# Wine Reviews Statistical Analysis

This is a brief introduction to summary statistics for the project of creating a model for determining the best value wine based on price and points from Wine Spectator reviews.

### Description of Variables

After performing data clean up, the dataset contains 120,975 observations of 12 variables. They are:

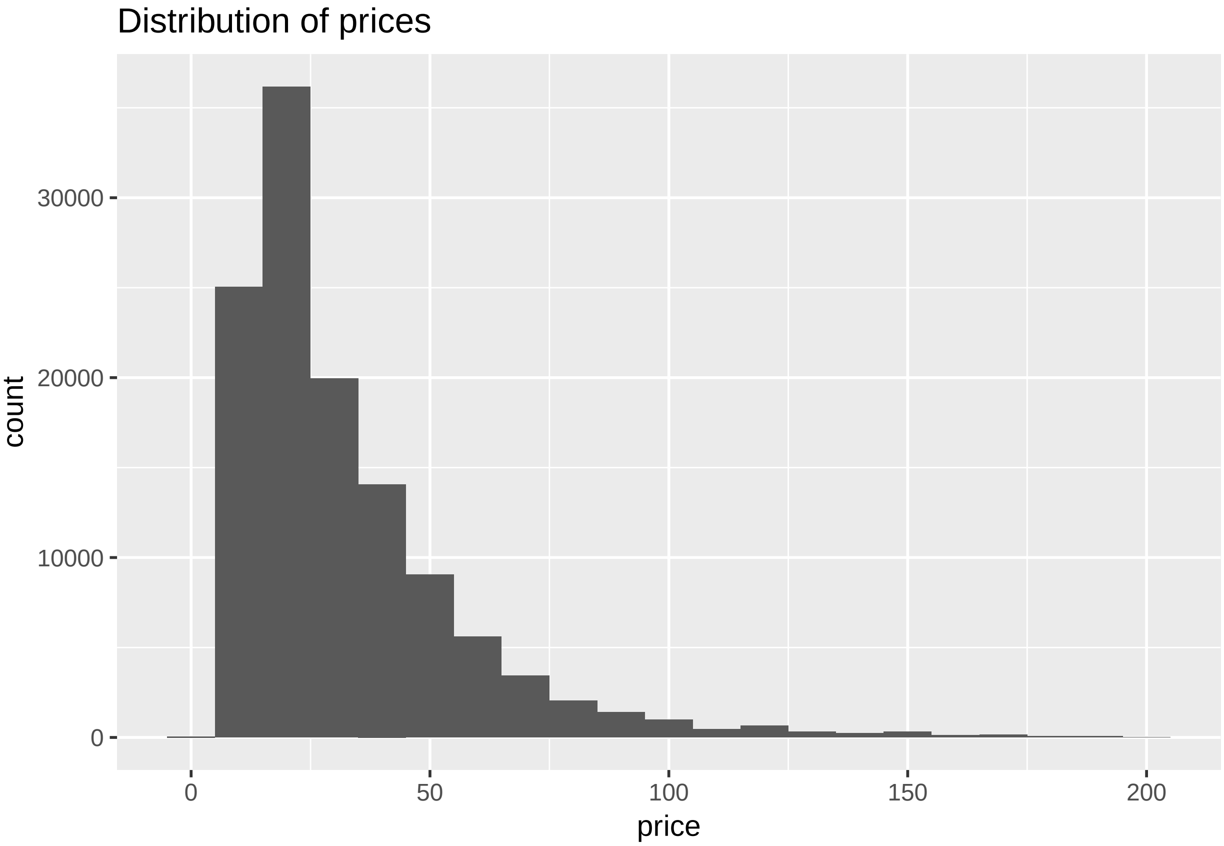
- **X** - unique identifier
- **country** - the country of origin of the wine
- **continent** - the countinent of origin of the wine
- **province** - the province of origin of the wine
- **region_1** - the wine growing region of the wine
- **winery** - the winery that created the wine
- **title** - the name of the wine
- **year** - the vintage of the wine
- **variety** - the variety of the wine
- **points** - the points awarded by a Wine Spectator reviewer
- **price** - the price in USD of the wine per bottle
- **value** - the points awarded to the wine per USD

### Summary Statistics

- All wine-producing continents are represented, with North America the most-represented continent with 54,589 observations. Asia is least-represented with 652 observations
- The wines reviewed come from 43 unique countries. The most-represented country is the United States with 54,265 observations and least-represented is Slovakia with 1.
- For American wines, province typically corresponds to state. The wine-producing powerhouse state of California is the most represented of the 423 worldwide provinces in the dataset.
- The Napa Valley of California (4475 observations) is the most commonly reviewed region of 1205 regions.
- There are 698 distinct varieties of wine with the popular Pinot Noir (12,787) and Chardonnay(11,080) as the most commonly occurring wine varieties.

## Price

- The distribution of price is right skewed as the median price is $25, mean price is $35.36, but there are a handful of prices that are far above average. The maximum price in the origindal dataset is $3300. The minimum price is $4. Outliers from $200 to $3300 have been removed to prevent skewing the final model.



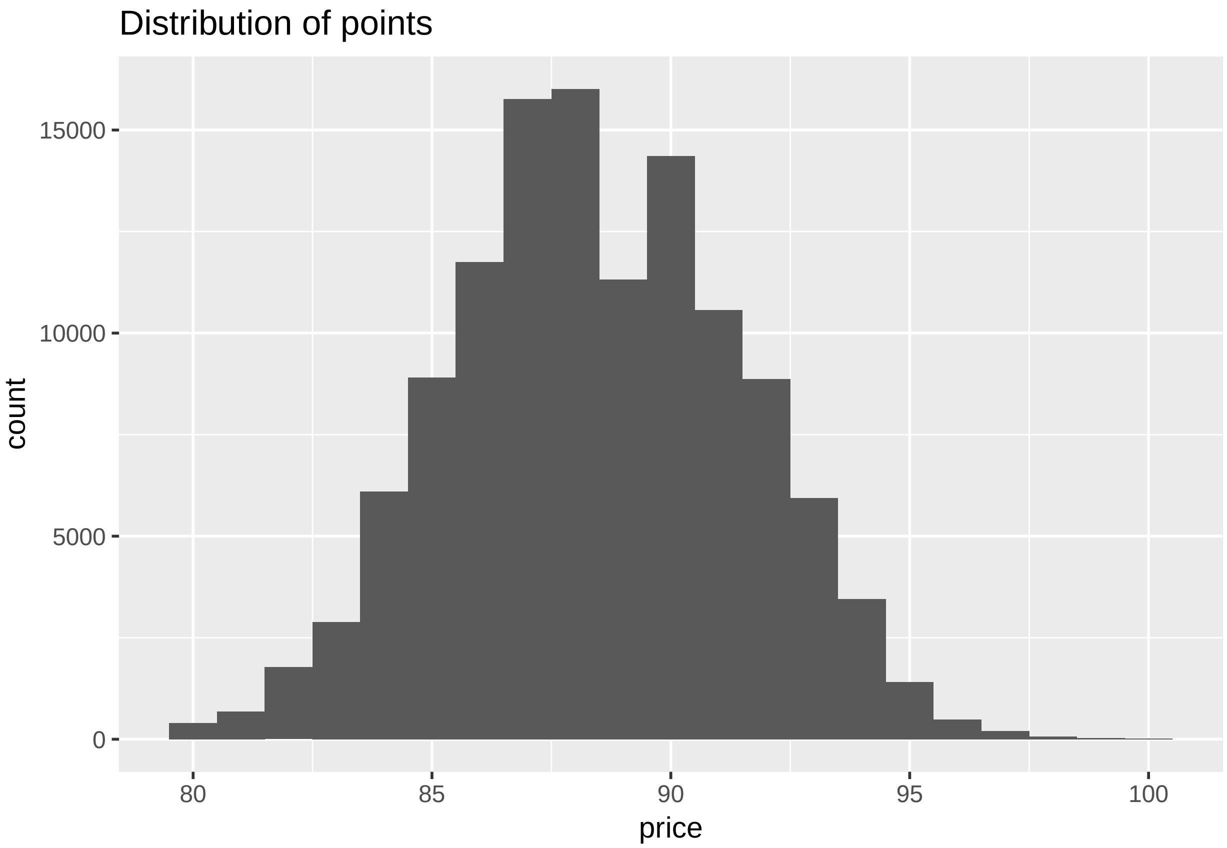
## Points

- Points are awarded by Wine Spectator on a scale of 0-100. This dataset only contains wines with point assignments of 80 and up. The table below gives explanations of the points.

| Score | Description |
|---|---|
| 80-84 | Good: a solid, well-made wine |
| 85-89 | Very good: a wine with special qualities |
| 90-94 | Outstanding: a wine of superior character and style |
| 95-100 | Classic: a great wine |

- The minimum points in this dataset is 80 and maximum 100.

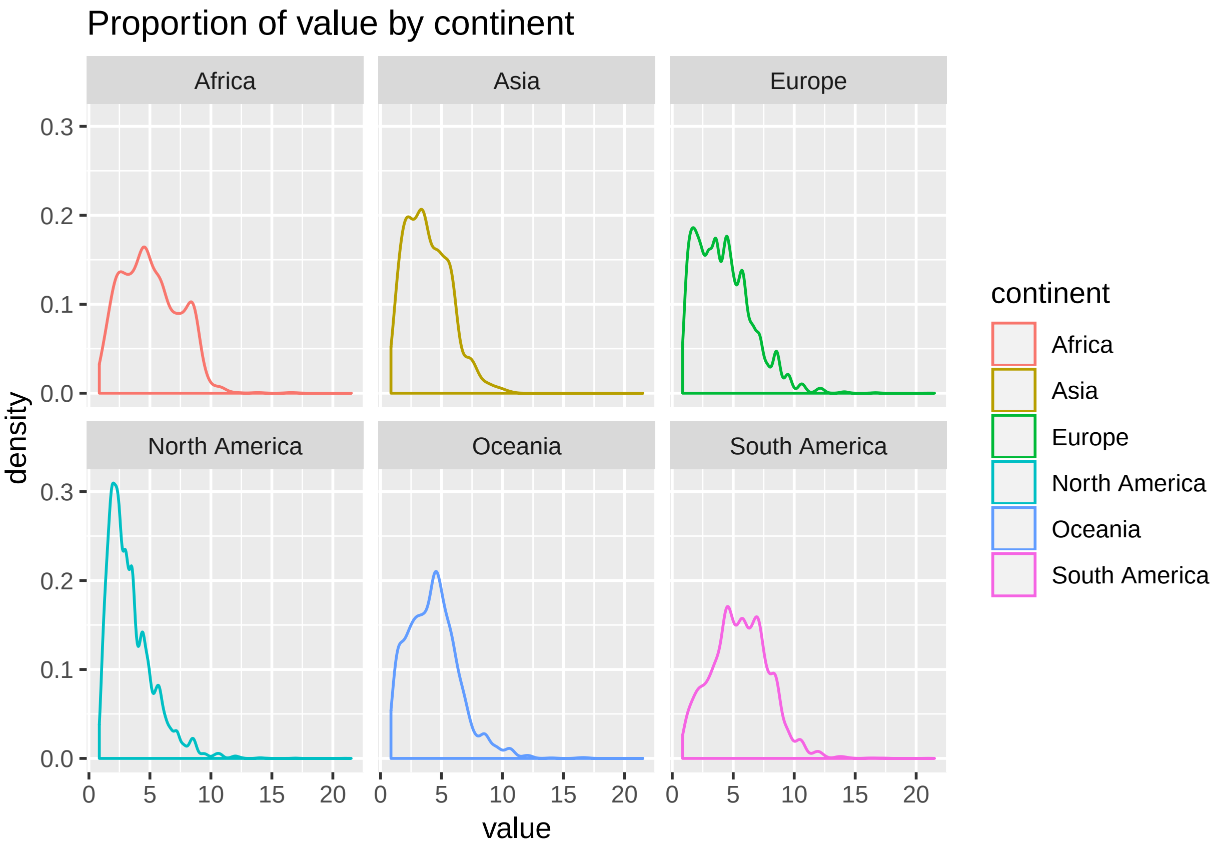
- The median points is 88 and mean points 88.42.



## Value

Value is a function of points/price. A higher value indicates that the wine is a good purchase. When working with a dataset with a drastically large range of prices (from 4 USD to 3300 USD), and a points scale that only varies by 20 points, value is an imperfect measure.

*I am planning on doing analysis of value further by looking at the IQR of value and focusing on the mid range to avoid massive skews from super expensive wines, but considering that this is a draft, for now I am just getting some thoughts down*


Proportion of value by continent
Africa
Asia
Europe
North America
Oceania
South America
density
0.0
0.1
0.2
0.3
value
0
5
10
15
20
continent
Africa
Asia
Europe
North America
Oceania
South America