Analysis on Wine Dataset for wine classification

Prepared By: Manisha Khatri

Under the guidance of:

Dr. Vineetha Menon

CS588:Intro to Big Data Computing

Agenda:

- Introduction
- Dataset Description
- Data visualization
- Data preprocessing
- Discussion of methods
- Results and analysis
- Conclusion

Introduction

- There are more than a 1000 varieties
- A million reviews
- And guess what, they are pricy too!
- How do we choose?
- Big data to rescue!

Dataset Description

Columns

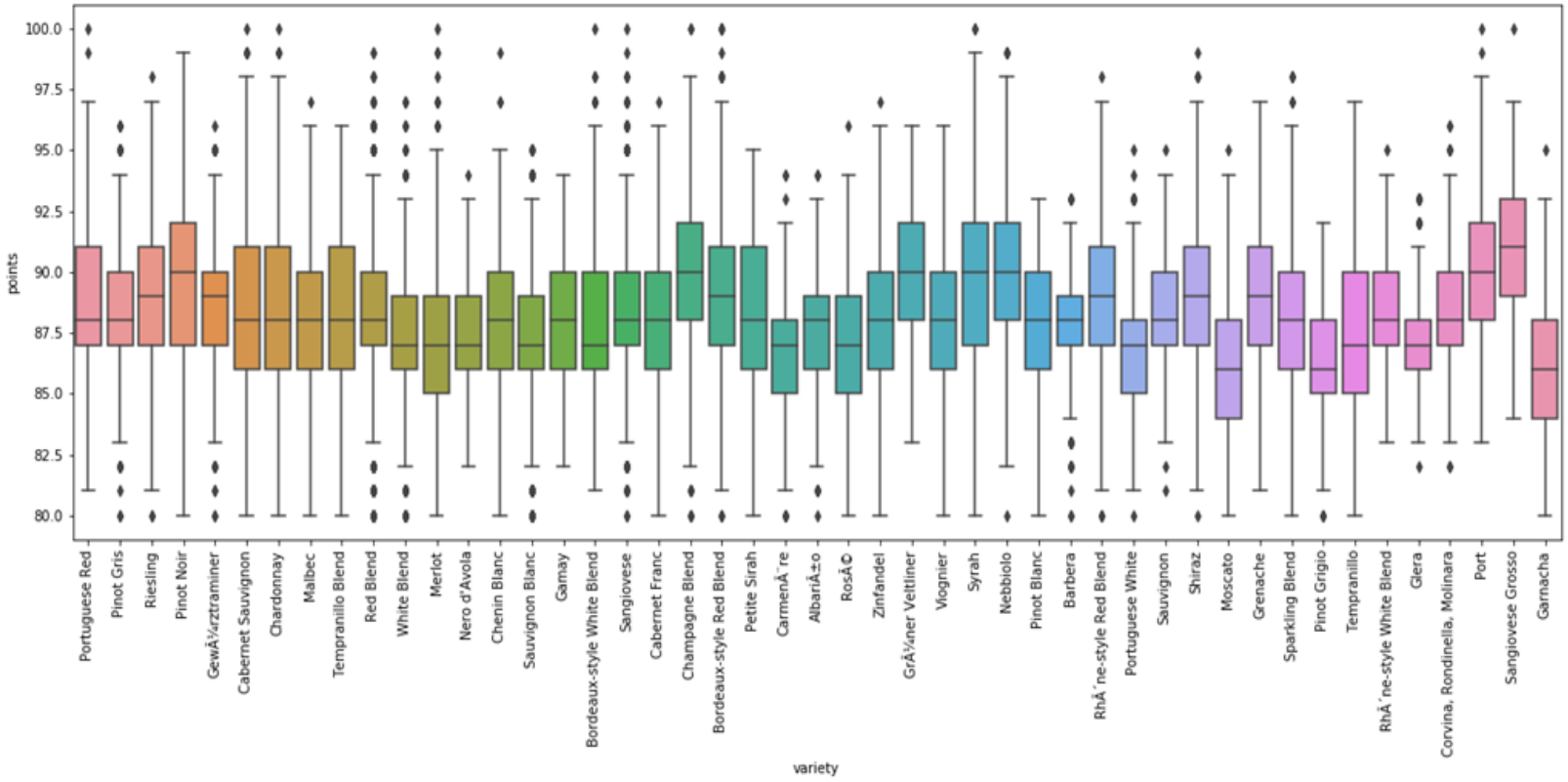
#	
A	country The country that the wine is from
A	description
A	designation The vineyard within the winery where the grapes that made the wine are from
#	points The number of points WineEnthusiast rated the wine on a scale of 1-100 (though they say they only post reviews for wines that score >=80)
#	price The cost for a bottle of the wine
A	province The province or state that the wine is from
A	region_1 The wine growing area in a province or state (ie Napa)
A	region_2 Sometimes there are more specific regions specified within a wine growing area (ie Rutherford inside the Napa Valley), but this value can sometimes be blank
A	taster_name
A	taster_twitter_handle
A	title The title of the wine review, which often contains the vintage if you're interested in extracting that feature
A	variety The type of grapes used to make the wine (ie Pinot Noir)
A	winery The winery that made the wine

Preprocessing

Features
- Description
- Points
- country
- variety
- title

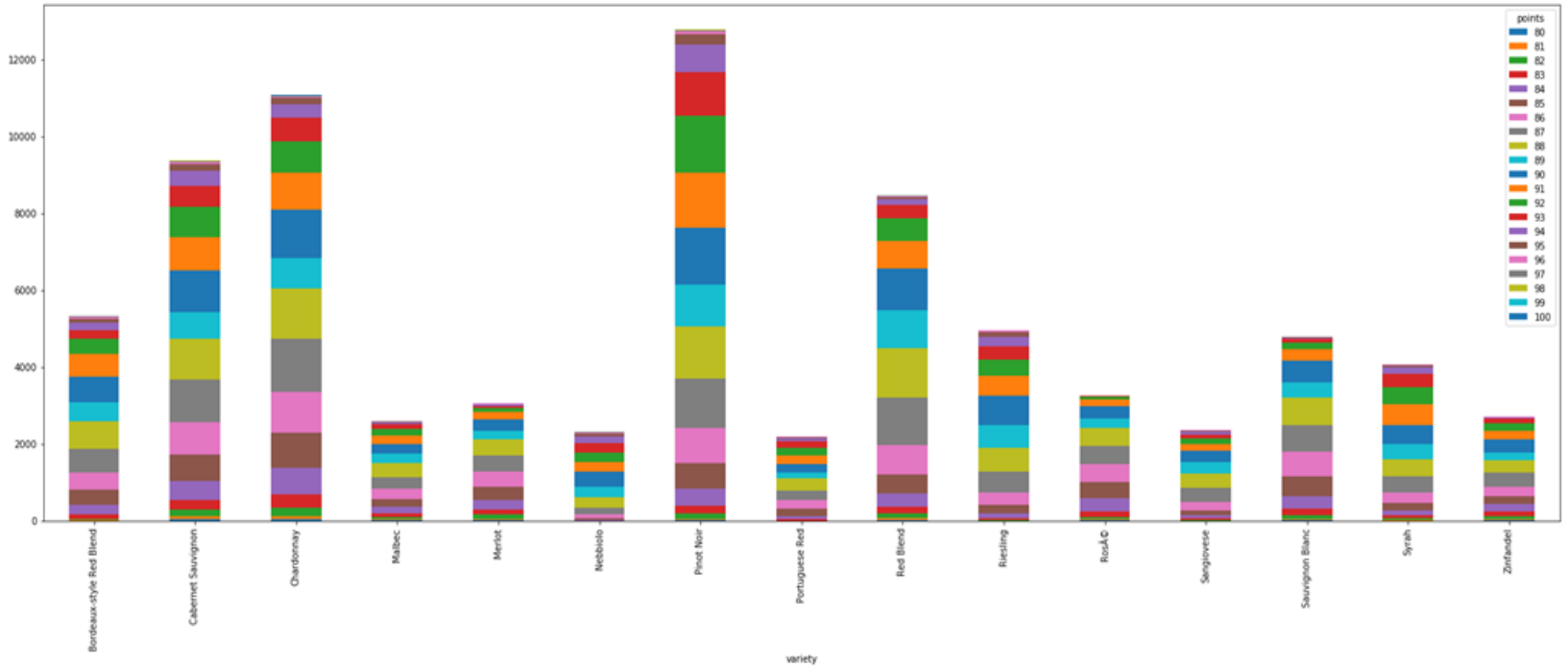
Source:-Wine Reviews https://www.kaggle.com/zynicide/wine-reviews

Data Visualizations



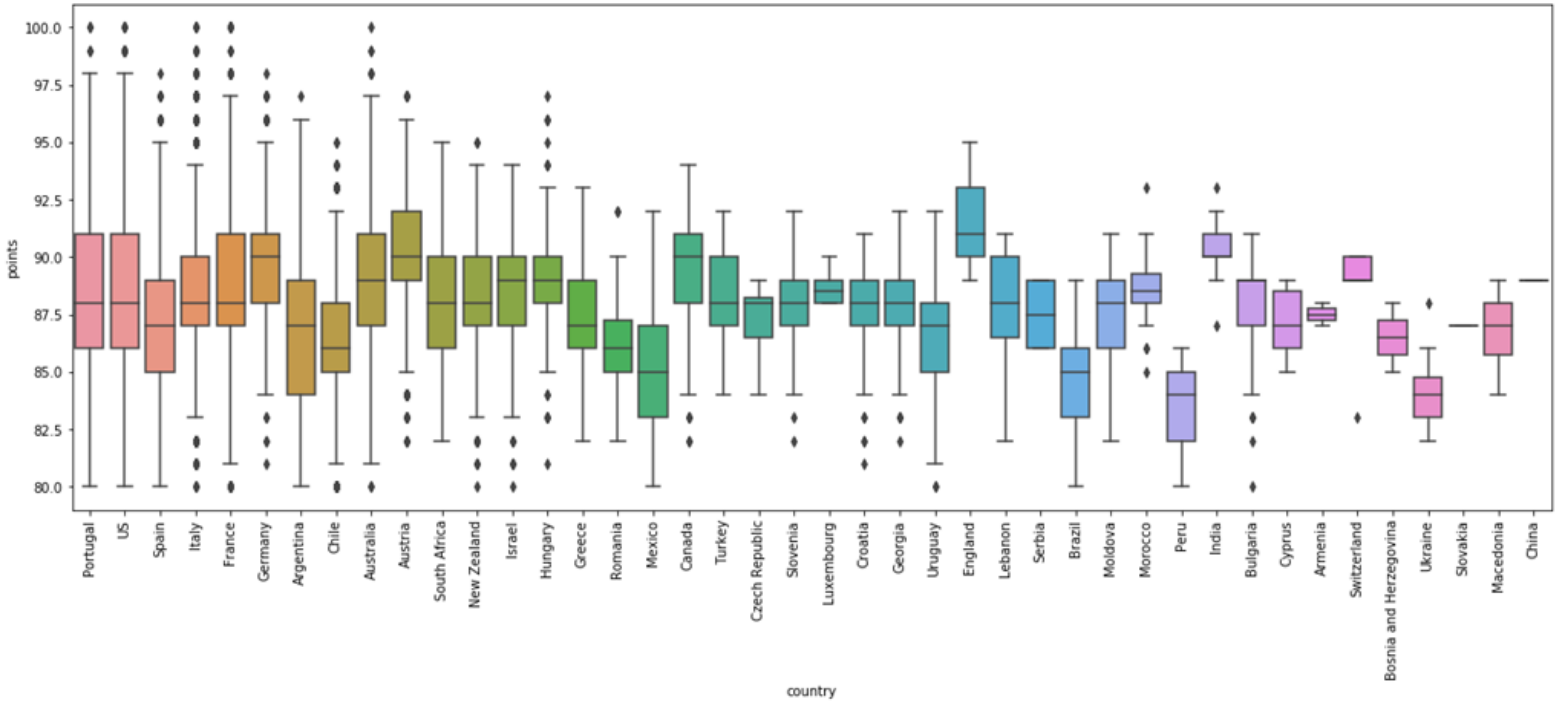
Variety Vs Points

Data Visualizations



Variety Vs Points

Data Visualizations



Country Vs Points

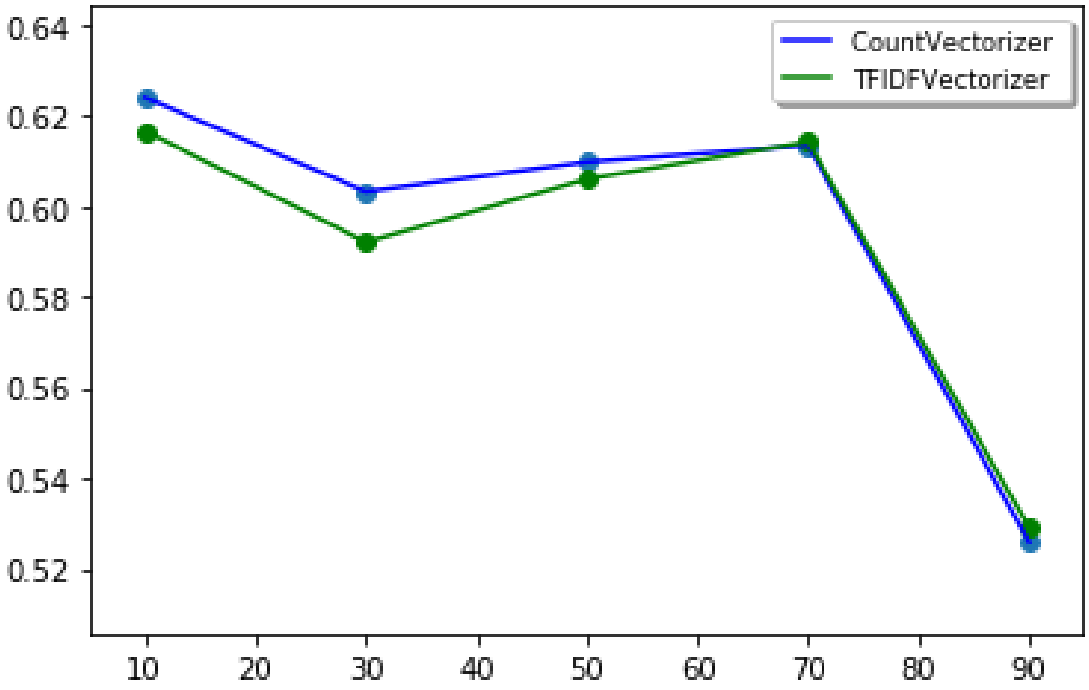
Data Preprocessing

- Why?
  - Real-world data can have missing values for important attributes, contain outliers or invalid data that can skew the results
  - Make it compatible for big data processing
- Steps :
  - Remove missing/duplicate values
  - Remove records with special characters
  - The countries with more than 3764 records and varieties with more than 5000 are considered
  - Create 4 class labels from points(“Bad”, “Good”, “Better”, “Best”)
  - Vectorize textual data

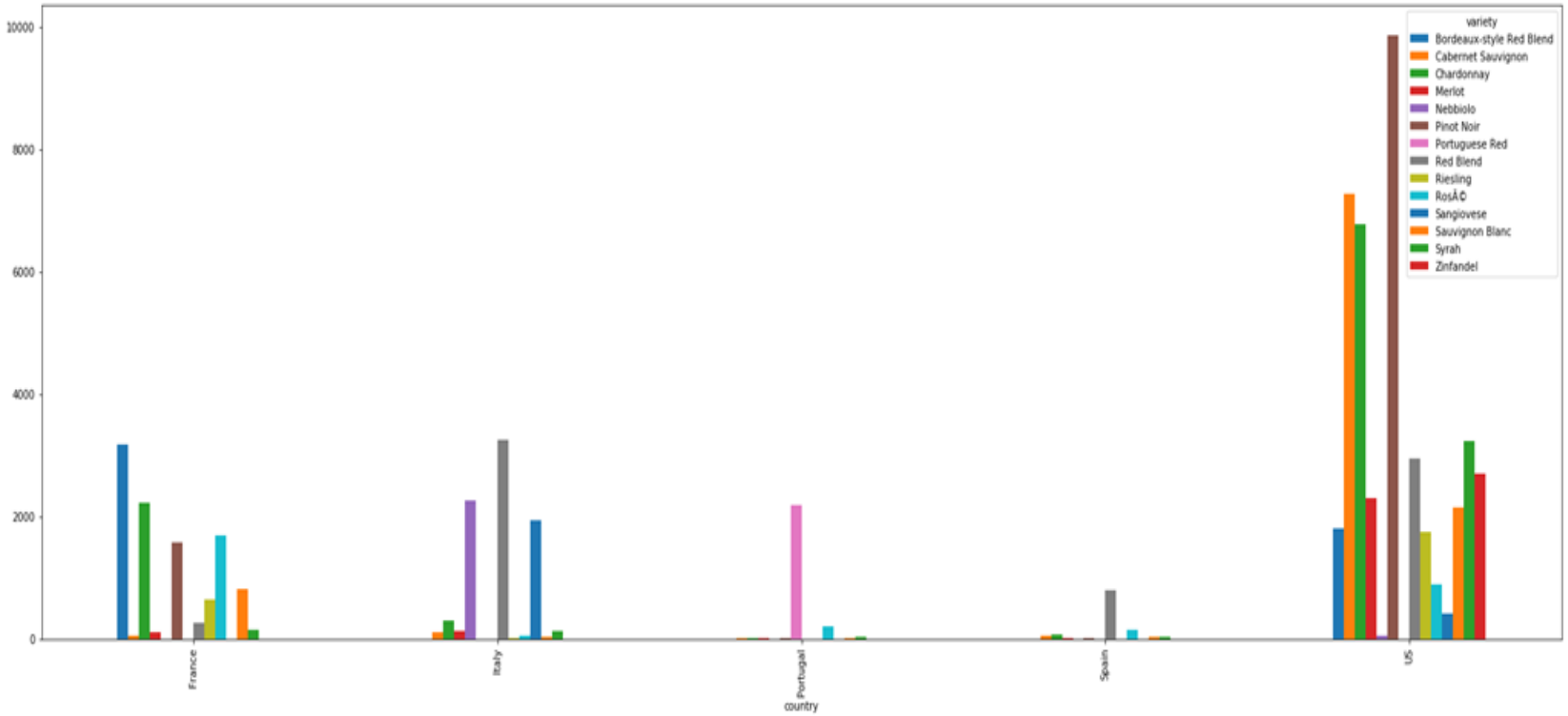
COUNTVECTORIZER VS TFIDFVECTORIZER

Countvectorizer: a simple way to both tokenize a collection of text documents and build a vocabulary of known words, but also to encode new documents using that vocabulary It counts the word frequencies in a document

TFIDFVectorizer: The value increases proportionally to count, but is offset by the frequency of the word in the corpus



Data Visualization



Country Vs Variety

Discussion of methods

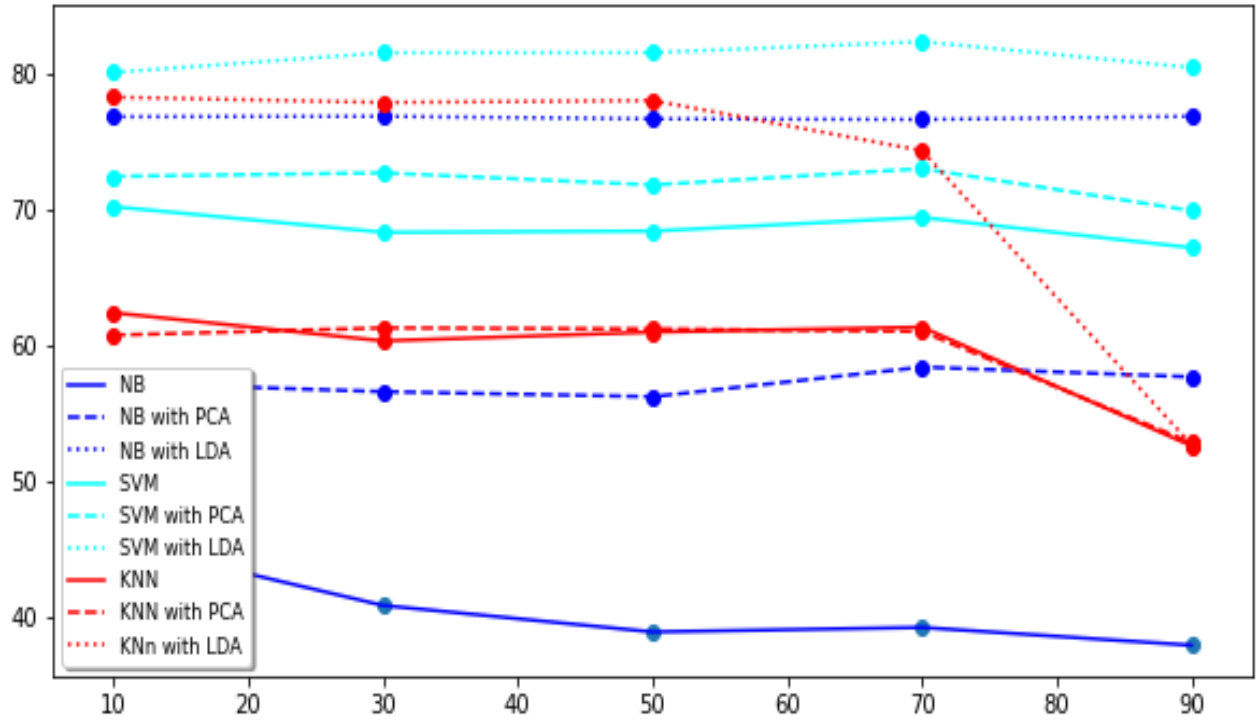
- **Dimensionality reduction:**
  - **PCA:** Find the directions of maximum variance in high-dimensional data and projects it onto a new subspace with equal or fewer dimensions than the original one
  - **LDA:** It finds a new feature space to project the data in order to maximize class separability and minimizes inter-class variability

Discussion of methods

- **Classification Algorithms:**
  - **K-Nearest Neighbor:** KNN (K-Nearest Neighbor) is a simple supervised classification algorithm that implements the k-nearest neighbors vote
  - **Support Vector machines:** The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space(N — the number of features) that distinctly classifies the data points.
  - **Naive Bayes classification:** It predicts the probability of each class based on the feature vector for text classification for continuous big data with a prior distribution of the probability

Results and analysis

- PCA: Explained variance ratio (first two components):- 9.91225863e-01, 7.93060405e-04



Conclusion

- Better performance after dimensionality reduction
- Performances are comparable for Naive bayes and KNN
- SVM gives the highest accuracy of classification of wine


Now You can sit and
enjoy while Big data
finds the best wine for
you!!
Cheers☺