Leveraging Data to Choose Wine With Confidence

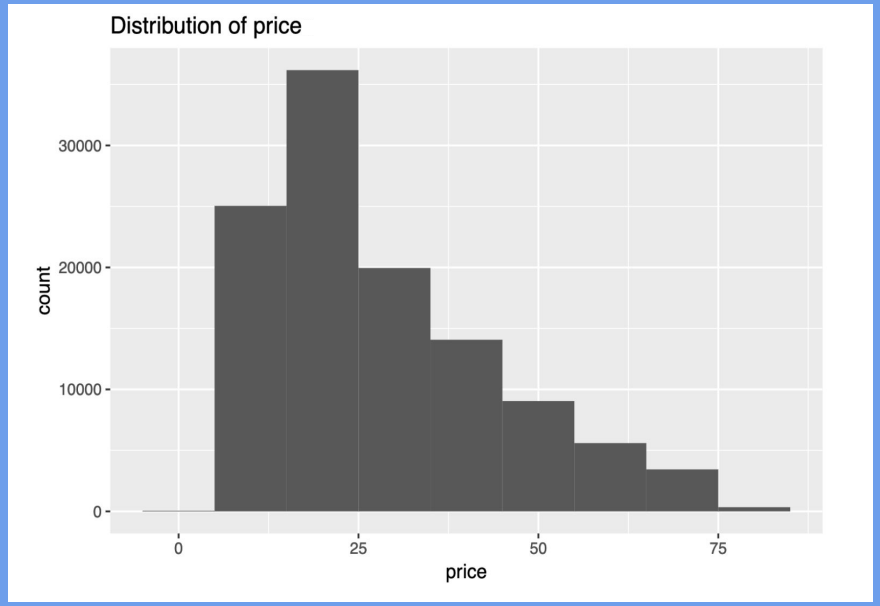
Zoë Bakker

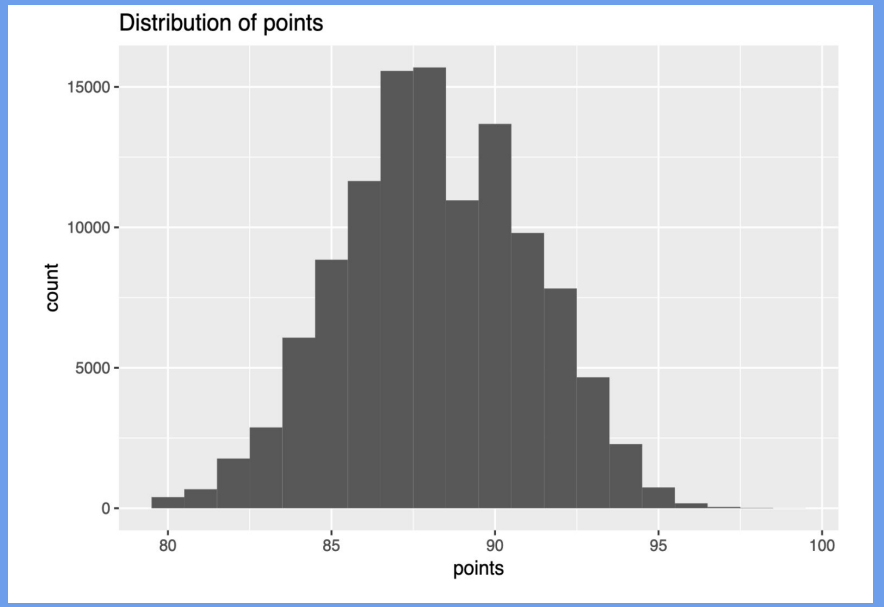
In 2016, Americans drank 4.24 billion bottles of wine - 13.3 bottles for every man, woman, and child.

Wrangling the Dataset

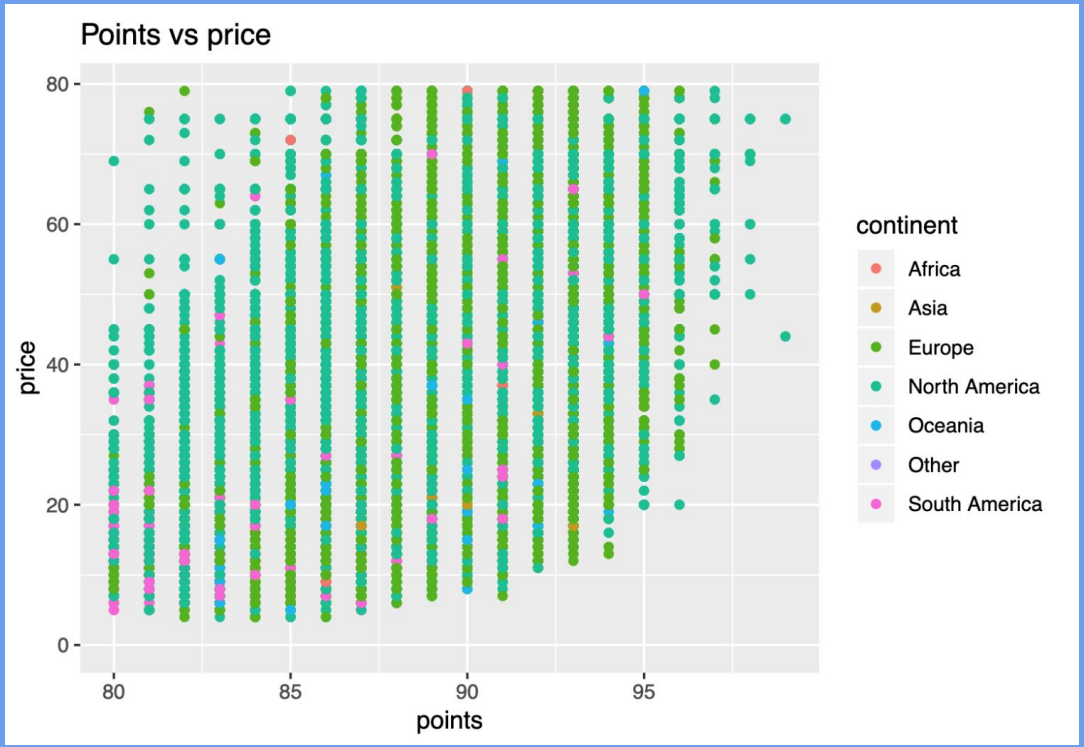
- Replacement of blank values
- Removal of observations where price is not recorded
- Removal of unnecessary and duplicate columns
- Addition of value, continent, and year columns

Price & Points



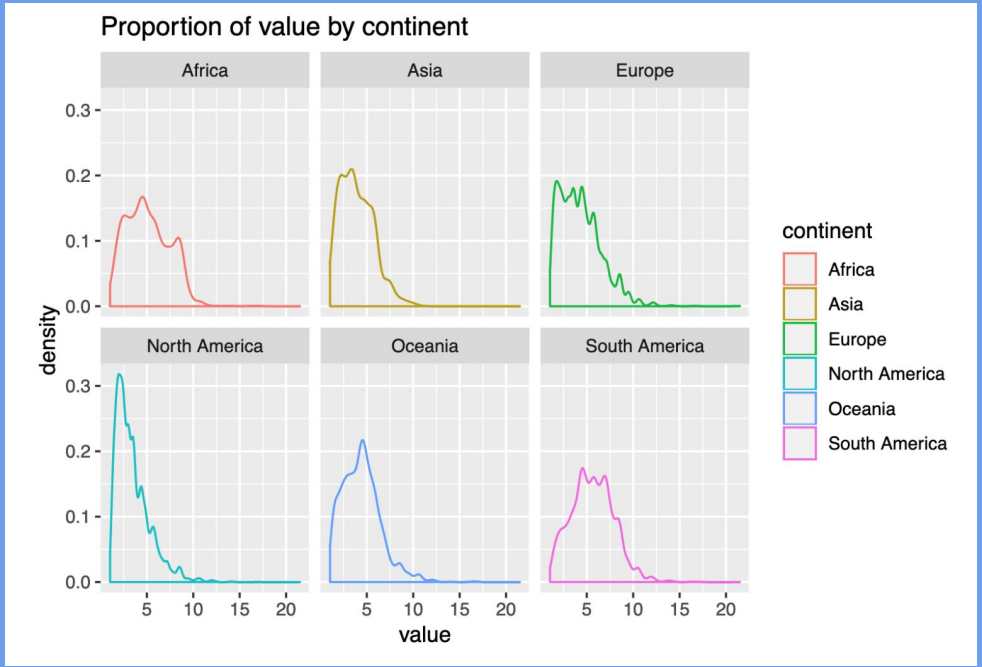


Price, Points & Value



Value Limitations

- Extreme price outliers
- Limited scale
- Inherent bias
- Personal taste



Choosing a Predictive Model

- Models Considered:
  - Linear Regression
  - Gradient Boosting Machine
  - Random Forest

Fitting a Linear Regression Model

Fitting the model to a training set

R-squared: 0.6126
Adjusted R-squared: 0.6045

~~Validating the model on a testing set~~

Cross validation model

R-squared: 0.6449
Adjusted R-squared: 0.6387
RMSE: 1.265

Improving the Linear Regression

Previous Model:

- R-squared: 0.6449
- Adjusted R-squared: 0.6387
- RMSE: 1.265

Modified Model:

- R-squared: 0.7904
- Adjusted R-squared: 0.7955
- RMSE: 0.9198

Future Improvements

- Red vs white analysis
- Food pairings
- Geographic-based analysis