Wine Reviews Along California State

Data Visualization Final Project

Fall Semester 2023/2024

Advisor: Dr. Federico Iuricich

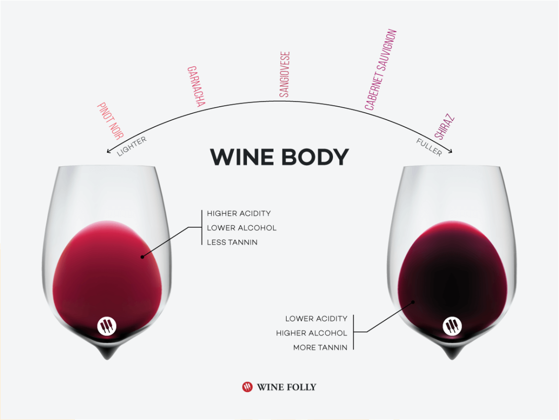
GITHUB Link:

https://github.com/sumanthnangineedi97/CPSC6030_Project/tree/main

Akhila Nangineedi	**C32054414**
Mohammad Alshurbaji	**C19803342**

Overview and Motivation:

This project has been done specifically to show the Wine Reviews a long California State! As Wine is an area of interest for almost all the people, if not all of them. It's very important to point out that the project was done by two teammates who have never drank Alcohol before! Yeah, that's right. So, this was one of the most important inspirations to go with this data.

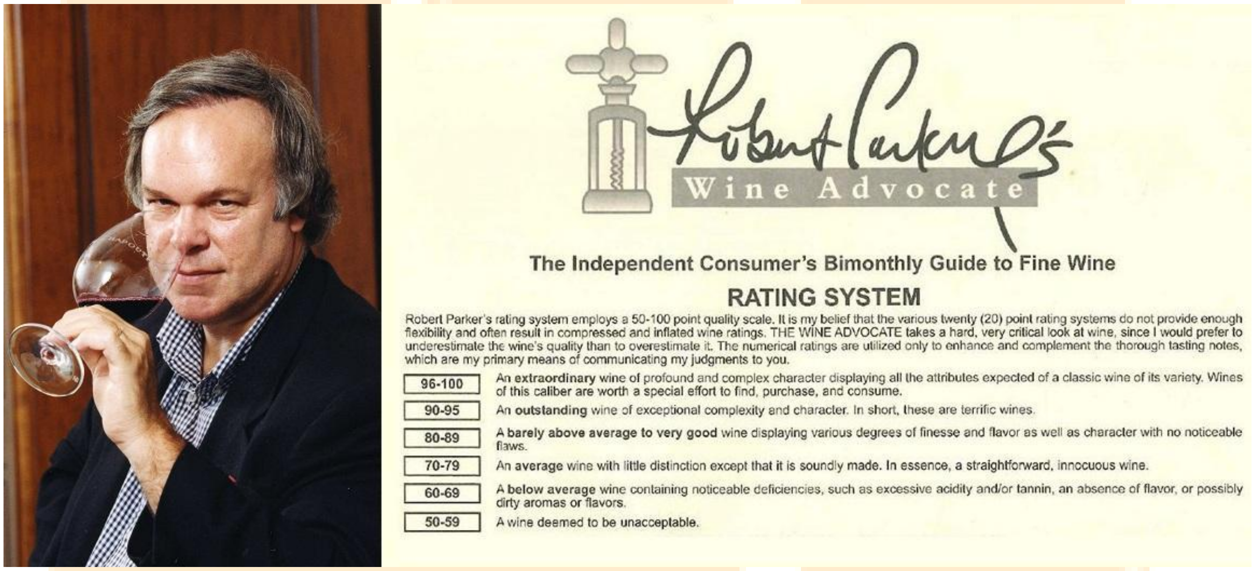


Generally, there are a lot of factors that affect the chosen wine to the selected one over others. As it's an interesting topic and everyone is wondering what I shall pick up for the Christmas Holy! Will this be okay for the guests? Is it expensive to get that wine? Hold on, what are the reviews about it? Oh, come on, what is the winery that produced that kind of wine? Is it authentic and worthy to buy or is it conventional? And a lot of questions are running in our heads. We got you covered! Yes, you've heard it right!

Our website got a lot of attention and attracted a huge number of people! That’s why we’ve done professional work to meet people's concerns. Basically, wines have been classified under different varieties! And that was our meaningful role factor.

Related Work:

As International students, who come from Jordan and India. We’ve looked up to the most interesting point for the Americans! And as the Christmas holy’s approaching! We’ve interviewed our friends about the activities that they used to do during this break. The most repetitive answer was sitting around with the family and drinking different kinds of Wines. We got inspired and excited, starting to do our research and noticed that there are different websites specifically designed for this topic, such as: Wine Advocate, Vivino, and Wine Spectator. That’s how we got our motivation.



Robert Parker's Wine Advocate

The Independent Consumer’s Bimonthly Guide to Fine Wine

RATING SYSTEM

Robert Parker's rating system employs a 50-100 point quality scale. It is my belief that the various twenty (20) point rating systems do not provide enough flexibility and often result in compressed and inflated wine ratings. THE WINE ADVOCATE takes a hard, very critical look at wine, since I would prefer to underestimate the wine's quality than to overestimate it. The numerical ratings are utilized only to enhance and complement the thorough tasting notes, which are my primary means of communicating my judgments to you.

96-100	An **extraordinary** wine of profound and complex character displaying all the attributes expected of a classic wine of its variety. Wines of this caliber are worth a special effort to find, purchase, and consume.
90-95	An **outstanding** wine of exceptional complexity and character. In short, these are terrific wines.
80-89	A **barely above average to very good** wine displaying various degrees of finesse and flavor as well as character with no noticeable flaws.
70-79	An average wine with little distinction except that it is soundly made. In essence, a straightforward, innocuous wine.
60-69	A **below average** wine containing noticeable deficiencies, such as excessive acidity and/or tannin, an absence of flavor, or possibly dirty aromas or flavors.
50-59	A wine deemed to be unacceptable.

Figure 3: Wine Advocate Website

Questions:

The initial project objective was to visualize the global distribution of wines, addressing questions such as: How many wines are present in each country and province? What are the pricing and points of each wine? And which wineries are leading in producing the most wines? As the project evolved, we refined our dataset to focus exclusively on the United States—specifically, the states of California, Oregon, Washington, and New York. Further narrowing our focus to California, known for its extensive variety of wines and wineries, led to the decision to showcase all wineries in our analysis. This deliberate choice was made to offer users a comprehensive overview of the diverse wine landscape within the state.

Revised questions guiding our analysis:

- What varieties are present in California?
- What are the Trends, outliers, and correlations between price and points of wines?
- Which wineries are present, and how many wines does each produce?

By exploring these aspects, we aim to provide users with a thorough understanding of the entire wine ecosystem in the state of California.

Data:

The Wine Reviews dataset (https://www.kaggle.com/datasets/zynicide/wine-reviews) was downloaded from Kaggle.com . The data in the dataset (winemag-data-130k-v2.csv) was collected from wineenthusiast.com in November 2017. It comprises 14 attributes with a total of 129,971 items. However, for the purpose of our analysis, we selected specific attributes deemed crucial, including Country, Province, Price, Points, Title, Variety, and Winery. Subsequently, records containing null values and duplicate entries were eliminated. Further refinement involved narrowing down the dataset to entries corresponding to the province of California. Within this subset, varieties with a lower count of wines were excluded. As a result of these processing steps, the dataset was ultimately reduced to 7 attributes, encompassing 16,475 distinct items.

Table 1: Data Explanotary

Data Explanotary	
	Wines
#Records	129,971
#Fined Records(in California State)	18,135
#Final Records	16,475
	Points
Range	80-99
	Prices
Range	4-750
Max	750
Minimum	4
Avg For California State	40.94
Standard Deviation	28.668

Exploratory Data Analysis:

As our data got reduced from 14 to 7 attributes to include the most important factors on our problem statements. We deduced that Varieties are going to be a meaningful factor and would affect all other attributes. We've selected the Varieties to be the filter for our website, and everything else relies on them.

Basically, our website consists of three different charts. Each of them addresses a specific question. The following diagram shows the relationship between them:

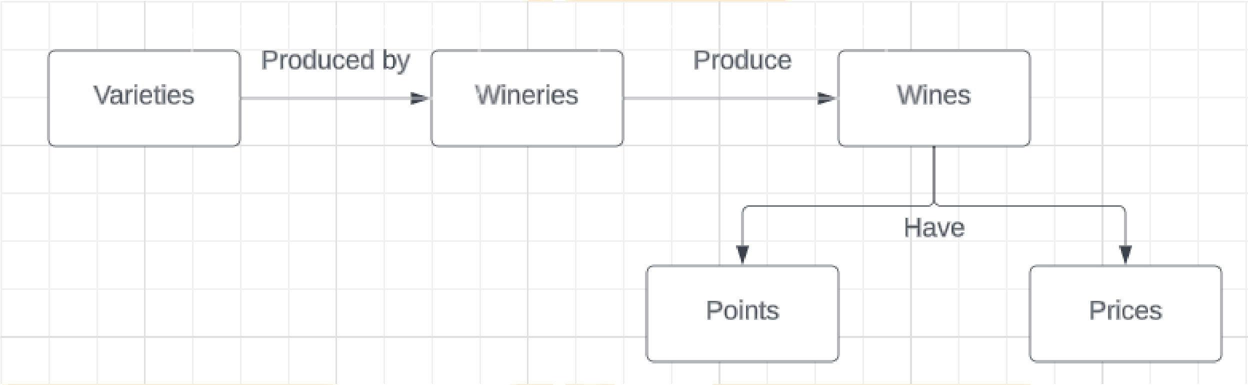


Figure 4: Relationship Between Charts

In order to express each of the relationships, we had to represent the following charts as follow:

To display the varieties presented in California, we employed a Tree-map to visually represent the distribution of wines based on most common varieties. Insights gained from this visualization included a comprehensive view of the quantity of wines for each variety, guiding subsequent analyses. The tree map's high-level overview informed the design of more detailed visualizations, emphasizing the essential attribute of wine variety.

Table 2: Decoding Tree Map

Chart 1: Tree-Map	
Attribute	Variety
Data Type	Categorical
Marks	Areas
Channels	Areas, Colors

For the second objective, we contemplated employing a side-by-side bar chart to display two categorical attributes (price and points) for each wine. However, this approach posed limitations in showcasing all wines, restricting the display to only a subset. As an alternative, we opted for a scatterplot, which provides scalability. The scatterplot facilitates the visualization of trends, outliers, distribution, and correlations, allowing for the identification of correlations and extremes among price and points. As a result, the scatter plot showed that there is no direct correlation between these two factors, which will be explained later.

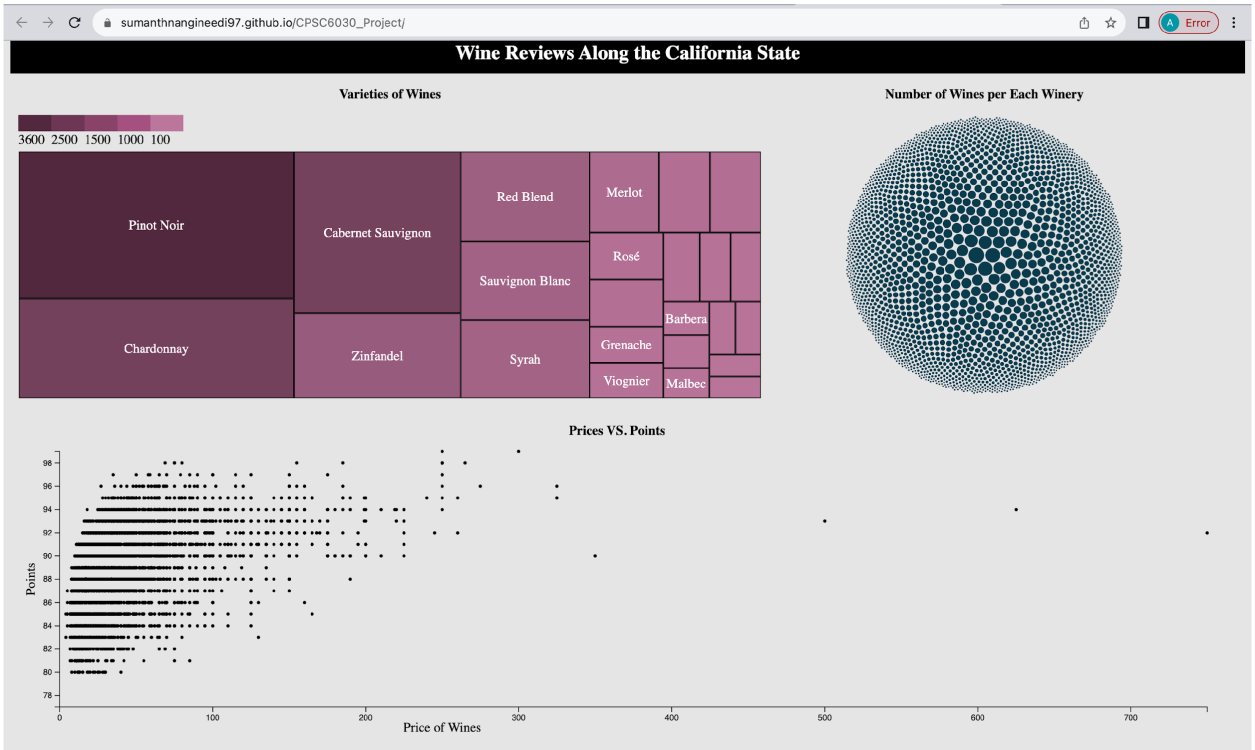
Table 3: Decoding Scatter Plot

Chart 3: ScatterPlot		
Attribute	Points	Price
Data	Quantitative	Quantitative
Marks	Points	Points
Channels	Position, Color	Position, Color

For the last one, we thought of the Sankey chart to show the distribution of wines along the varieties and winery but since the dataset has a significant number of wineries and varieties, which means a very high cardinality! it limited us to display on the top varieties and wineries. Later we thought of showing only top wineries in the form of a bar chart but it was also not scalable then the professor’s suggestion was considered. We packed a bubble chart with which we are able to display all the wineries and the number of wines each produce with their size.

Table 4: Decoding Bubble Chart

Chart 2: Bubble Chart	
Attribute	Winery
Data	Categorical
Marks	Areas
Channels	Areas, Position



Design Evolution:

As good designers and data scientists, there is no way to obtain your first design from the first attempt. Always, the conceptual design needs to be modified, actually It might be re-obtained from scratch again, as what happened in our case. Obviously, it's the conceptual design, to just show you the first perception of your data. The following chart shows our Time-Line process and how we evolve over time. It's very important to point out that our final design deviated from the reported one before.

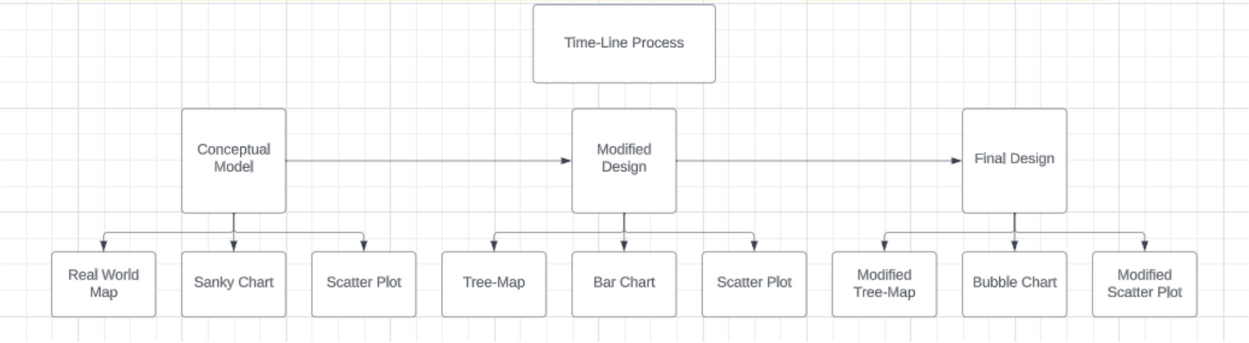
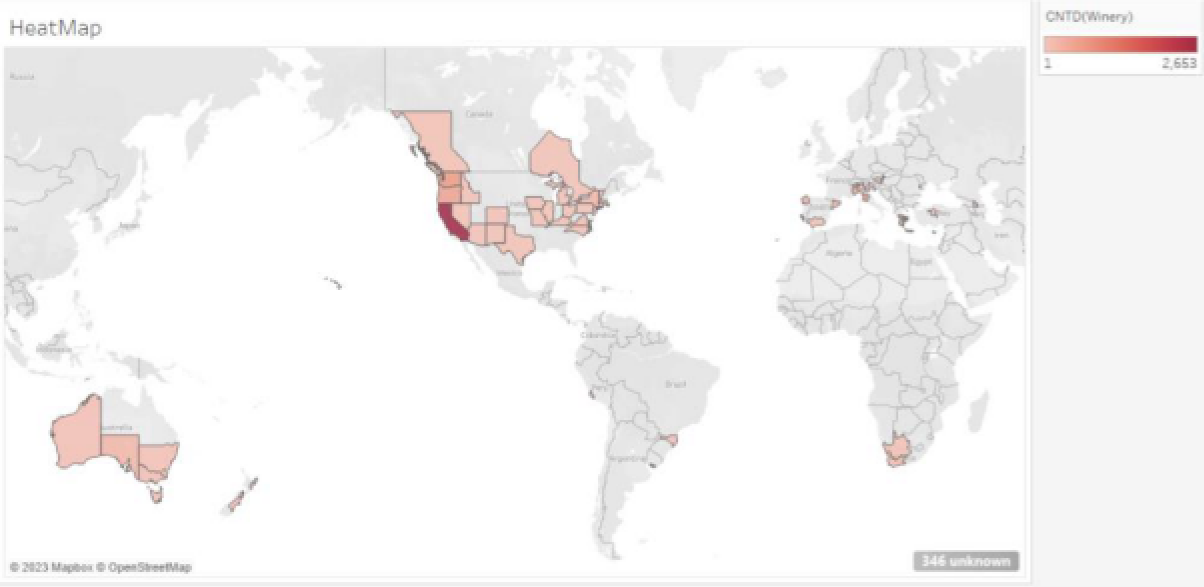
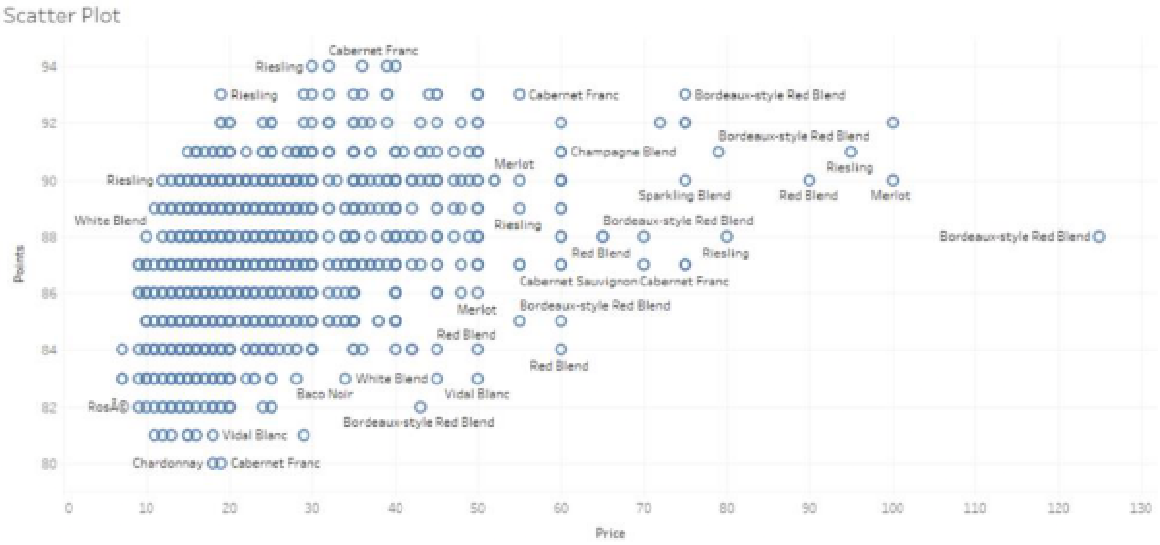
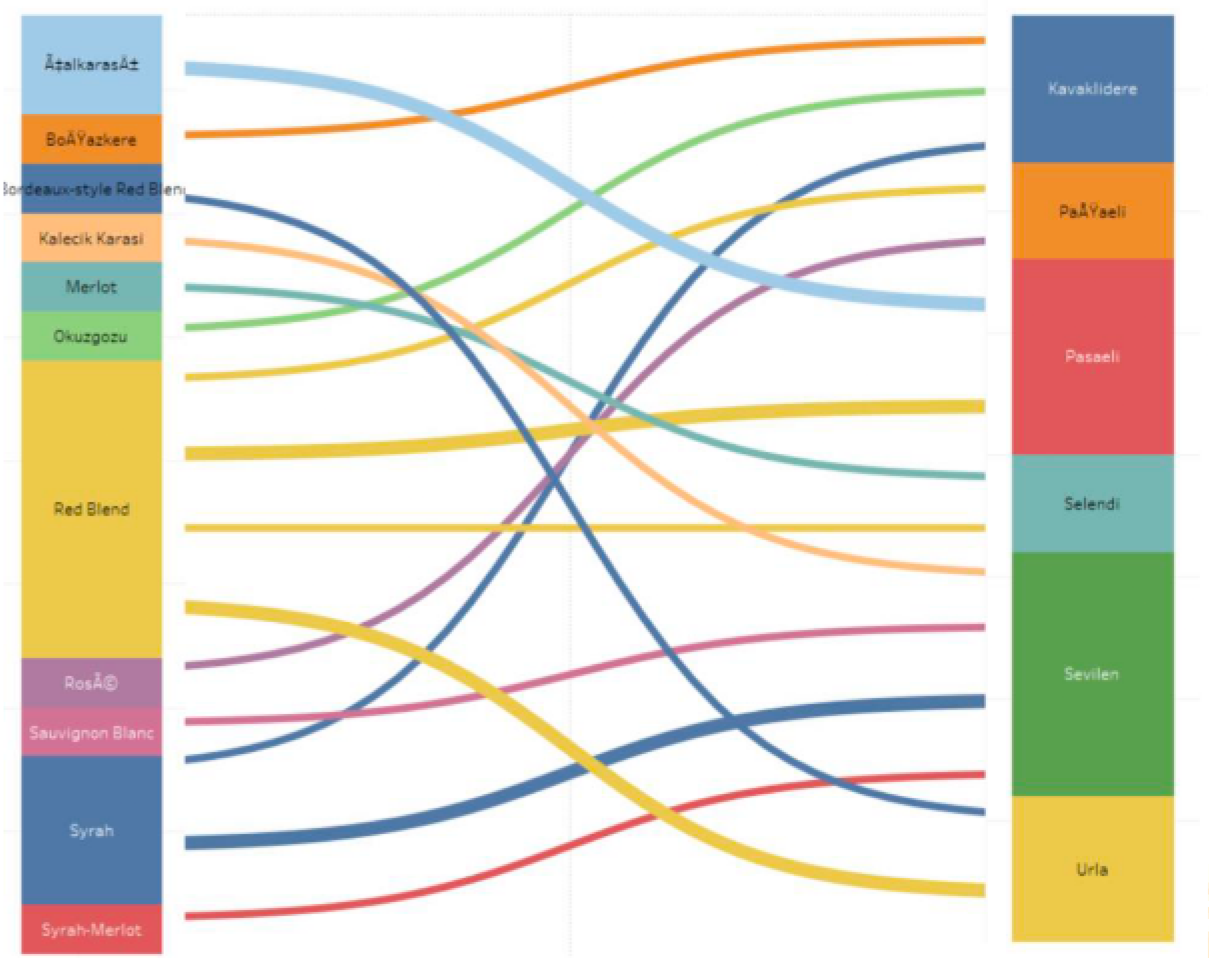


Figure 5: Time-Line Process

As shown above, our first design is the conceptual model, which consisted of a Real-World Map, Sanky Chart, and Scatter Plot. As shown in the Dashboard below:

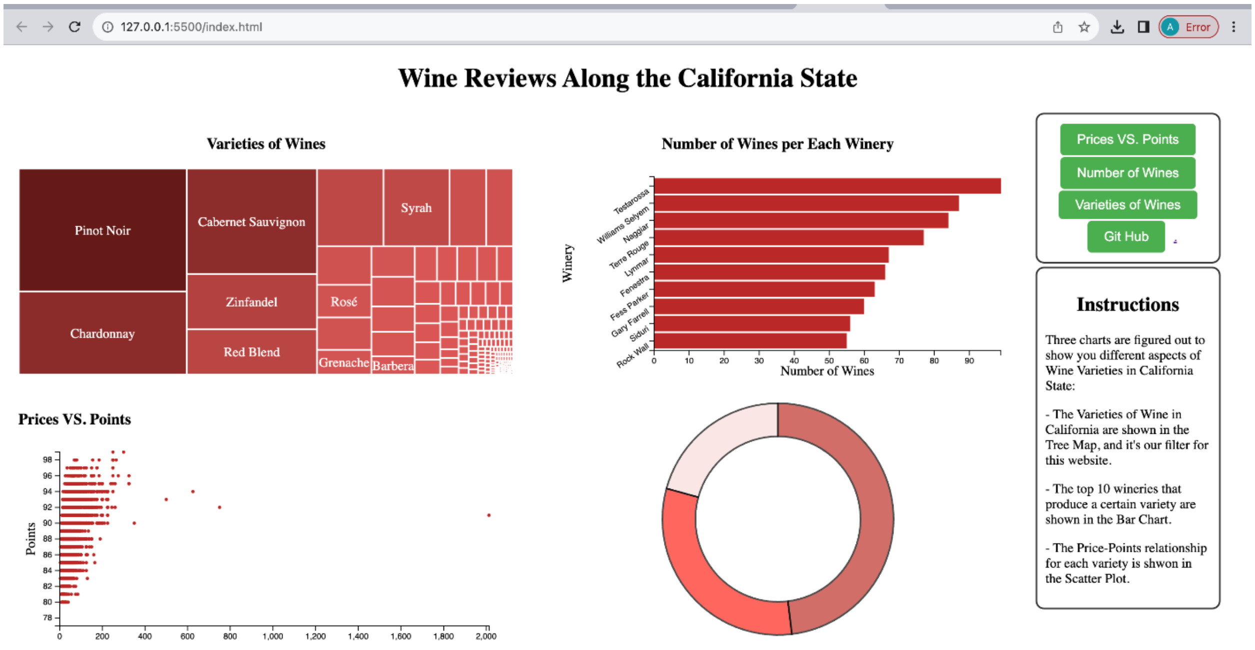

HeatMap
CNTD(Winery)
1
2,653
© 2023 Mapbox © OpenStreetMap
346 unknown


Scatter Plot
Cabernet Franc
Riesling
Riesling
Cabernet Franc
Bordeaux-style Red Blend
Bordeaux-style Red Blend
Champagne Blend
Riesling
Merlot
Riesling
Red Blend
Merlot
Sparkling Blend
White Blend
Riesling
Bordeaux-style Red Blend
Bordeaux-style Red Blend
Red Blend
Riesling
Cabernet Sauvignon
Cabernet Franc
Merlot
Bordeaux-style Red Blend
Red Blend
Red Blend
White Blend
Baco Noir
Vidal Blanc
RosÃ©
Bordeaux-style Red Blend
Vidal Blanc
Chardonnay
Cabernet Franc
94
92
90
88
86
84
82
80
Points
0
10
20
30
40
50
60
70
80
90
100
110
120
130
Price



Clearly, the dashboard above shows a lot of messy points. Such as, the Real-World map took a huge amount of space for a useless amount of information. Most of the map is empty since our major data exists in California rather than the other countries. Holy shit, the Sanky chart is a headache, the cardinality is so high, that's why there is no way to include it in our study.

Moving on, our second design evolved in a fast way after our perception evolved with the Class Lectures. Basically, the Real World Map the whole country turned into a Tree-Map for California State, as our data plays a major part in this state. In addition to that, the Sanky chart becomes a Bar chart which is way better to present the addressed attributes.



Wine Reviews Along the California State

Instructions

Three charts are figured out to show you different aspects of Wine Varieties in California State:

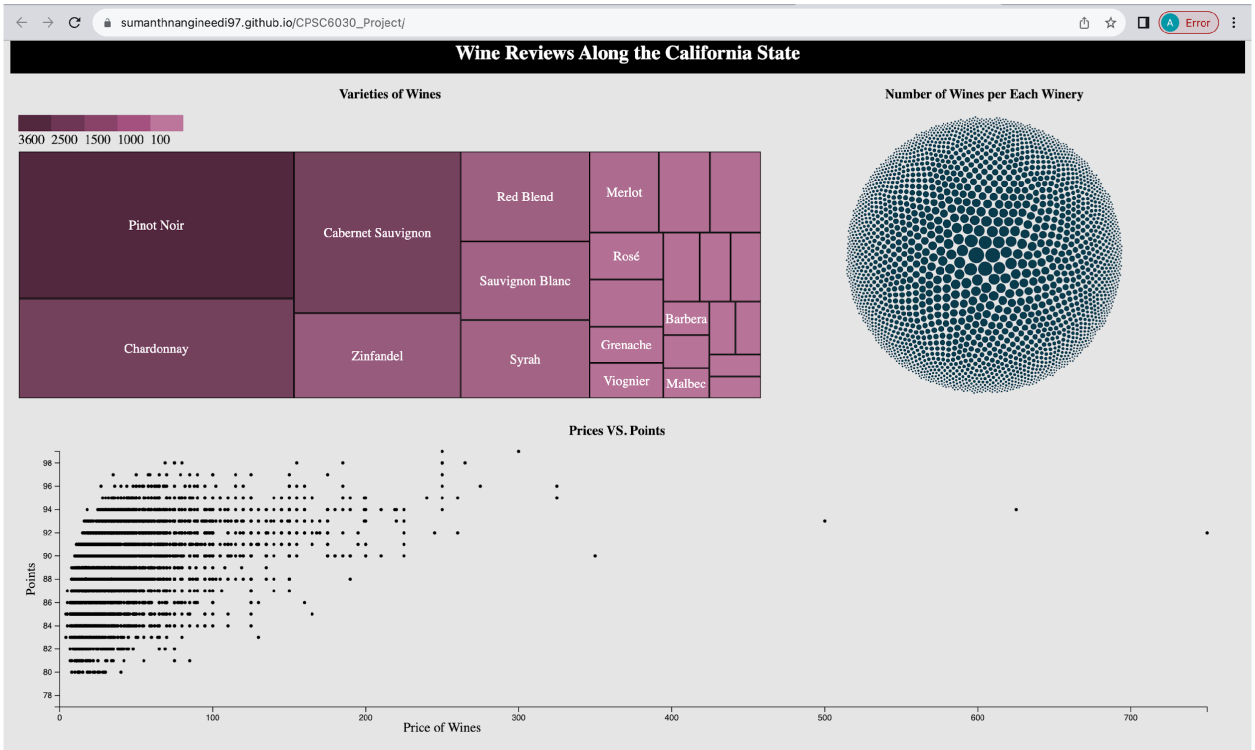
- The Varieties of Wine in California are shown in the Tree Map, and it's our filter for this website.

- The top 10 wineries that produce a certain variety are shown in the Bar Chart.

- The Price-Points relationship for each variety is shwon in the Scatter Plot.

From a bird’s point of view, it looks good and totally presents our problem statements. But when you pay close attention, you’ll notice a few limited and strict things while walking through the whole data. By getting different recommendations from the professor, and by doing a Brainstorming! We reached our final design.

Finally, our final design was considered to be the final view of our website. The dashboard below shows the incredible design ever for the Wine people concerns:



Considered visualizations include treemaps for illustrating variety distribution, scatterplots for displaying price-points of each wine, and packed bubble charts for showcasing wineries' distribution.

Treemap for Varieties in California:

The treemap visualization serves the purpose of illustrating the distribution of wines based on varieties in California. It allows users to effortlessly extract information about the quantity of wines for each variety, offering a comprehensive view of the diverse wine landscape in the state. The ease of interpretation is high, making it an effective tool for users to grasp the variety distribution with confidence.

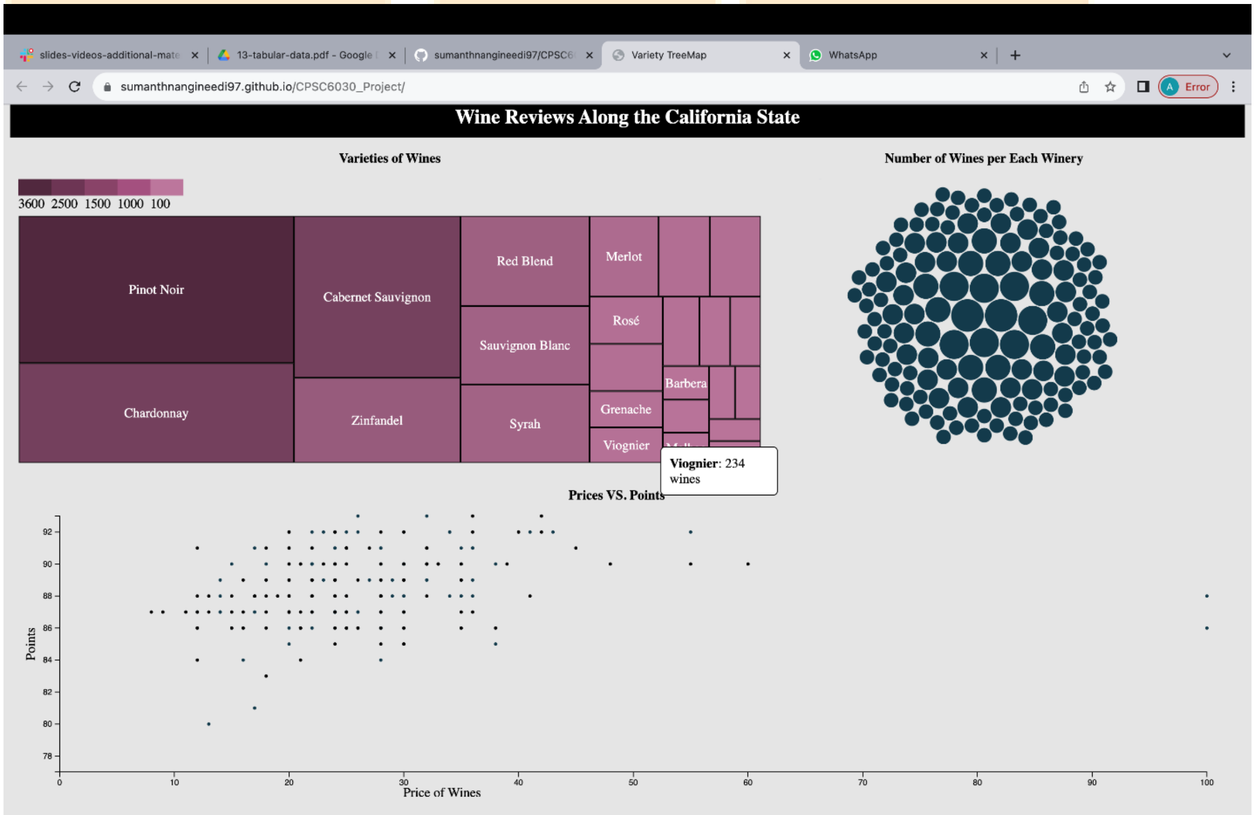
Scatterplot for Price and Points:

The scatterplot is designed to showcase the relationship between wine price and points. Users can easily discern trends, outliers, and correlations between these two attributes. Extracting information from this visualization is straightforward, as patterns and relationships are visually apparent. Confidence in the presented information is reasonably high, given the direct observation capabilities of the scatterplot.

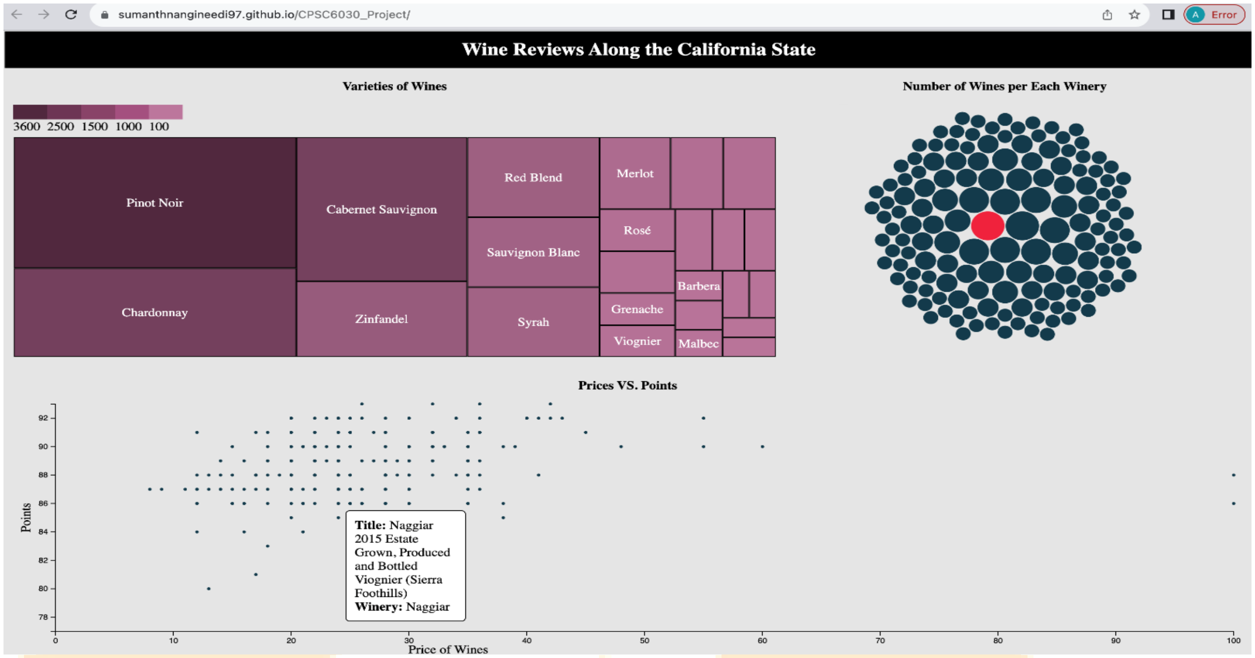
Packed Bubble Chart for Wineries and Wines:
The packed bubble chart provides a visual representation of the distribution of wines among wineries in California. It overcomes scalability issues faced by other visualizations, displaying all wineries and the number of wines each produces. Extracting information from this chart is manageable, although due to the number of bubbles, closer inspection may be required for specific details. Confidence in the information is reasonably high, particularly for identifying top wineries and their production.

Implementation:

The Treemap serves as the main interactive visualization, providing a visual representation of the distribution of wine varieties across California. The intent is to offer users an at-a-glance overview of the diversity in wine types. By clicking on a specific variety (rectangle), users trigger an update in the Scatterplot and Packed Bubble Chart, allowing them to delve deeper into the details associated with the selected variety. This interactive feature aims to facilitate a seamless exploration of specific subsets within the dataset, enhancing user engagement.

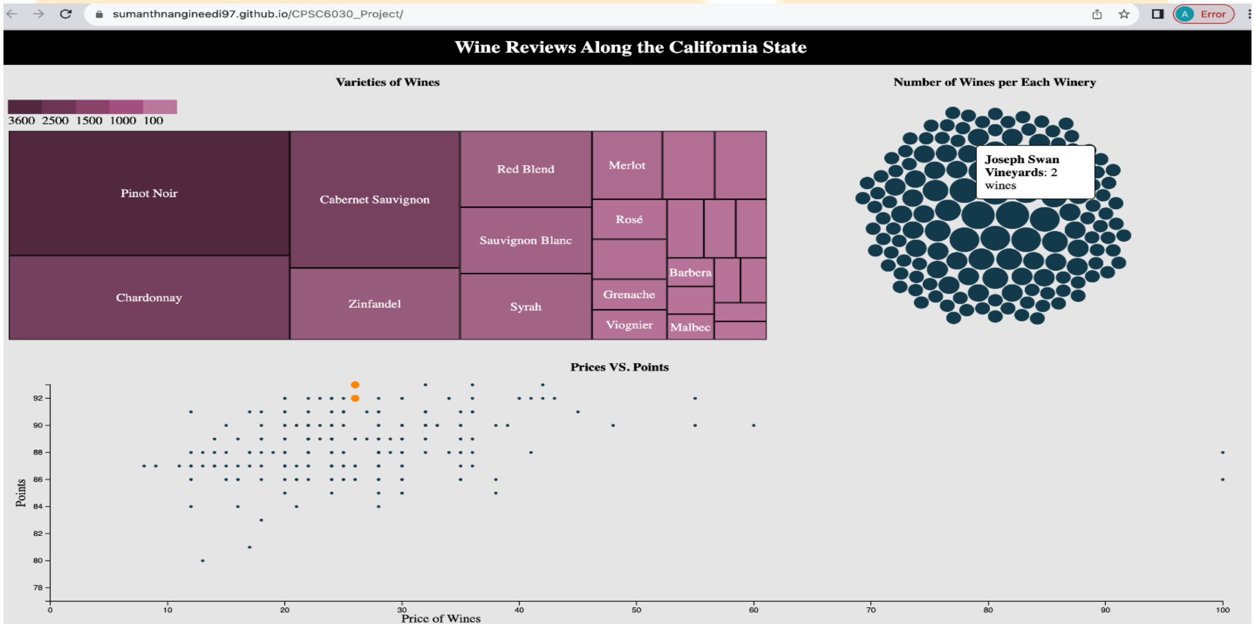


The Scatterplot is designed to showcase the relationship between price and points for each wine. Its interactive functionality comes into play when a user clicks on a particular data point. This action prompts a change in the color of the corresponding winery bubble in the Packed Bubble Chart, establishing a connection between individual wines and their respective wineries. The intent is to allow users to investigate specific wines in the context of their associated wineries, providing a more nuanced understanding of the dataset.



Title: Naggiar 2015 Estate Grown, Produced and Bottled Viognier (Sierra Foothills)
Winery: Naggiar

The Packed Bubble Chart visualizes the distribution of wines among different wineries in California. Clicking on a winery bubble triggers changes in the Scatterplot, modifying the color of wines associated with that winery and increasing the radius of their data points. This interactive feature aims to highlight and explore wines produced by specific wineries, offering users a comprehensive view of the relationship between wineries and individual wines. The intent is to facilitate dynamic exploration and analysis, encouraging users to uncover insights from various facets of the dataset.



Joseph Swan Vineyards: 2 wines

Evaluation:

The treemap visualization provided valuable insights into the distribution of wine varieties in California. By presenting a visual representation of the quantity of wines for each variety, users could easily grasp the diversity within the state. This visualization effectively answered the question of what varieties are present in California, offering a comprehensive view of the wine landscape.

The scatterplot, designed to showcase the relationship between wine price and points, facilitated the identification of trends, outliers, and correlations between these two attributes. This visualization likely answered questions related to the pricing and points of each wine, providing users with a straightforward understanding of the distribution and relationships within the dataset.

The packed bubble chart, visualizing the distribution of wines among different wineries in California, offered insights into the production landscape. Users could identify top wineries and the number of wines each produced. The interactive features, such as clicking on winery bubbles to update the scatterplot and vice versa, aimed to enhance user exploration. While managing a large dataset, this visualization allowed users to uncover insights into the relationships between wineries and individual wines.

In terms of enhancements, it would be beneficial to establish interactions from the scatter plot and bubble chart to the treemap. Given that the size of numerous wineries in the bubble chart is relatively small for effective interaction, allocating more space for the bubble chart could enhance usability. Additionally, for the scatterplot, implementing geometric distortion techniques like Fisheye could be considered to make room for details in the focus regions.