UT Data Analysis and Visualization Bootcamp

Project 4, Group 7

WineInsight: Predicting Ratings & Recommending Wines for Beginners

Lakshmi Adithya, Alyssandra Calhoun, Bianca Torres, Sabrina Martin, Jose Moncada

Welcome to our presentation

We are excited to share our project with you!

Presented for the UT Data Analysis and Visualization

Bootcamp

Project 4, Group 7

Team Members: Lakshmi Abbaraju, Alyssandra

Calhoun, Bianca Torres, Sabrina Martin, Jose Moncada.

Let us take you through our insights — sip by sip!.

01

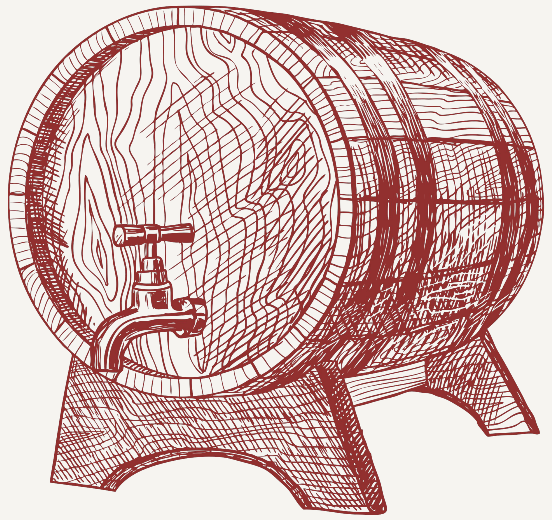
PROJECT OVERVIEW

Objective

- Analyze wine data to predict ratings and recommend beginner-friendly wines.
- Use machine learning to uncover patterns and enhance the wine selection experience.

Motivation

- Simplify wine choices for enthusiasts.
- Leverage data insights to guide wine preferences.



02

GOALS

GOALS

Data Cleaning & Transformation

Handle missing values.
Extract key features (e.g., vintage year).
Categorize ratings and wine styles.

Exploratory Data Analysis (EDA)

Visualize rating distribution, price trends, and variety patterns

Predictive Modeling

Use machine learning to forecast wine ratings.

Recommendation System

Suggest wines based on beginner preferences.

03

DATA CLEANING & TRANSFORMATION

Dataset Overview

- Source: **winemag-data-130k-v2.csv**
  - From Kaggle: Wine Reviews Dataset
  - Size: ~130,000 entries
- Key Features:
  - Country, Province, Region
  - Wine Variety, Winery, Taster Name
  - Points (Rating), Price
  - Wine Title (includes year and descriptions)

Data Cleaning & Transformation

- Handling Missing Values:
  - Fill or drop null entries for key features.
- Feature Engineering:
  - Extract vintage year from wine titles.
  - Categorize points into quality labels (e.g., excellent, good, average).
  - Classify wines into styles (red, white, rosé).
- Data Export:
  - Save cleaned dataset as **winemag-data-clean.csv**

Before

- Original CSV: **winemag-data-130k-v2.csv**

wine_no	country	description	designation	points	price	province	region_1	region_2	taster_nam	taster_twitt	title	variety	winery
0	Italy	Aromas incl	VulkÃ Bian	87		Sicily & Sarc	Etna		Kerin Oâ€™	@kerinokee	Nicosia 201	White Blen	Nicosia
1	Portugal	This is ripe	Avidagos	87	15	Douro			Roger Voss	@vossroger	Quinta dos	Portuguese	Quinta dos
2	US	Tart and snappy, the fla		87	14	Oregon	Willamette	Willamette	Paul Gregut	@paulgwin	Rainstorm 2	Pinot Gris	Rainstorm
3	US	Pineapple r	Reserve Lat	87	13	Michigan	Lake Michigan Shore		Alexander Peartree		St. Julian 20	Riesling	St. Julian

After

- Clean CSV: **winemag-data-clean.csv**

country	points	price	province	region_1	taster_nam	title	variety	winery	vintage_yea	style	rating_category
Portugal	87	15	Douro	Douro	Roger Voss	Quinta dos	Portuguese	Quinta dos	2011	red	good
US	87	14	Oregon	Willamette	Paul Gregut	Rainstorm 2	Pinot Gris	Rainstorm	2013	unknown	good
US	87	13	Michigan	Lake Michig	Alexander P	St. Julian 20	Riesling	St. Julian	2013	white	good
US	87	65	Oregon	Willamette	Paul Gregut	Sweet Chee	Pinot Noir	Sweet Chee	2012	red	good

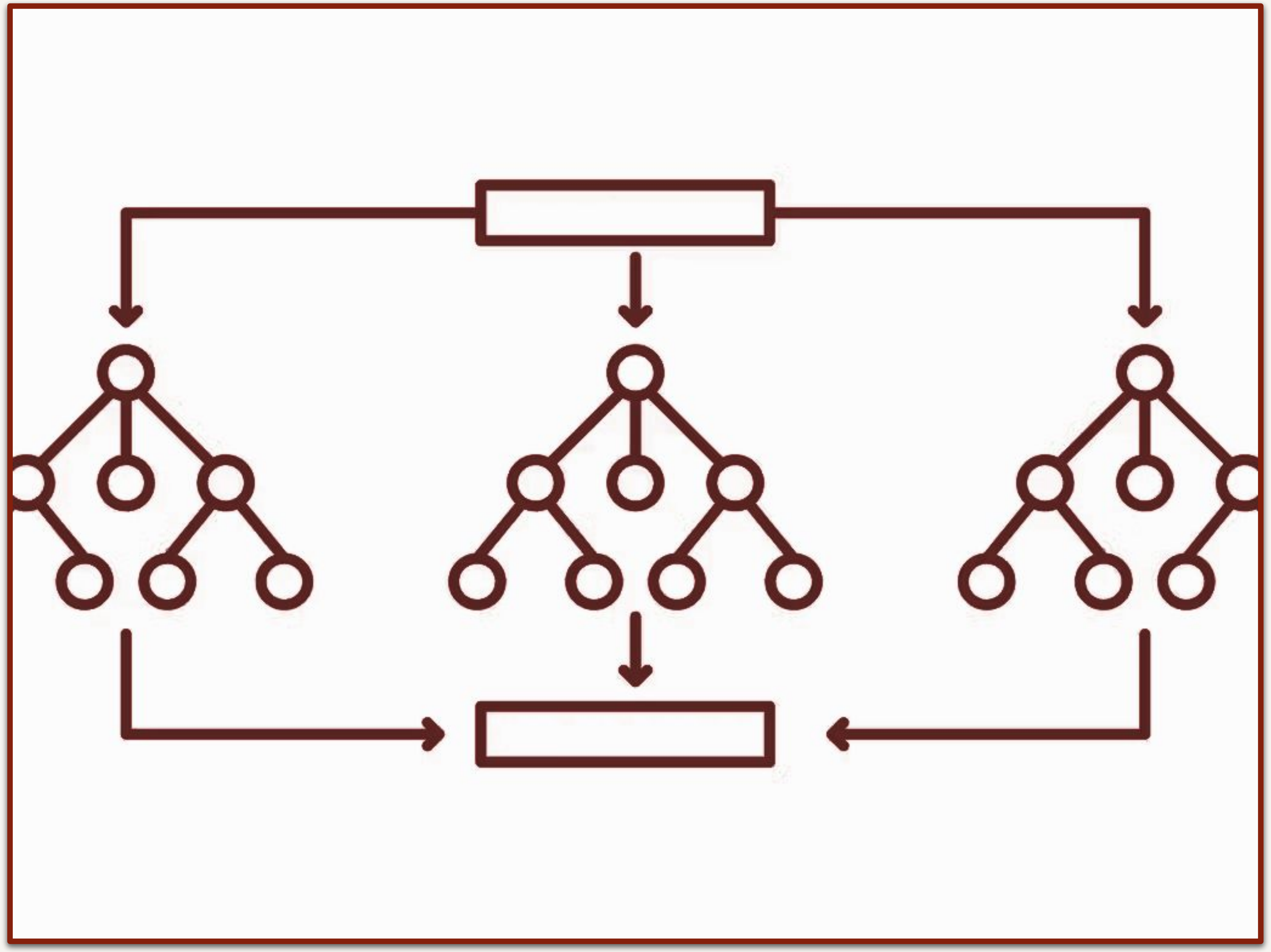
We processed the raw dataset and introduced three new features:

- Extracted the year from the title column.
- Determined a red or white classification using details from the variety, description, and title columns.
- Categorized ratings into five groups: below average, average, good, very good, and excellent.

04

MACHINE LEARNING

MACHINE LEARNING MODEL



Random Forest Classifier

- Random Forest Classifier model chosen to predict classifications in this large dataset.
- We decided to use 250 decision trees in this model

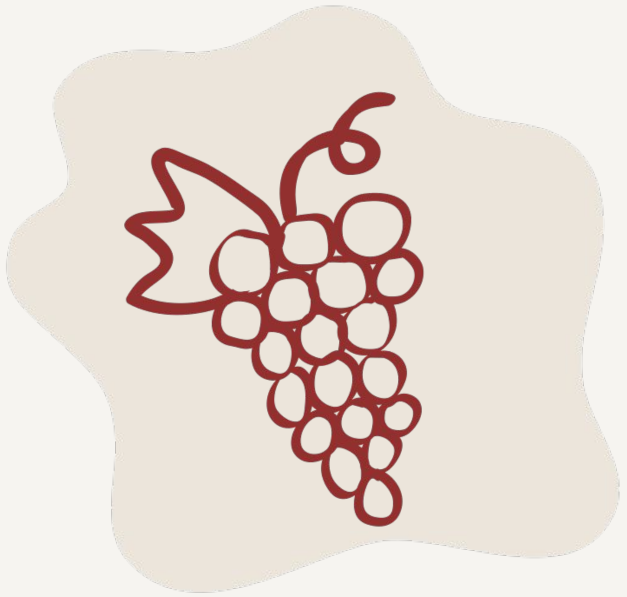
THE PROCESS

Clean Dataframe

Import dataframe

View value counts and info

Drop ‘unknown’ values and ‘title’ & ‘winery’ columns



OneHotEncoder

Convert ‘country’, ‘region’, ‘variety’, ‘style’ and ‘rating_category’ columns to binary values.

Data Frame is now ready for machine learning!

Fit Model

Train model on ‘rating_category_good’ column

Use sklearn to calculate feature importances

Use model to predict points_category_excellent’ column

Profit

Create and display confusion matrix

Analyze the results and accuracy score

Go out and buy an ‘excellent’ bottle of wine

RESULTS

Confusion Matrix

	Predicted 0	Predicted 1
Actual 0	3581	2434
Actual 1	2391	3961

```
Accuracy Score : 0.9949866580415623
----------------------------------------
Classification Report
              precision    recall  f1-score   support

         0.0       0.99      1.00      1.00     12117
         1.0       0.99      0.76      0.86       250

    accuracy                           0.99     12367
   macro avg       0.99      0.88      0.93     12367
weighted avg       0.99      0.99      0.99     12367

----------------------------------------
r2 Score
0.746883221919617
```

Importances

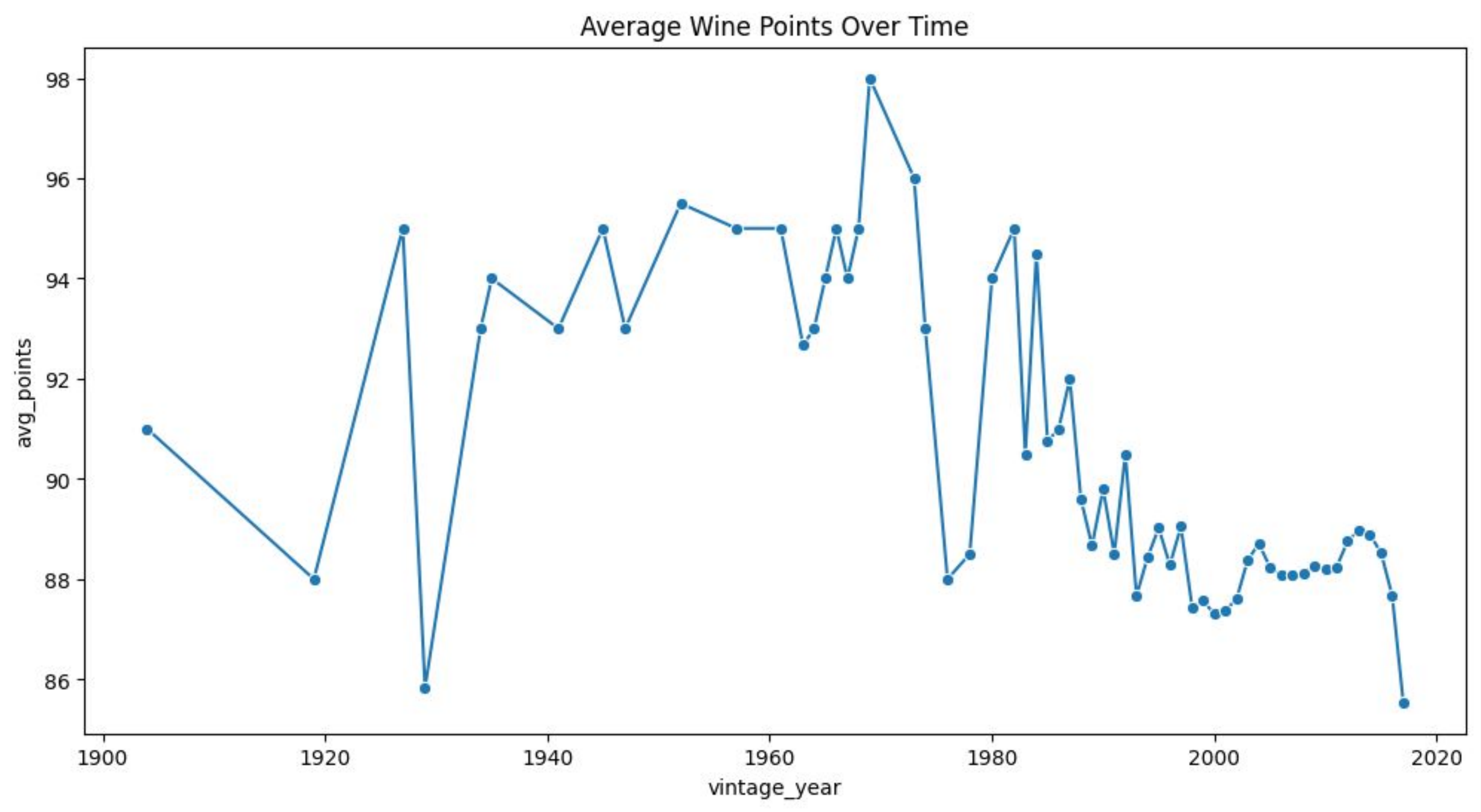
- Top 3 most important features:
  - Price (28%)
  - Vintage Year (19%)
  - Points (19%)

The Confusion Matrix

- The model scored 99% accuracy to predict the qualities that mark an 'rating_category_excellent' wine
- Worth noting is the model's ability to predict the true negative with a 76%

05

VISUALIZATIONS



Average Wine Points Over Time
avg_points
vintage_year
98
96
94
92
90
88
86
1900
1920
1940
1960
1980
2000
2020

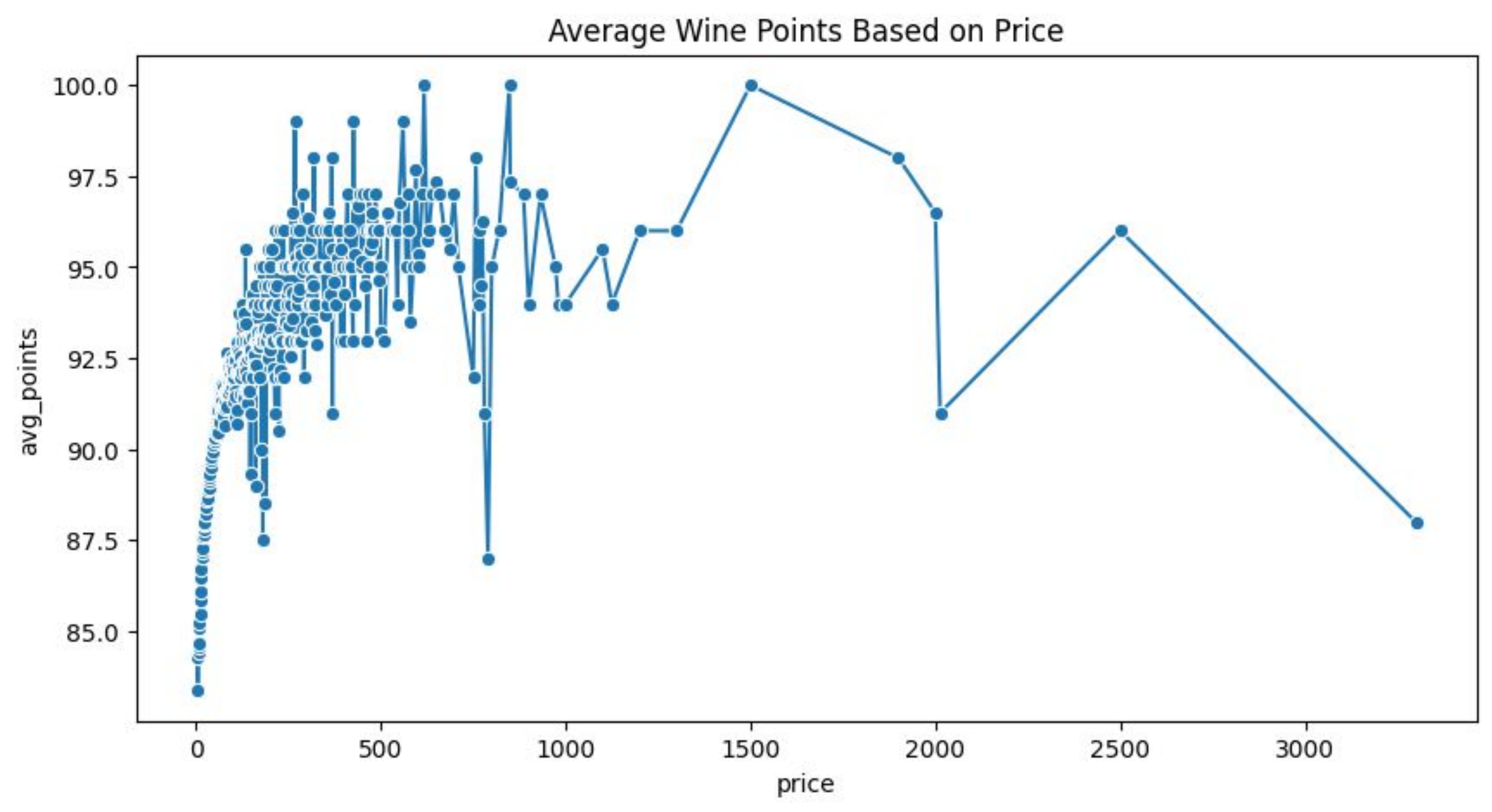

Average Wine Points Based on Price
avg_points
100.0
97.5
95.0
92.5
90.0
87.5
85.0
0
500
1000
1500
2000
2500
3000
price

TABLEAU - DASHBOARD

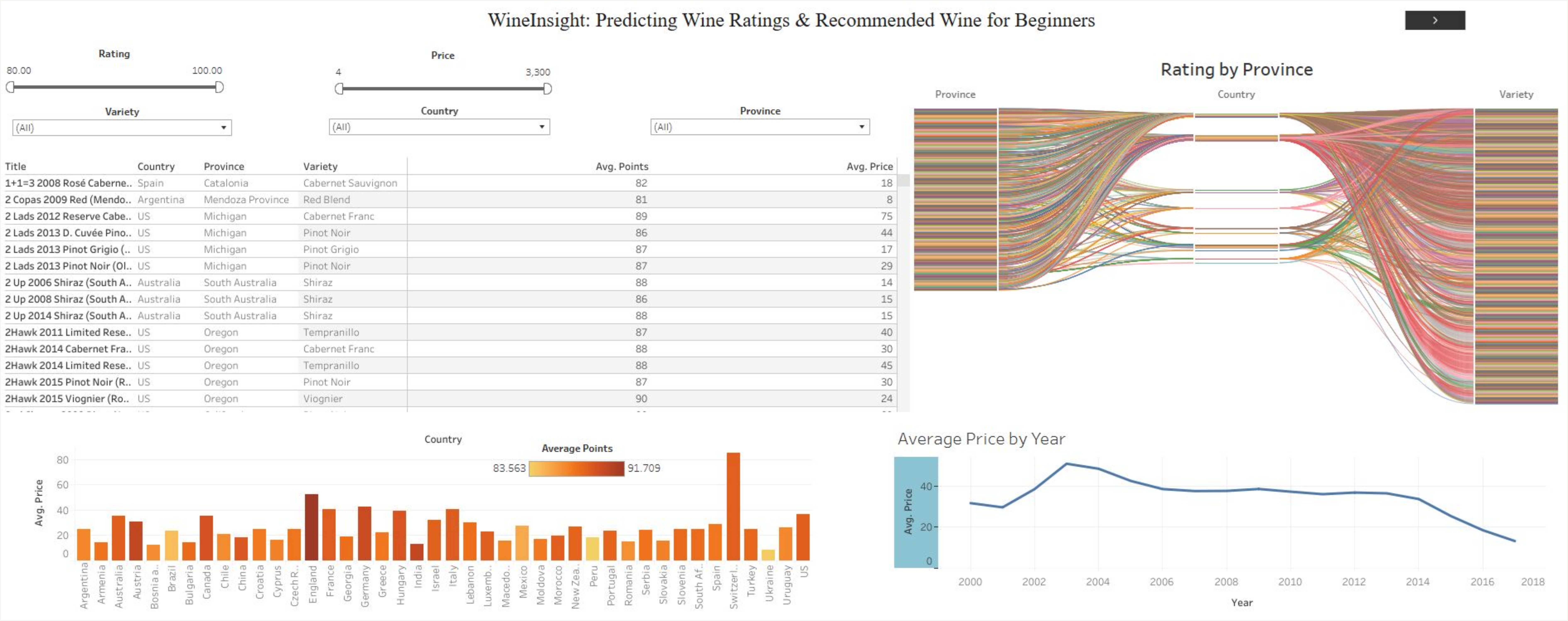


TABLEAU - STORY

Wine Insight Story

Wine by Variety | Average Ratings by Country | Wine TreeMap

Variety: (All)

Title	Country	Province	Variety	Avg. Points	Avg. Price
1+1=3 2008 Rosé Caberne..	Spain	Catalonia	Cabernet Sauvignon	82	18
2 Copas 2009 Red (Mendo..	Argentina	Mendoza Province	Red Blend	81	8
2 Lads 2012 Reserve Cabe..	US	Michigan	Cabernet Franc	89	75
2 Lads 2013 D. Cuvée Pino..	US	Michigan	Pinot Noir	86	44
2 Lads 2013 Pinot Grigio (..	US	Michigan	Pinot Grigio	87	17
2 Lads 2013 Pinot Noir (Ol..	US	Michigan	Pinot Noir	87	29
2 Up 2006 Shiraz (South A..	Australia	South Australia	Shiraz	88	14
2 Up 2008 Shiraz (South A..	Australia	South Australia	Shiraz	86	15
2 Up 2014 Shiraz (South A..	Australia	South Australia	Shiraz	88	15
2Hawk 2011 Limited Rese..	US	Oregon	Tempranillo	87	40
2Hawk 2014 Cabernet Fra..	US	Oregon	Cabernet Franc	88	30
2Hawk 2014 Limited Rese..	US	Oregon	Tempranillo	88	45
2Hawk 2015 Pinot Noir (R..	US	Oregon	Pinot Noir	87	30
2Hawk 2015 Viognier (Ro..	US	Oregon	Viognier	90	24
2nd Chance 2009 Pinot No..	US	California	Pinot Noir	90	29
2nd Chance 2009 Swiss Cl..	US	California	Pinot Noir	87	48
2nd Chance 2009 Syrah (S..	US	California	Syrah	91	32
2Plank 2013 Cold Smoke ..	US	California	Viognier-Chardonnay	84	25
2Plank 2013 Grenache-Sy..	US	California	Grenache-Syrah	91	28
2Plank 2013 Solstice Viog..	US	California	Viognier	88	34
2Plank 2013 Zinfandel (A..	US	California	Zinfandel	92	42
2Plank 2014 Bonsall Cabe..	US	California	Cabernet Sauvignon	91	52
2Plank 2014 Zinfandel (A..	US	California	Zinfandel	84	42
3 Badge Beverage 2014 G..	US	California	Chardonnay	84	30
3 Ball 2012 Zinfandel (Cali..	US	California	Zinfandel	88	15
3 Horse Ranch Vineyards ..	US	Idaho	Syrah-Mourvèdre	83	25
3 Horse Ranch Vineyards ..	US	Idaho	Syrah-Mourvèdre	87	25
3 Horse Ranch Vineyards ..	US	Idaho	Syrah	88	19
3 Horse Ranch Vineyards ..	US	Washington	Merlot	89	21
3 Horse Ranch Vineyards ..	US	Idaho	Pinot Gris	88	15
3 Horse Ranch Vineyards ..	US	Idaho	Cabernet Sauvignon-..	86	26
3 Horse Ranch Vineyards ..	US	Idaho	Rosé	88	15
3 Horse Ranch Vineyards ..	US	Washington	Bordeaux-style Red ..	86	26
3 Horse Ranch Vineyards ..	US	Idaho	Cabernet Sauvignon	87	23
3 Horse Ranch Vineyards ..	US	Idaho	Merlot	86	23
3 Horse Ranch Vineyards ..	US	Idaho	Rosé	84	15