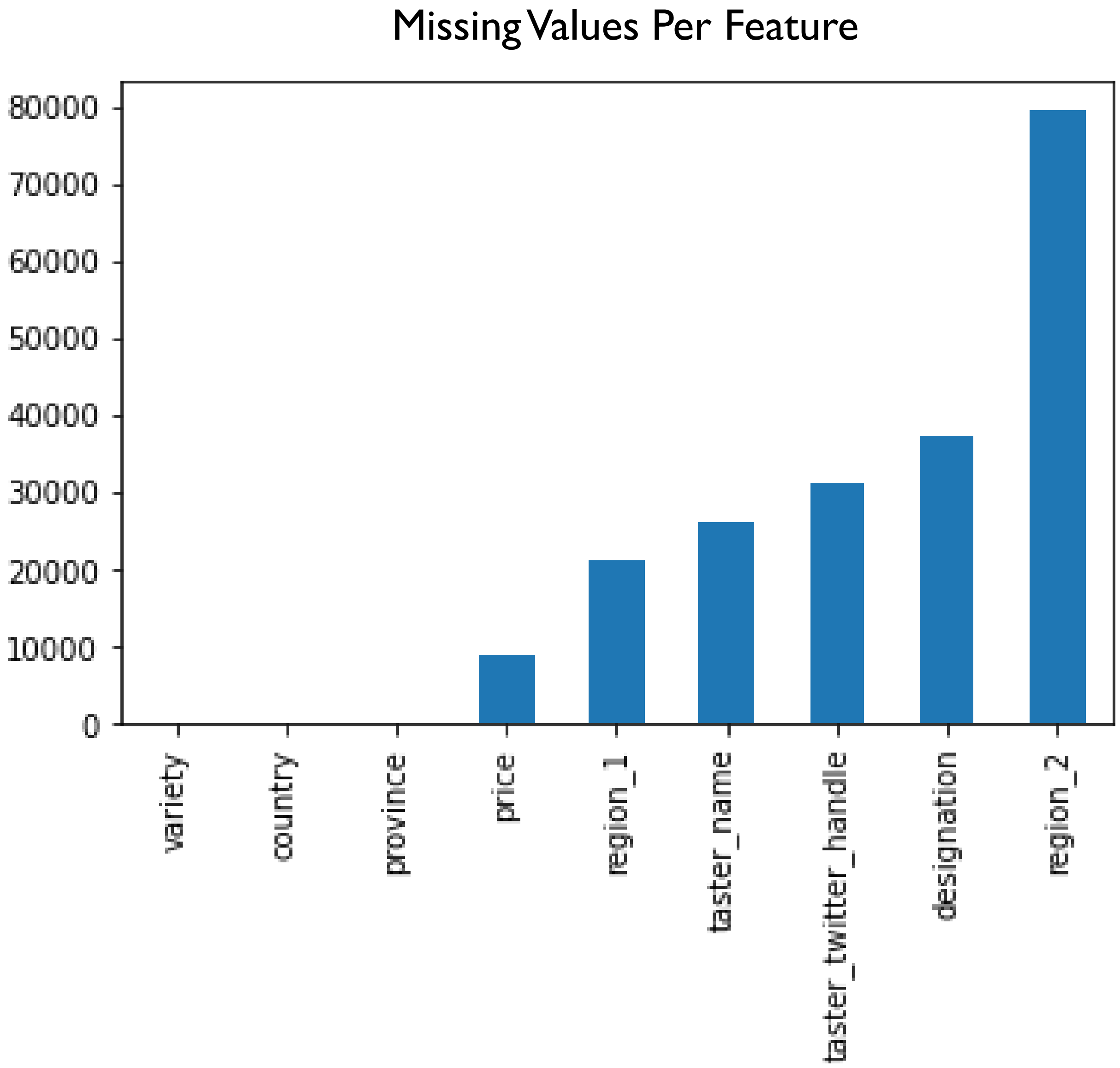
FCA Data Science Assessment Center Task
“Wine Enthusiast”

Tariq Hawili

DATA PRE-PROCESSING APPROACH

- Each problem is tackled with a different copy of the dataset.
  - Irrelevant features are dropped (depending on the problem).
  - Bad/non-sensical examples are removed from the dataset (e.g., 150 points out of 100)
  - Rows with missing values (after irrelevant column dropping) are removed entirely rather than imputed.



Relevant Variables	Variable Descriptions
designation	The vineyard within the winery where the grapes that made the wine are from
points	The number of points WineEnthusiast rated the wine on a scale of 1-100 (though they say they only post reviews for wines that score >=80)
winery	The winery that made the wine

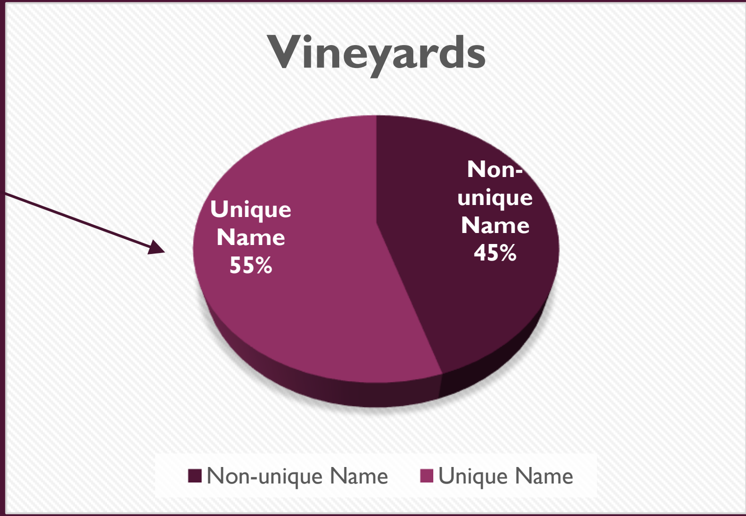
WHICH VINEYARD PRODUCES THE BEST WINE?

Relevant files:

best_Vineyard.py
wine_W_V.csv

Important to Note

- A designation (vineyard) will appear, on average, 2.45 times in the dataset.
- 45% of the data is labeled with a vineyard name that is shared by at least one other winery.
- Data shape after pre-processing: (130,147, 14) ⟶ (92,506, 3)



WHICH VINEYARD PRODUCES THE BEST WINE?

Complicating Factor:

Many wineries share the same name for their vineyards.

Solution:

1. Find the average number of times a vineyard appears in the dataset (call this 'N').
2. Store and sort the vineyards by mean points.
3. Remove all vineyards with count less than 4 * N.
4. Iterate through the list. The first vineyard that does not belong to more than one winery is the best.

Vineyard	Count	Unique Wineries
Reserve	2009	687
Estate	1322	460
Reserva	1259	402
Riserva	698	348
Estate Grown	621	188

THE BEST VINEYARD

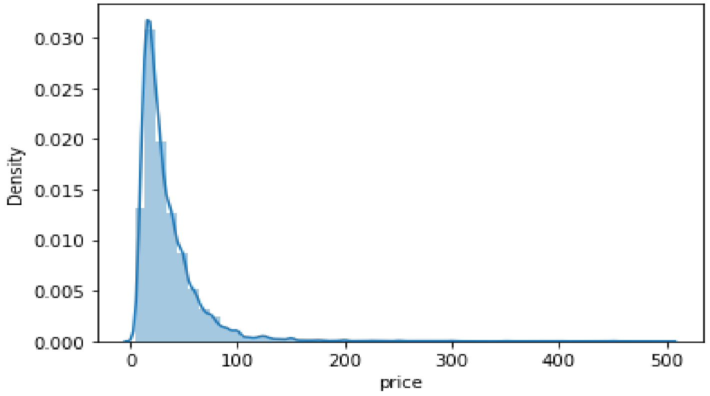
"Clos Saint Urbain Rangen de Thann Grand Cru"
Winery: Domaine Zind-Humbrecht
Bottles of wine in the dataset (count) : **11.0**
Average points awarded to its wine bottles: **95.36**

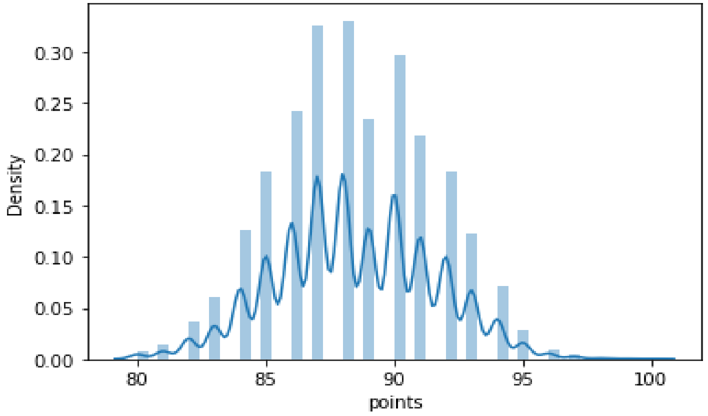
Relevant Variables	Variable Descriptions
points	The number of points WineEnthusiast rated the wine on a scale of 1-100 (though they say they only post reviews for wines that score >=80)
price	The cost of a bottle of wine
title	The title of the wine review
variety	The type of grapes used to make the wine (ie Pinot Noir)

Important to Note

- Price is skewed to the right:
  - Average = $35.36
  - Median = $25.00
  - Standard Deviation = $41.02

- Points are normally distributed.
  - Average = 88.42
  - Median = 88
  - Standard Deviation = 3.04





TOP 3 (RECOMMENDED) WINES

<u>Relevant files</u>:

top_3_wines.py
wine_top_3_wines.csv

DATA SHAPE AFTER PRE-PROCESSING:

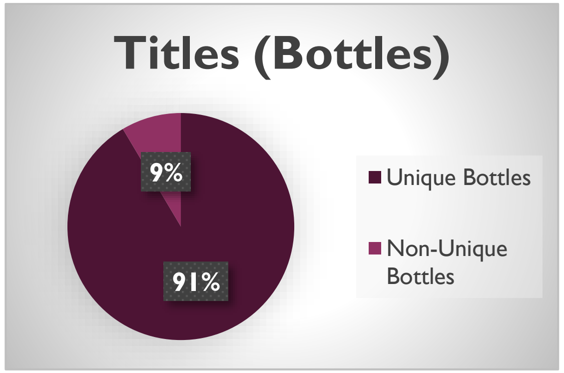
(130,147, 14) → (120,974, 4)

TOP 3 RECOMMENDED WINES

- **Objectives:**
  - **Low Price:** We want to recommend a bottle that isn't too expensive (within 0.5 standard deviations of the mean).
  - **Variety:** No two bottles can be of the same variety—we don't want to be boring.
  - **Points:** We want to recommend a bottle with the highest number of points given that the price and variety constraints are satisfied.
- **Solution:**
  1. Sort the bottles by points.
  2. Initialize an empty list "best_bottles."
  3. Iterate through the sorted bottles:
     1. If price is within 0.5 standard deviations of the mean, and—
     2. If a bottle of the same title or variety is not already in best_bottles, append the bottle to best_bottles.
     3. Once we have 3 bottles, stop.

Top 3 Recommended Wines

Points	Price	Title	Variety
99	$44.00	Failla 2010 Estate Vineyard Chardonnay (Sonoma Coast)	Chardonnay
98	$55.00	Gramercy 2010 Lagniappe Syrah (Columbia Valley (WA))	Syrah
98	$50.00	Pirouette 2008 Red Wine Red (Columbia Valley (WA))	Bordeaux-style Red Blend



This means that only thinking about those bottles that have a significant sample size is not feasible. We should consider all bottles equally, regardless of their presence in the data.

91% of the titles (bottles) in the dataset are unique.

Relevant Variables	Variable Descriptions
country	The country that the wine is from
points	The number of points WineEnthusiast rated the wine on a scale of 1-100 (though they say they only post reviews for wines that score >=80)
price	The cost of a bottle of wine
taster_name	The name of the taster of the wine
variety	The type of grapes used to make the wine (ie Pinot Noir)

Important to Note

- country has 42 unique values (categorical)
- price has 381 unique values (numerical)
- taster_name has 19 unique values (categorical)
- variety has 653 unique values (categorical)
- One-hot encoding performs poorly when applied to categorical variables with > 15 unique features.

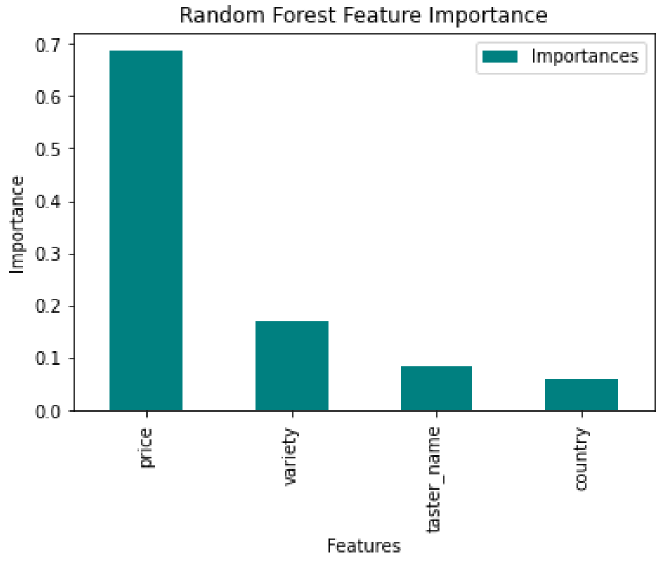
MOST IMPORTANT FEATURES

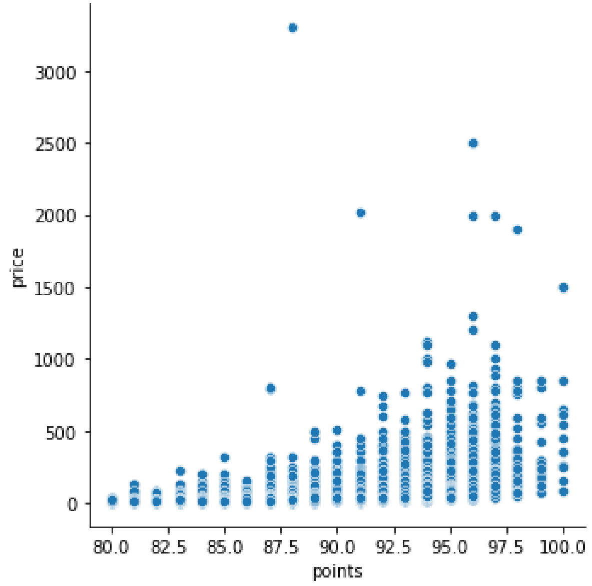
Relevant files:

Wine_topFactors.py
wine_top_factors.csv

DATA SHAPE AFTER PRE-PROCESSING:

(130,147, 14) → (196,420, 5)





FEATURE RANKING

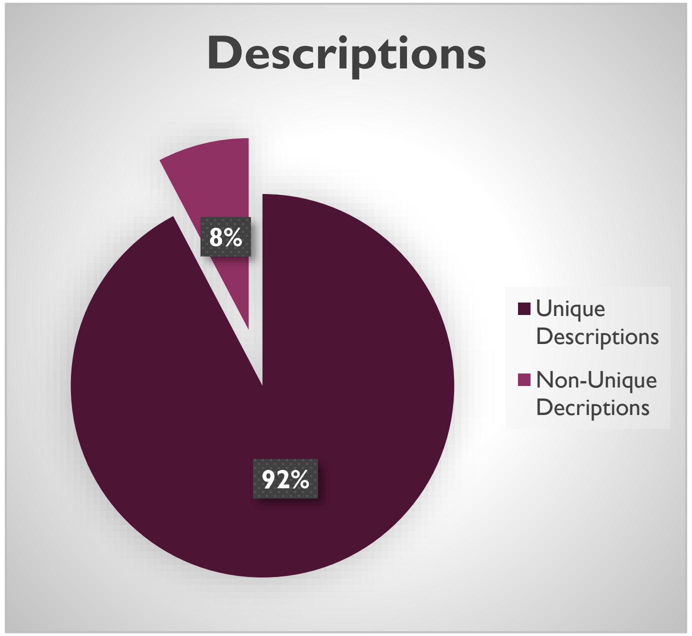
- <u>Solution</u>:
  1. Split the data into features and target.
  2. Transform the categorical feautres using LabelEncoder. One-hot encoding would be a very poor option due to the number of unique categorical variables we are dealing with.
  3. Fit a random forest regressor to the data.
  4. Calculate the importances using the RandomForestRegressor.

Feature Name	Importance
price	0.671292
variety	0.188343
taster_name	0.082739
country	0.057626

Variable	Description
description	A description of the wine
points	The number of points WineEnthusiast rated the wine on a scale of 1-100 (though they say they only post reviews for wines that score >=80)
variety	The type of grapes used to make the wine (ie Pinot Noir)

Important to Note

- There are some descriptions that are non-unique (duplicates), but this may not have been a problem since the final ordering was unaffected by removing duplicates.



"DRY" AND "CITRUS"

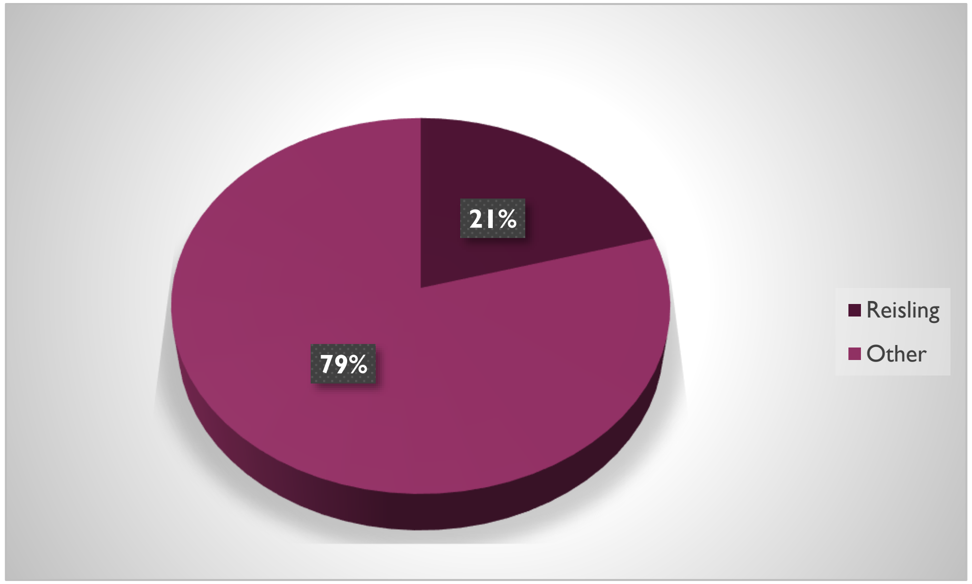
Relevant files:

dry_citrus_wine.py
wine_dry_citrus.csv

DATA SHAPE AFTER PRE-PROCESSING:

(130,147, 14) → (119,954, 3)

DRY AND CITRUS VARIETIES



- Number of "dry" and "citrus" varieties: 1630
  - **<u>Riesling: 423</u>**
    - **The next 4…**
      - **Sauvignon Blanc: 161**
      - **Chardonnay: 147**
      - **Sparkling Blend: 128**
      - **Champagne Blend: 77**
  - **Solution:**
    1. **Iterate through descriptions. If "dry" and "citrus" in description, then label as True.**
    2. **Select only those labeled True.**
    3. **Order by count.**
    4. **Recommend the variety at index 0.**

TOOLS/LIBRARIES USED

- pandas
- Visual Studio Code and Spyder
- Python 3.11
- Seaborn
- Matplotlib
- Scikit-learn
- Microsoft Excel, Powerpoint, Word