Exploring Patterns and Trends in Wine Quality Through Data Analysis and MapReduce Approach

Justin Bowers, Dong Jun Woun

Background & Motivation

Motivation:

- Identify the best quality of wine for the best price.
- Find common qualities of highly rated wines
- Subjectivity in wine tasting demand objective data analysis for a wide range of reviews

Dataset:

- Wine Reviews is a popular dataset that contains analysis of approximately 130 thousand wine reviews from WineEnthusiast. It contains the name, quality points, description, and regional data.

MapReduce:

- Process datasets through distributed computing
- Map transforms the data in to a new shape.
- Reduce applies functions to create a new dataset.

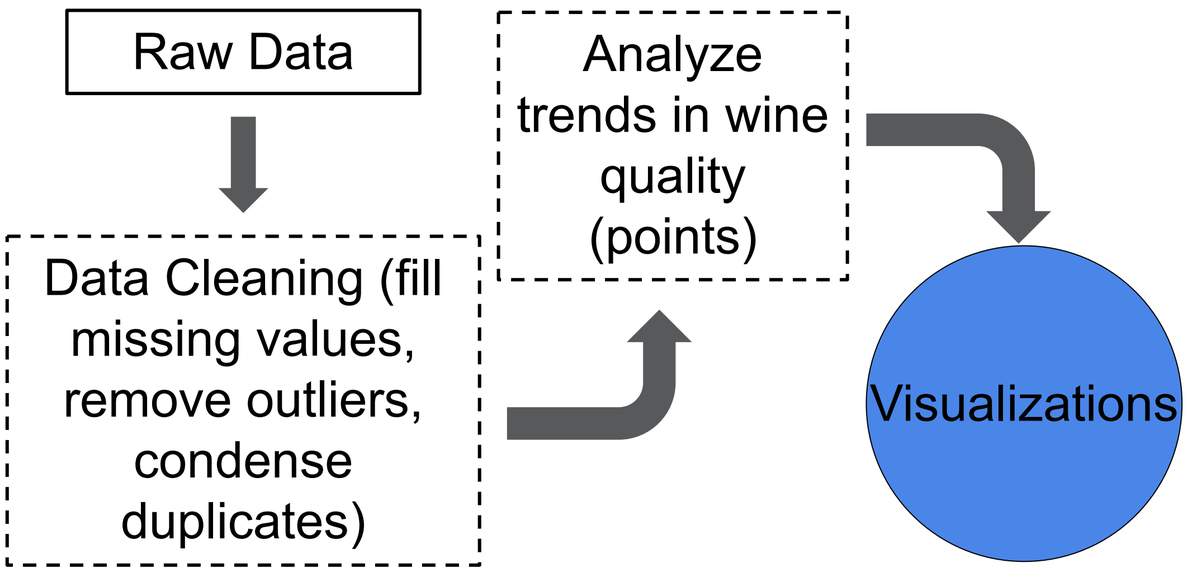
Data Analysis:

- Identify trends in data based on statistical methods.

Data Cleaning:

- Fill missing data points, condense duplicate data points, and remove outlying data points.

Workflow



MapReduce On Description Of WIne

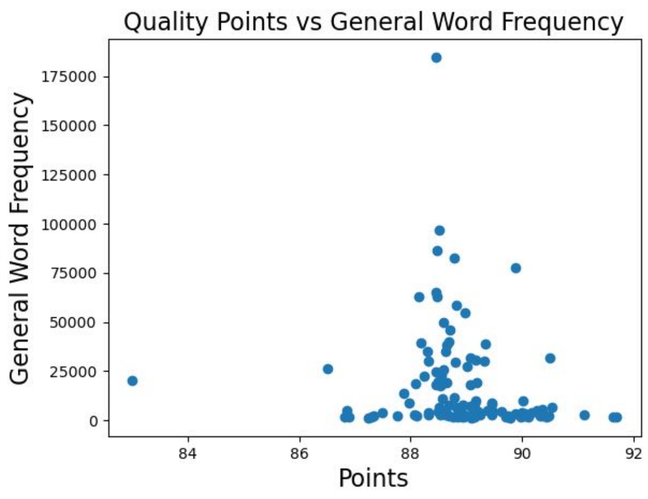
- **Preprocess :** Read the data (tabular)
- **Map :** Map each word by key value
- **Reduce :** Pair of frequency words and the word
- **Classify** : Classify words noun, verb, and adjective
- **Plot :** Plot the points and the words frequency

4 Most Common Adjective:

1. On : 46108
2. Ripe : 27383
3. Fresh: 17543
4. Soft : 13674

4 Most Common Verbs:

1. Is : 96848
2. Bodied : 11558
3. Long : 8606
4. Balanced: 8510



Quality and Price Trends Through Pandas

- Averaged the points of each country of origin to determine which produced the highest quality wines.

Top 5 Countries

```
Number of unique countries: 44
country
England     91.581081
India       90.222222
Austria     90.101345
Germany     89.851732
Canada      89.369650
```

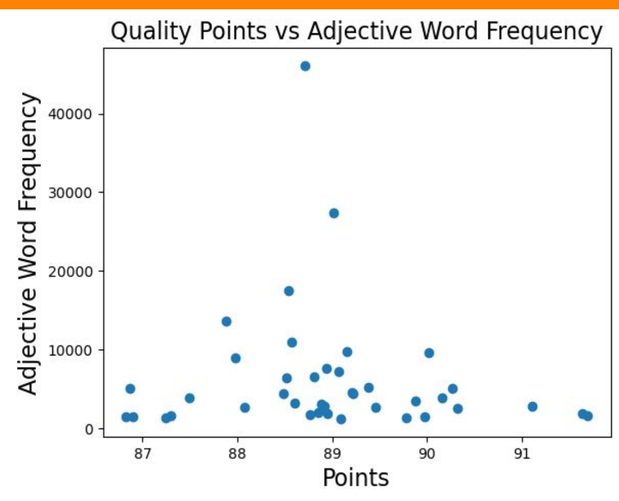
- Created a quality-to-price ratio for each unique wine to identify any potential trends

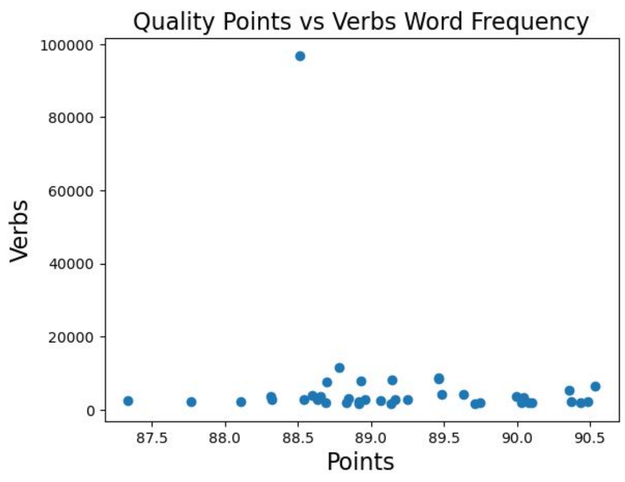
Results

Mean & Standard Deviation

General : **Mean :** 88.91
Standard Deviation : 1.07
Adjective: **Mean :** 88.96
Standard Deviation : 1.17
Verb: **Mean :** 89.20
Standard Deviation : 0.78

	80-90	90-100
count	87260.000000	13632.000000
mean	26.868356	36.152867
std	16.554934	19.963320
min	4.000000	8.000000
25%	15.000000	22.000000
50%	22.000000	32.000000
75%	35.363389	44.000000
max	155.000000	155.000000





- Higher rated wines (90-100 points) are slightly more expensive in general than lower rated wines (80-90).
- 15 counties were greater than average, and only one was greater than one Std. deviation above average.
- Generally the mean points did not differ significantly based on the adjective and verb usage

Conclusion and Future Work

- Verbs had a slightly higher mean but it is not statistically significant. Wine description does provide a robust conclusion about wine quality
- The price of entry into exceptional wine is slightly higher, but price is not indicative of quality.
- Develop a better process to classify non-english words as there were many Italian and French words that could not be classified.

THE UNIVERSITY OF TENNESSEE KNOXVILLE

This project was created for the **Introduction Data Mining and Analytics (COSC 426) taught at** The University of Tennessee Knoxville. Thank you to Dr. **Michela Taufer** for her instruction and support throughout the course, enabling this project.

Dataset used: https://www.kaggle.com/datasets/zynicide/wine-reviews