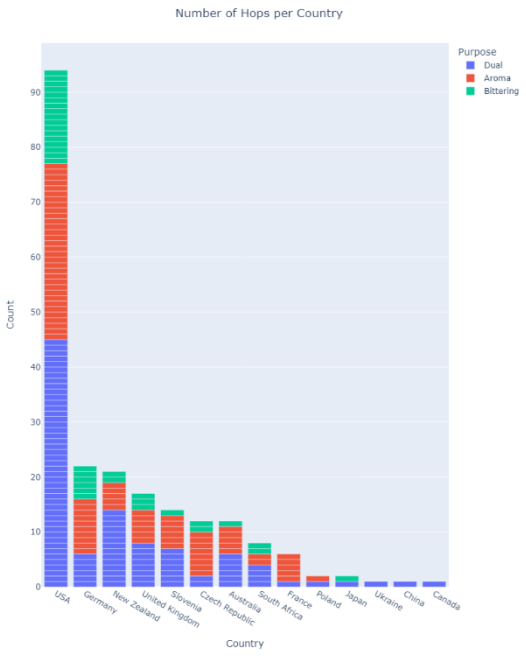
It is often necessary to create new features for a more in-depth analysis. For this research, the following new features were engineered.

- **Boolean Aroma Tags**: The aroma profile tags were scraped for each hop along with the rest of the features. It was saved into a list. To make use of this data in visualizations and analysis, it was converted into Boolean columns, with True/False entries.
- **Grouping of data frames for plots** – A few plots show the relationship between brewing values and continent or oil concentration and continent. This required to group the dataframe by 'Continent' and sum() or mean() the columns of each group to engineer new features.
- **Continent**: The hop country of origin was scraped with the rest the features. It was observed that these data could be group into larger regions. This will look cleaner for visualizations and work better for classification models. Therefore, a new column with the name Continent was created.
- **Alpha-beta Ratio:** converted string representation of a ratio expression (x:y) to an evaluated integer(z).

Visualizations, Classifier Models & Conclusions

Number of Hops Per Country

This graphic shows the count of hops for the countries in the dataset. Quite clearly, the USA has developed the most hops out of all the countries, followed by Germany and New Zealand. Europe would come in second after USA. The colors in the plot represent the use purpose of the hop. Some hops are used specifically for bittering, while others are used more for their aromatic properties. Some hops give the best of both worlds and have dual use. Countries like USA and New Zealand have mostly dual-use hops, while European countries such as Germany and Czech Republic have mostly aromatic hops. As hop techniques improved, the focus has shifted from using hops as a bittering agent to using hops for their aroma characteristics. As it can be seen in the graphic, most hops have either a dual purpose or aromatic purpose.

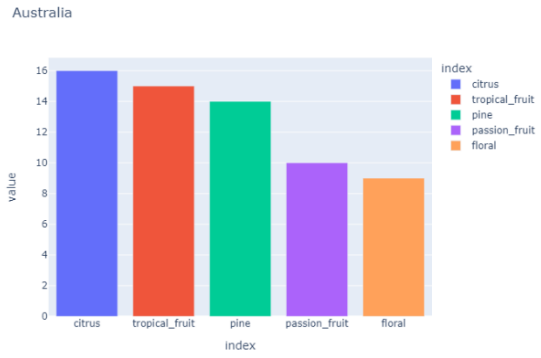


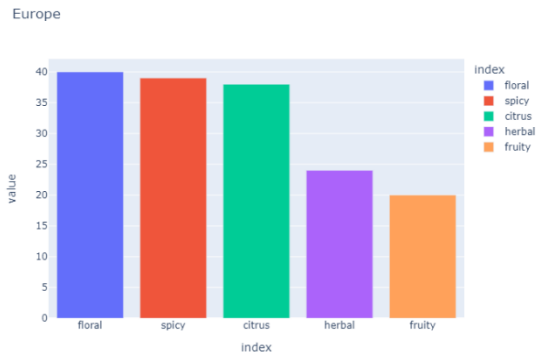
Conclusion 1: Oil Concentration Effect on Aroma Tags

What is the relationship between the oil concentrations and aroma tags of hops from the same continent?

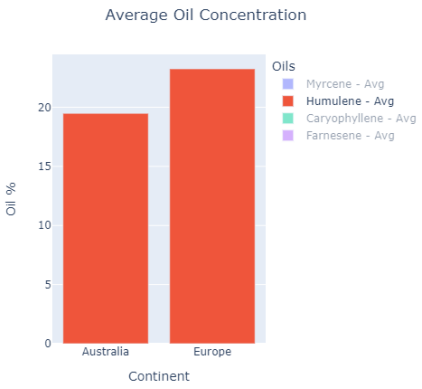
There exists a relationship between the oil percentages in the hops, to their aromas, and the continent they are from. Each hop has a unique composition of oils that affect its aromatic properties. In order to study the relationship between these oils and the continent of the hop, the following sets of plots were created. The first set presents the top 5 most used aroma tags for a specific continent. Each continent has a combination of tags that make it unique. The second set of plots relate the aroma tags to the oils they are associated with.

Australia/New Zealand vs. Europe





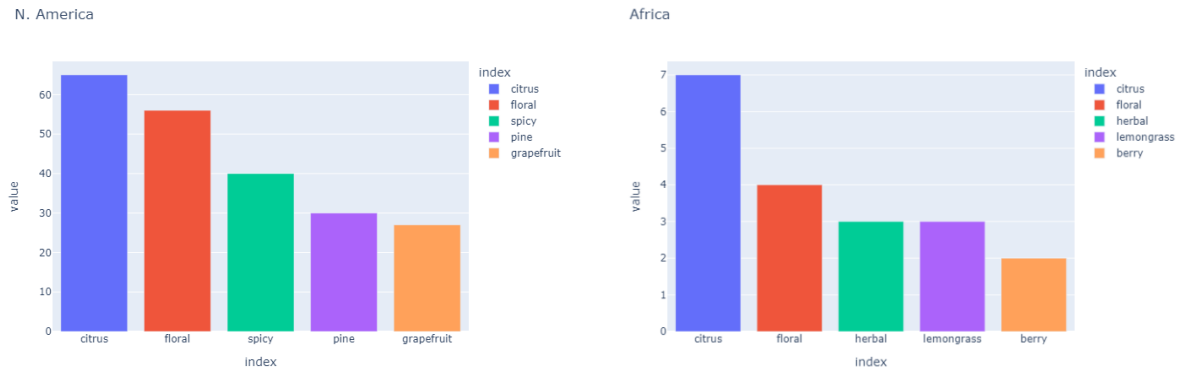
The aroma tags most used to describe Australian and New Zealand hops are citrus, tropical fruit, pine, passion fruit, and floral. This is quite different than the aroma tags most used to describe European hops, which are floral, spicy, citrus, herbal, and fruity. The only tags shared by both continents are citrus and floral. While citrus is the most used in Australia/NZ, it is third in Europe. European hops are most described as floral, which is the fifth most used tag in Australian/NZ hops. Some European hops are described as spicy, herbal, and fruity, which do not appear in the Australia/NZ top five. On the other hand, Australian/NZ hops are described to have passion fruit, tropical fruit, and pine, which are not present in the European top five aromas.



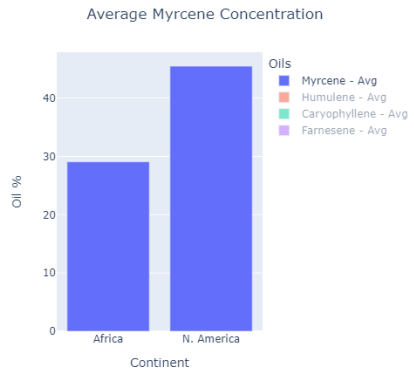
Humulene is mostly associated with woody and pine aromas; however, hops with higher amounts tend to be more floral, herbal, and black pepper(spicy) in character. These tags align much more with the European top five, than the Australia/NZ top five. Thus, it is expected that the humulene percentages in European hops would be greater. The plot to the right presents the

conclusion that based on this data, the average percentages of humulene are in fact greater in European hops.

North America vs. South Africa

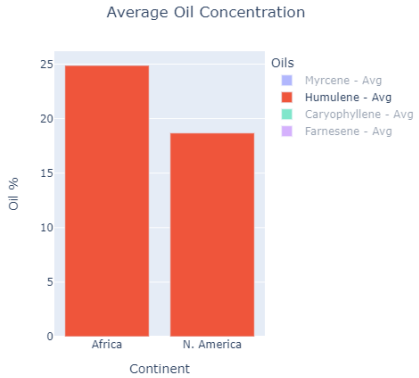


While both North America and South Africa share citrus and floral as their top two, the rest of the tags are different. North American hops have spicy, pine, and grapefruit which are not in the top five of South African hops. On the other hand, South African hops are described to have herbal, lemongrass, and berry aroma tags which are not present in the American top five.



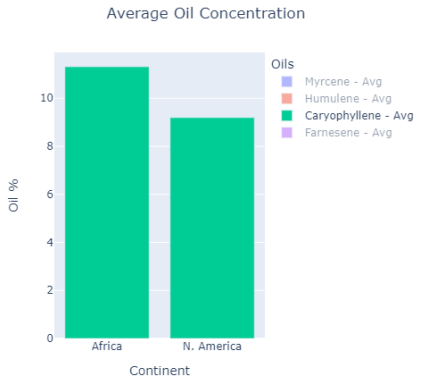
Myrcene is mostly associated with citrusy and resinous-pine aromas. North American hops have pine in their top five while South African do not. The expectation is that North American hops have a higher myrcene percentages as compared to South African hops. Based on this data, they do. It can be observed in the plot to the left.

Higher humulene concentrations lend to floral, herbal, and black pepper(spicy) aromas, which align more so with South African hops than North American hops. The expectation is the South African hops would have higher percentages of humulene. Based on this data, it does. I can be seen in the plot to the right.



Caryophyllene is mostly associated with black pepper, spiciness, and herbal aromas. This is also aligned more with the top five South

African aromas tags. As expected, the plot to the right shows that the average percentages of Caryophyllene are higher in South African hops than North American hops.

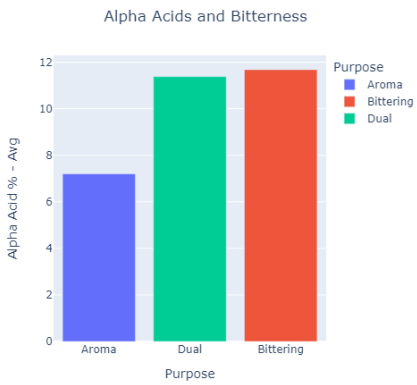


While some continents have a tag or two in common, their blueprint appears to be different. It can be concluded that a relationship exists between the oil concentrations and the aroma tags, and that hops from each continent have some uniqueness to them.

Conclusion 2: Bitterness in Hops

What is the relationship between the brewing values of hops and the brewing purpose?

There exists a relationship between the brewing value alpha acids to the purpose of the hop. The alpha acids in hops are related to the bitterness it provides. The hops in the dataset were labeled with one of three purposes: Aroma, Bittering, and Dual-purpose. The plot to the right presents this relationship.



As expected, the average alpha acid percentages of hops used for Aroma is significantly lower than those used for Bittering with Dual-purpose. The highest average is of Bittering hops.

Conclusion 3: Classification Models

Could a model be developed to accurately predict the origin and purpose of a hop based on the brewing values and oil concentrations?

Performing exploratory analysis helped build better intuition into the various features of the dataset and shed light on how diverse hops can be. Most notably, the shared characteristics of hops among a particular region or purpose showed potentially interesting relationships that could benefit from additional studies. To that extent, machine learning models were explored to delve deeper into these relationships and develop even stronger conclusions to the research questions. Taking this a step further, a predictive tool was created to classify hops in order to output a practical application for this project.

As this is a classification problem, tree-based ensemble algorithms (Random-Forest & XG-Boost) were explored in an effort to predict the region or purpose of a hop. These methods are known for their robustness in the industry,

as both rely on multiple weak learners that are combined into a strong learner. However, bagging methods such as Random-Forest build these models in parallel while boosting methods such as XG-Boost build sequentially. Both techniques are proven to be viable choices for multi-level classification scenarios, as with this study.

While the exploratory analysis gave the proper insight into which attributes to classify, further feature-engineering was required to prepare the data for the model. The outcome variables for both studies (region and purpose) were encoded to transform the categorical variables into appropriate numeric form for the models. Dummy-vectors were created for the predictor categorical variables that included the various aroma tags. Afterwards, this data was partitioned to create a training and testing set to train and evaluate the models. Although the final accuracy rates fluctuated with splitting, both methods seemed to classify hops based on region, or purpose, at an approximate rate of 70%.

It can be concluded that a model could be developed to predict the origin and brewing purpose of a hop; however, additional studies are needed to improve accuracy. Although this was a respectable result, the exploratory work seemed to promise more. There are several reasons that could be contributing to this. Most importantly, although our dataset has a large number of features, the total number of hops is only 304. While this data was scraped from a reliable and extensive hops database, there is an innate issue in the continuous data that is hard to overcome while training these models: the high variability of the brew values. As the database also claims, individual hop measurements can differ substantially depending on many unaccounted factors.

To compensate for this in future studies, exploring statistical methods that are designed for smaller-level studies seems to be an attractive option at this juncture. Either way, modeling the dataset helped understand the data on a deeper level, and output a tool that can be strengthened in the future.

Appendix 1

Source: https://beermaverick.com/the-science-behind-identifying-hop-aromas/

Floral

- Aromas: elderflower, chamomile blossom, lily of the valley, jasmine, apple blossom, rose, geranium, carnation, lily, lilac, lavender, osmanthus.
- Compounds Responsible: rose oxide, geraniol, geraniol acetate, citronellol, neral

Citrus

- Aromas: grapefruit, orange, lime, lemon, bergamot, lemon grass, ginger, tangerine, pomelo
- Compounds Responsible: alpha-terpineol, limonene, linalool, citral, decanal

Tropical/Sweet Fruits

- Aromas: banana, watermelon, honeydew melon, peach, apricot, passion fruit, lychee, dried fruit, plum, pineapple, cherry, kiwi, mango, guava
- Compounds Responsible: 2-methylpropyl hexanoate, ethyl 2-methylpropanoate, sec-amyl acetate, ethyl caproante, ethyl 3-methylbutanoate

Stone/Green Fruits

- Aromas: pear, apple, quince, gooseberry, white wine grapes
- Compounds Responsible: decanal, cis-3-dexenal, d-3-carene, 2-dodecanone, hexyl 2-methyl-propanoate

Berries & Currant

- Aromas: cassis, blueberry, raspberry, blackberry, strawberry, red currant, black currant, wild strawberry, cranberry, mulberry
- Compounds Responsible: beta ionone, 4-mercapto-4-methylpentan-2-one, ethyl 3-methylbutanoate, raspberry ketone, p-metha-8-thiol-3-one

Cream & Caramel

- Aromas: butter, chocolate, yogurt, honey, cream, caramel, toffee, coffee, tonka bean, vanilla, coconut
- Compounds Responsible: methyl decanoate, Y-nonalactone, vanillin, phenylacetic acid

Woody Aromatic

- Aromas: tobacco, cognac, barrique, leather, woodruff, incense, myrrh, resin, cedar, pine, earthy
- Compounds Responsible: humulene, alpha-pinene, beta-pinene, farnesene, carvacrol, beta-caryophyllene

Menthol

- Aromas: mint, lemon balm, sage, camphor, menthol, wine yeast, eucalyptus
- Compounds Responsible: carvone, terpinen-4-ol, camphene

Herbal

- Aromas: marjoram, tarragon, dill, parsley, basil, fennel, cilantro, rosemary, thyme, green tea, black tea, mate tea, oregano
- Compounds Responsible: myrcene, humulene, epoxide, p-cymene, cis-b-ocimene, thymol

Spicy

- Aromas: lovage, pepper, chili, curry, juniper, aniseed, licorice, fennel seeds, clove, cinnamon, gingerbread, coriander seeds, nutmeg
- Compounds Responsible: beta-caryophyllene, eugenol, 2-isopropyl-3-methoxypyrazine, beta-eudesmol

Grassy

- Aromas: fresh cut grass, hay, tomato leaves, green pepper, nettle, cucumber, bamboo leaves
- Compounds Responsible: E, Z-2, 6-nonadienal, cis-3-hexenol, trans-2-hexenal

Vegetal

- Aromas: celery, leek, onion, artichoke, garlic, wild garlic, radish
- Compounds Responsible: diallyl sulphide, dimethyl disulfide, s-methylthiohexanoate

Sources

- Beer Maverick, https://beermaverick.com/
- Wikipedia, “Hops”, https://en.wikipedia.org/wiki/Hops
- BarthHaas Group, “HOPSESSED® - The language of hop flavors” https://www.barthhaas.com/en/world-of-flavor/hopsessed
- WINNING-HOMEBREW, https://winning-homebrew.com/noble-hops.html