Angelo, Dumaliang, Gambao, Tan, Villanueva
CS 129.1 - A

Big Data Final Project

Big data problem:

- Given the list of wine reviews globally, how do we determine which country offers the **best** wine?

Description of data source

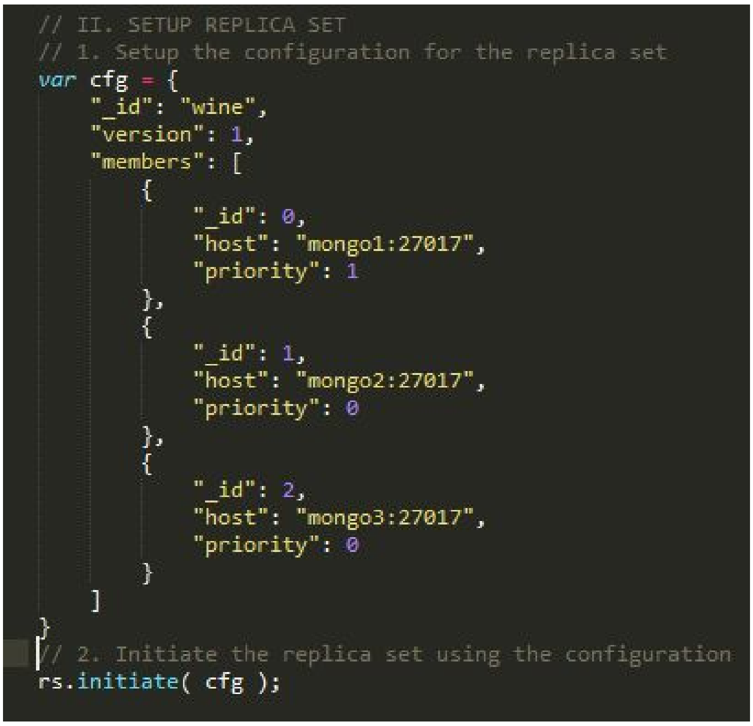
The dataset to be used for the project will come from Kaggle. The list of wine reviews enumerates 150,930 different and unique wines. The said list includes the wine's country of origin, description, designation, points, price, region, variety, and winery. In order to determine which countries offer the best wine, the following **positive** keywords will be used.

- Juicy
- Rich
- Ripe
- Delicious
- Balanced
- Fresh
- Impressive
- Smooth
- Silky
- Opulent
- Beautiful
- Gorgeous
- Wonderful
- Exceptional
- Polished
- Top-notch
- Sophisticated
- Perfect
- Decadent
- Remarkable
- Oaky
- Fruity
- Crisp
- Soft
- Noble

Procedure on how to obtain the dataset

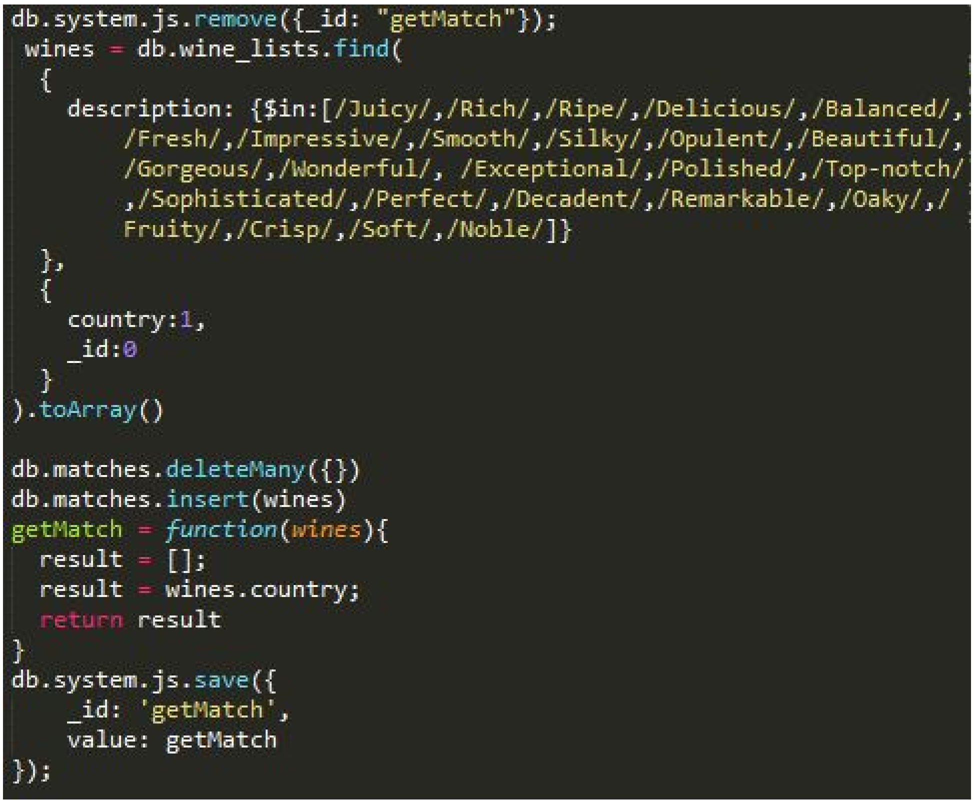
1. Go to https://www.kaggle.com/zynicide/wine-reviews
2. Download the JSON file readily available.

Replicate Sets Configuration

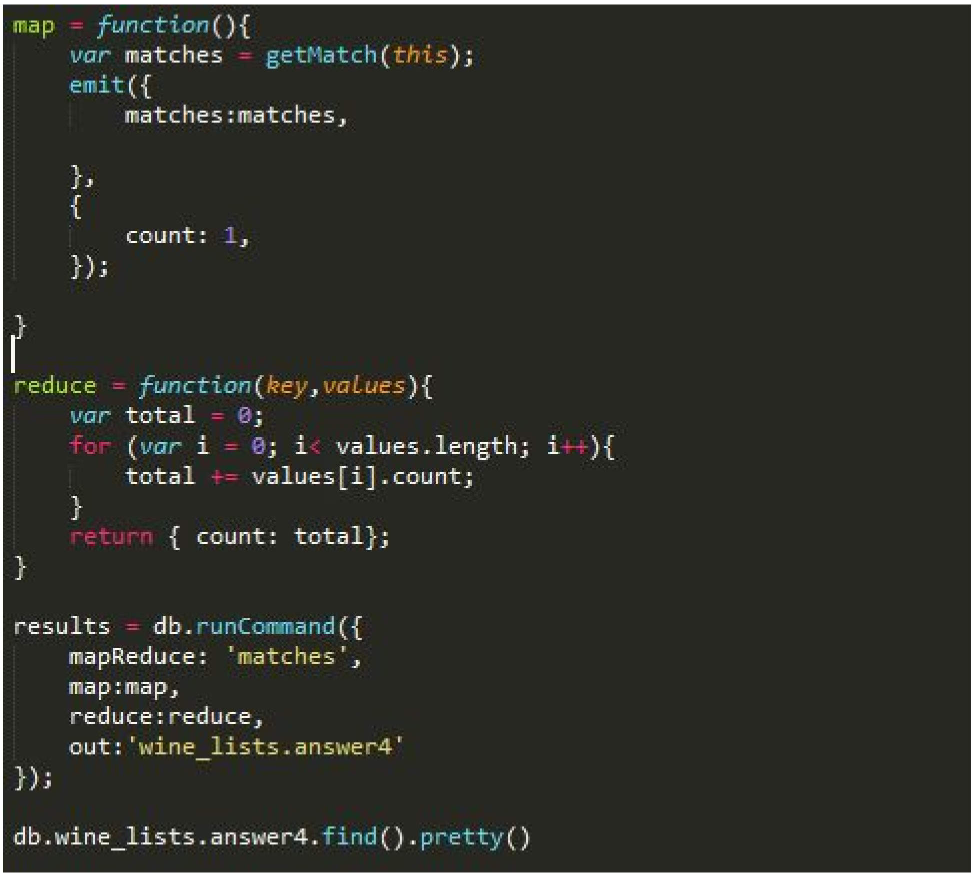


```
// II. SETUP REPLICA SET
// 1. Setup the configuration for the replica set
var cfg = {
    "_id": "wine",
    "version": 1,
    "members": [
        {
            "_id": 0,
            "host": "mongo1:27017",
            "priority": 1
        },
        {
            "_id": 1,
            "host": "mongo2:27017",
            "priority": 0
        },
        {
            "_id": 2,
            "host": "mongo3:27017",
            "priority": 0
        }
    ]
}
// 2. Initiate the replica set using the configuration
rs.initiate( cfg );
```

MapReduce Functions

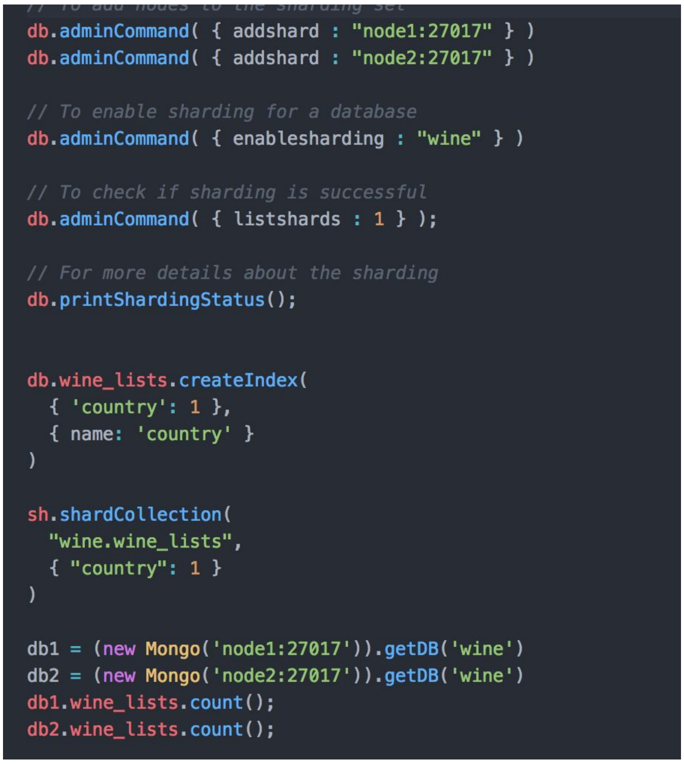


```
db.system.js.remove({_id: "getMatch"});
 wines = db.wine_lists.find(
  {
    description: {$in:[/Juicy/,/Rich/,/Ripe/,/Delicious/,/Balanced/,
        /Fresh/,/Impressive/,/Smooth/,/Silky/,/Opulent/,/Beautiful/,
        /Gorgeous/,/Wonderful/, /Exceptional/,/Polished/,/Top-notch/
        ,/Sophisticated/,/Perfect/,/Decadent/,/Remarkable/,/Oaky/,/
        Fruity/,/Crisp/,/Soft/,/Noble/]}
  },
  {
    country:1,
    _id:0
  }
).toArray()

db.matches.deleteMany({})
db.matches.insert(wines)
getMatch = function(wines){
  result = [];
  result = wines.country;
  return result
}
db.system.js.save({
    _id: 'getMatch',
    value: getMatch
});
```



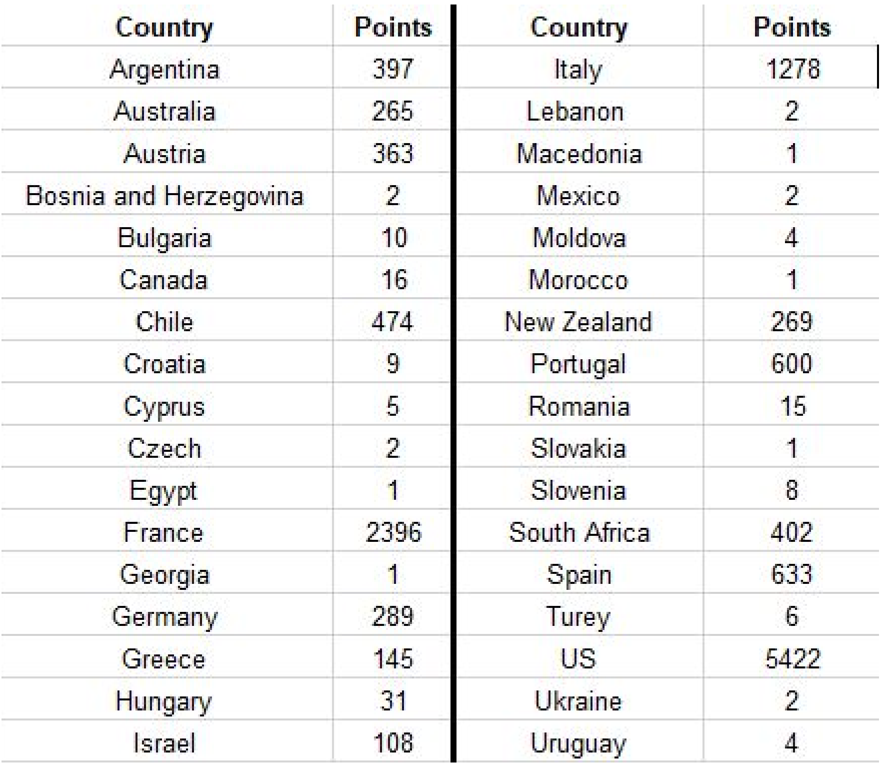
```
map = function(){
    var matches = getMatch(this);
    emit({
        matches:matches,

    },
    {
        count: 1,
    });

}
reduce = function(key,values){
    var total = 0;
    for (var i = 0; i< values.length; i++){
        total += values[i].count;
    }
    return { count: total};
}

results = db.runCommand({
    mapReduce: 'matches',
    map:map,
    reduce:reduce,
    out:'wine_lists.answer4'
});

db.wine_lists.answer4.find().pretty()
```

Sharding Set Configuration



```
// To add nodes to the sharding set
db.adminCommand( { addshard : "node1:27017" } )
db.adminCommand( { addshard : "node2:27017" } )

// To enable sharding for a database
db.adminCommand( { enablesharding : "wine" } )

// To check if sharding is successful
db.adminCommand( { listshards : 1 } );

// For more details about the sharding
db.printShardingStatus();


db.wine_lists.createIndex(
  { 'country': 1 },
  { name: 'country' }
)

sh.shardCollection(
  "wine.wine_lists",
  { "country": 1 }
)

db1 = (new Mongo('node1:27017')).getDB('wine')
db2 = (new Mongo('node2:27017')).getDB('wine')
db1.wine_lists.count();
db2.wine_lists.count();
```

Discussion of Results



Country	Points	Country	Points
Argentina	397	Italy	1278
Australia	265	Lebanon	2
Austria	363	Macedonia	1
Bosnia and Herzegovina	2	Mexico	2
Bulgaria	10	Moldova	4
Canada	16	Morocco	1
Chile	474	New Zealand	269
Croatia	9	Portugal	600
Cyprus	5	Romania	15
Czech	2	Slovakia	1
Egypt	1	Slovenia	8
France	2396	South Africa	402
Georgia	1	Spain	633
Germany	289	Turey	6
Greece	145	US	5422
Hungary	31	Ukraine	2
Israel	108	Uruguay	4

Table 1: Tally of Points Per Country

The table above shows the results after doing the MapReduce function. A point is counted for every instance of description with at least 1 of the aforementioned keywords. The country with the most number of points will be considered the country offering the best wine. For example, the review below gets a point given the presence of at least one of the keywords mentioned above.

> ***"Ripe** aromas of fig, blackberry and cassis are softened and sweetened by a slathering of **oaky** chocolate and vanilla. This is full, layered, intense and cushioned on the palate, with **rich** flavors of chocolaty black fruits and baking spices. A toasty, everlasting finish is heady but ideally **balanced.** Drink through 2023."*

With this, the country with the most number of points, therefore offering the best wine in this context, is USA having 5422 points. Following USA is France with 2396 points, and Italy with 1278 points.

Data Visualization

In order to give a visual understanding of the mapreduce results, data visualizations were made using Google sheets.

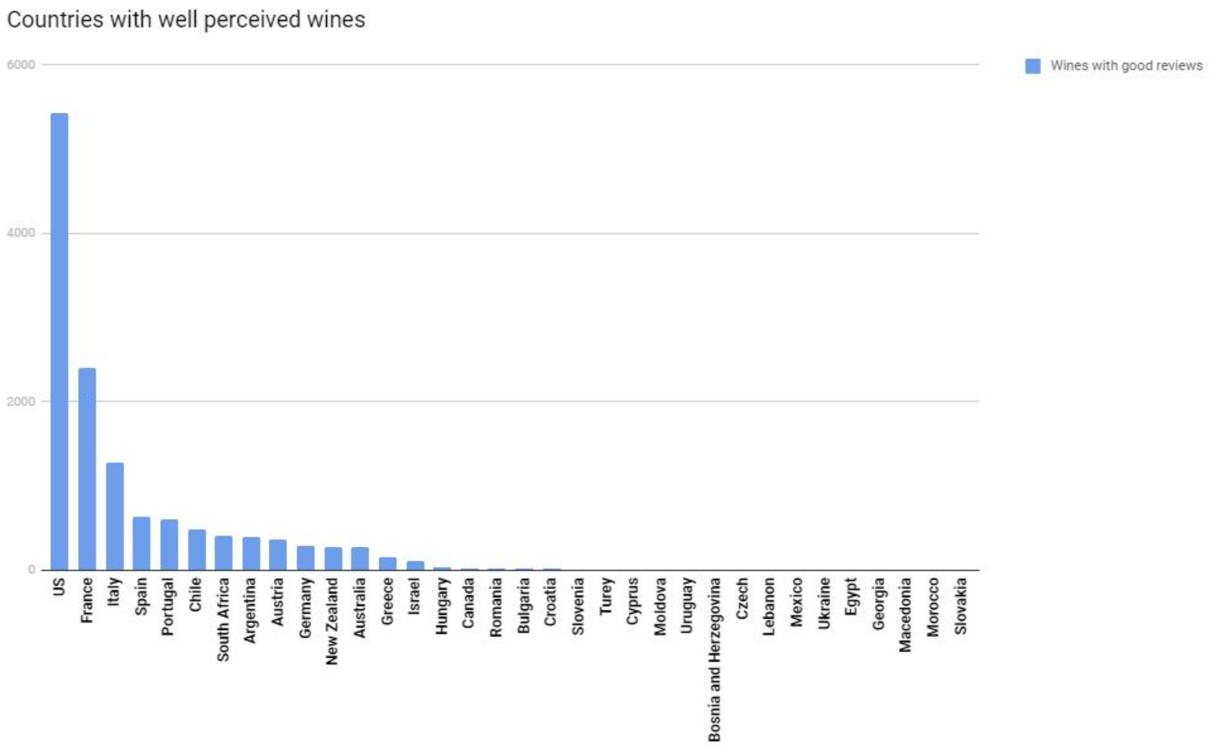


Figure 1.0 *Bar graph of countries with well perceived wines*

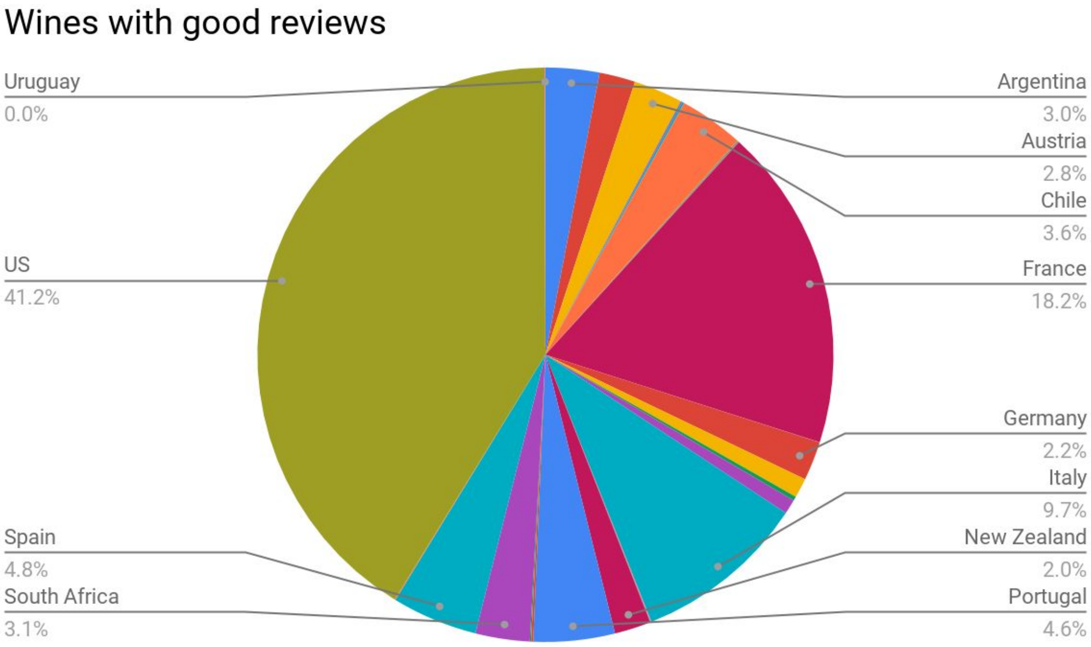


Figure 2.0 *Pie chart of countries with well perceived wines*