GR5243 APPLIED DATA SCIENCE

Data Analysis of Wines Reviews

Group 5

Jianping Mu
Nannan Wang
Qingyang Zhong
Han Gao

Contents

1. Basic Introduction

(1)Wine And Wine market

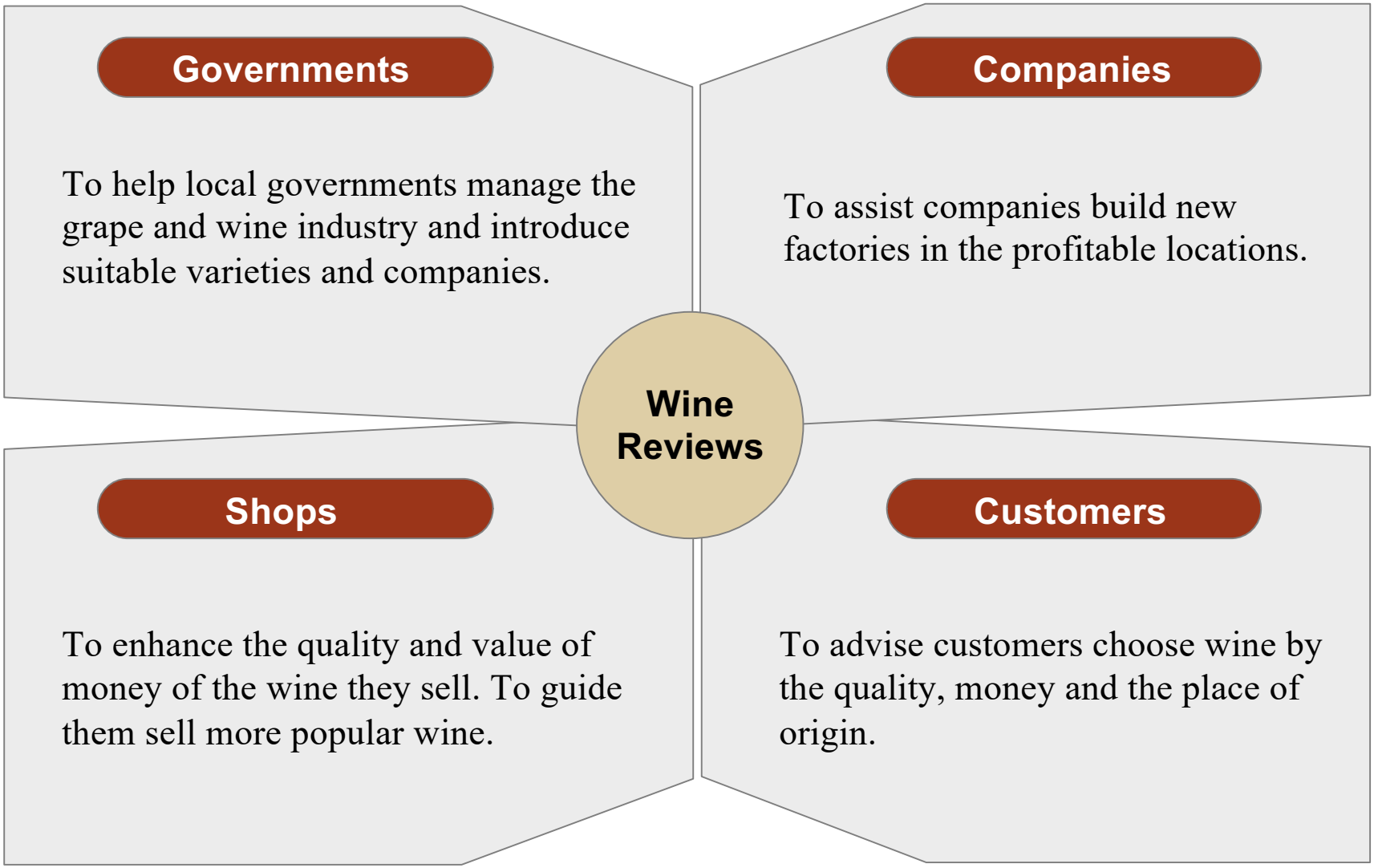
Wine is an alcoholic beverage made with the fermented juice of grapes, but they are different than what you'll find at the grocery store. Technically, wine can be made with any fruit (i.e. apples, cranberries, plums, etc) but most wines are made with wine grapes.

The quality of grapes decides the quality and price of wine.

Wine market is huge and growing steadily.U.S. consumers quaffed $32 billion worth of wine in 2017, and that figure is expected to reach $43 billion by 2022, an annual growth rate of more than 6%.Premium is the place to be, the "fine and premium" category (over $10 a bottle) has been almost bubbly.

1. Basic Introduction

(2). Target of our research



2.Descriptive Statistics

(1). Data preview

	country	description	points	price	province	variety	winery
0	US	This tremendous 100% varietal wine hails from ...	96	235	California	Cabernet Sauvignon	Heitz
1	Spain	Ripe aromas of fig, blackberry and cassis are ...	96	110	Northern Spain	Tinta de Toro	Bodega Carmen Rodríguez
2	US	Mac Watson honors the memory of a wine once ma...	96	90	California	Sauvignon Blanc	Macauley
3	US	This spent 20 months in 30% new French oak, an...	96	65	Oregon	Pinot Noir	Ponzi
4	France	This is the top wine from La Bégude, named aft...	95	66	Provence	Provence red blend	Domaine de la Bégude
5	Spain	Deep, dense and pure from the opening bell, th...	95	73	Northern Spain	Tinta de Toro	Numanthia
6	Spain	Slightly gritty black-fruit aromas include a s...	95	65	Northern Spain	Tinta de Toro	Maurodos
7	Spain	Lush cedary black-fruit aromas are luxe and of...	95	110	Northern Spain	Tinta de Toro	Bodega Carmen Rodríguez
8	US	This re-named vineyard was formerly bottled as...	95	65	Oregon	Pinot Noir	Bergström
9	US	The producer sources from two blocks of the vi...	95	60	California	Pinot Noir	Blue Farm

The data are collected from Kaggle, after we removing all NA values we finally get 258144 data.

2.Descriptive Statistics

(2). Wordclouds of 3 price levels

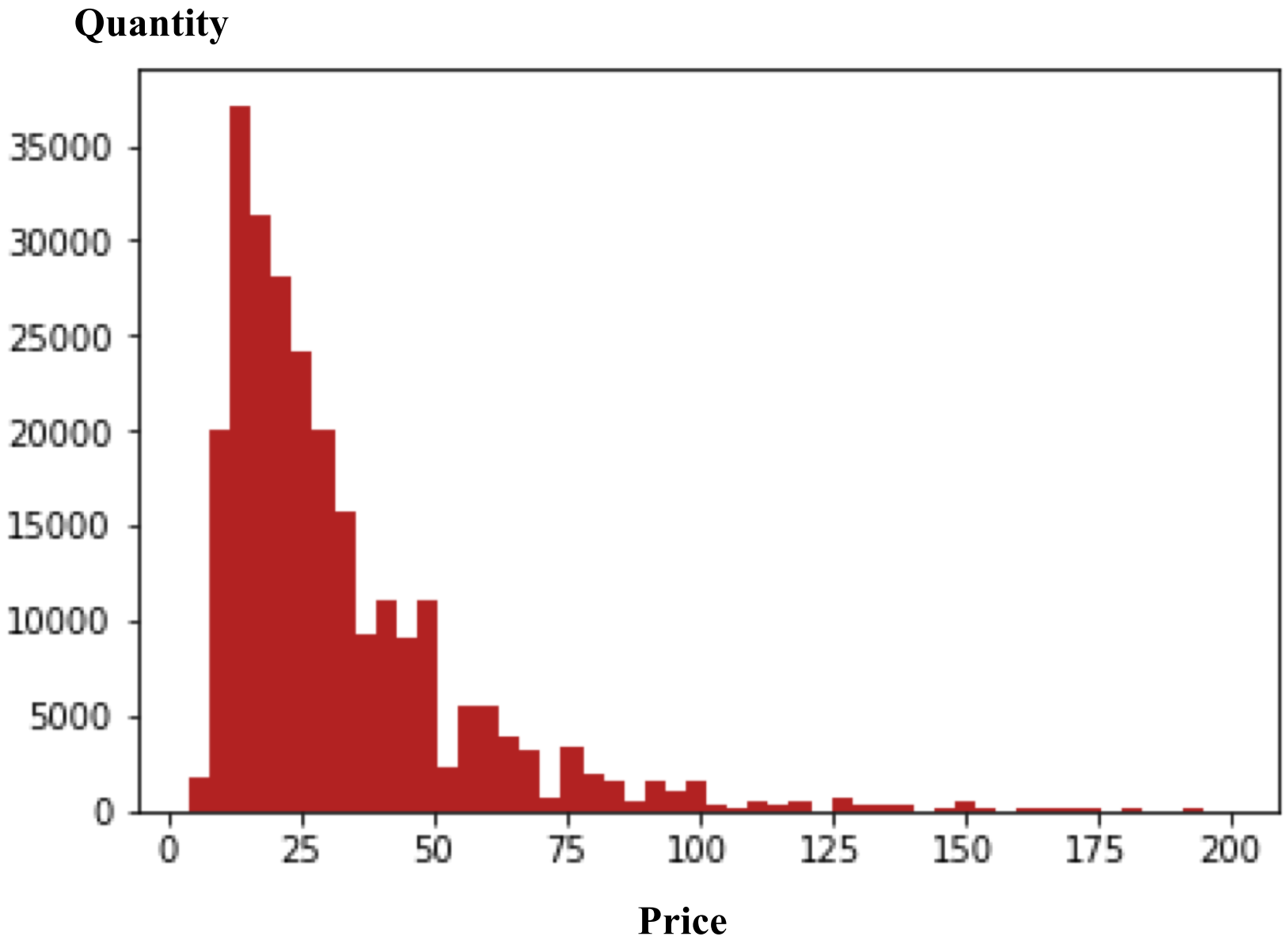
Low Price

Medium Price

High Price

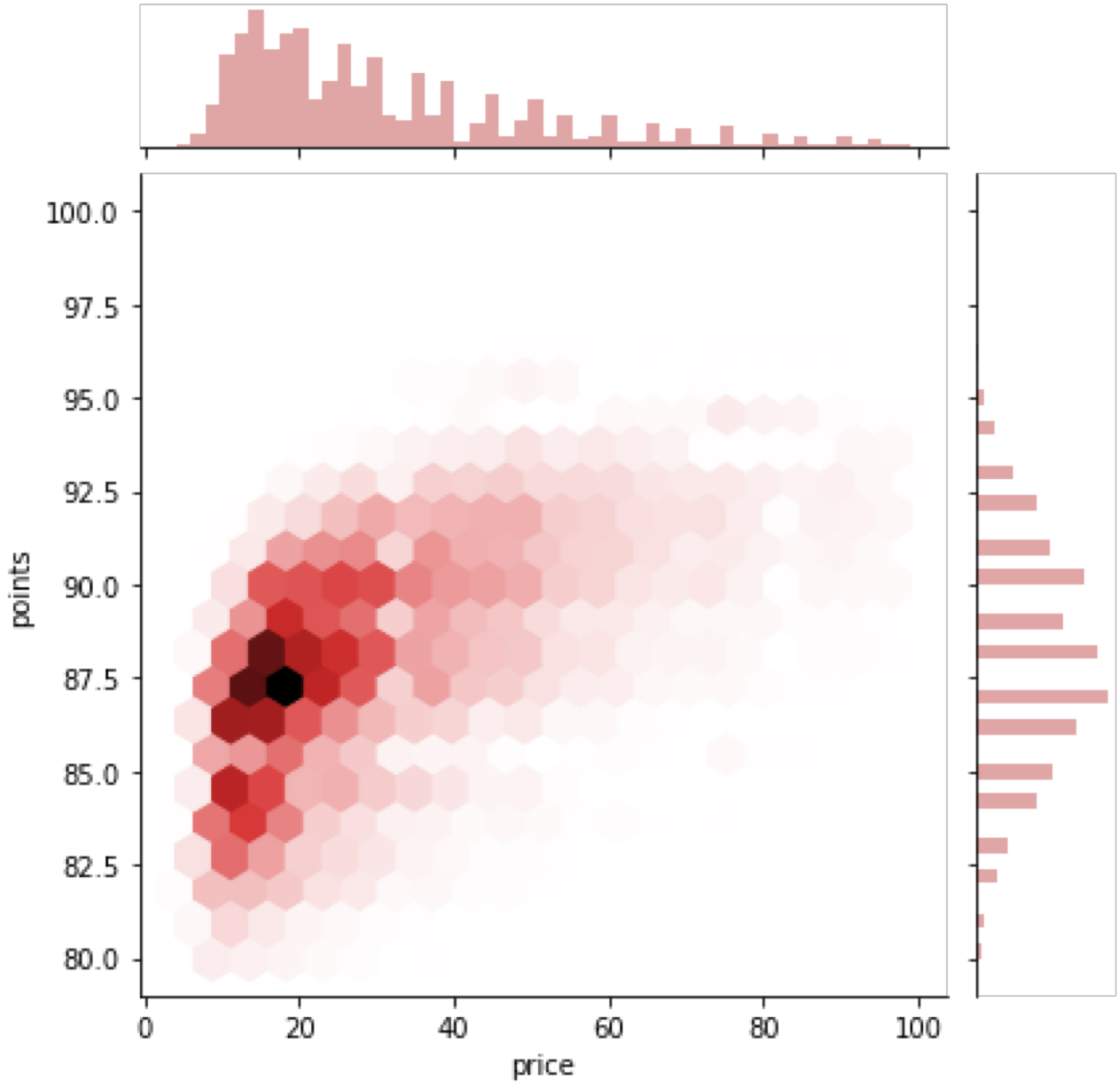
2.Descriptive Statistics

(3). Distribution of wine price



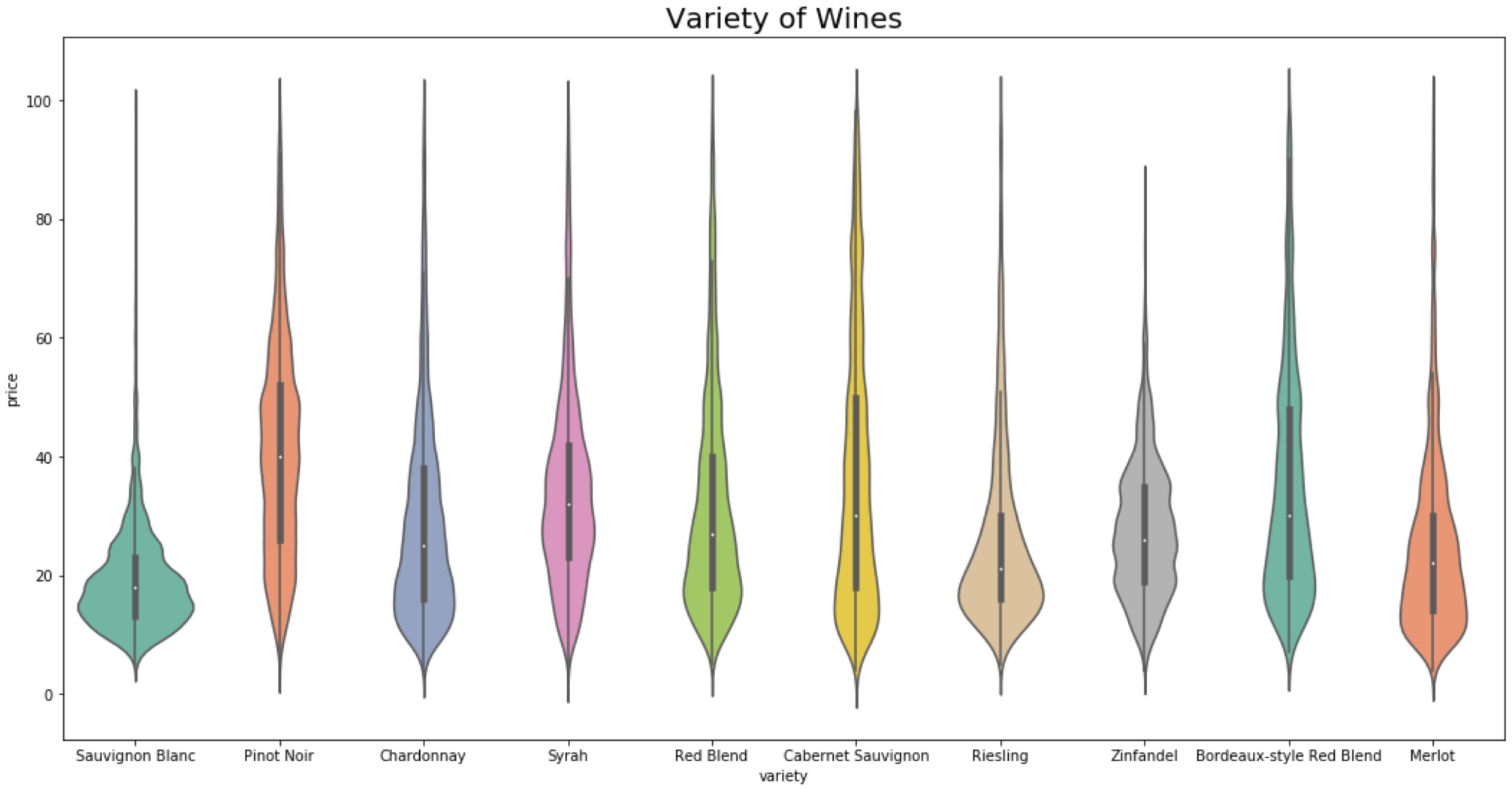
2.Descriptive Statistics

(4). Distribution of wine price & points



2.Descriptive Statistics

(5). Price of wine by variety

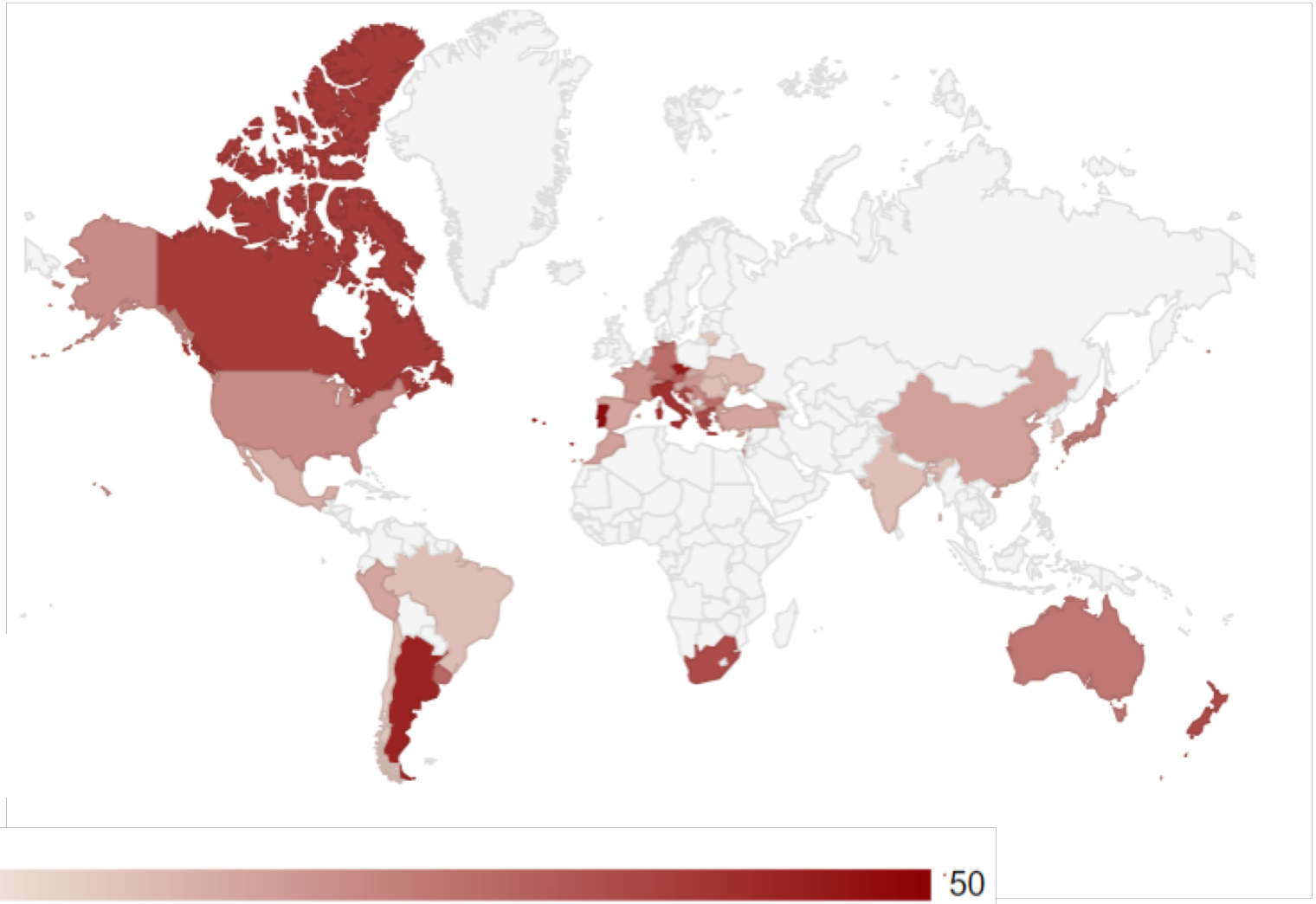


2.Descriptive Statistics

(6) Price of wine by country

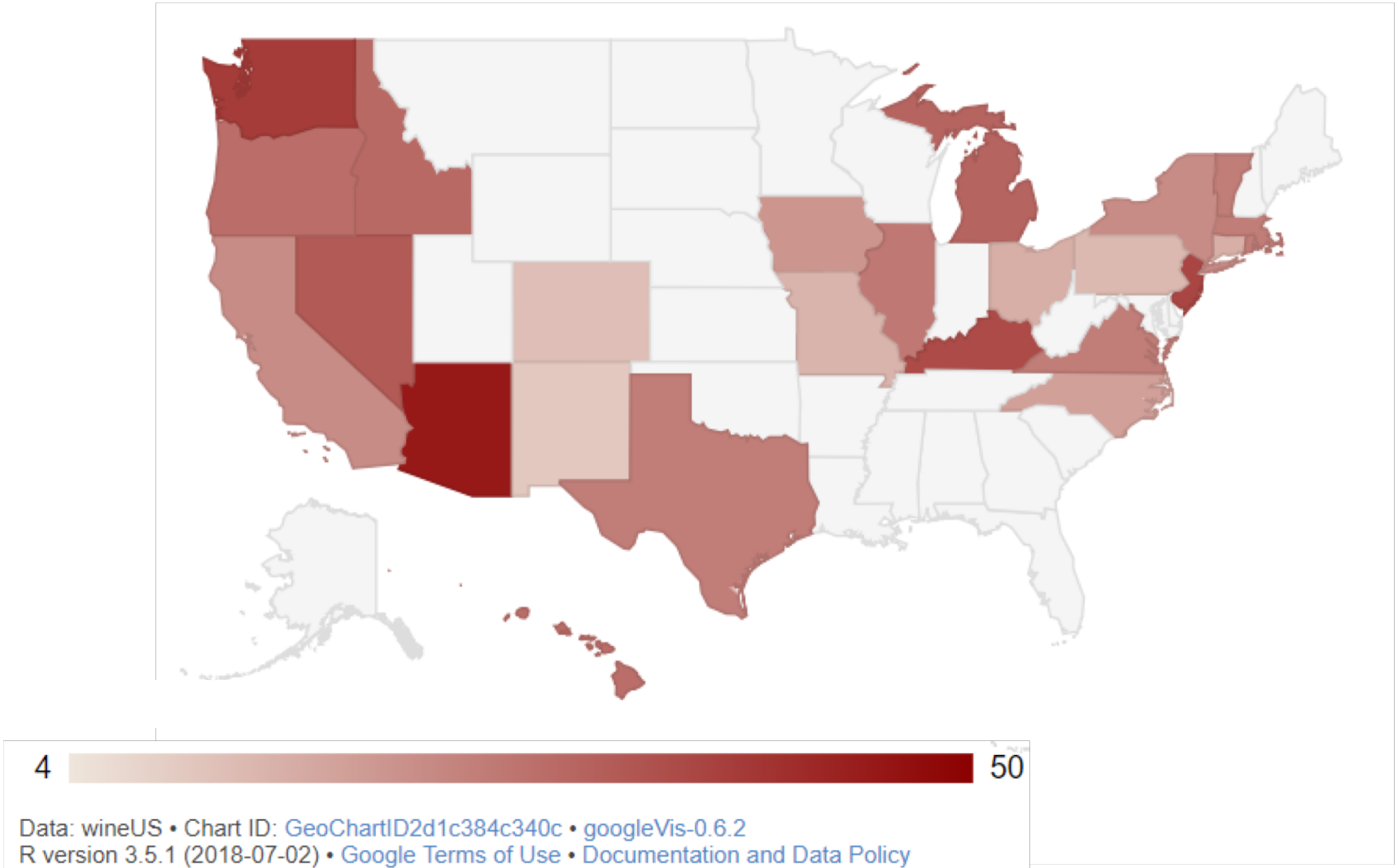
2.Descriptive Statistics

(9) Price shown in Worldmap



2.Descriptive Statistics

(10) Price shown in USmap



2.Descriptive Statistics

(11) Data Processing (Text ⟹ Number)

Hashing Vectorizer

	A	B	C	D	E	F	G	H	I	J	K	L
1	feature1	feature2	feature3	feature4	feature5		feature20	feature21	feature22	feature23	feature24	price
2	-0.1066	0	0.1066	0.1066	-0.4264		0.1066	-0.57735	-0.57735	0	0.57735	235
3	-0.12039	0	0.36116	0.24077	-0.24077		0	-0.37139	-0.55709	0.74278	0	110
4	0	0.212	0	0.212	0		0	-0.57735	0	0.57735	-0.57735	90
5	0	0.26941	0.0898	0	-0.17961	...	-0.0898	0	0	-1	0	65
6	-0.49614	-0.12403	0.12403	-0.24807	0		0	0	0	0.5547	-0.83205	66
7	0	0.12599	0.37796	0.12599	-0.37796		0.12599	-0.22942	-0.68825	0.68825	0	73
8	0.11785	0.2357	0.11785	0	-0.11785		0.11785	0	-0.5547	0.83205	0	65
9	0	0	0.50508	0	-0.20203		0.10102	-0.37139	-0.55709	0.74278	0	110
10												
11			...						...			
12	0.17408	0.17408	0.34816	0	0		0	0	0	-1	0	65
13	-0.23094	0.11547	0.34641	0	-0.34641		-0.11547	0	-0.8165	-0.40825	-0.40825	60
14	-0.10483	-0.10483	0.31449	0.10483	-0.10483		-0.31449	0	1	0	0	80
15	0	0	0.254	-0.381	0.127	...	0.254	0	0.57735	-0.57735	-0.57735	48
16	-0.19803	0.29704	0.29704	0	0.09901		0.09901	0	0.57735	-0.57735	-0.57735	48
17	-0.09245	0	0.3698	-0.09245	0.09245		0	-0.40825	-0.40825	0	0.8165	90
18	-0.18334	0.09167	0.18334	0.18334	-0.45835		-0.09167	0	0	-1	0	185

Description: feature1 - feature 20

Country/Province/Variety/Winery: feature 21 - feature 24

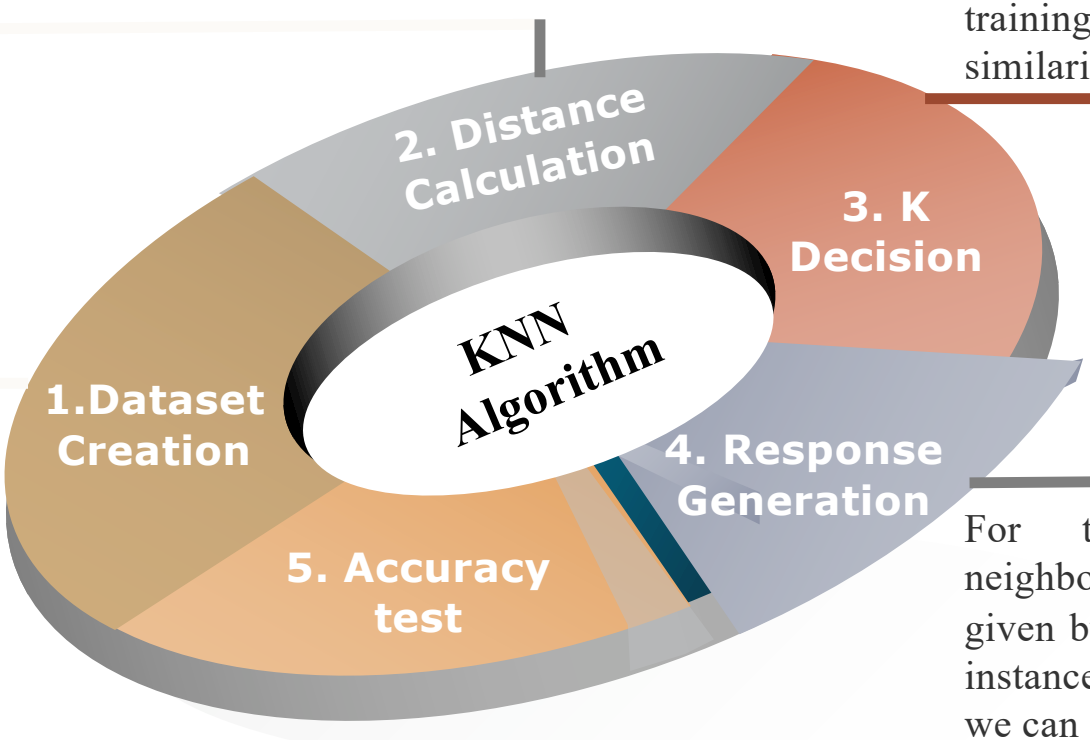
3.Algorithm

(1).K-nearest neighbors algorithm

Calculate the distance between two data instances within the training set, using Euclidean distance measure

Locate k most similar data instances in the training set with the similarity measures

Split the data into a training set (to make predictions) and a test set (to evaluate model accuracy)



For the most similar neighbors of a test instance given by k most similar data instances in the training set, we can generate a response

Test the accuracy of predictions

KNN method has to run every data in the dataset, so it is very time consuming.

3.Algorithm

(1).K-nearest neighbors algorithm

```
from sklearn.neighbors import KNeighborsClassifier
classifier = KNeighborsClassifier(n_neighbors=3)
classifier.fit(X_train, y_train)
```

```
KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski',
           metric_params=None, n_jobs=1, n_neighbors=3, p=2,
           weights='uniform')
```

Take the response of the two nearest points for each predictor and compare it with the test set.

Output Result

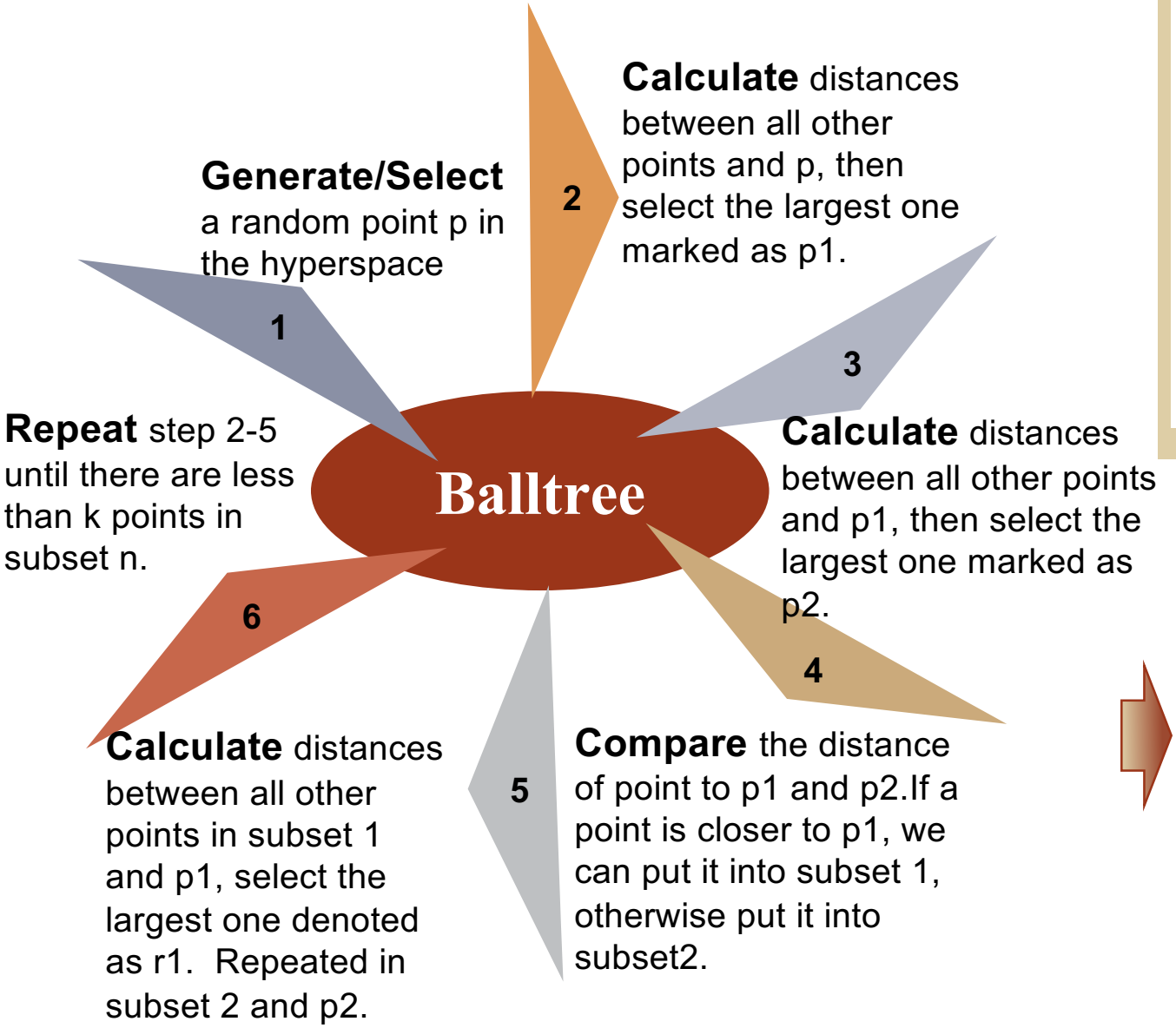
	precision	recall	f1-score	support
4	0.42	0.67	0.52	12
5	0.18	0.34	0.24	50
6	0.19	0.44	0.27	139
7	0.24	0.38	0.29	424
8	0.25	0.44	0.32	838
9	0.25	0.44	0.32	1248
avg / total	0.28	0.27	0.27	85188

When the KNN model is applied to the test set, the average precision is only 28%.

The relevance between features is not high, resulting in poor adaptability of KNN and 28% average precision.

3.Algorithm

(2). Balltree



- Advantage: Higher running speed than other algorithm
- Disadvantage: Calculating and understanding the meaning are not easy.

Finally we will get a tree with many leaves or external nodes.

3.Algorithm

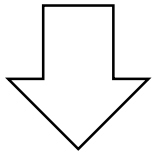
(2). Balltree

We classify the data into four classes through balltree.

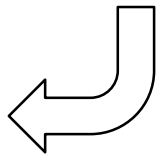
Code

The quantities of clusters is depend on ourselves, here is four. Considering the factors like description, combined of others, price, origin, grape varieties etc.

```
tree = BallTree(data, leaf_size = 4)
dist, ind = tree.query(data[1:2], k = 65436)
print(ind)
```



```
[[     1 257857  89109 ...  40706 178640 191598]]
```



By calculating **Euclidean distance** between Wine No.1 and other wines, it reaches the lowest value between itself, and second lowest value to Wine No.257857, which means Wine No.1 is most similar to Wine No.257857 except itself.

Output

```
0.8426659435588526
Time used: 1274.5424852640008 seconds!
```

4. Conclusion

1. Different price levels have different features with respect to target customers
2. American Wines are expensive!
3. Arizona's wines are generally more expensive than the other places
4. KNN did poor on classifying the groups of wines
5. Balltree performs good, it is hard to understand the meaning of each leaf of the result of Balltree Algorithm

Thanks !

Q&A