POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Systems and Methods for Big and Unstructured Data Project

Author(s): **Francesco Di Giore 10776567**
Federico De Introna 10796946

Group Number: **1**

Academic Year: 2023-2024

Contents

1 | Introduction

1.1. Problem Specification

The purpose of this project is to design, store and query data on a NoSQL database among the three main types we have studied during the course: Graph, Document and Information Retrieval databases.
The open dataset we decided to load into our database contains information about different types of wines, it includes a lot of data on specific wines, characterised by the winery, the variety of the wine, the price, and so on; all the features will be described and explained in the following sections. In particular, all the wines have a score, which provides an idea of the quality of that wine. So, this dataset could be useful to all the people who are doubtful on what wine to choose for a launch or a dinner, according also to the relevance of that event. Our aim is to provide, with the chosen dataset, a strong support to this decision, because the person could refer to the dataset to find the suitable wine for his event, taking into account also his preferences.
To achieve our purpose, we decided to load the dataset into **ElasticSearch**, a search engine based on the concept of relevance, assigning to each returned document a score according to the parameters specified by the user in the queries. Moreover, it allows full-text search, allowing to find specific words in a description, or whatever is specified as a text, and, through **Kibana**, a data visualization dashboard for ElasticSearch, the user can organise, aggregate, filter and represent data in many different ways, giving a general perspective on the information in the dataset.
To sum-up, the main reason of having chosen Elasticsearch as our database is the possibility to match the preferences of the user and return only the wines that fulfill the specified conditions, building also an implicit ranking by assigning a relevance score to each result, being very fast and intuitive to use and a good tool to find an appropriate solution to the problem faced by the user, in our case the choice of a suitable wine for an important event. Secondary, ElasticSearch could be used also for statistical and analytical purposes, by representing the results of the queries with graphs and aggregated diagrams.

2 | Dataset

2.1. Description

The used dataset can be found on **Kaggle**, called **winemag**.
It contains infromation about a large variety of wines, characterised by 14 fields, briefly described and explained:

- **column1**: an increasing number identifying the specific record;
- **country**: the country of production of the specific wine;
- **description**: a brief and technical description of the features of the wine;
- **points**: score assigned to the wine by a magazine, or by a taster if present;
- **price**: the price of the wine;
- **designation**: the informal name assigned to the wine by the producer;
- **province**: the area where the wine is produced;
- **region1**: the region where the wine is produced;
- **region2**: a more precise area with respect to the region where the wine is produced;
- **taster_name**: the name and surname of the taster who has tasted and reviewed the wine;
- **taster_twitter_handle**: the twitter reference of the taster;
- **title**: the official name of the wine;
- **variety**: the type of the wine;
- **winery**: the name of the winery that has produced the wine;

These attributes provide a complete description of a wine, allowing the user to personally assess the quality of the wine according to variety, winery, points, description and price and then evaluate its suitability for the user's event.

2.2. Data Schema

In Elasticsearch, when the data is imported, it is asked to the user if he wants to personally define a *mapping* of the imported data or ElasticSearch does the job done for him. The *mapping* of the data is the schema, or model, of the data themselves and it defines the types of each field of the data. The two types of mapping are:

- **dynamic mapping**: ElasticSearch does the mapping by itself trying to infer the structure of the documents;
- **explicit mapping**: the user explicitly defines the type of each fields during the creation of an index.

In our database, we decided to keep and use the dynamic mapping specified by Elastic-Search, because it coincides to the mapping we would have defined and it is adherent to our purpose.
The following picture shows the mapping of the data:

(a) The first part of the mapping.

(b) The second part of the mapping.

Figure 2.1: It shows the full mapping of the data.

Here there are some explanations and details about the mapping just defined:

- column1, points and price are defined as numbers, because their values are numerical: in particular, column1 and points are **long**, while price is **double** because it admits fractional values (such as 28.30);
- country, province, region_2, taster_name, taster_twitter_handle, variety and winery are defined as **keyword** because we want to match those features exactly; for example, one could want to search the variety "Pinot Noir" produced in "Italy" by the winery "Testarossa", and the result must match exactly each of these parameters;
- description, designation, region_1 and title are dined as **text** because we want to perform text search on them; for example, one could want to find a wine that is "spicy" and "fresh", so by text match he can find all the wines whose description includes these two words;

3 | Queries

In this chapter, the 20 requested queries are presented. For each query, the aim, the text and the result will be highlighted. For the result, unfortunately, only very few documents can be shown due to practical issues (taking a screenshot of the entire result is impossible). In the section Extra, the complete result of each query will be shown through dashboards created using Kibana, so the results in this section are very partial.

3.1. Query 1: Wines produced in Italy

The first query to be proposed is a simple query involving an exact match with a keyword field using a **term** query. The aim of the query is to find all the wines produced in Italy.

```
//wines produced in Italy   (1)
GET /wine_index/_search
{
  "query": {
    "term": {
      "country": {
        "value": "Italy"
      }
    }
  }
}
```

Figure 3.1: The text of the query.

```
{
  "took": 24,
  "timed_out": false,
  "_shards": {
    "total": 1,
    "successful": 1,
    "skipped": 0,
    "failed": 0
  },
  "hits": {
    "total": {
      "value": 10000,
      "relation": "gte"
    },
    "max_score": 1.8943729,
    "hits": [
      {
        "_index": "wine_index",
        "_id": "ZLivwIwBUQJRwN2MlRd9",
        "_score": 1.8943729,
        "_source": {
          "column1": 0,
          "country": "Italy",
          "taster_name": "Kerin O'Keefe",
          "taster_twitter_handle": "@kerinokeefe",
          "description": "Aromas include tropical fruit, broom, brimstone and dried herb. The palate isn't overly expressive,
            offering unripened apple, citrus and dried sage alongside brisk acidity.",
          "title": "Nicosia 2013 Vulkà Bianco  (Etna)",
          "points": 87,
          "province": "Sicily & Sardinia",
          "variety": "White Blend",
          "designation": "Vulkà Bianco",
          "winery": "Nicosia",
          "region_1": "Etna"
        }
      },
      {
        "_index": "wine_index",
        "_id": "arivwIwBUQJRwN2MlRd9",
        "_score": 1.8943729,
        "_source": {
          "column1": 6,
          "country": "Italy",
          "taster_name": "Kerin O'Keefe",
          "taster_twitter_handle": "@kerinokeefe",
          "description": "Here's a bright, informal red that opens with aromas of candied berry, white pepper and savory herb that
            carry over to the palate. It's balanced with fresh acidity and soft tannins.",
          "title": "Terre di Giurfo 2013 Belsito Frappato (Vittoria)",
          "points": 87,
          "province": "Sicily & Sardinia",
          "variety": "Frappato",
          "price": 16,
          "designation": "Belsito",
          "winery": "Terre di Giurfo",
          "region_1": "Vittoria"
```


Figure 3.2: The result of the query.

3.2. Query 2: Specific range of price

This query provides wines whose price is within a specific range. It shows how to search for a wine setting a desired range of price using a **range** query. This query returns all the wines whose price is within 1000 and 10000 euros.

```
//wines with a price within a certain range    (2)
GET /wine_index/_search
{
  "query": {
    "range": {
      "price": {
        "gte": 1000,
        "lte": 10000
      }
    }
  }
}
```

Figure 3.3: The text of the query.

```
{
  "took": 11,
  "timed_out": false,
  "_shards": {
    "total": 1,
    "successful": 1,
    "skipped": 0,
    "failed": 0
  },
  "hits": {
    "total": {
      "value": 16,
      "relation": "eq"
    },
    "max_score": 1,
    "hits": [
      {
        "_index": "wine_index",
        "_id": "erivwIwBUQJRwN2MlR2A",
        "_score": 1,
        "_source": {
          "column1": 1558,
          "country": "France",
          "taster_name": "Roger Voss",
          "taster_twitter_handle": "@vossroger",
          "description": "A massive wine for Margaux, packed with tannins and ripe fruit. It has more Cabernet Sauvignon than usual, giving
            intense black currant flavors with enticing acidity balanced by the sweetness of the fruit. Ripe swathes of this opulent fruit are
            also elegant and structured.",
          "title": "Château Margaux 2009  Margaux",
          "points": 98,
          "province": "Bordeaux",
          "variety": "Bordeaux-style Red Blend",
          "price": 1900,
          "winery": "Château Margaux",
          "region_1": "Margaux"
        }
      },
      {
        "_index": "wine_index",
        "_id": "grivwIwBUQJRwN2MlR2A",
        "_score": 1,
        "_source": {
          "column1": 1566,
          "country": "France",
          "taster_name": "Roger Voss",
          "taster_twitter_handle": "@vossroger",
          "description": "Such a generous and ripe wine, with a dark core of tannins surrounded by opulent fruit. Black fruits, coffee, very
            concentrated flavors, a powerhouse of structure and richness. The warmth of the wine is palpable, as is the aging potential.",
          "title": "Château La Mission Haut-Brion 2009  Pessac-Léognan",
          "points": 97,
          "province": "Bordeaux",
          "variety": "Bordeaux-style Red Blend",
          "price": 1100,
          "winery": "Château La Mission Haut-Brion",
          "region_1": "Pessac-Léognan"
        }
      },
```


Figure 3.4: The result of the query.

3.3. Query 3: Number of wines with perfect score for each country

This query provides the number of wines which have achieved the maximum score possible (100) for each country. First, there is a pre-filtering phase, where only the wines with a perfect score are kept, then with **aggs**, aggregation is performed on "country" to return the count of the wines for each country.

```
//Number of wines with a perfect score for each country  (3)
GET /wine_index/_search/
{
  "query": {
    "range": {
      "points": {
        "gte": 100,
        "lte": 100
      }
    }
  },

  "size": 0,
  "aggs": {
    "countries_with_best_wines": {
      "terms": {
        "field": "country"
      }
    }
  }
}
```

Figure 3.5: The text of the query.

```
{
  "took": 6,
  "timed_out": false,
  "_shards": {
    "total": 1,
    "successful": 1,
    "skipped": 0,
    "failed": 0
  },
  "hits": {
    "total": {
      "value": 19,
      "relation": "eq"
    },
    "max_score": null,
    "hits": []
  },
  "aggregations": {
    "countries_with_best_wines": {
      "doc_count_error_upper_bound": 0,
      "sum_other_doc_count": 0,
      "buckets": [
        {
          "key": "France",
          "doc_count": 8
        },
        {
          "key": "Italy",
          "doc_count": 4
        },
        {
          "key": "US",
          "doc_count": 4
        },
        {
          "key": "Portugal",
          "doc_count": 2
        },
        {
          "key": "Australia",
          "doc_count": 1
        }
      ]
    }
  }
}
```


Figure 3.6: The result of the query.

3.4. Query 4: Number of reviews for each taster

This query provides the number of reviews for each taster. With **aggs**, aggregation is performed on "taster_name" to return the count of the reviews (or reviewed wines) for each taster.

```
//Number of reviews per taster  (4)
GET /wine_index/_search/
{
  "size": 0,
  "aggs": {
    "reviews_per_taster": {
      "terms": {
        "field": "taster_name"
      }
    }
  }
}
```

Figure 3.7: The text of the query.

```
{
  "took": 1,
  "timed_out": false,
  "_shards": {
    "total": 1,
    "successful": 1,
    "skipped": 0,
    "failed": 0
  },
  "hits": {
    "total": {
      "value": 10000,
      "relation": "gte"
    },
    "max_score": null,
    "hits": []
  },
  "aggregations": {
    "reviews_per_taster": {
      "doc_count_error_upper_bound": 0,
      "sum_other_doc_count": 8196,
      "buckets": [
        {
          "key": "Roger Voss",
          "doc_count": 25514
        },
        {
          "key": "Michael Schachner",
          "doc_count": 15134
        },
        {
          "key": "Kerin O'Keefe",
          "doc_count": 10776
        },
        {
          "key": "Virginie Boone",
          "doc_count": 9537
        },
        {
          "key": "Paul Gregutt",
          "doc_count": 9532
        },
        {
          "key": "Matt Kettmann",
          "doc_count": 6332
```

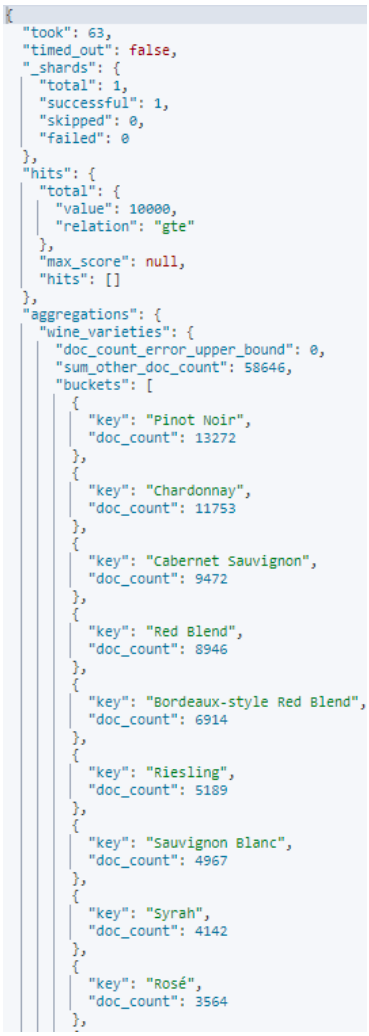

Figure 3.8: The result of the query.

3.5. Query 5: Number of wines for each variety

This query is similar to the previous one, it provides the number of wines for each variety. With **aggs**, aggregation is performed on "variety" to return the count of the wines for each variety.

```
//Number of wines per wine variety  (5)
GET /wine_index/_search/
{
  "size": 0,
  "aggs": {
    "wine_varieties": {
      "terms": {
        "field": "variety"
      }
    }
  }
}
```

Figure 3.9: The text of the query.



```
{
  "took": 63,
  "timed_out": false,
  "_shards": {
    "total": 1,
    "successful": 1,
    "skipped": 0,
    "failed": 0
  },
  "hits": {
    "total": {
      "value": 10000,
      "relation": "gte"
    },
    "max_score": null,
    "hits": []
  },
  "aggregations": {
    "wine_varieties": {
      "doc_count_error_upper_bound": 0,
      "sum_other_doc_count": 58646,
      "buckets": [
        {
          "key": "Pinot Noir",
          "doc_count": 13272
        },
        {
          "key": "Chardonnay",
          "doc_count": 11753
        },
        {
          "key": "Cabernet Sauvignon",
          "doc_count": 9472
        },
        {
          "key": "Red Blend",
          "doc_count": 8946
        },
        {
          "key": "Bordeaux-style Red Blend",
          "doc_count": 6914
        },
        {
          "key": "Riesling",
          "doc_count": 5189
        },
        {
          "key": "Sauvignon Blanc",
          "doc_count": 4967
        },
        {
          "key": "Syrah",
          "doc_count": 4142
        },
        {
          "key": "Rosé",
          "doc_count": 3564
        },
```


Figure 3.10: The result of the query.

3.6. Query 6: Number of wines for each Italian province

This query returns the number of wines for each Italian province. First, a query to filter the wines according to the country is performed, with "Italy" as the desired country, and then through **aggs**, the result is aggregated by province, returning for each province the number of wines produced in it.

```
//Italian provinces with the most reviews  (6)
GET /wine_index/_search/
{
  "query": {
    "term": {
      "country": "Italy"
    }
  },
  "size": 0,
  "aggs": {
    "most_common_province": {
      "terms": {
        "field": "province"
      }
    }
  }
}
```

Figure 3.11: The text of the query.


```
{
  "took": 16,
  "timed_out": false,
  "_shards": {
    "total": 1,
    "successful": 1,
    "skipped": 0,
    "failed": 0
  },
  "hits": {
    "total": {
      "value": 10000,
      "relation": "gte"
    },
    "max_score": null,
    "hits": []
  },
  "aggregations": {
    "most_common_province": {
      "doc_count_error_upper_bound": 0,
      "sum_other_doc_count": 0,
      "buckets": [
        {
          "key": "Tuscany",
          "doc_count": 5897
        },
        {
          "key": "Piedmont",
          "doc_count": 3728
        },
        {
          "key": "Veneto",
          "doc_count": 2716
        },
        {
          "key": "Northeastern Italy",
          "doc_count": 2138
        },
        {
          "key": "Sicily & Sardinia",
          "doc_count": 1797
        },
        {
          "key": "Southern Italy",
          "doc_count": 1349
```


Figure 3.12: The result of the query.

3.7. Query 7: Number of wines per variety and winery, for each country

This query returns the number of wines considering each variety and each winery separately for each country. That is, first it aggregates according to country, then inside this aggregation, it performs another aggregation on two separate fields (variety and winery), returning the number of wines for each of these two fields given the country of the previous aggregation.

```
//number of wines per variety and winery, for each country   (7)
GET /wine_index/_search
{
  "size": 0,
  "aggs": {
    "wine_per_country": {
      "terms": {
        "field": "country"
      },
      "aggs": {
        "count_per_variety": {
          "terms": {
            "field": "variety"
          }
        },
        "count_per_winery": {
          "terms": {
            "field": "winery"
          }
        }
      }
    }
  }
}
```

Figure 3.13: The text of the query.

```
{
  "took": 17,
  "timed_out": false,
  "_shards": {
    "total": 1,
    "successful": 1,
    "skipped": 0,
    "failed": 0
  },
  "hits": {
    "total": {
      "value": 10000,
      "relation": "gte"
    },
    "max_score": null,
    "hits": []
  },
  "aggregations": {
    "wine_per_country": {
      "doc_count_error_upper_bound": 0,
      "sum_other_doc_count": 5323,
      "buckets": [
        {
          "key": "US",
          "doc_count": 54503,
          "count_per_winery": {
            "doc_count_error_upper_bound": 0,
            "sum_other_doc_count": 53002,
            "buckets": [
              {
                "key": "Testarossa",
                "doc_count": 218
              },
              {
                "key": "Williams Selyem",
                "doc_count": 211
              },
              {
                "key": "Chateau Ste. Michelle",
                "doc_count": 194
              },
              {
                "key": "Columbia Crest",
                "doc_count": 159
              },
```


Figure 3.14: The result of the query.

3.8. Query 8: Statistics about the price for each variety

This query returns some statistics about the price of a specific variety of wine. It aggregates with respect to variety and then performs another aggregation, not with **terms** but with **stats**, on price, returning the **count, min, max, avg, sum** of the prices of that specific variety. It's a query useful for analytic purposes or to know some information on the price of a variety of wine for a future purchase.

```
// statistics about price for each variety   (8)
GET /wine_index/_search
{
  "size": 0,
  "aggs": {
    "variety_with_insight_on_price": {
      "terms": {
        "field": "variety"
      },
      "aggs": {
        "statistics_about_price": {
          "stats": {
            "field": "price"
          }
        }
      }
    }
  }
}
```

Figure 3.15: The text of the query.


```
{
  "took": 9,
  "timed_out": false,
  "_shards": {
    "total": 1,
    "successful": 1,
    "skipped": 0,
    "failed": 0
  },
  "hits": {
    "total": {
      "value": 10000,
      "relation": "gte"
    },
    "max_score": null,
    "hits": []
  },
  "aggregations": {
    "variety_with_insight_on_price": {
      "doc_count_error_upper_bound": 0,
      "sum_other_doc_count": 58646,
      "buckets": [
        {
          "key": "Pinot Noir",
          "doc_count": 13272,
          "statistics_about_price": {
            "count": 12787,
            "min": 5,
            "max": 2500,
            "avg": 47.528896535543915,
            "sum": 607752
          }
        },
        {
          "key": "Chardonnay",
          "doc_count": 11753,
          "statistics_about_price": {
            "count": 11080,
            "min": 4,
            "max": 2013,
            "avg": 34.52202166064982,
            "sum": 382504
          }
        },
```


Figure 3.16: The result of the query.

3.9. Query 9: Statistics about price and points for each variety

This query is similar to the previous one because it returns some statistics on points and price of the wines produced in a specific country. So, the first aggregation is performed on country, then the following internal aggregation is performed on price and points, separately. This aggregation is a **stats** aggregation.

```
//query to find for each country, statistics about points and prices   (9)
GET /wine_index/_search
{
  "size": 0,
  "aggs": {
    "group_by_country": {
      "terms": {
        "field": "country"
      },
      "aggs": {
        "stats_about_points": {
          "stats": {
            "field": "points"
          }
        },
        "stats_about_price": {
          "stats": {
            "field": "price"
          }
        }
      }
    }
  }
}
```

Figure 3.17: The text of the query.

```
{
  "took": 3,
  "timed_out": false,
  "_shards": {
    "total": 1,
    "successful": 1,
    "skipped": 0,
    "failed": 0
  },
  "hits": {
    "total": {
      "value": 10000,
      "relation": "gte"
    },
    "max_score": null,
    "hits": []
  },
  "aggregations": {
    "group_by_country": {
      "doc_count_error_upper_bound": 0,
      "sum_other_doc_count": 5323,
      "buckets": [
        {
          "key": "US",
          "doc_count": 54503,
          "stats_about_points": {
            "count": 54503,
            "min": 80,
            "max": 100,
            "avg": 88.56378547969837,
            "sum": 4826992
          },
          "stats_about_price": {
            "count": 54264,
            "min": 4,
            "max": 2013,
            "avg": 36.573953265516735,
            "sum": 1984649
          }
        },
        {
          "key": "France",
          "doc_count": 22093,
          "stats_about_points": {
            "count": 22093,
```


Figure 3.18: The result of the query.

3.10. Query 10: Wines with high score, grouped by taster and then variety

This query returns the number of wines with an high score are reviewed by each taster, according to the variety. First, a pre-filtering phase is done, keeping only those wines with a score greater or equal than 95. Then the result is aggregated by taster_name, and for each taster, another aggregation is performed on the variety of the wine.

```
//Wines with a high score, grouped first by taster then variety   (10)
GET /wine_index/_search/
{
  "query": {
    "range": {
      "points": {
        "gte": 95
      }
    }
  },
  "size": 0,
  "aggs": {
    "high_score_reviews_per_taster": {
      "terms": {
        "field": "taster_name"
      },
      "aggs": {
        "variety": {
          "terms": {
            "field": "variety"
          }
        }
      }
    }
  }
}
```

Figure 3.19: The text of the query.


```
{
  "took": 25,
  "timed_out": false,
  "_shards": {
    "total": 1,
    "successful": 1,
    "skipped": 0,
    "failed": 0
  },
  "hits": {
    "total": {
      "value": 2416,
      "relation": "eq"
    },
    "max_score": null,
    "hits": []
  },
  "aggregations": {
    "high_score_reviews_per_taster": {
      "doc_count_error_upper_bound": 0,
      "sum_other_doc_count": 40,
      "buckets": [
        {
          "key": "Roger Voss",
          "doc_count": 707,
          "variety": {
            "doc_count_error_upper_bound": 0,
            "sum_other_doc_count": 43,
            "buckets": [
              {
                "key": "Bordeaux-style Red Blend",
                "doc_count": 184
              },
              {
                "key": "Chardonnay",
                "doc_count": 121
              },
              {
                "key": "Pinot Noir",
                "doc_count": 106
              },
              {
                "key": "Champagne Blend",
                "doc_count": 60
              },
              {
                "key": "Port",
                "doc_count": 53
              },
              {
                "key": "Bordeaux-style White Blend"
                "doc_count": 51
              },
              {
                "key": "Portuguese Red",
                "doc_count": 45
              },
              {
                "key": "Riesling",
                "doc_count": 21
```


Figure 3.20: The result of the query.

3.11. Query 11: Wines with a tarragon taste, preferably from Italy

This query returns all the wines whose description contains the word "tarragon" and assigns a high score to those wines that come from Italy. The filter is done with a **must**, implying that the feature specified inside it must be necessarily owned by the wine, while the preference to be from Italy is expressed by the **should** part of the **bool** query.

```
//Wines with a tarragon taste, preferably from Italy   (11)
GET /wine_index/_search
{
  "query": {
    "bool": {
      "must": [
        {
          "match": {
            "description": "tarragon"
          }
        }
      ],
      "should": [
        {
          "term": {
            "country": "Italy"
          }
        }
      ]
    }
  }
}
```

Figure 3.21: The text of the query.

```
{
  "took": 1,
  "timed_out": false,
  "_shards": {
    "total": 1,
    "successful": 1,
    "skipped": 0,
    "failed": 0
  },
  "hits": {
    "total": {
      "value": 179,
      "relation": "eq"
    },
    "max_score": 8.417324,
    "hits": [
      {
        "_index": "wine_index",
        "_id": "M7qvwIwBUQJRwN2M3wE2",
        "_score": 8.417324,
        "_source": {
          "column1": 125394,
          "country": "Italy",
          "province": "Sicily & Sardinia",
          "variety": "White Blend",
          "taster_name": "Kerin O'Keefe",
          "taster_twitter_handle": "@kerinokeefe",
          "description": "Made with 70% Catarratto and 30% Viognier, this has aromas suggesting honeysuckle, yellow stone fruit and a savory
            whiff of tarragon. The soft, round palate shows pear, citrus and a hint of apricot while a note of bitter almond closes the finish
            .",
          "winery": "Gregorio De Gregorio",
          "region_1": "Terre Siciliane",
          "title": "Gregorio De Gregorio 2015 White (Terre Siciliane)",
          "points": 88
        }
      },
      {
        "_index": "wine_index",
        "_id": "s7ivwIwBUQJRwN2MtOFH",
        "_score": 7.8385177,
        "_source": {
          "column1": 51794,
          "country": "Serbia",
          "province": "Fruška Gora",
          "variety": "Portuguiser",
          "price": 15,
          "taster_name": "Jeff Jenssen",
          "taster_twitter_handle": "@worldwineguys",
          "description": "This Serbian 100% Portuguiser offers scents of dried cherry, dried herbs and fresh tarragon, and follows up with
            tastes of red plum and dried strawberry.",
          "winery": "Agrina",
          "title": "Agrina 2014 Portuguiser (Fruška Gora)",
          "points": 86
        }
      },
```


Figure 3.22: The result of the query.

3.12. Query 12: Wines reviewed by Kerin O'Keefe, preferably Cabernet Sauvignon by Mezzacorona

This query returns all the wines reviewed by Kerin O'Keefe and assigns a high score to those wines that are Cabernet Sauvignon produced by the winery Mezzacorona. The filter on the taster_name is done with a **must**, while the further preferences are specified by a **should**.

```
//query to search some wines in particular   (12)
GET /wine_index/_search
{
  "query": {
    "bool": {
      "must": [
        {"term": {
          "taster_name": {
            "value": "Kerin O'Keefe"
          }
        }}
      ],
      "should": [
        {"term": {
          "variety": "Cabernet Sauvignon"
        }},
        {"term": {
          "winery": {
            "value": "Mezzacorona"
          }
        }}
      ]
    }
  }
}
```

Figure 3.23: The text of the query.


```
{
  "took": 26,
  "timed_out": false,
  "_shards": {
    "total": 1,
    "successful": 1,
    "skipped": 0,
    "failed": 0
  },
  "hits": {
    "total": {
      "value": 10000,
      "relation": "gte"
    },
    "max_score": 13.344155,
    "hits": [
      {
        "_index": "wine_index",
        "_id": "vbivwIwBUQJRwN2MmCea",
        "_score": 13.344155,
        "_source": {
          "column1": 4185,
          "country": "Italy",
          "taster_name": "Kerin O'Keefe",
          "taster_twitter_handle": "@kerinokeefe",
          "description": "Aromas of cassis, underbrush and toast lead the nose on this simple but well-made red. The easy-drinking palate
            delivers black currant, raspberry, coffee and a toasted note alongside polished tannins. It's made to be enjoyed young so drink
            soon.",
          "title": "Mezzacorona 2013 Cabernet Sauvignon (Vigneti delle Dolomiti)",
          "points": 87,
          "province": "Northeastern Italy",
          "variety": "Cabernet Sauvignon",
          "price": 9,
          "winery": "Mezzacorona",
          "region_1": "Vigneti delle Dolomiti"
        }
      },
      {
        "_index": "wine_index",
        "_id": "ULivwIwBUQJRwN2MlR-B",
        "_score": 10.72526,
        "_source": {
          "column1": 2028,
          "country": "Italy",
          "taster_name": "Kerin O'Keefe",
          "taster_twitter_handle": "@kerinokeefe",
          "description": "Aromas of green apples, peach and citrus carry over to palate. The juicy flavors are accompanied by lively acidity
            that makes this wine very food friendly.",
          "title": "Mezzacorona 2013 Pinot Grigio (Vigneti delle Dolomiti)",
          "points": 87,
          "province": "Northeastern Italy",
          "variety": "Pinot Grigio",
          "price": 9,
          "winery": "Mezzacorona",
          "region_1": "Vigneti delle Dolomiti"
        }
      },
```

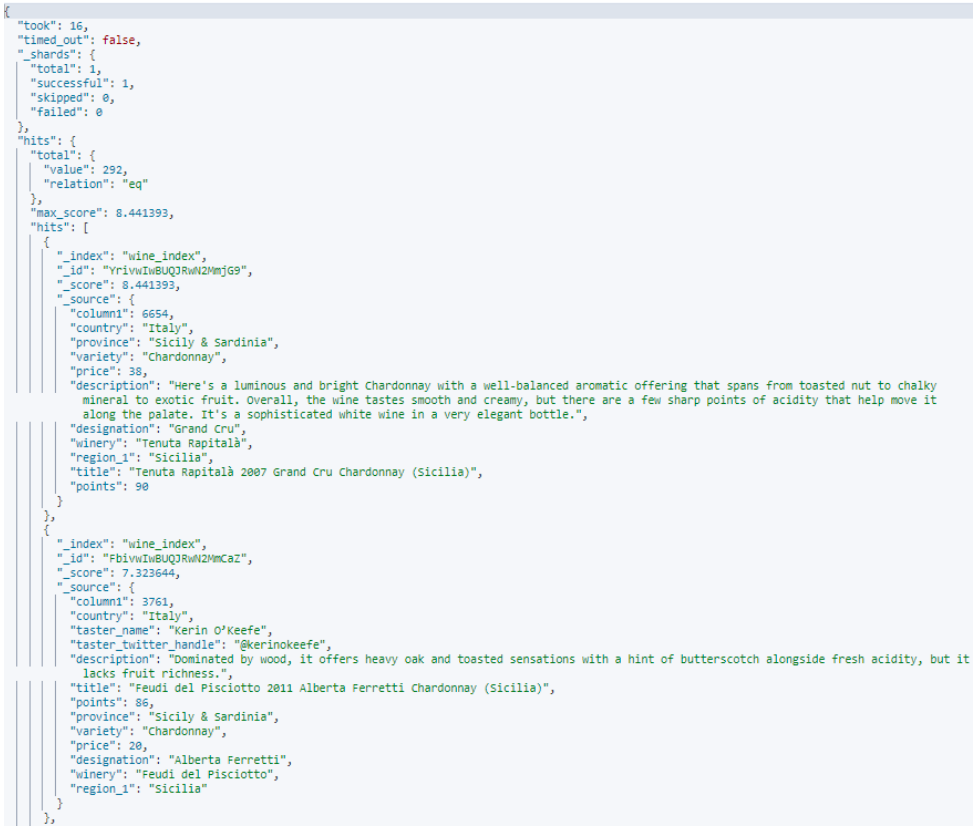

Figure 3.24: The result of the query.

3.13. Query 13: An Italian cheap Chardonnay, preferably with a specific taste

This query returns the wines whose variety is Chardonnay and they are produced in Italy. Moreover, they are also cheap, the price is less than 50 euros. The top results are influenced by the **should** clause, specifying that the wine is better if it is sharp, heavy or balanced. In this query, the **must_not** and the **filter** clauses: **must_not** is used to exclude those wines that satisfy the specified condition, in this case having a price greater or equal than 50 euros, ignoring the score; **filter** is used to filter the results according to the condition without influencing the score of the results. Finally, inside the should clause, the **operator** parameter is used to specify that the words in the match query are in **or**.

```
//a query to search an Italian cheap Chardonnay   (13)
GET /wine_index/_search
{
  "query": {
    "bool": {
      "should": [
        {"match": {
          "description": {
            "query": "sharp, heavy, balanced",
            "operator": "or"
          }
        }}
      ],
      "must": [
        {"term": {
          "variety": {
            "value": "Chardonnay"
          }
        }}
      ],
      "filter": [
        {"term": {
          "country": "Italy"
        }}
      ],
      "must_not": [
        {"range": {
          "price": {
            "gte": 50
          }
        }}
      ]
    }
  }
}
```

Figure 3.25: The text of the query.



```
{
  "took": 16,
  "timed_out": false,
  "_shards": {
    "total": 1,
    "successful": 1,
    "skipped": 0,
    "failed": 0
  },
  "hits": {
    "total": {
      "value": 292,
      "relation": "eq"
    },
    "max_score": 8.441393,
    "hits": [
      {
        "_index": "wine_index",
        "_id": "YrivwIwBUQJRwN2MmjG9",
        "_score": 8.441393,
        "_source": {
          "column1": 6654,
          "country": "Italy",
          "province": "Sicily & Sardinia",
          "variety": "Chardonnay",
          "price": 38,
          "description": "Here's a luminous and bright Chardonnay with a well-balanced aromatic offering that spans from toasted nut to chalky
            mineral to exotic fruit. Overall, the wine tastes smooth and creamy, but there are a few sharp points of acidity that help move it
            along the palate. It's a sophisticated white wine in a very elegant bottle.",
          "designation": "Grand Cru",
          "winery": "Tenuta Rapitalà",
          "region_1": "Sicilia",
          "title": "Tenuta Rapitalà 2007 Grand Cru Chardonnay (Sicilia)",
          "points": 90
        }
      },
      {
        "_index": "wine_index",
        "_id": "FbivwIwBUQJRwN2MmCaZ",
        "_score": 7.323644,
        "_source": {
          "column1": 3761,
          "country": "Italy",
          "taster_name": "Kerin O'Keefe",
          "taster_twitter_handle": "@kerinokeefe",
          "description": "Dominated by wood, it offers heavy oak and toasted sensations with a hint of butterscotch alongside fresh acidity, but it
            lacks fruit richness.",
          "title": "Feudi del Pisciotto 2011 Alberta Ferretti Chardonnay (Sicilia)",
          "points": 86,
          "province": "Sicily & Sardinia",
          "variety": "Chardonnay",
          "price": 20,
          "designation": "Alberta Ferretti",
          "winery": "Feudi del Pisciotto",
          "region_1": "Sicilia"
        }
      },
```


Figure 3.26: The result of the query.

3.14. Query 14: Wines perfect for summer not from the Us, preferably from Italy and with a high score

This query returns the wines not produced in the US whose description includes the words "citrus, summer, light, fresh", preferably from Italy and with a high score. In particular, in the **should** clause, the **boost** parameter is used to boost the score of the returned wine by 2 if its points are greater or equal to 85, highlighting the more importance given by the user to the points of the wine.

```
//Wines perfect for summer not from the Us, preferably from Italy and with a high score  (14)
GET /wine_index/_search
{
  "query": {
    "bool": {
      "must": [
        {
          "match": {
            "description": {
              "query": "citrus, fresh, summer, light",
              "operator": "and"
            }
          }
        }
      ],
      "should": [
        {
          "term": {
            "country": "Italy"
          }
        },
        {
          "range": {
            "points": {
              "gte": 85,
              "boost": 2
            }
          }
        }
      ],
      "must_not": [
        {
          "term": {
            "country": {
              "value": "US"
            }
          }
        }
      ]
    }
  }
}
```

Figure 3.27: The text of the query.

```
{
 "took": 28,
 "timed_out": false,
 "_shards": {
  "total": 1,
  "successful": 1,
  "skipped": 0,
  "failed": 0
 },
 "hits": {
  "total": {
   "value": 14,
   "relation": "eq"
  },
  "max_score": 17.230844,
  "hits": [
   {
    "_index": "wine_index",
    "_id": "-7ivwIwBUQJRwN2MlRuA",
    "_score": 17.230844,
    "_source": {
     "column1": 1175,
     "country": "Austria",
     "taster_name": "Roger Voss",
     "taster_twitter_handle": "@vossroger",
     "description": "Fresh, crisp, citrus and floral flavors make a light, easy Grüner Veltliner, a great summer drink of
       light white fruits and green berries. Screwcap.",
     "title": "Schloss Halbturn 2006 Koenigsegg Velt. 1 Grüner Veltliner (Burgenland)",
     "points": 86,
     "province": "Burgenland",
     "variety": "Grüner Veltliner",
     "price": 12,
     "designation": "Koenigsegg Velt. 1",
     "winery": "Schloss Halbturn"
    }
   },
   {
    "_index": "wine_index",
    "_id": "-7ivwIwBUQJRwN2Mt_rp",
    "_score": 17.169312,
    "_source": {
     "column1": 58266,
     "country": "Italy",
     "province": "Veneto",
     "variety": "Garganega",
     "price": 15,
     "description": "Here's an easy-drinking Soave Classico that opens with fresh fruit, citrus and a dusty touch of
       mineral at the back. Crisp and bright, it would pair with light summer meals.",
     "winery": "Coffele",
     "region_1": "Soave Classico",
     "title": "Coffele 2011  Soave Classico",
     "points": 86
    }
```


Figure 3.28: The result of the query.

3.15. Query 15: Wines with a high score, preferably from Italy and cheap

This query returns the wines whose points are above 83, providing a different level of boosting depending on what condition is satisfied by the wine: if the wine is Italian, its score is boosted by 1.5; if its price is below 5 euros, it is boosted by 4; if its price is between 5 and 30 euros, it is boosted by 2. This shows how to express the preferences of the user using different level of boosting, giving a high boost to those wines that respects the most important condition to the user, and vice-versa. Note that all the returned wines satisfy the condition in the **must** clause, which is the sufficient and necessary condition to be included in the result.

```
//wines with a high score, preferably from Italy  and cheap (15)
GET /wine_index/_search
{
  "query": {
    "bool": {
      "must": [
        {
          "range": {
            "points": {
              "gte": 83
            }
          }
        }
      ],
      "should": [
        {
          "term": {
            "country": {
              "value": "Italy",
              "boost": 1.5
            }
          }
        },
        {
          "range": {
            "price": {
              "lte": 5,
              "boost": 4
            }
          }
        },
        {
          "range": {
            "price": {
              "gt": 5,
              "lt": 30,
              "boost": 2
            }
          }
        }
      ]
    }
  }
}
```

Figure 3.29: The text of the query.

```
{
  "took": 35,
  "timed_out": false,
  "_shards": {
    "total": 1,
    "successful": 1,
    "skipped": 0,
    "failed": 0
  },
  "hits": {
    "total": {
      "value": 10000,
      "relation": "gte"
    },
    "max_score": 7.8415594,
    "hits": [
      {
        "_index": "wine_index",
        "_id": "w7ivwIwBUQJRwN2MmCOX",
        "_score": 7.8415594,
        "_source": {
          "column1": 3167,
          "country": "Italy",
          "province": "Veneto",
          "variety": "Glera",
          "price": 5,
          "description": "Packaged in a pocket-friendly 187-ml bottle, this is fragrant and simple, with tonic, creamy foam and
            bright aromas of peach, honeydew and jasmine.",
          "designation": "Mini",
          "winery": "Anna Spinato",
          "region_1": "Prosecco",
          "title": "Anna Spinato NV Mini  (Prosecco)",
          "points": 86
        }
      },
      {
        "_index": "wine_index",
        "_id": "JrmvwIwBUQJRwN2M0alr",
        "_score": 7.8415594,
        "_source": {
          "column1": 102853,
          "country": "Italy",
          "province": "Southern Italy",
          "variety": "Primitivo",
          "price": 5,
          "description": "Definitely not Zinfandel-like, despite its Primitivo origins; a rather nondescript wine, as a matter of
            fact. Light cherry overtones are noted on the nose and palate. The tannins are mouth-drying. This is certainly not a
            good example of how far Puglia has come with this grape. However, it would work fine with a slice or two of pizza.",
          "winery": "Terrale",
          "region_1": "Puglia",
          "title": "Terrale 1998 Primitivo (Puglia)",
          "points": 83
        }
      },
```


Figure 3.30: The result of the query.

3.16. Query 16: Wines (not Rosè) using boosting query, preferably from France

This query returns the wines that are not Rosè and they are "soft, delicate or white", preferably from France. So, if the wine is from France, the score of the result increases. Moreover, if the wine's description includes "spicy, dry or strong", the score of the wine is decreased through a **boosting** feature: in the **must** clause a **boosting** clause is included and it is divided into a **positive** branch and a **negative** one. The wines matching the conditions in the positive branch are evaluated positively, while if the wine matches also the conditions in the negative branch, its score is reduced by multiplying it with a number (between 0 and 1) specified at the end by the parameter **negative_boost**. This kind of feature helps to underline the possible dissatisfaction of the user if some conditions are matched by afflicting negatively the score of the results. Finally, a **should** and a

must_not clauses are included to express the preference and the unwanted feature, respectively.

```
//Wines (not Rosè) using boosting query, preferably from France   (16)
GET /wine_index/_search
{
  "query": {
    "bool": {
      "must": [
        {
          "boosting": {
            "positive": {
              "match": {
                "description": "soft, white, delicate"
              }
            },
            "negative": {
              "match": {
                "description": "spicy, dry, strong"
              }
            },
            "negative_boost": 0.5
          }
        }
      ],
      "should": [
        {
          "term": {
            "country": "France"
          }
        }
      ],
      "must_not": [
        {
          "term": {
            "variety": {
              "value": "Rosé"
            }
          }
        }
      ]
    }
  }
}
```

Figure 3.31: The text of the query.

```
{
  "took": 222,
  "timed_out": false,
  "_shards": {
    "total": 1,
    "successful": 1,
    "skipped": 0,
    "failed": 0
  },
  "hits": {
    "total": {
      "value": 10000,
      "relation": "gte"
    },
    "max_score": 12.030886,
    "hits": [
      {
        "_index": "wine_index",
        "_id": "b7ivwIwBUQJRwN2MqXoh",
        "_score": 12.030886,
        "_source": {
          "column1": 25357,
          "country": "Italy",
          "province": "Veneto",
          "variety": "Glera",
          "price": 30,
          "description": "This is a soft and delicate sparkler with aromas of white flower, jasmine, talc powder and soft peach or
            nectarine. Creamy bubbles make for a round, delicate mouthfeel.",
          "designation": "46 Parallelo Extra Dry",
          "winery": "Il Colle",
          "region_1": "Conegliano Valdobbiadene Prosecco Superiore",
          "title": "Il Colle NV 46 Parallelo Extra Dry  (Conegliano Valdobbiadene Prosecco Superiore)",
          "points": 87
        }
      },
      {
        "_index": "wine_index",
        "_id": "JbivwIwBUQJRwN2Mp3YA",
        "_score": 11.947951,
        "_source": {
          "column1": 24259,
          "country": "France",
          "taster_name": "Joe Czerwinski",
          "taster_twitter_handle": "@JoeCz",
          "description": "The Grenache Blanc (60%) gives this wine its delicate aromas of white flowers, while the Clairette (40%)
            provides backbone and pineapple fruit. Drink this relatively soft, mouthfilling white over the next few months.",
          "title": "Ferraton Pere et Fils 2010 Samorëns White (Côtes du Rhône)",
          "points": 87,
          "province": "Rhône Valley",
          "variety": "Rhône-style White Blend",
          "price": 14,
          "designation": "Samorëns",
          "winery": "Ferraton Pere et Fils",
          "region_1": "Côtes du Rhône"
        }
      }
```


Figure 3.32: The result of the query.

3.17. Query 17: Wines, preferably Red Blend ones, not by Siduri, from the US, with specific characteristics

This query returns the wines that are not produced by the winery Siduri, they are produced in the US and they are "fruity, warm, tasty or amber". If they are Red Blend, the score increases. Looking at the query, there is a pre-filter phase, where all these conditions are specified: a **match**and a **term** query inide the **filter** clause for the country and description; a **must_not** clause to exclude the winery Siduri; a **should** clause to increase the score if the wine is from the US. After this pre-filtering phase, there is an aggregation phase, where the wines are aggregated according to their variety and for each variety, statistics about the price are displayed. Note that the **size** parameter of the aggregation

is set to 20, so 20 wines are returned alongside the results of the aggregation.

```
//Wines, preferably Red Blend ones, not by Siduri, from the US, with specific characteristics   (17)
GET /wine_index/_search
{
  "size": 20,
  "query": {
    "bool": {
      "filter": [
        {"term": {
          "country": "US"
        }},
        {"match": {
          "description": {
            "query": "fruity, warm, tasty, amber",
            "operator": "or"
          }
        }}
      ],
      "must_not": [
        {"term": {
          "winery": {
            "value": "Siduri"
          }
        }}
      ],
      "should": [
        {"term": {
          "variety": {
            "value": "Red Blend"
          }
        }}
      ]
    }
  },
  "aggs": {
    "price_per_variety": {
      "terms": {
        "field": "variety"
      },
      "aggs": {
        "stats_about_price": {
          "stats": {
            "field": "price"
          }
        }
      }
    }
  }
}
```

Figure 3.33: The text of the query.

```
{
  "took": 232,
  "timed_out": false,
  "_shards": {
    "total": 1,
    "successful": 1,
    "skipped": 0,
    "failed": 0
  },
  "hits": {
    "total": {
      "value": 3905,
      "relation": "eq"
    },
    "max_score": 2.6760259,
    "hits": [
      {
        "_index": "wine_index",
        "_id": "SbivwIwBUQJRwN2MlRh9",
        "_score": 2.6760259,
        "_source": {
          "column1": 229,
          "country": "US",
          "description": "Not clear what the varietal blend is, but it's like a nice, easy southern Rhône wine. Soft and fruity, it's rich in jammy
            blackberry, cherry, Mandarin orange, chocolate and anise spice flavors.",
          "title": "Benessere 2005 Costa Del Sol Red (Napa Valley)",
          "points": 85,
          "province": "California",
          "variety": "Red Blend",
          "price": 18,
          "designation": "Costa Del Sol",
          "winery": "Benessere",
          "region_1": "Napa Valley",
          "region_2": "Napa"
        }
      },
      {
        "_index": "wine_index",
        "_id": "N7ivwIwBUQJRwN2MlRyA",
        "_score": 2.6760259,
        "_source": {
          "column1": 1235,
          "country": "US",
          "taster_name": "Matt Kettmann",
          "taster_twitter_handle": "@mattkettmann",
          "description": "Very roasted notes of coffee, warm leather, tobacco leaves, sagebrush and stewed blackberries show on the nose of this
            blend of 70% Tempranillo, 20% Grenache and 10% Cabernet Sauvignon. The roasted quality carries through the sip, with more coffee and
            char tones alongside dried plums, crushed gravel and dusty tannins.",
          "title": "Bodega de Edgar 2014 Toro de Paso Red (Central Coast)",
          "points": 92,
          "province": "California",
          "variety": "Red Blend",
          "price": 48,
          "designation": "Toro de Paso",
          "winery": "Bodega de Edgar",
          "region_1": "Central Coast",
          "region_2": "Central Coast"
        }
      },
```


Figure 3.34: The result of the query.

3.18. Query 18: OR with minimum_should_match

This query shows how to perform OR in ElasticSearch: use **minimum_should_match**. In fact, this query involves a **bool** query with only a **should** clause. In this clause, two **bool** queries are included: the first query involves a **must** clause, searching for fresh wines produced by winery Rainstorm, and a **should** clause, improving the score if the wine is produced in Oregon. In the other query, a **filter** clause searches for all the Pinot Gris whose price is between 85 and 95 euros, a **should** clause to search for wines with points between 80 and 98 and a **must_not** for the wines not reviewed by taster Matt Kettmann. At the end of the query, there is the **minimum_should_match** which specifies how many conditions, at least, must be satisfied inside the should of the most external bool clause (so the first should, which contains the two other bool queries).

```
//OR query with the minimum_should_match    (18)
GET /wine_index/_search
{
  "query": {
    "bool": {
      "should": [
        {"bool": {
          "must": [
            {"term": {
              "winery": {
                "value": "Rainstorm"
              }
            }},
            {"match": {
              "description": "fresh"
            }}
          ],
          "should": [
            {"term": {
              "province": {
                "value": "Oregon"
              }
            }}
          ]
        }},
        {"bool": {
          "filter": [
            {"term": {
              "variety": "Pinot Gris"
            }},
            {"range": {
              "price": {
                "gte": 85,
                "lte": 95
              }
            }}
          ],
          "should": [
            {"range": {
              "points": {
                "gte": 80,
                "lte": 98
              }
            }}
          ],
          "must_not": [
            {"term": {
              "taster_name": {
                "value": "Matt Kettmann"
              }
            }}
          ]
        }}
      ],
      "minimum_should_match": 1
    }
  }
}
```

Figure 3.35: The text of the query.

```
{
  "took": 3,
  "timed_out": false,
  "_shards": {
    "total": 1,
    "successful": 1,
    "skipped": 0,
    "failed": 0
  },
  "hits": {
    "total": {
      "value": 7,
      "relation": "eq"
    },
    "max_score": 14.853968,
    "hits": [
      {
        "_index": "wine_index",
        "_id": "IbmvwIwBUQJRwN2MvCGb",
        "_score": 14.853968,
        "_source": {
          "column1": 68032,
          "country": "US",
          "taster_name": "Paul Gregutt",
          "taster_twitter_handle": "@paulgwine ",
          "description": "The second vintage of this Rainstorm rosé sports fruit flavors of watermelon and strawberry, with hints
            of fresh cracker in both the aroma and the finish. But it's the fresh fruit that's driving the flavor bus here, with
            some added length and detail as it winds down, perhaps due in part to the native yeast fermentation.",
          "title": "Rainstorm 2015 Silver Linings Pinot Noir Rosé (Oregon)",
          "points": 88,
          "province": "Oregon",
          "variety": "Rosé",
          "price": 17,
          "designation": "Silver Linings Pinot Noir",
          "winery": "Rainstorm",
          "region_1": "Oregon",
          "region_2": "Oregon Other"
        }
      },
      {
        "_index": "wine_index",
        "_id": "vLivwIwBUQJRwN2Mt_no",
        "_score": 14.574098,
        "_source": {
          "column1": 57947,
          "country": "US",
          "taster_name": "Paul Gregutt",
          "taster_twitter_handle": "@paulgwine ",
          "description": "Spicy cranberry/raspberry fruit is at the heart of this bright, quaffable and value-priced wine. There's
            a light dusting of pepper also, but it's the fresh fruit that stands out.",
          "title": "Rainstorm 2013 Pinot Noir (Oregon)",
          "points": 87,
          "province": "Oregon",
          "variety": "Pinot Noir",
          "price": 17,
```


Figure 3.36: The result of the query.

3.19. Query 19: Fuzzy query

this query shows a peculiar feature of ElasticSearch: **fuzzy** queries, or queries with a level of fuzzyness. Text type allows to perform match queries to find those documents whose text field has some words specified in the query, so the text match is done some words, not on the entire text. This operation cannot be done with keyword type fields. But, using fuzzy queries, a few errors on the sought-after word are admitted, and this error is specified by the parameter **fuzziness**: The measure of the error is the Levenshtein distance, i.e. the minimum number of single-character edits (insertions, deletions or substitutions) required to change one word into the other. So, if, for example, the user doesn't remember perfectly the name of a taster, he can use a fuzzy query to search for a similar name, despite the taster_name field is mapped as a keyword. Unfortunately, the maximum fuzziness admitted is only 2, meaning that the word included in the query

must be almost equal to the wanted one, only very few differences are permitted.
Looking at the proposed query, two fuzzy queries are used to retrieve those wines reviewed by Virginia Boone and, preferably, whose designation is Mountain Covèe. Other conditions are specified inside the must clause and the should one: price between 10 and 200 euros and, preferably, produced by Souverain winery. At the end several parallel and nested aggregations are done: first, it groups by variety and country, separately; then, for each variety, it groups by winery and, for each winery, it also groups by price; while, for each country, it groups by region_2. Note that the size is set to 0, so only the number of wines per aggregation is returned.

```
//Fuzzy query  (19)
GET /wine_index/_search
{
  "size": 0,
  "query": {
    "bool": {
      "must": [
        {"fuzzy": {
          "taster_name": {
            "fuzziness": 2,
            "value": "Virginia Bone"
          }
        }},
        {"range": {
          "price": {
            "gte": 10,
            "lte": 200
          }
        }}
      ],
      "should": [
        {"fuzzy": {
          "designation": {
            "fuzziness": 2,
            "value": "Mountain cove"
          }
        }},
        {"term": {
          "winery": {
            "value": "Souverain",
            "boost": 2
          }
        }}
      ]
    }
  },
  "aggs": {
    "group_by_variety": {
      "terms": {
        "field": "variety"
      },
      "aggs": {
        "winery_for_variety": {
          "terms": {
            "field": "winery"
          },
          "aggs": {
            "price_per_variety_per_winery": {
              "stats": {
                "field": "price"
              }
            }
          }
        }
      }
    },
    "group_by_country": {
      "terms": {
        "field": "country"
      },
      "aggs": {
        "region_2_per_country": {
          "terms": {
            "field": "region_2"
          }
        }
      }
    }
  }
}
```

Figure 3.37: The text of the query.

```
{
  "took": 3,
  "timed_out": false,
  "_shards": {
    "total": 1,
    "successful": 1,
    "skipped": 0,
    "failed": 0
  },
  "hits": {
    "total": {
      "value": 9448,
      "relation": "eq"
    },
    "max_score": null,
    "hits": []
  },
  "aggregations": {
    "group_by_country": {
      "doc_count_error_upper_bound": 0,
      "sum_other_doc_count": 0,
      "buckets": [
        {
          "key": "US",
          "doc_count": 9448,
          "region_2_per_country": {
            "doc_count_error_upper_bound": 0,
            "sum_other_doc_count": 0,
            "buckets": [
              {
                "key": "Sonoma",
                "doc_count": 4290
              },
              {
                "key": "Napa",
                "doc_count": 2835
              },
              {
                "key": "Napa-Sonoma",
                "doc_count": 524
              },
              {
                "key": "Sierra Foothills",
                "doc_count": 413
              },
              {
                "key": "Central Valley",
                "doc_count": 353
              },
              {
                "key": "Central Coast",
                "doc_count": 248
              },
              {
                "key": "North Coast",
```


Figure 3.38: The result of the query.

3.20. Query 20: Recapitulatory query to show the main features of ElasticSearch

This query returns the wines that are Chardonnay, with a score greater or equal than 80, not produced by the winery Clos Troteligotte, preferably from Italy and withspecific characteristics ("spicy, fresh or bright"). A **must** clause with a **term** query is used to state that the wine must be a Chardonnay. A **filter** clause with a **range** query specifies that the points must be greater or equal to 80, without affecting the final score. A **must_not** clause excludes the wines produced by the winery Clos Troteligotte. A **should** clause expresses the preferences of the user by boosting specific conditions: "spicy" with a boost of 4, "fresh" with 3, "bright" with 2 and "Italy" with 1.5. This underlines the possibility of the user to state directly what are his preferences and exploits the concept of relevance to obtain the most relevant and suitable results for him. Then, a final aggregation is performed: it aggregates by winery and price in "parallel" (so separately), then for each winery, another aggregation is performed, on the points, returning with the **stats** aggregation, statistics on the points of the wines belonging to that winery; while, for the aggregation on the price, a descending order of the results is specified by the **order** parameter. Note that the **size** of the aggregation is set to 20 (20 wines are returned).

```
//Recapitulatory query that shows the main features of ElasticSearch  (20)
GET /wine_index/_search
{
  "size": 20,
  "query": {
    "bool": {
      "must": [
        {"term": {
          "variety": {
            "value": "Chardonnay"
          }
        }}
      ],
      "filter": [
        {"range": {
          "points": {
            "gte": 80
          }
        }}
      ],
      "must_not": [
        {"term": {
          "winery": {
            "value": "Clos Troteligotte"
          }
        }}
      ],
      "should": [
        {"term": {
          "country": {
            "value": "Italy",
            "boost": 1.5
          }
        }},
        {"match": {
          "description": {
            "query": "spicy",
            "boost": 4
          }
        }},
        {"match": {
          "description": {
            "query": "fresh",
            "boost": 3
          }
        }},
        {"match": {
          "description": {
            "query": "bright",
            "boost": 2
          }
        }}
      ]
    }
  },
  "aggs": {
    "wines_per_winery": {
      "terms": {
        "field": "winery"
      },
      "aggs": {
        "stats_about_score": {
          "stats": {
            "field": "points"
          }
        }
      }
    },
    "wines_per_price": {
      "terms": {
        "field": "price",
        "order": {
          "_key": "desc"
        }
      }
    }
  }
}
```

Figure 3.39: The text of the query.

```
{
  "took": 79,
  "timed_out": false,
  "_shards": {
    "total": 1,
    "successful": 1,
    "skipped": 0,
    "failed": 0
  },
  "hits": {
    "total": {
      "value": 10000,
      "relation": "gte"
    },
    "max_score": 33.32409,
    "hits": [
      {
        "_index": "wine_index",
        "_id": "X7mvwIwBUQJRwN2MuxWd",
        "_score": 33.32409,
        "_source": {
          "column1": 65022,
          "country": "South Africa",
          "taster_name": "Susan Kostrzewa",
          "taster_twitter_handle": "@suskostrzewa",
          "description": "Bright, fresh fruit mingled with a spicy toast give this wine a joyous but serious edge. Fresh but complex, with ripe fruit and spicy
            flavors and a lingering finish. Pair with richer seafood dishes, poultry.",
          "title": "Thelema 2007 Sutherland Chardonnay (Elgin)",
          "points": 88,
          "province": "Elgin",
          "variety": "Chardonnay",
          "price": 20,
          "designation": "Sutherland",
          "winery": "Thelema"
        }
      },
      {
        "_index": "wine_index",
        "_id": "GLmvwIwBUQJRwN2M07o8",
        "_score": 26.959726,
        "_source": {
          "column1": 107191,
          "country": "Italy",
          "province": "Tuscany",
          "variety": "Chardonnay",
          "price": 25,
          "description": "L'Erta is a Tuscan expression of Chardonnay that opens with a bright, golden color and spicy aromas of candied fruit, citrus, apricot
            pear, honey and a touch of spicy saffron. The wine sports a bold, saturated style.",
          "designation": "L'Erta",
          "winery": "Vigliano",
          "region_1": "Toscana",
          "title": "Vigliano 2010 L'Erta Chardonnay (Toscana)",
          "points": 88
        }
      },
      {
        "_index": "wine_index",
        "_id": "zbivwIwBUQJRwN2MrrBS",
        "_score": 26.457775,
        "_source": {
          "column1": 39275,
          "country": "US",
          "taster_name": "Virginie Boone",
          "taster_twitter_handle": "@vboone",
          "description": "Bright and alive in crisp pear and fresh citrus, this is a lovely, spicy Chardonnay, integrated and pleasing on the palate. The juicy
            fruit flavors delve into a creamy texture that lingers through to the finish.",
          "title": "Black Stallion 2012 Limited Release Chardonnay (Los Carneros)",
          "points": 91,
          "province": "California",
          "variety": "Chardonnay",
          "price": 35,
          "designation": "Limited Release",
```


Figure 3.40: The result of the query.

4 | Extras

In this chapter, some extra work is presented. This work is focused on the use of the visualization tool named **Kibana**. Through this tool, intuitive and impactful dashboards could be created, representing in a user_friendly the results of the queries. Note that the dashboard was created to be interactive, so to fully take advantage of it, the user's interaction with Kibana is needed.

4.1. Dashboard with simple statistics

The first dashboard to be presented is a simple dashboard with multiple panels, each containing a specific diagram.

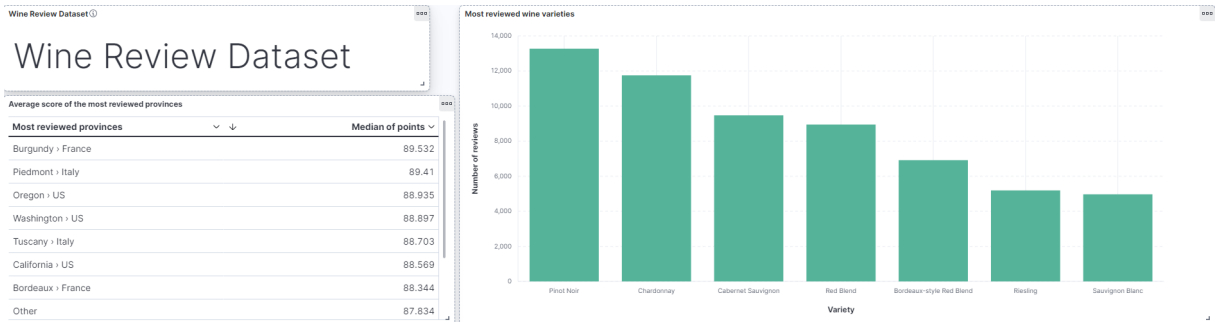


Figure 4.1: Top part of the dashboard.

Figure 4.2: Middle part of the dashboard.

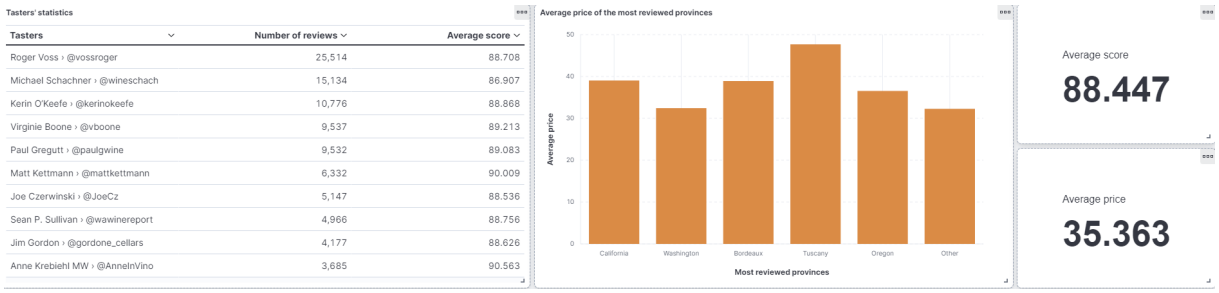


Figure 4.3: Bottom part of the dashboard.

In Figure 4.1, the name of the dataset is included, then the results of two queries are represented: in the left part, there is a ranking about the most reviewed provinces, with the median of the points scored by the wines produced in that province. It helps to find immediately what are the most relevant provinces for the production of wine. On the right, there is a histogram representing the most reviewed variety of wines: on the horizontal axis, there are the varieties of the wines, while on the vertical axis, there are the values representing the number of reviewed wines.

In Figure 4.2, on the left part there is another ranking showing the most expensive wines, identified by their variety and their winery and classified according to their price in a descending order. On the right, there is a pie chart showing information about the top 5 most reviewed countries. The user, at a glance, can know what are the countries mainly involved in the worldwide production of wine.

In Figure 4.3, on the left there is a ranking about the tasters, identified by their name and twitter_handle. Each taster is ranked according to the number of reviews he wrote and the average of the score he has assigned. The order could be ascending or descending for one of the two values (in the figure is descending for the number of reviews). In the middle, there is a histogram representing the average price of the most reviewed provinces: on the horizontal axis, there are the most reviewed provinces, while on the vertical axis, there are the values representing the average price. Finally, on the right, there are two small panles containing two values: the average points of the wines and their average price.

This dashboard is useful because the user has, at a glance, the main information that can support his decision, for example, of purchasing a specific wine. Moreover, these diagrams are more readable than the pure results of the query and improve the process of collecting information by summarizing, through specific queries, what is inside the dataset.

4.2. Dashboard: Query 1-10

The following dashboard is a representation of the queries from 1 to 10 previously described. The queries are represented by including a filter based on the query in the panel or by simply representing the result of that query. The complete dashboard is the following:

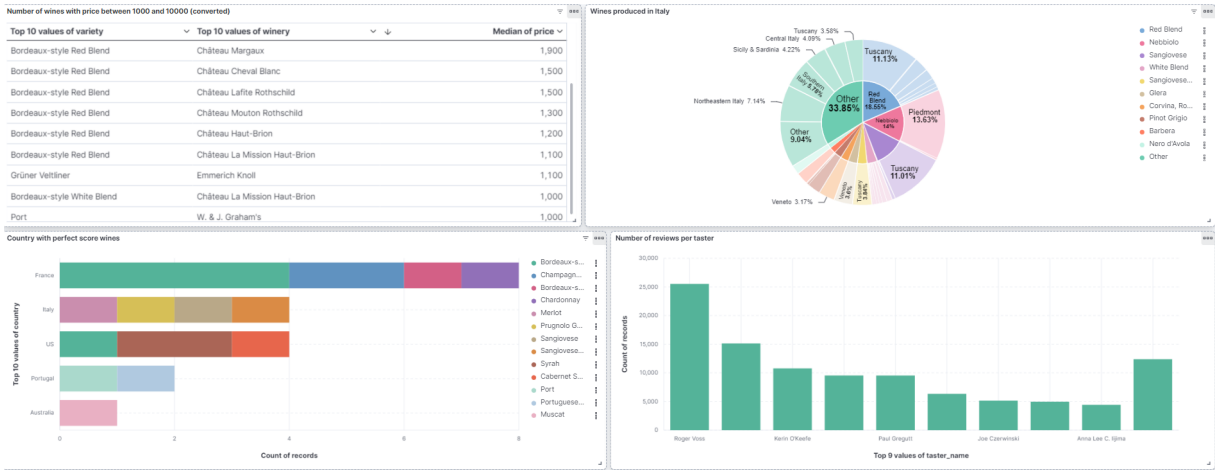


Figure 4.4: Top part of the dashboard.

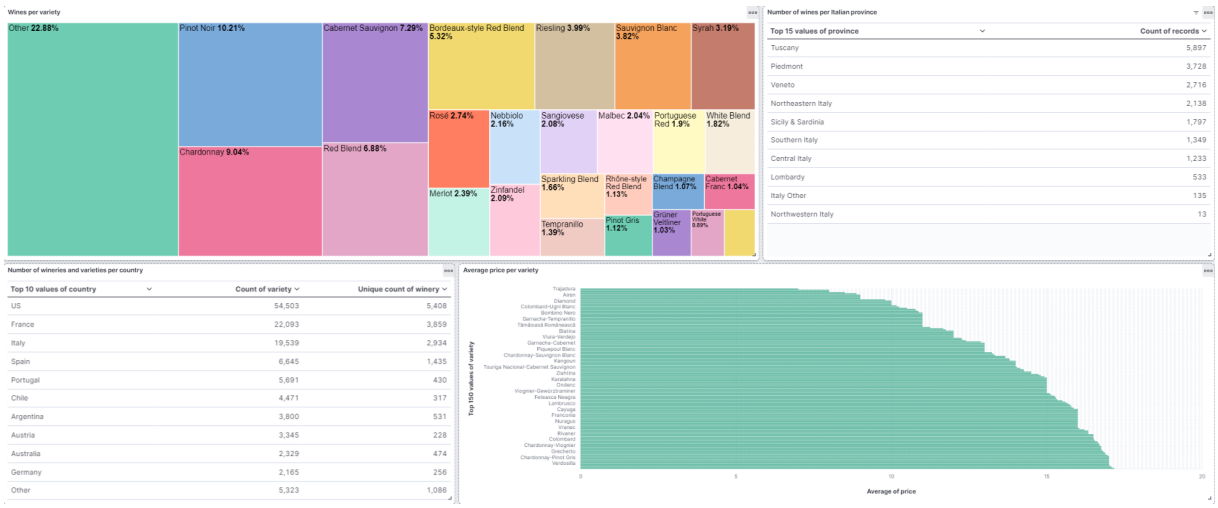


Figure 4.5: Middle part of the dashboard.

Figure 4.6: Bottom part of the dashboard.

In the following subsections, each diagram is explained, by pointing out what represents and why.

4.2.1. Dashboard: Query 1

The query labeled as query 1 is represented by the following diagram:

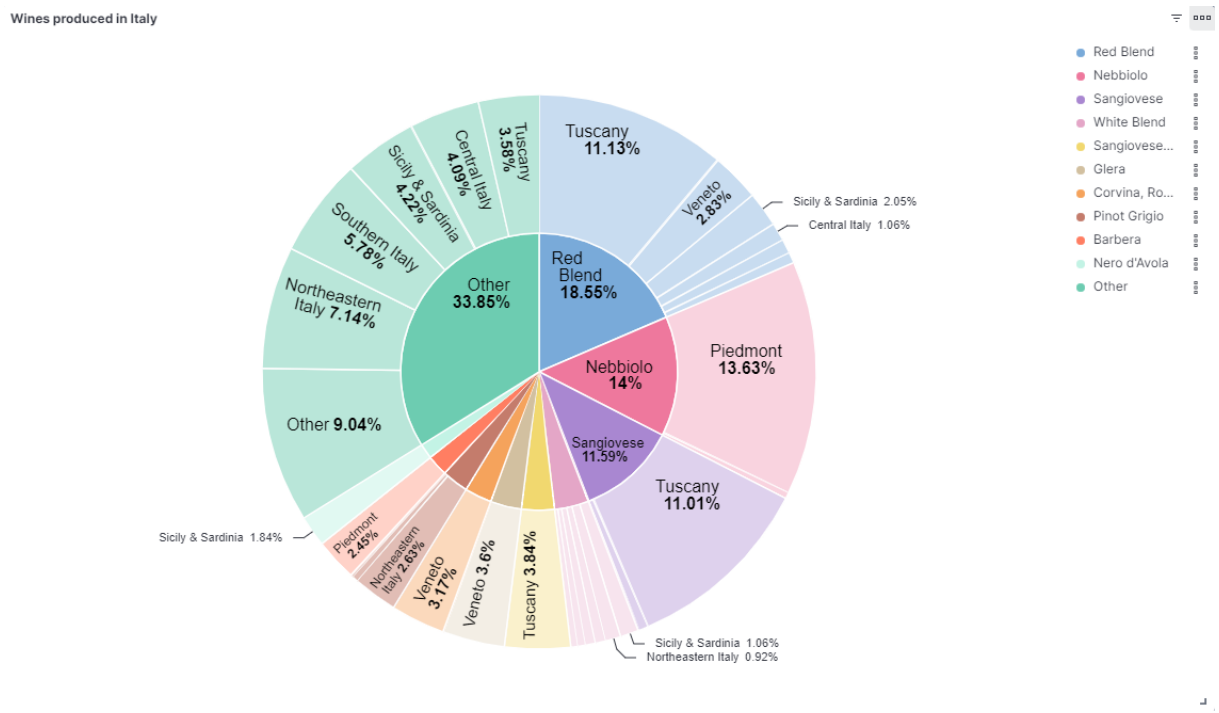


Figure 4.7: Representation of query 1.

It shows what are the wines produced in Italy by including the variety and the provinces of those wines inside a pie chart. We chose a pie chart because it allows to include more fields alongside the metric applied to the numerical field, i.e. the count of the wines satisfying the specified condition. It is very useful for statistics purposes rather than simple decision

ones because it is immediate to identify what are the most productive provinces and what are the varieties of wines mostly produced.

4.2.2. Dashboard: Query 2

The query labeled as query 2 is represented by the following diagram:

Number of wines with price between 1000 and 10000 (converted)

Top 10 values of variety	Top 10 values of winery	Median of price
Bordeaux-style Red Blend	Château les Ormes Sorbet	3,300
Bordeaux-style Red Blend	Château Pétrus	2,250
Chardonnay	Blair	2,013
Pinot Noir	Domaine du Comte Liger-Belair	2,000
Bordeaux-style Red Blend	Château Margaux	1,900
Bordeaux-style Red Blend	Château Cheval Blanc	1,500
Bordeaux-style Red Blend	Château Lafite Rothschild	1,500
Bordeaux-style Red Blend	Château Mouton Rothschild	1,300
Bordeaux-style Red Blend	Château Haut-Brion	1,200
Bordeaux-style Red Blend	Château La Mission Haut-Brion	1,100
Grüner Veltliner	Emmerich Knoll	1,100
Bordeaux-style White Blend	Château La Mission Haut-Brion	1,000
Port	W. & J. Graham's	1,000

Figure 4.8: Representation of query 2.

We chose a simple table to represent this query because it allows to put two different features to represent a specific wine, including then the price (the median of the price, because for numerical fields, only specific metrics are allowed). It is useful when the user wants to buy a valuable wine for a relevant event, for example. This diagram can speed up the decision of the user.

4.2.3. Dashboard: Query 3

The query labeled as query 3 is represented by the following diagram:

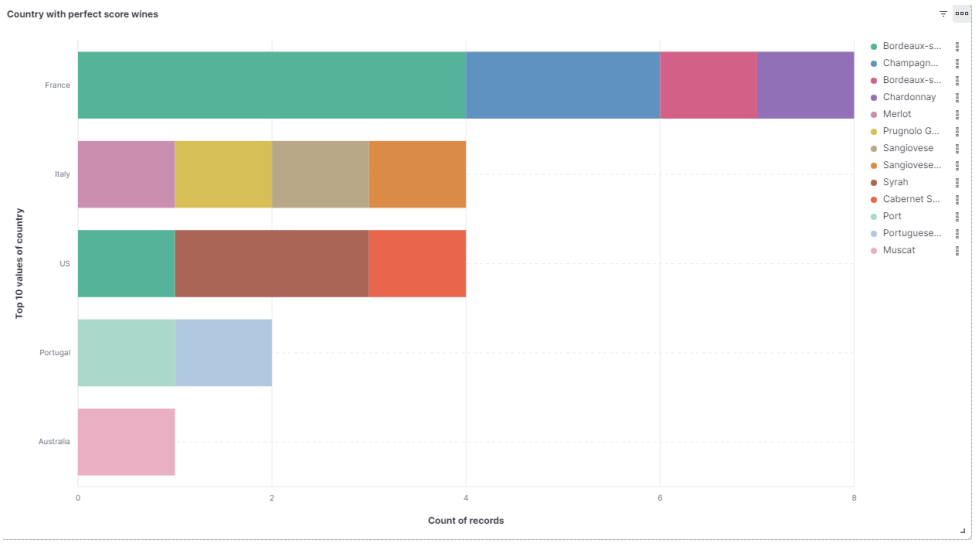


Figure 4.9: Representation of query 3.

It shows how many wines which have achieved a perfect score (100) are produced in the countries. On the vertical axis, there are the countries, while on the horizontal axis, there are the numerical values. This diagram underlines that if you want a relevant and valuable wine, probably you have to buy a French wine.

4.2.4. Dashboard: Query 4

The query labeled as query 4 is represented by the following diagram:

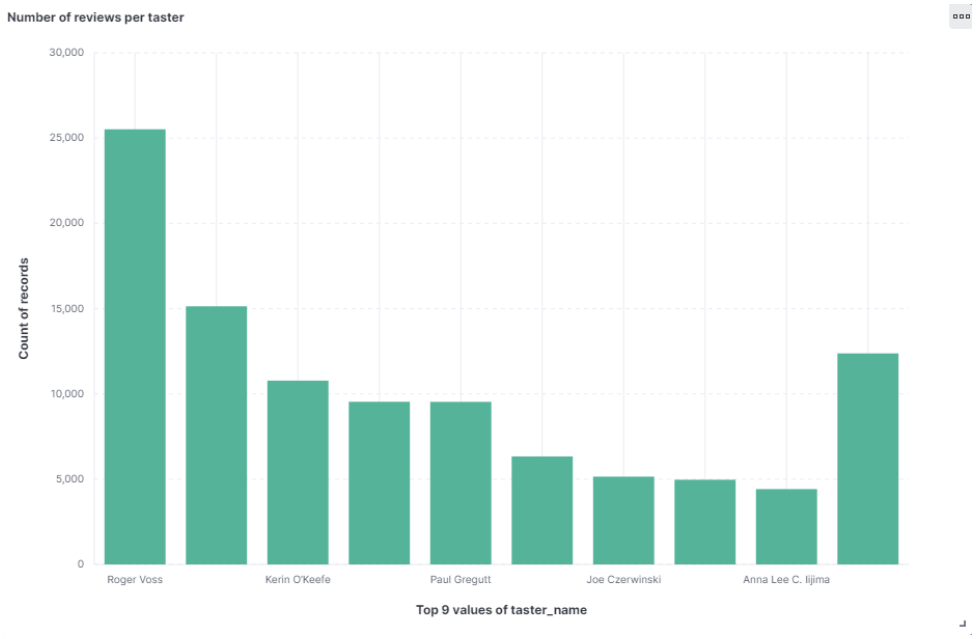


Figure 4.10: Representation of query 4.

This diagram highlights the number of reviews written by each taster. According to this diagram, if the user wants to read some reviews, he will read mostly reviews by Roger Voss, who is probably more influential than the others, based on how many reviews he has written, and so he has received a lot of requests of tasting wines.

4.2.5. Dashboard: Query 5

The query labeled as query 5 is represented by the following diagram:

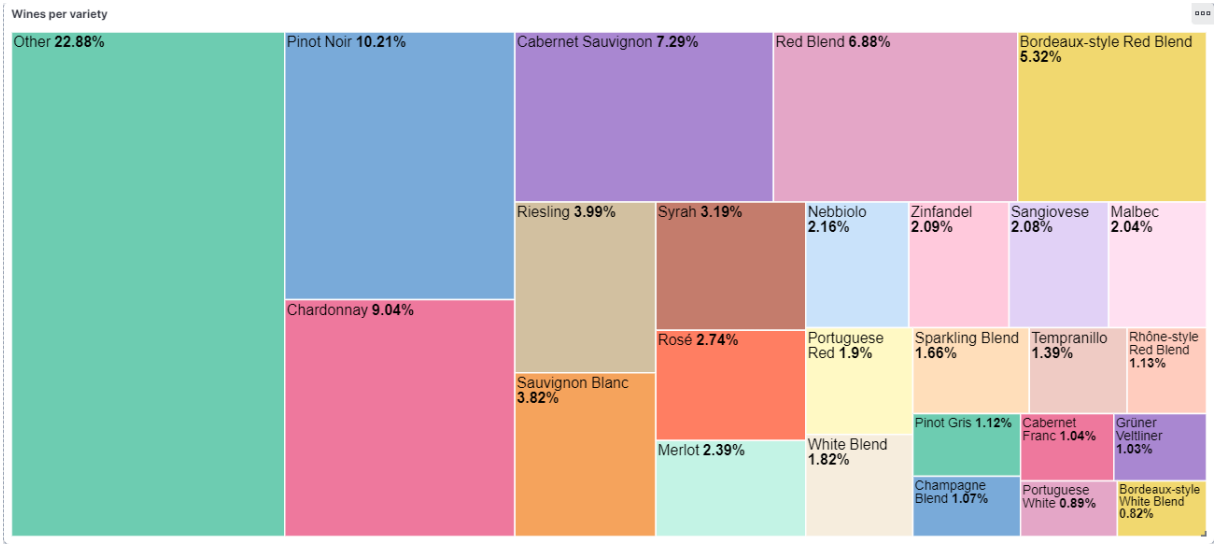


Figure 4.11: Representation of query 5.

It shows what are the varieties of wines that have a significant number of reviewed wines. We chose this kind of diagrams because it is able to provide at a glance the predominant varieties using different shapes. It is useful when the user wants to buy a specific wine and by looking to this diagram, he can know if the chosen wine has been reviewed sufficiently or it isn't much requested.

4.2.6. Dashboard: Query 6

The query labeled as query 6 is represented by the following diagram:

Number of wines per Italian province

Top 15 values of province	Count of records
Tuscany	5,897
Piedmont	3,728
Veneto	2,716
Northeastern Italy	2,138
Sicily & Sardinia	1,797
Southern Italy	1,349
Central Italy	1,233
Lombardy	533
Italy Other	135
Northwestern Italy	13

Figure 4.12: Representation of query 6.

It is a quite simple table to show, as a ranking, what are the most reviewed Italian provinces. So, the user, during his process of deciding which wine to buy, if the wine is Italian, he can look at this ranking and choose accordingly the province of production.

4.2.7. Dashboard: Query 7

The query labeled as query 7 is represented by the following diagram:

Number of wineries and varieties per country

Top 10 values of country	Count of variety	Unique count of winery
US	54,503	5,408
France	22,093	3,859
Italy	19,539	2,934
Spain	6,645	1,435
Portugal	5,691	430
Chile	4,471	317
Argentina	3,800	531
Austria	3,345	228
Australia	2,329	474
Germany	2,165	256
Other	5,323	1,086

Figure 4.13: Representation of query 7.

It is another table because it allows to include two different metrics on different fields, in this case it provides the number of varieties and wineries for each country. It could be useful for analytics purposes or for making a decision based on how many types of wines are produced in that country, selecting maybe a country which, producing nowadays a lot of wines, could be more familiar with the production of wines and those wines could have a high quality, but it is not guaranteed.

4.2.8. Dashboard: Query 8

The query labeled as query 8 is represented by the following diagram:

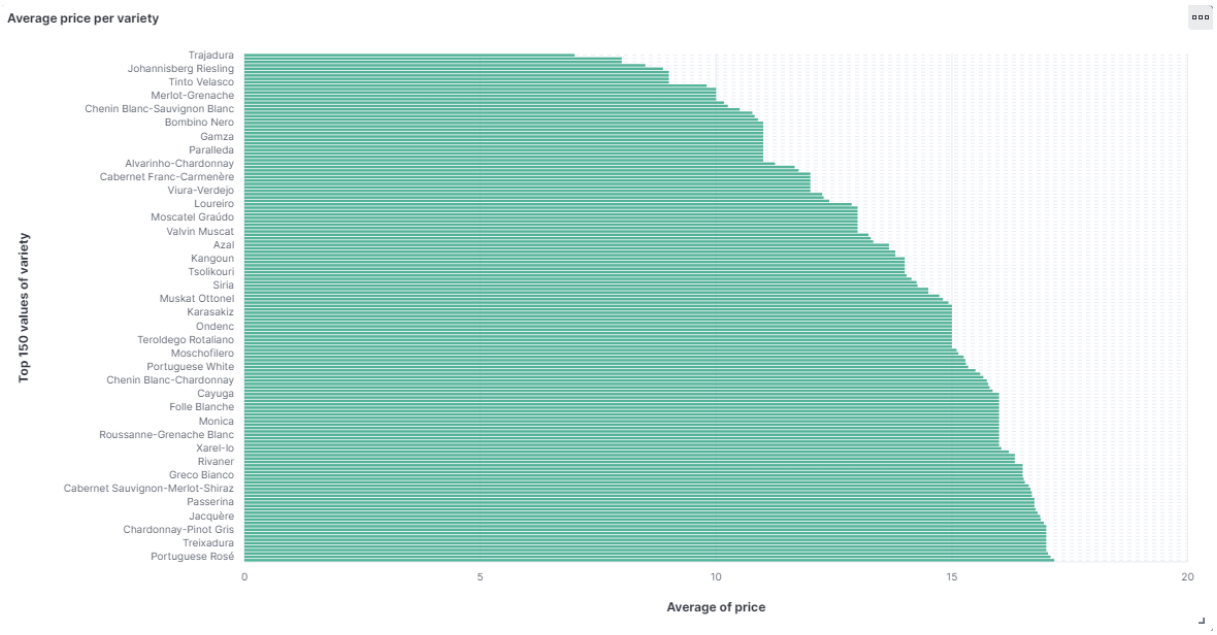


Figure 4.14: Representation of query 8.

This diagram provides the average price for each variety, so the user can refer to it to find the most suitable type of wine according to his economic availability.

4.2.9. Dashboard: Query 9

The query labeled as query 9 is represented by the following diagram:

Price and points per country

Top 20 values of country	Median of points	Median of price
England	91	50
Austria	90	24.672
Canada	90	30
India	90	12
Germany	89.917	26.444
Bulgaria	89	13
China	89	18
Hungary	89	25
Israel	89	25

Figure 4.15: Representation of query 9.

This table is fundamental after the decision of the variety of the wine to be purchased be-

cause the user must choose the country of production of the wine and having immediately available the medians of points and prices for each country makes the decision faster.

4.2.10. Dashboard: Query 10

The query labeled as query 10 is represented by the following diagram:

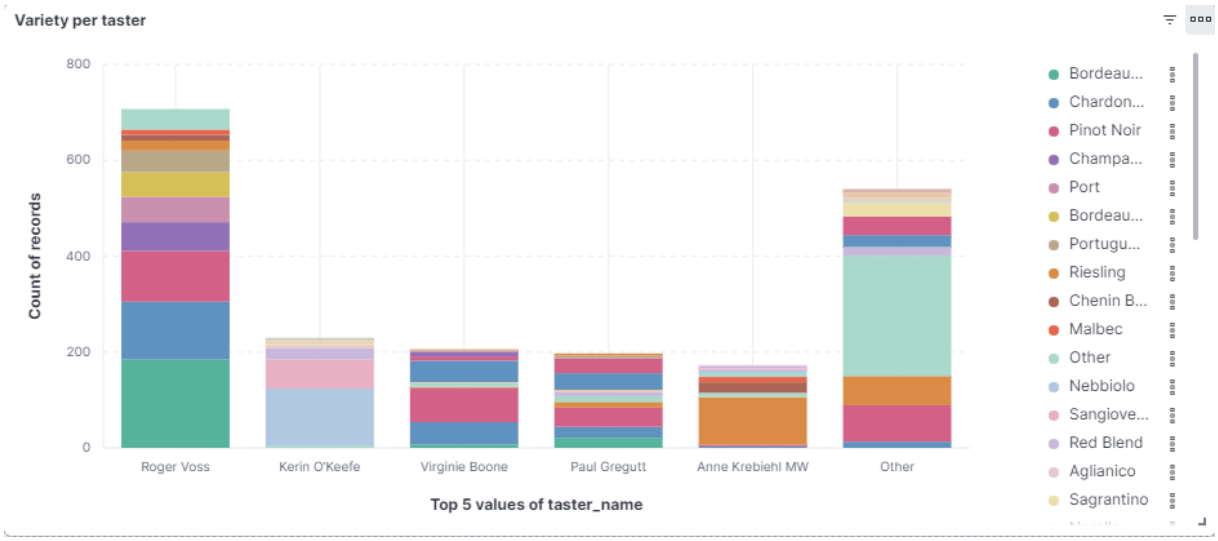


Figure 4.16: Representation of query 10.

This vertical shaped diagram shows how many high-scored wines each taster has reviewed, split by variety. It is impactful and well-structured because it mixes up two different features, providing to the user both the number of reviews and what is the variety mostly reviewed by that taster.