Ericka Houle
Alisha Hill
Diana Melendez
October 11, 2021

Lady Wine Geeks
Project 3 - Group 1
METHODOLOGY

INSPIRATION & REASONING

Our inspiration comes from the allure of Napa Valley and the desire to create a vacation atmosphere in the comfort of your own home. By analyzing the different varieties of wines based on wine reviews from our dataset we aim to identify a highly rated wine by province in a given country. The data was retrieved from Wine Enthusiast Magazine (June, 2017, https://www.winemag.com). This dataset will allow for an analysis of country, province, winery, type, variety and points on review.

DESIGN CHOICES

Our dashboard includes a vertical navigation bar on the left side of the screen to toggle between the different pages which include our overview, "about us" page, main visualizations, final report and works cited. The color scheme is a reference to wine type; the color merlot for red wine and the color champagne for white wine. We chose images related to wine throughout our presentation and website.

DATA SOURCES

Our goal was to identify a data source that provided locations and values. Having one or the other would make getting our data more complex. Our main dataset is based on Wine Reviews from a Wine Enthusiast Magazine web scrape referenced in this document. Along with our main dataset our group also obtained previously formatted data from a 2019 SMU Data Science Boot Camp. We retrieved this formatted data from the referenced project in GitHub.

HYPOTHESES

Going into this project we predicted that white wine from California would be the most popular type of wine. This is because we had thought that white wine as it can be sweeter, less dry and easier on the pallet. To further explore our data set we wanted to answer the following questions:

- What wine type is the most popular, red or white wine?
- Which region has the highest production?
- What region will have the highest point value based on the reviews?

CLEANING DATA

We loaded the main dataset with 150 thousand customer reviews. Columns were dropped and renamed to get the data as flat as possible. We dropped any and all entries with null values. This action decreased our data set about 20%. However, we felt this did not have an adversary effect on our analysis.

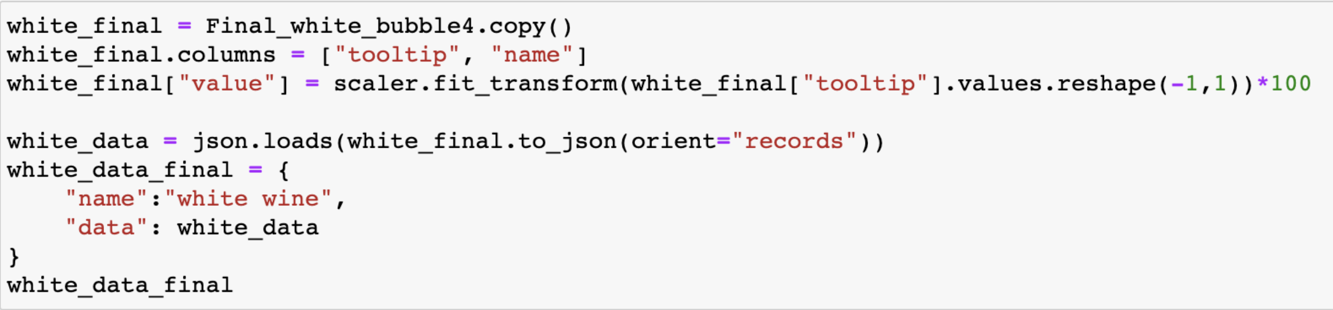
The next step was to read in a data set retrieved from our additional data source. The Wine_Reds.csv was used to merge with our data set to fill in the wine types according to their research. Additional columns were renamed and dropped to keep the data flatten. Code was written to convert the “type” column from a “boolean” true and false to a “string” red wine and white wine.

After which, one last merge took place. This merge came from the additional datasource of Red_Province.csv used to add the lat and longs to our main dataframe. This action reduced time in having to scrape them ourselves.

DATA TRANSFORMATION

During our research into the visualizations we wanted to create we learned that each library we would use would require the data to be formatted differently. This proved to have the greatest challenge when it came to the Highcharts “Split Packed Bubble Chart” and the Observable “Sunburst Chart”.

We decided to first tackle the “Split Packed Bubble Chart” we needed to create a dictionary that contained the “name” which was the type of wine and “data” which consisted of another dictionary. The data dictionary consisted of “name” which was the province and “value” which was the average point value in that particular province.

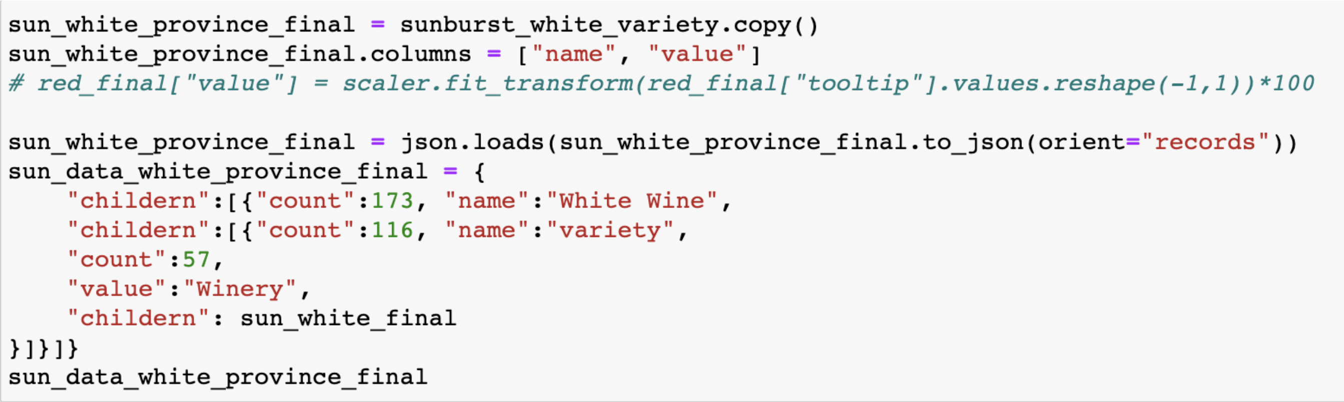


```
white_final = Final_white_bubble4.copy()
white_final.columns = ["tooltip", "name"]
white_final["value"] = scaler.fit_transform(white_final["tooltip"].values.reshape(-1,1))*100

white_data = json.loads(white_final.to_json(orient="records"))
white_data_final = {
    "name":"white wine",
    "data": white_data
}
white_data_final
```

To avoid having the “Split Packed Bubble Chart” bubbles the same exact size, additional statistical coding was used to manipulate the data. We scaled the value column to normalize between min and max. The Min Max Scaler took our Z-score and base in a positive range to eliminate the negative z-score values. This action resulted, a noticeable size difference in closely related data points.

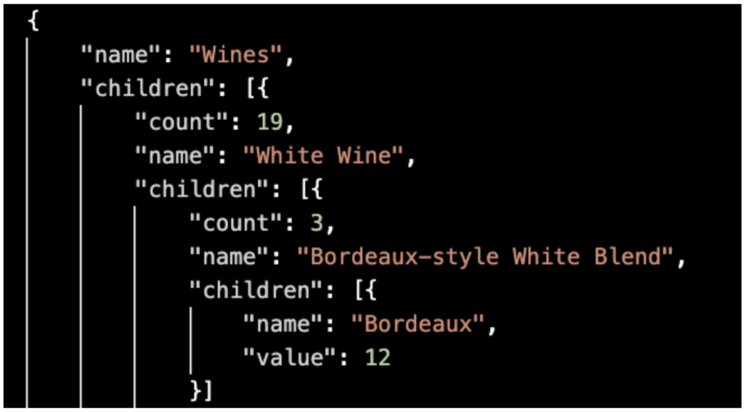
When it came to the “Sunburst Chart” we took the same approach to format the data as we used on the bubble chart. This would format the children dictionary into “count” which was the number of wineries in that province and “name” of the province. However upon viewing the visualization it was noticed that this only provided one level of detail which was the outer ring.



```
sun_white_province_final = sunburst_white_variety.copy()
sun_white_province_final.columns = ["name", "value"]
# red_final["value"] = scaler.fit_transform(red_final["tooltip"].values.reshape(-1,1))*100

sun_white_province_final = json.loads(sun_white_province_final.to_json(orient="records"))
sun_data_white_province_final = {
    "childern":[{"count":173, "name":"White Wine",
    "childern":[{"count":116, "name":"variety",
    "count":57,
    "value":"Winery",
    "childern": sun_white_final
}]}]}
sun_data_white_province_final
```

This allowed us to see what the final output would need to look like in order to differentiate the data at each level of the sunburst chart. In order to save time we had decided to manually make these updates in the csv file due to the time constraints we had on this project.



```
{
    "name": "Wines",
    "children": [{
        "count": 19,
        "name": "White Wine",
        "children": [{
            "count": 3,
            "name": "Bordeaux-style White Blend",
            "children": [{
                "name": "Bordeaux",
                "value": 12
            }]
```

CONCLUSIONS

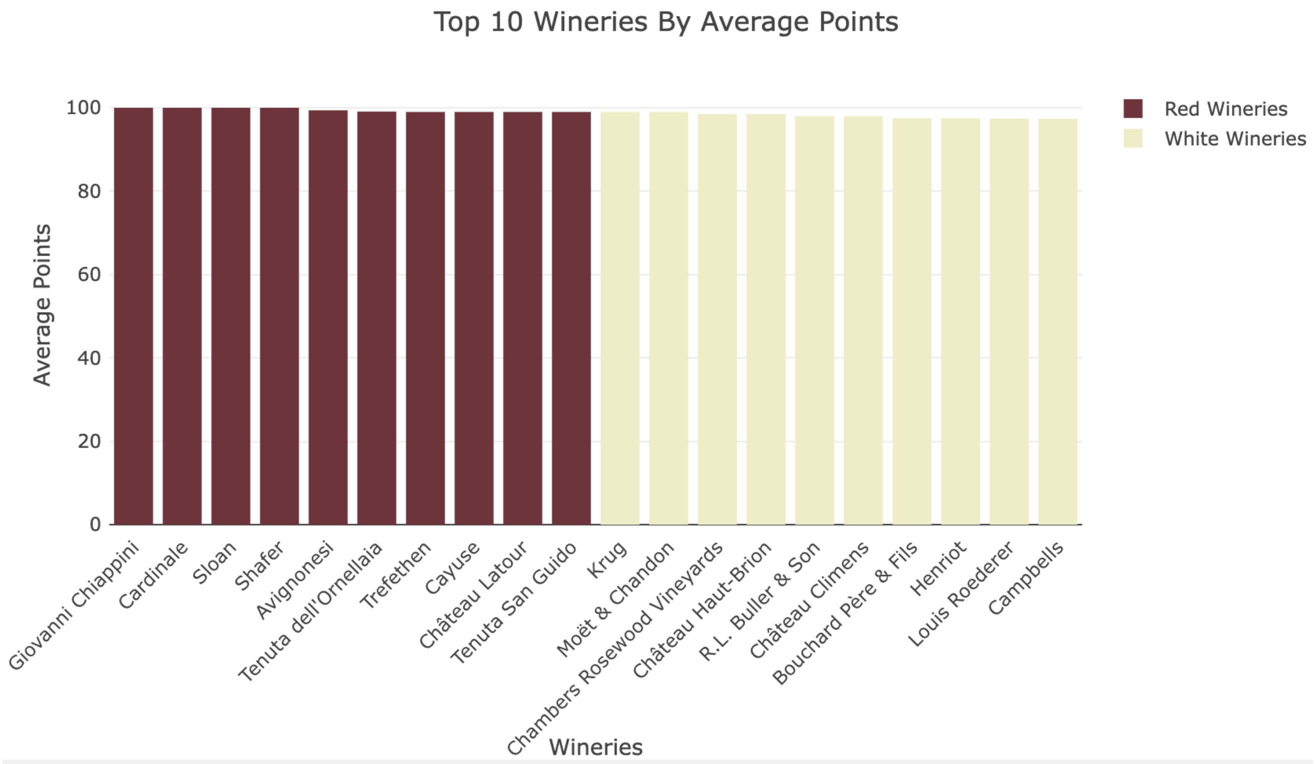
We determined that when it comes to wine enthusiasts most prefer red wine over white wine. We were able to identify the top varieties of both red and white wines.
Chardonnay is the most popular white wine variety which is primarily produced in both Burgundy and California. For the top red wine we found it to be Syrah which is produced in Washington, Tuscany and South Australia.

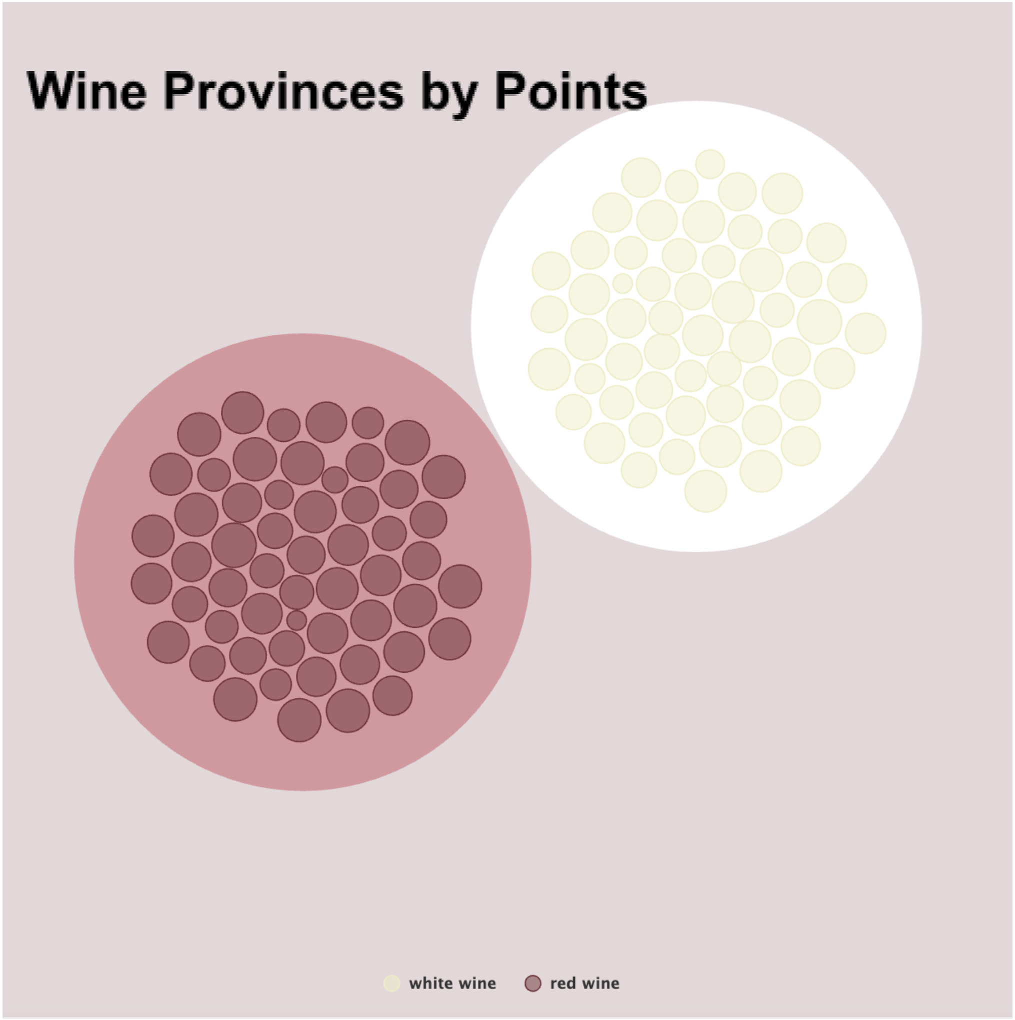
LIMITATIONS

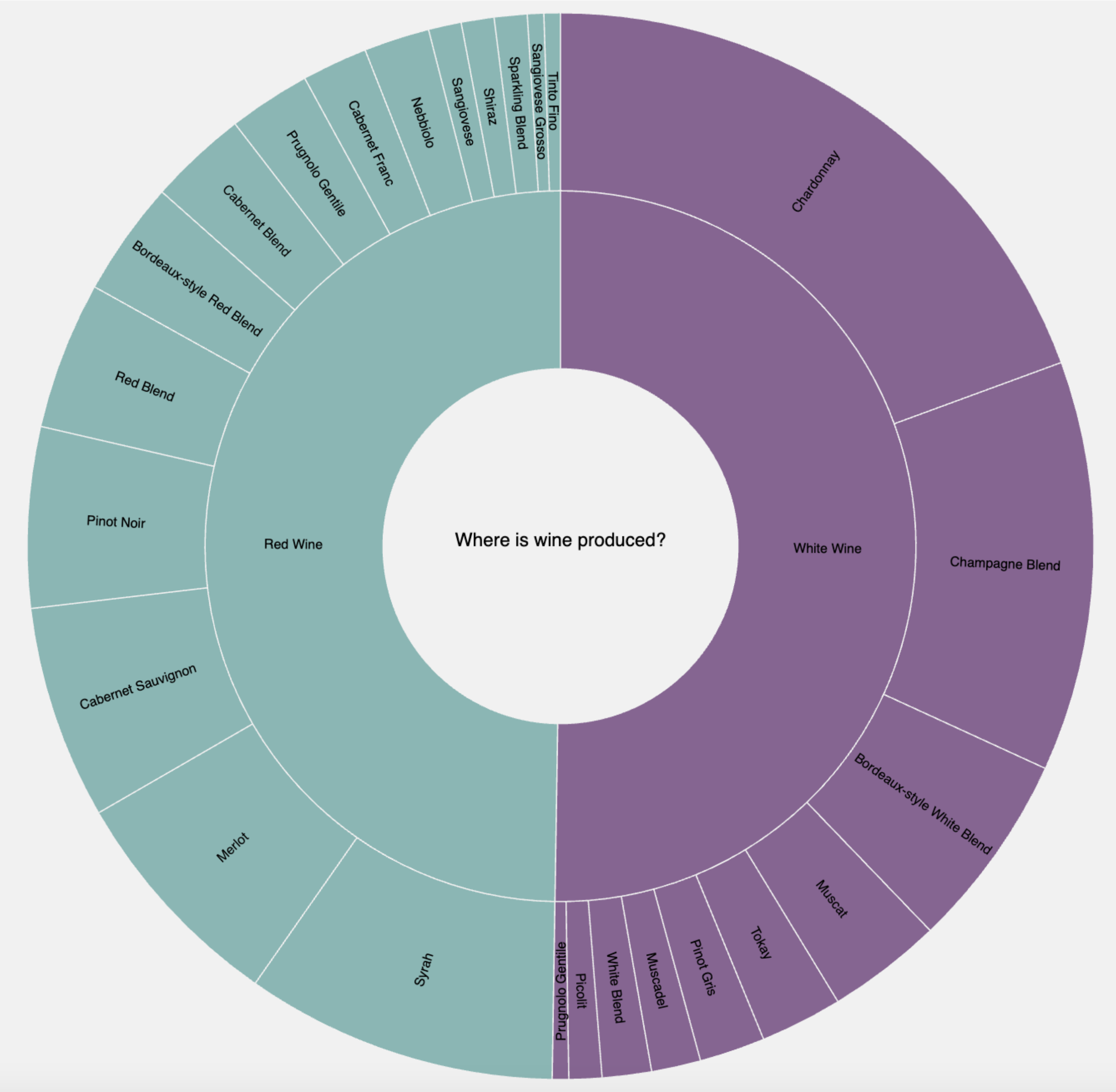
The data was limited to 2017 from a web scraping of the site “Wine Enthusiast.” Our focus was on the top 100 wines based on the points accumulated by customer reviews. A limitation in the data is there isn’t an easily identifiable unique identifier. The unique identifier was a long description of the wine. Not ideal when creating visualizations. Each visualization required specific data formatting, in turn rendering too much time on data formation. This timely process took away from the actual interactive dashboard.

FUTURE WORK

Future work would ideally include an analysis of the popularity of varieties in California due to the state’s high production of wine. Additionally, examine the popularity of organic wines in California. Equally important, scraping the parent website to gather the actual wine name instead of the description.

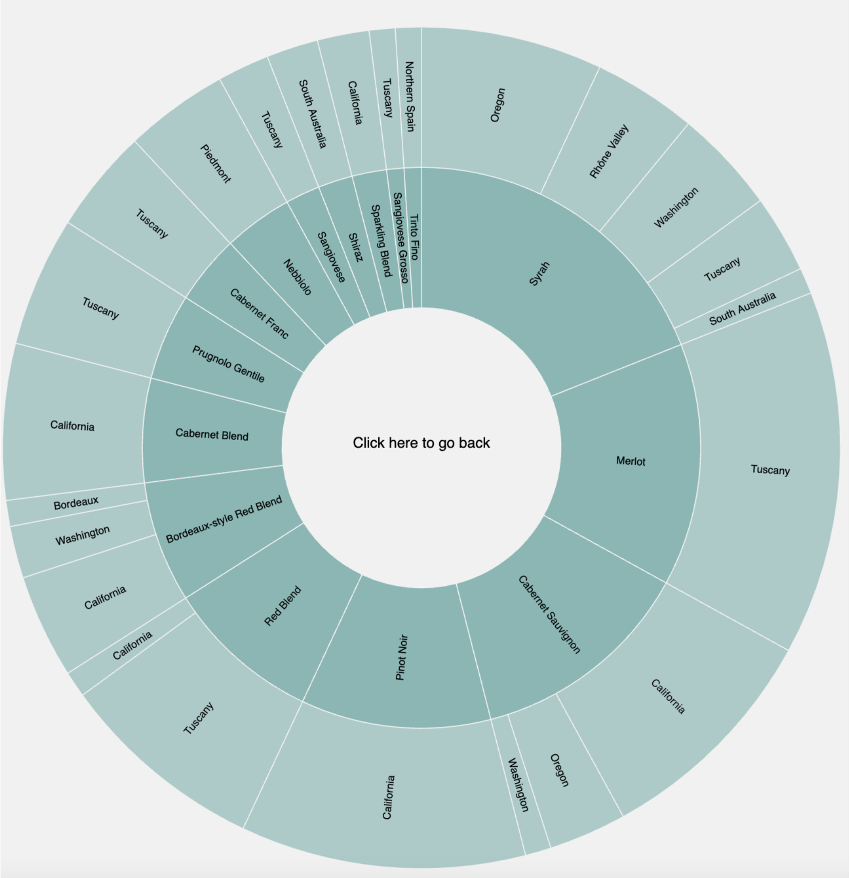

Top 10 Wineries By Average Points
Average Points
100
80
60
40
20
0
Red Wineries
White Wineries
Giovanni Chiappini
Cardinale
Sloan
Shafer
Avignonesi
Tenuta dell'Ornellaia
Trefethen
Cayuse
Château Latour
Tenuta San Guido
Krug
Moët & Chandon
Chambers Rosewood Vineyards
Château Haut-Brion
R.L. Buller & Son
Château Climens
Bouchard Père & Fils
Henriot
Louis Roederer
Campbells
Wineries


Wine Provinces by Points
white wine
red wine

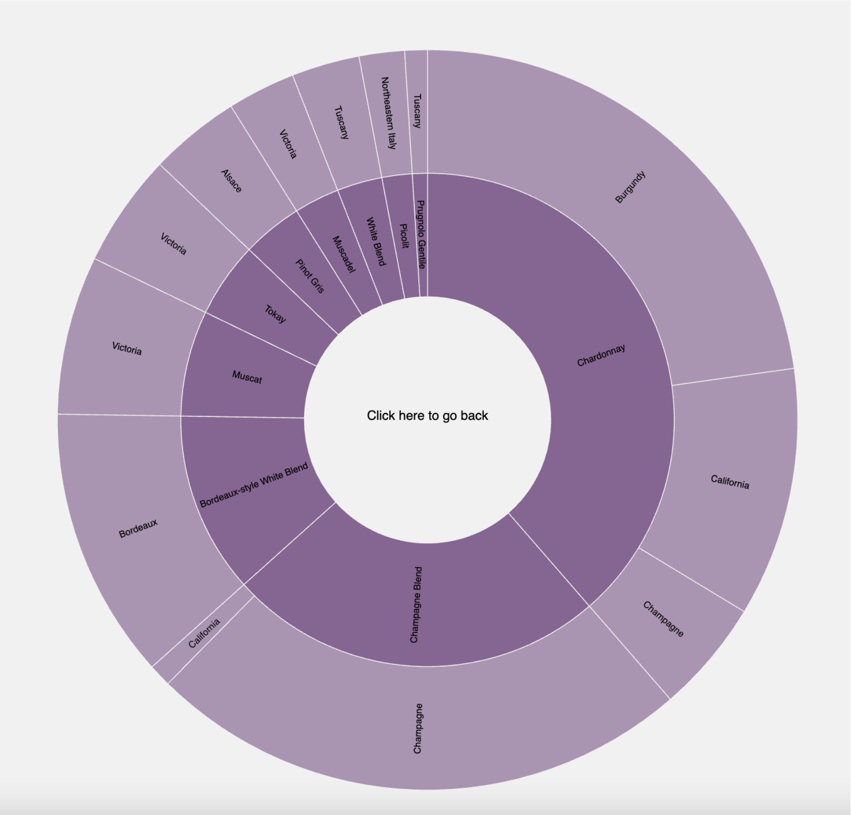


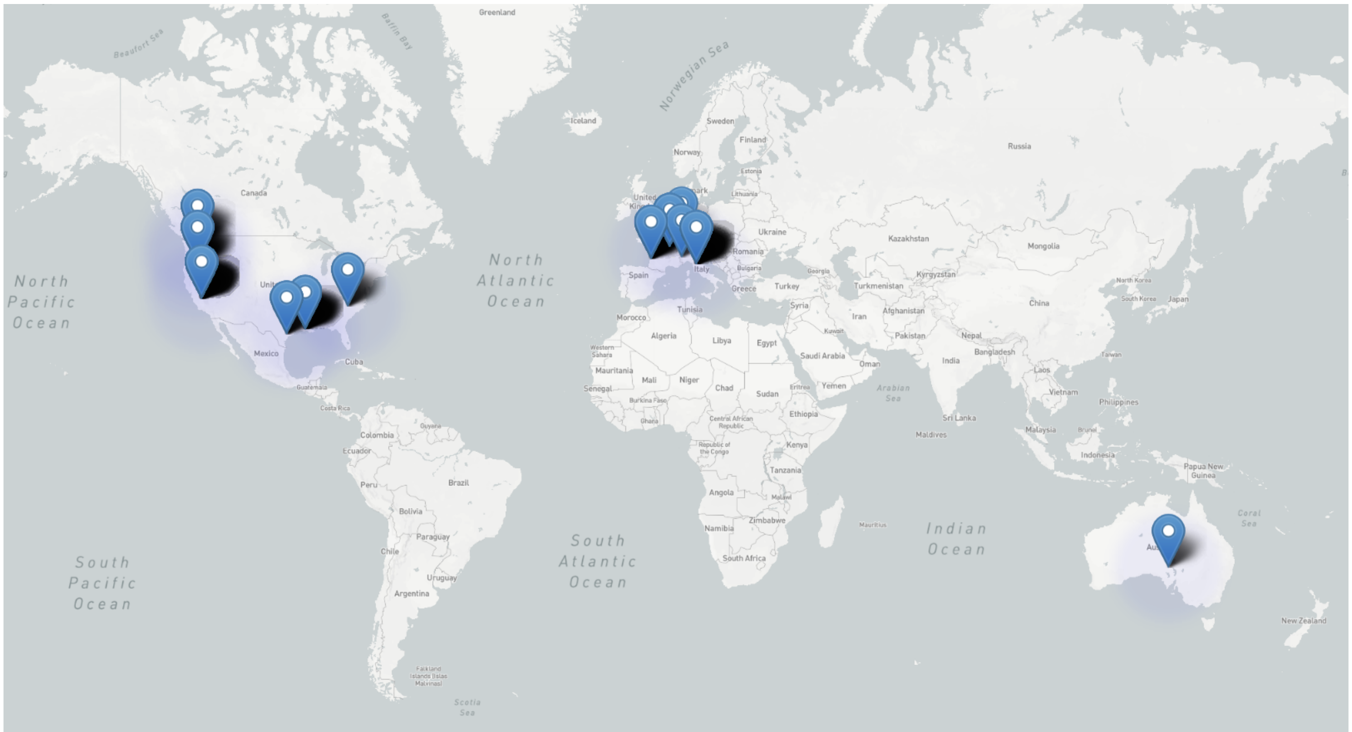
Wine Varieties by Wine Type (Red or White Wine)

Red Wine Varieties



White Wine Varieties





Heat Layer and Marker Map for Top 100 Red and White Wines

WORKS CITED

1. **Wine Reviews (Kaggle) - Main Data Set**
   https://www.kaggle.com/zynicide/wine-reviews ,
   https://www.kaggle.com/zynicide/wine-reviews?select=winemag-data_first150k.csv
1. **GitHub Project- Referenced for wine varieties.**
   https://github.com/06S197GT/SMU-Project-1
2. **Packed Bubble Chart - Inspiration and code reference.**
   https://www.highcharts.com/demo/packed-bubble-split
   https://jsfiddle.net/gh/get/library/pure/highcharts/highcharts/tree/master/samples/highcharts/demo/packed-bubble-split
3. **San Francisco Crime Class Activity - Starter code for Heat Layer May and Marker Map.**
   https://smu.bootcampcontent.com/SMU-Coding-Bootcamp/smu-dal-data-pt-06-2021-u-c/-/tree/master/01-Lesson-Plans/15-Mapping-Web/2/Activities/02-Evr_CrimeHeatmap/Solved%20-%20Booth%20-%20Bootstrap
4. **Sunburst Chart - Inspiration and code reference. -**
   https://observablehq.com/@shuaihaofzny/sunburst-map-of-nyc-crime-data-in-2018