US Beer Production Visualisation

Name: ***Om Anish Dalal***

# Introduction

The objective is to investigate the US Beer Production Dataset (2008-2019), analyze the data and displaying in a variety of different visualisations. The US Beer Production data comes originally from Alcohol and Tobacco Tax and Trade Bureau (TTB) **[1]**. The data on the website is given in PDF’s and spreadsheets which is harder to extract data into a csv file as a whole.

The data was extracted into 4 different files and uploaded onto github to do the analysis. The process of learning how to extract data from PDFs is a bit complex with some R libraries available.

**Dataset Source and variables information:** https://github.com/rfordatascience/tidytuesday/tree/master/data/2020/2020-03-31

The dataset consists of 4 files:

[beer\_states.csv](https://github.com/rfordatascience/tidytuesday/blob/master/data/2020/2020-03-31/beer_states.csv)  
beer\_taxed.csv  
brewer\_size.csv  
brewing\_materials.csv

This report has the analysis of 3 of the files provided excluding the beer\_taxed.csv file. Analysis on the beer\_taxed.csv was not done as it was hard to understand what data was presented based on the domain knowledge.

# Beer States Dataset (beer\_states.csv)

This file contains the data of total barrels produced and type of production for years 2008-2019 per state. This set has 1,872 observations.

# Brewer Size Dataset (brewer\_size.csv)

This file contains the data of the number of brewers, total barrels, taxable removals, and total shipped based on year and brewer size. This set has 137 observations.

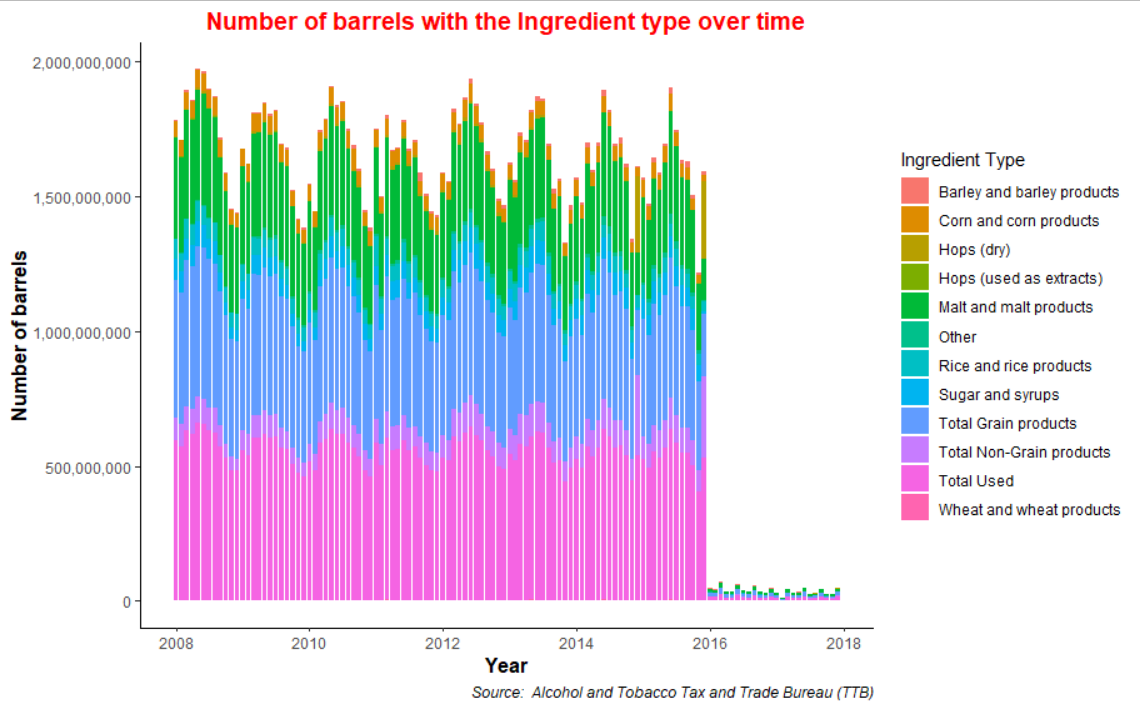
# Brewing Materials Dataset (brewing\_materials.csv)

This file contains the data of the material type and amount of brewing materials used in production per year. This set has 1440 observations.

# Analysis on the Brewing Materials Data set

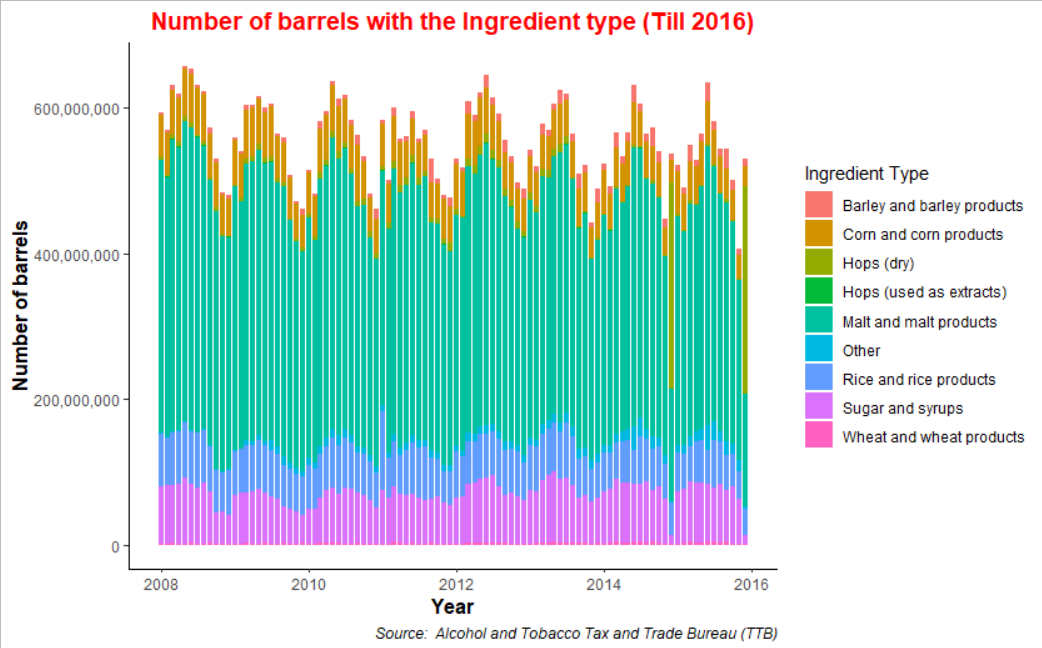
# Number of Barrels with the Ingredient Type over Time

Let’s have a look at the current number of barrels for this year/month for each Ingredient type over the years. To look at the changes over time in graph, year and month data have been combined into one variable with 1st date of every month. This was done using the 'ymd' function from the 'lubridate' library to convert the string into Date format.



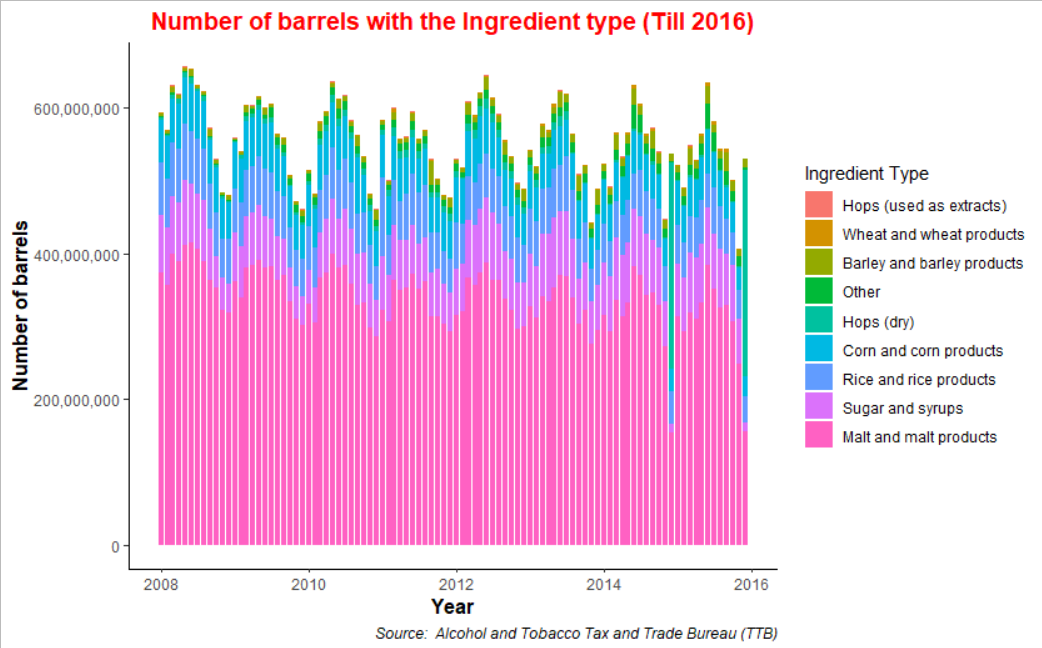
For all the graphs, formatting the theme and labels, the reference was taken from **[2]**. Reference for formatting the continuous numeric values with commas e.g., 2,00,000 was taken from **[3]**.

In the above column graph, ‘Total Used’ and ‘Total Grain Products’ takes up a lot of space in the graph and it’s hard to determine the distribution of the other materials. Also, it seems like the data from year 2016 is not available in the 'brewing\_materials.csv' dataset for some reason.

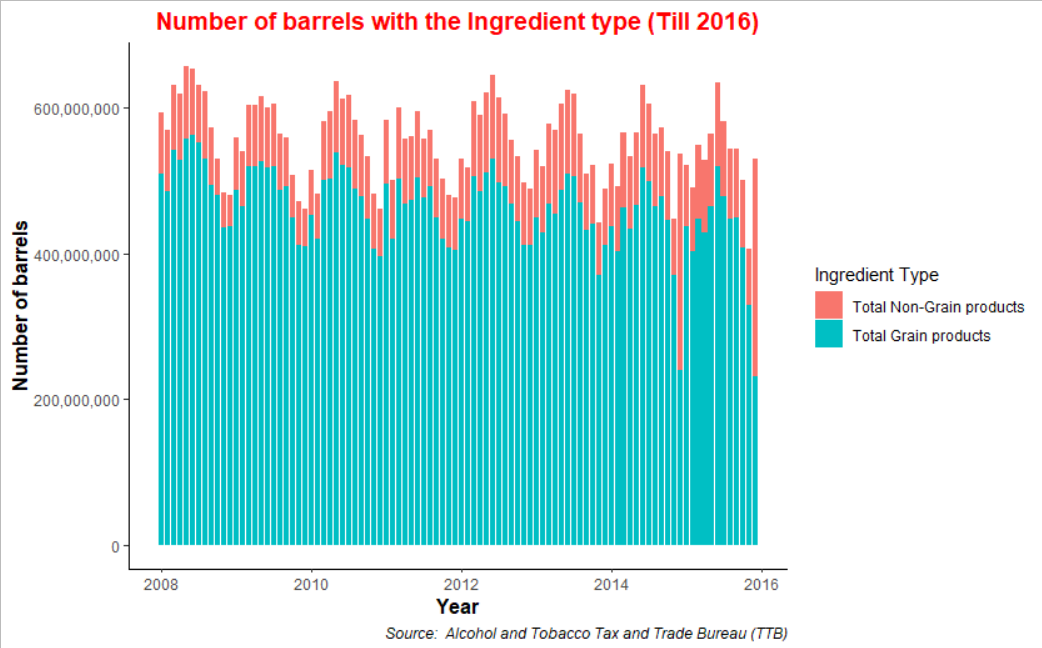
We filter out the years after 2016 from the 'brewing\_materials.csv' dataset. Let’s look at the above graph after filtering out the 'Total Non-Grain Products, Total Grain products and Total Used' types. We used filter function and remove the data containing the word ‘total’ from the ‘material\_type’ variable using the ‘str\_detect’ function.

It is evident that ‘Malt and malt products’ were the most used ingredient in the beer production followed by ‘Sugar and syrups’. Something seems off with some month data for the year 2015 as the ‘Malt and malt products’ data is not visible.

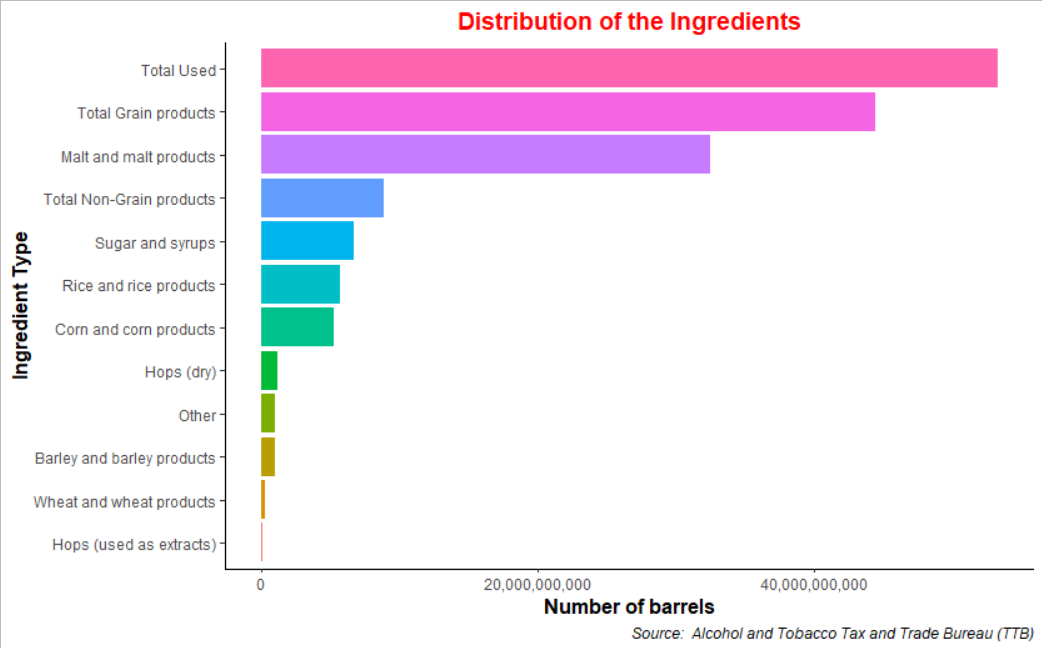
Reordering the 'type' based on the values of 'month\_current' value, below graph shows that ‘Malt and malt products’ are the most used ingredients in the beer production. ‘reorder’ function was used.



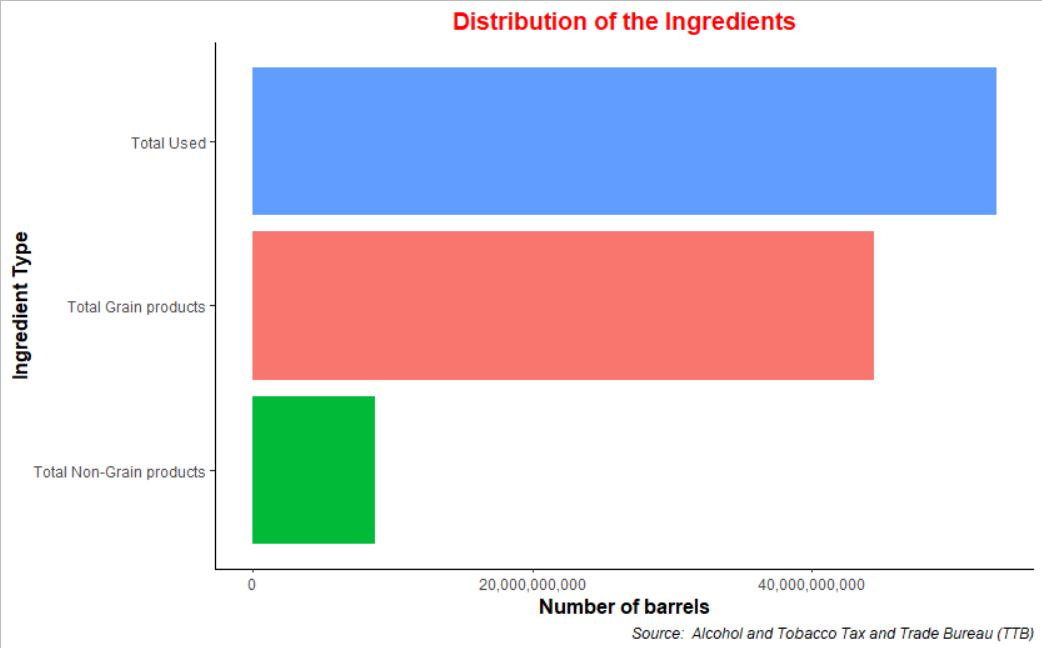
Considering only the "Total Non-Grain products" and "Total Grain products", below graph shows Grain products were dominant in the beer production process.



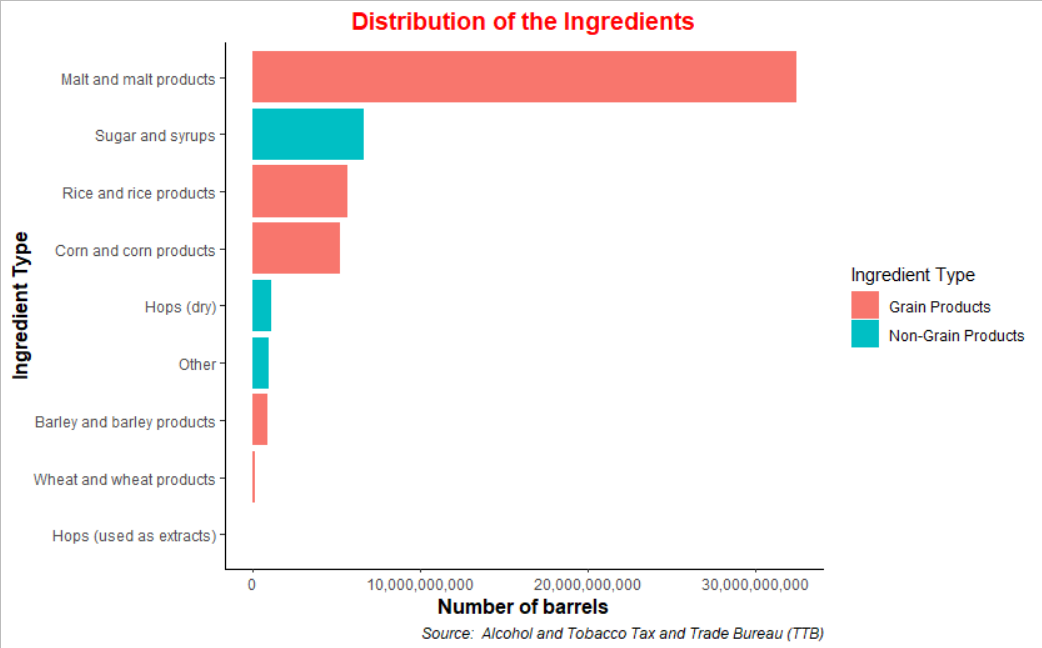
# Distribution of the Ingredients used in Beer Production

Below graph gives the clear picture of the distribution of the Ingredients used in the beer production. Reordering the 'type' based on the values of 'month\_current' value, ‘Malt and malt products’ are the most used ingredients in the beer production.

In the distribution of Ingredients considering only the "Total Non-Grain products" and "Total Grain products" below, ‘Total Grain Products’ are mostly used in the beer production.

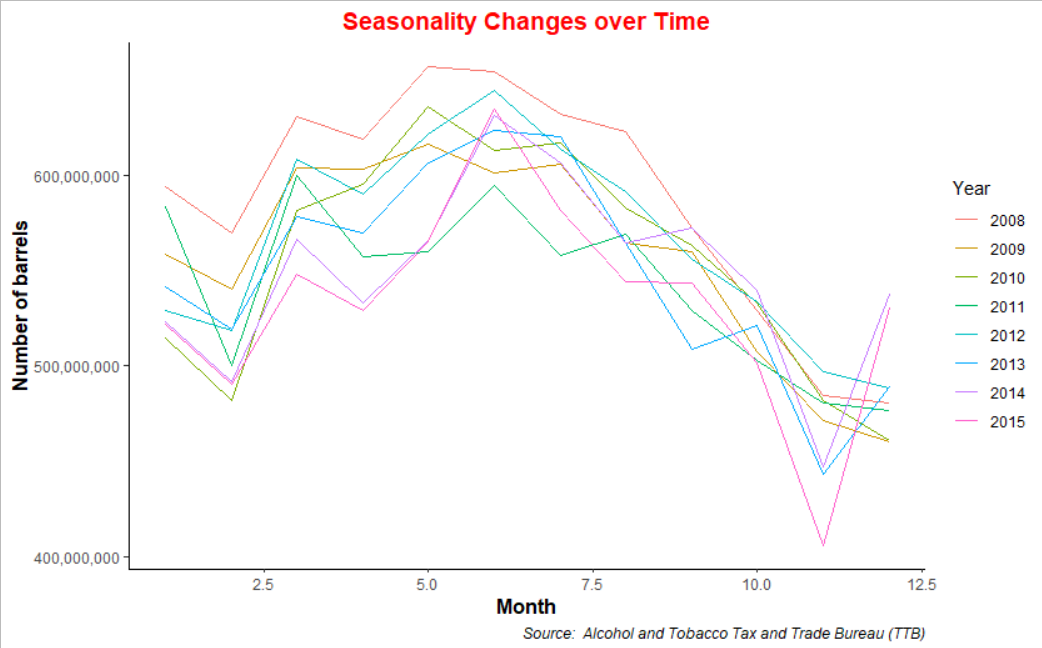


Filtering out the 'Total Non-Grain Products, Total Grain products and Total Used' types, ‘Rice and rice products’ are contribute more after malt products for the ‘Grain Products’ type.



# Seasonality over the years in the number of barrels

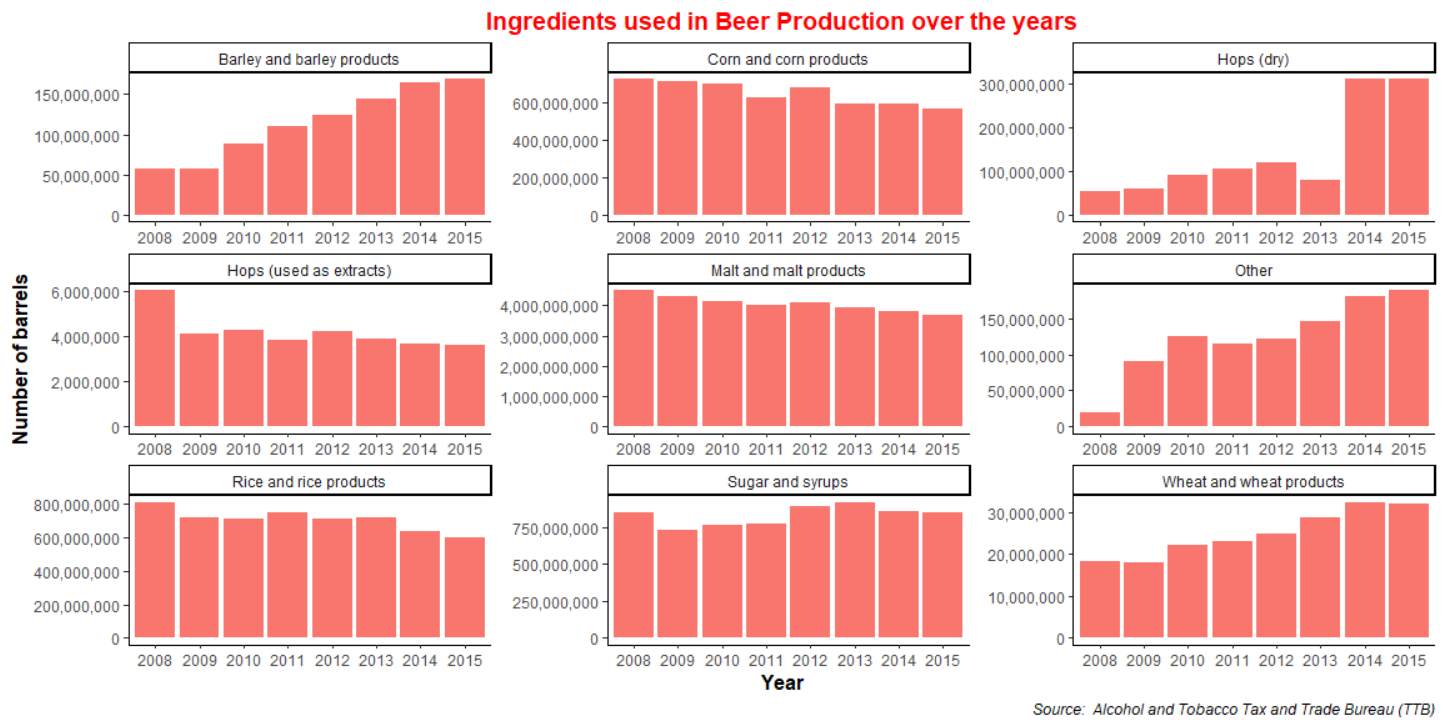
Below shows the Line Trend Chart to see the seasonality over the years in the number of barrels considering 'Total Used' material type. A sharp downward trend is visible for the month of November over the years in the number of barrels produced. Higher number of barrels are produced in the summer.



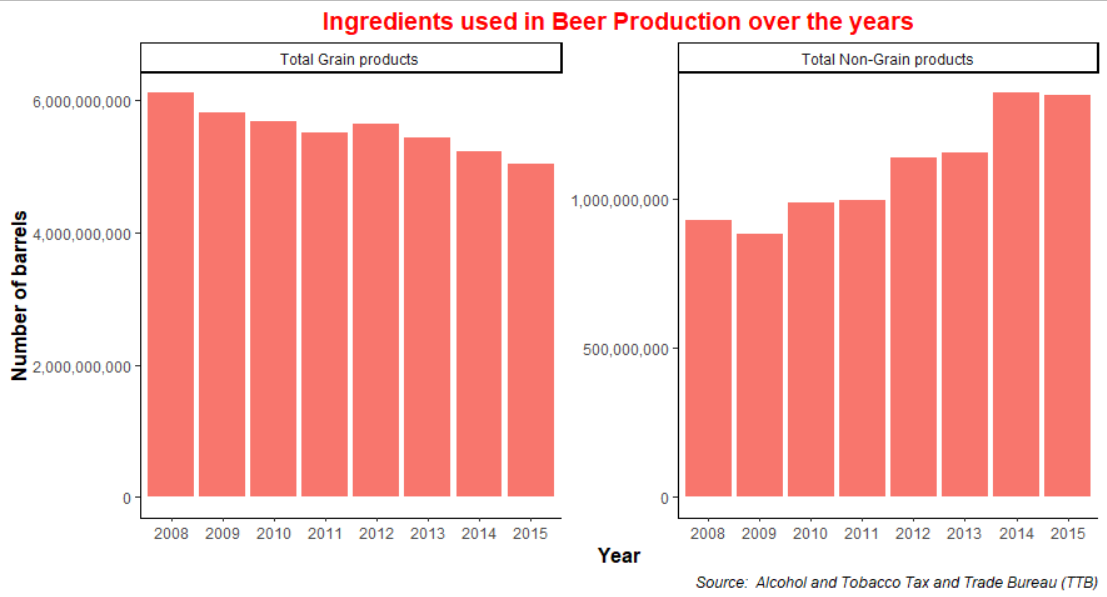
# Facet view of the Ingredient types over the years

Below graph shows the individual ingredients used in the production over the years. scales = "free" is included in the facet\_wrap function to not have the fixed scaling for the number of barrels.

‘Barley and barley products’ in the production have increased linearly over the years. ‘Wheat and wheat products’ follow the same trend. ‘Hops(dry)’ usage increased significantly in the year 2014 and 2015.



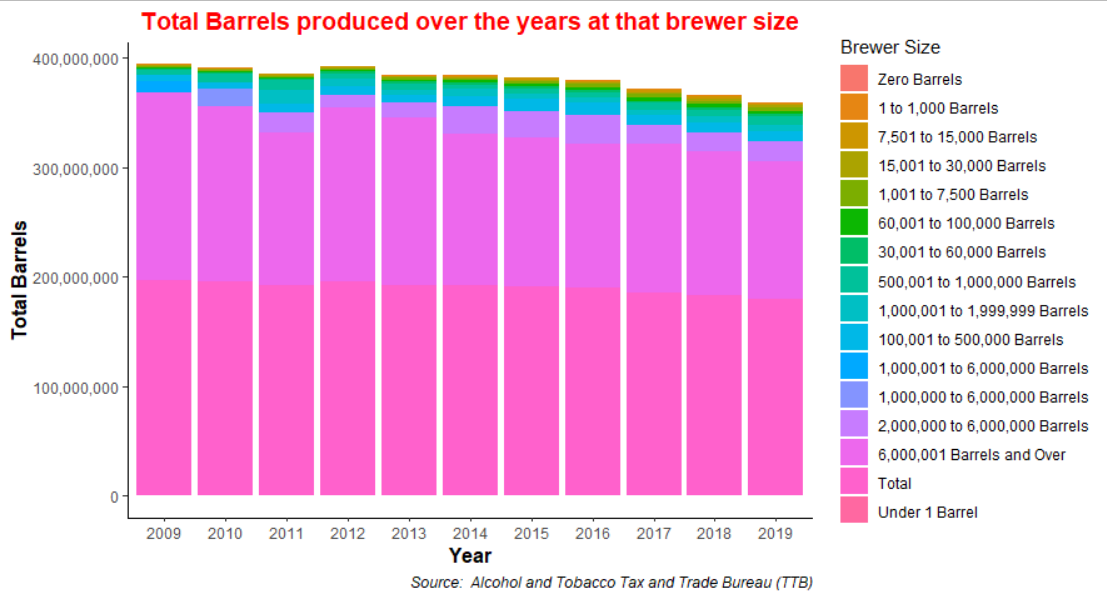
Filtering out the 'Total Non-Grain Products and Total Grain products' types, ‘Grain products’ in the production has been decreasing over the years while Non-Grain products have been increasing but the number of barrels are also to be considered as Grain Products produced are in much larger quantity.



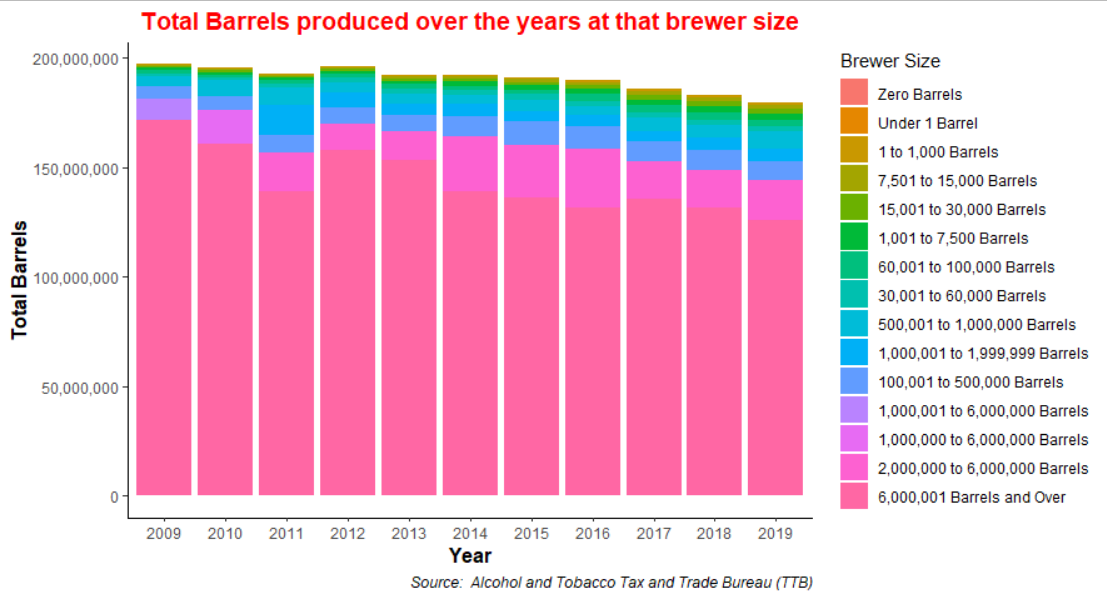
# Analysis on the Brewer Size Data set

This file contains the data of the number of brewers, total barrels, taxable removals, and total shipped based on year and brewer size.

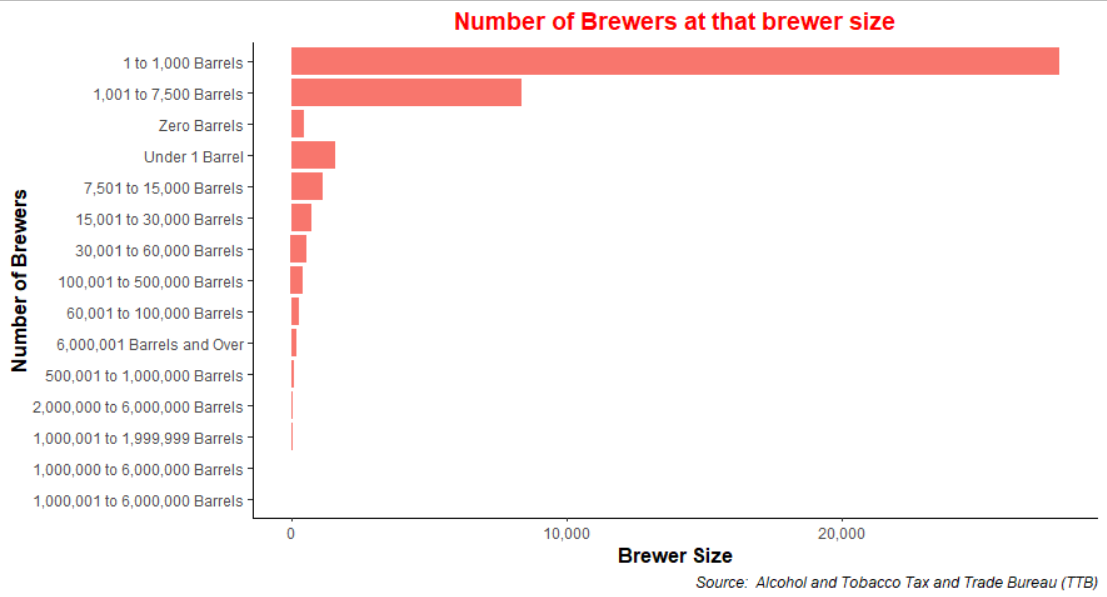
# Total Barrels produced over the years at that brewer size

Brewer sizes with ‘6000001 Barrels and Over’ produced the most barrels.

After removing 'Total' value from the 'brewer\_size' variable, removing 'NA' values from 'total\_barrels' variable (there is one record with ‘NA’ value) and reordering the 'brewer\_size' based on the values of 'total\_barrels' value, below graph is produced.



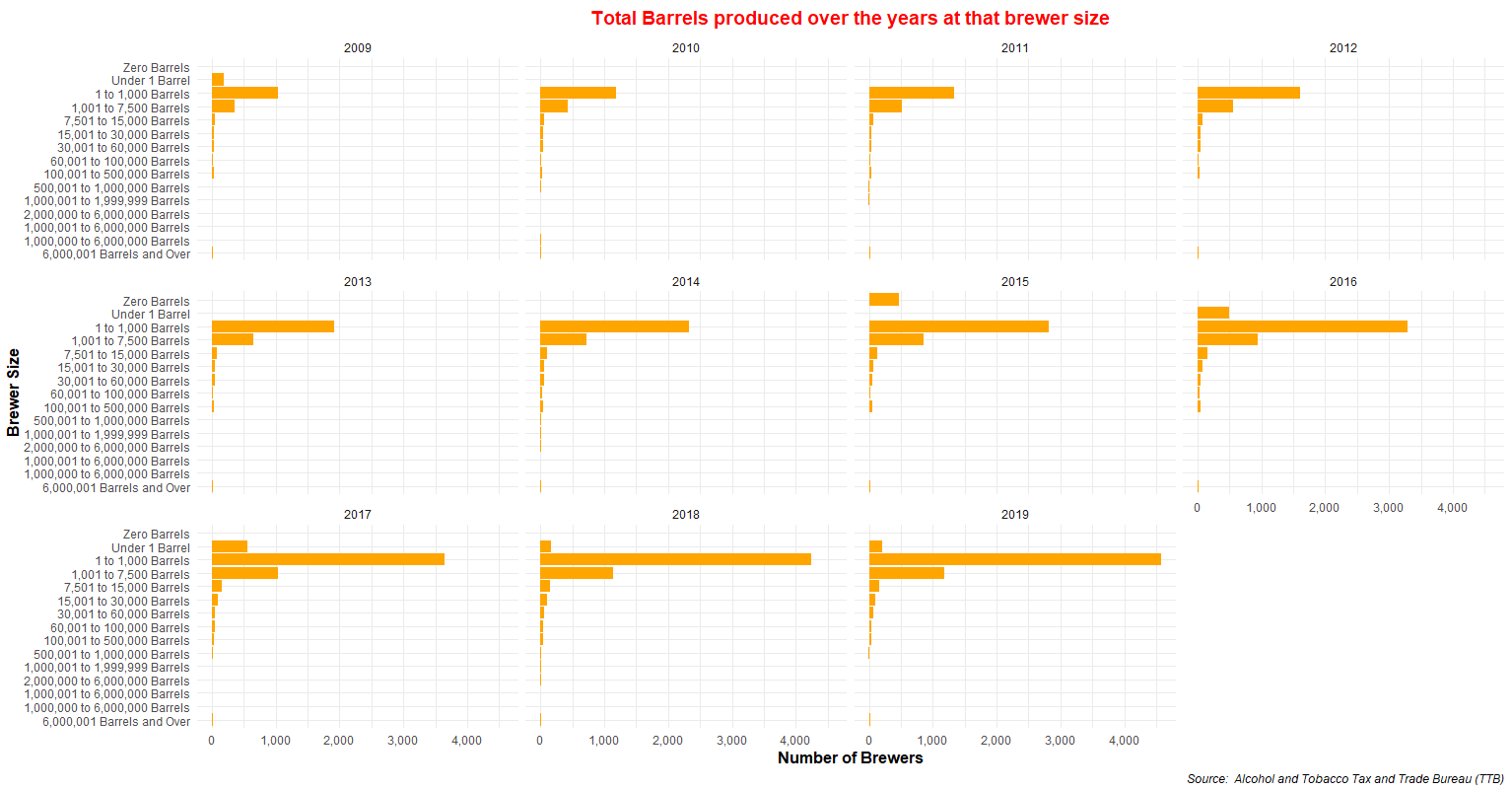
Let’s have a look at the number of brewers at that brewer size. Highest number of brewers produce ‘1 to 1000 Barrels’. There are 174 brewers producing ‘6000001 Barrels and Over’.



# Total Barrels produced over the years at that brewer size **[4]**

Concept Reference **[4]**

Grouping by year and ranking using row\_number() to get the data in descending order and applying 'facet\_wrap' on 'year' to show the brewer sizes through various years, we can see that brewers producing ‘1 to 1000 Barrels’ increased significantly over the years followed by brewers producing ‘1001 to 7500 Barrels’.



Animating Total Barrels produced over the years at that brewer size using 'gganimate' library.

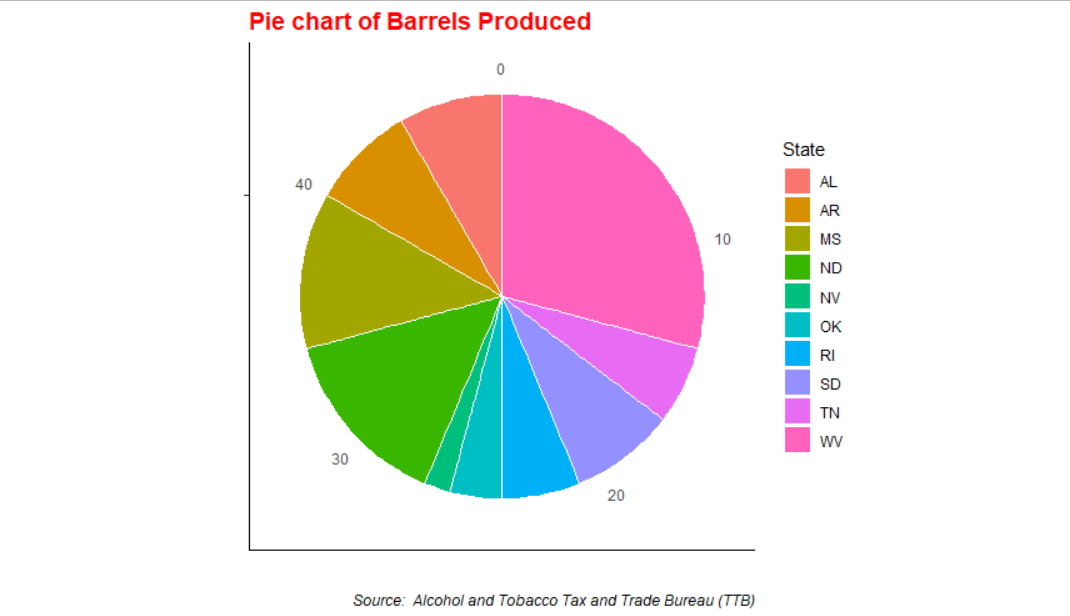
**Please open this gif**



# Analysis on the Beer States Data set

# Pie Chart of the States ranked by Barrels produced

Ranking was not effective here as I gave rank <=4 but 10 states were visible. But the data in the pie chart reflects correctly the barrels produced per state. West Virginia (WV) has produced the most barrels.



# Beer Produced by State **[5]**

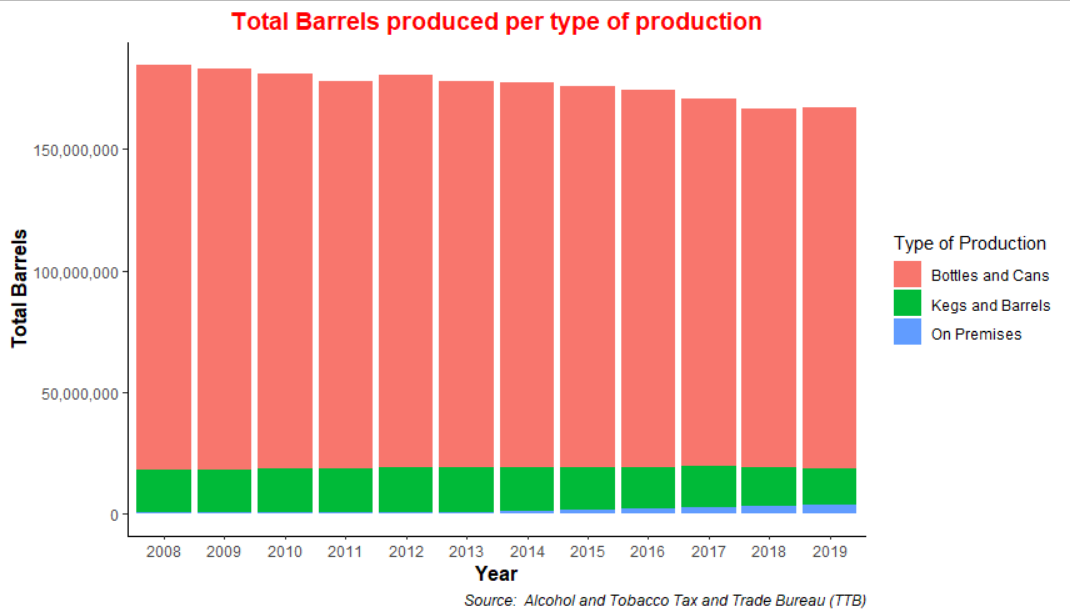
Concept Reference **[5]**

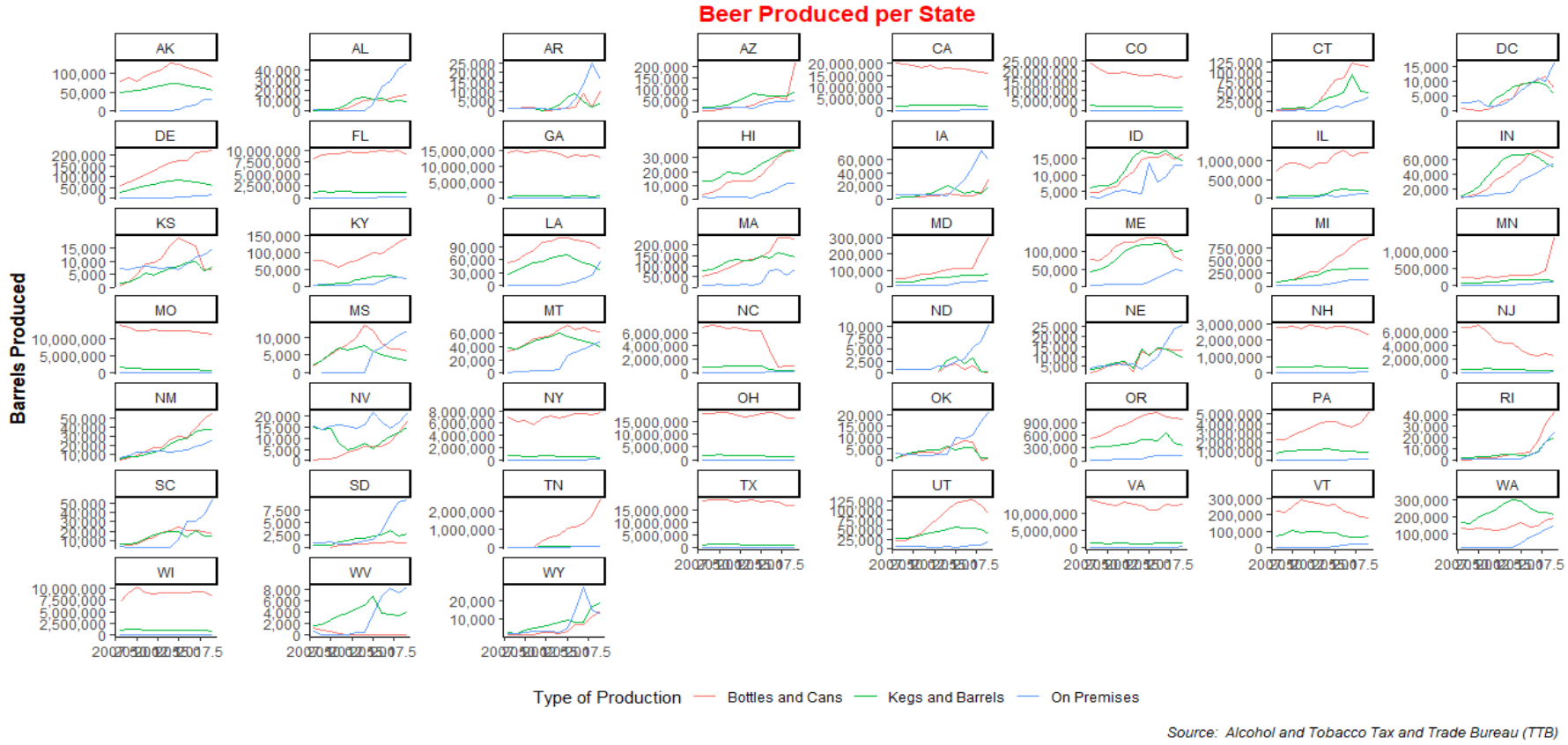
I was trying to achieve the animated graph produced in the link above but was not successful enough. Reordering was not successful based on the increasing values of Total Barrels produced to the top.

**Please open this gif**



# Beer Produced by each State showing the trend by type of production

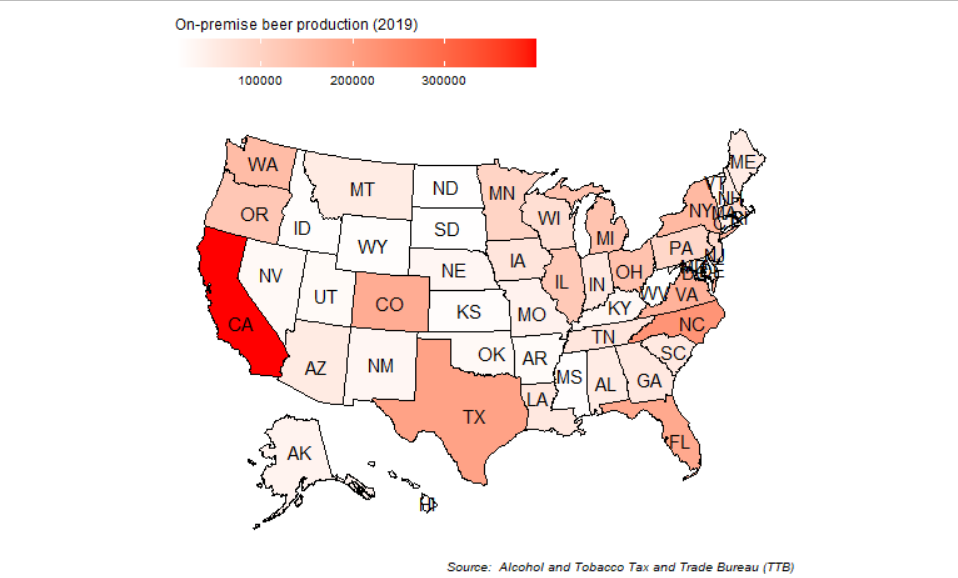
Column graph showing the Total Barrels produced per type of production over the years. Filtering the 'total' value from the 'state' variable, ‘Bottles and Cans’ type of production had most barrels.

Applying 'facet\_wrap' on 'state' to show the Total Barrels produced by each state per type of production. California (CA) and Colorado (CO) are heavy producing states and we can see the decrease in ‘Bottles and Cans’ production over years as other states could be increasing the production. Tennessee (TN) started producing more ‘Bottles and Cans’

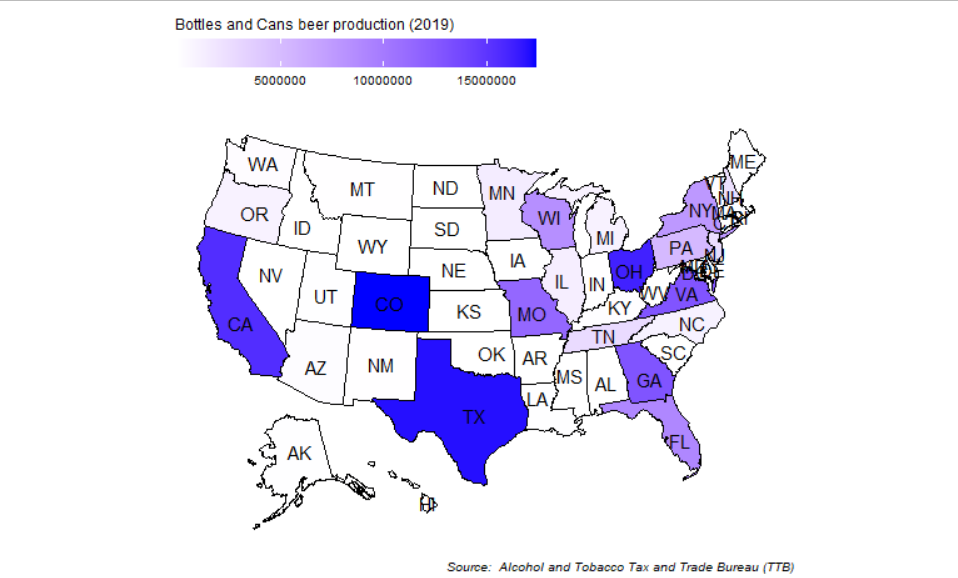
**Bigger size image**



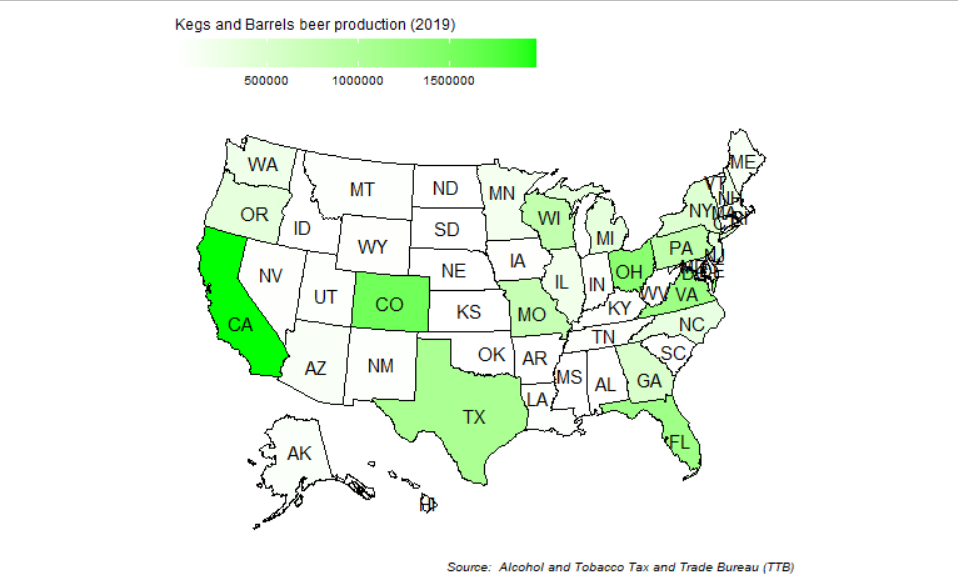
# Beer Produced by each State showing the trend by type of production

'On-Premise' beer production type for the year 2019. California (CA) has the highest production On-Premise.

Bottles and Cans' beer production type for the year 2019. CA, TX, CO, OH States have higher production of Bottles and Cans.



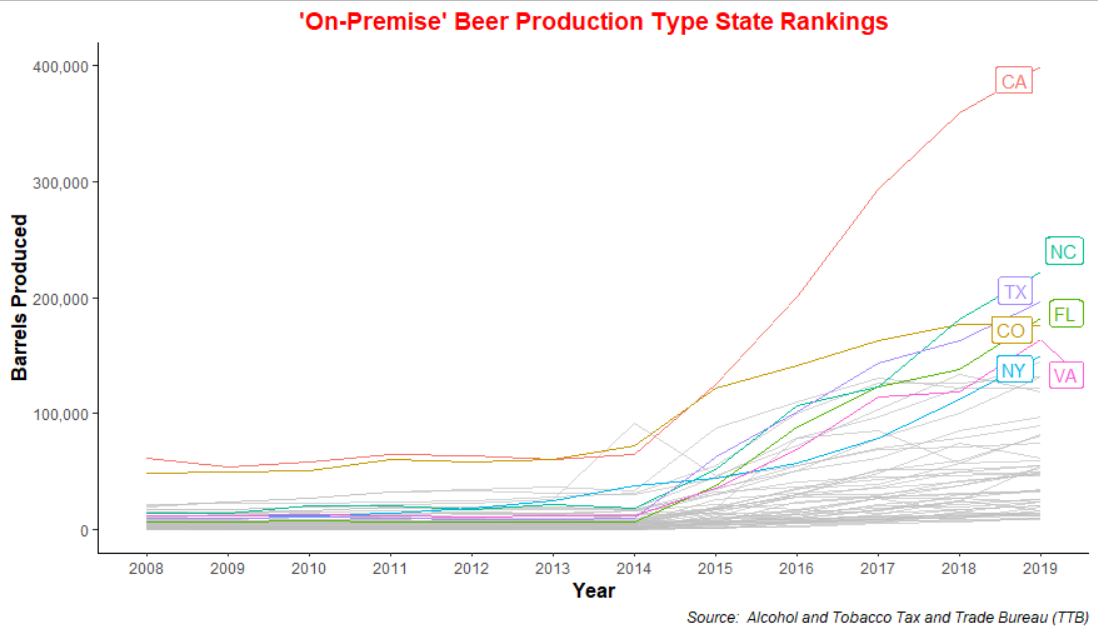
'Kegs and Barrels' beer production type for the year 2019. CA, CO has higher production of Kegs and Barrels.



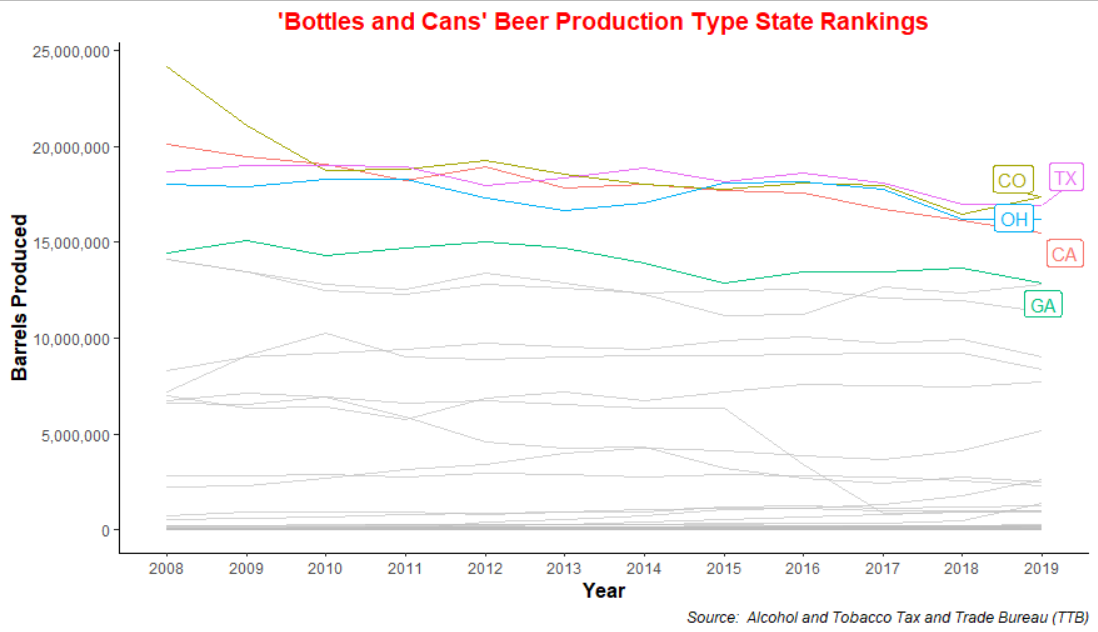
# Beer Production in Top US States ranking per each type of production over the years

In the below graphs, ‘gghighlight’ library is used to highlight particular data above a value. **[6]**

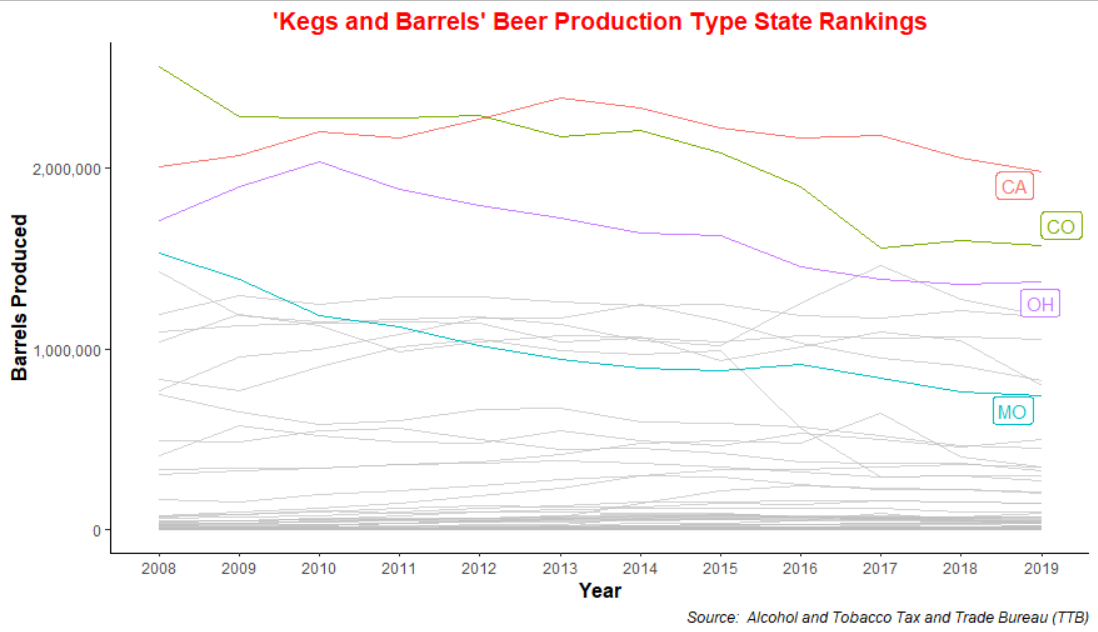
'On-Premise' beer production type. CA has higher production followed by NC, TX States.



'Bottles and Cans' beer production type

'Bottles and Cans' beer production type. CA, TX, CO, OH States have higher production of Bottles and Cans.

'Kegs and Barrels' beer production type. CA, CO, OH has higher production of Kegs and Barrels.



# References

Dataset Source: <https://github.com/rfordatascience/tidytuesday/tree/master/data/2020/2020-03-31>

[1] “Alcohol and Tobacco Tax and Trade Bureau (TTB)”

<https://www.ttb.gov/beer/statistics>

[2] Formatting the graph (theme and labels)

<https://viz-ggplot2.rsquaredacademy.com/labels.html>  
  
[3] Formatting the continuous numeric values with commas e.g., 2,00,000

<https://stackoverflow.com/questions/37713351/formatting-ggplot2-axis-labels-with-commas-and-k-mm-if-i-already-have-a-y-sc>

[4] Reference:

<https://rww.science/2020/04/01/a-quick-tidytuesday-on-beer-breweries-and-ingredients/>

[5] Reference:

<https://www.christopheryee.org/blog/tidytuesday-beer-production/>

[6] Reference for ‘gghighlight’ library

https://cran.r-project.org/web/packages/gghighlight/vignettes/gghighlight.html