Database Management Final Project Report

Alice Jiang
Andrew Huffman
Kai Le Becker
Kaidi Meng
Visar Ajroja

Dec. 11th, 2019

Project Report

Alice Jiang, Andrew Huffman, Kaidi Meng, Kai Le Becker, & Visar Ajroja

Introduction

Wine tasters undergo extensive training and spend years perfecting their palate. In this project, we will use data-driven analysis to identify reviewer bias in wines, how prices affect ratings of wine, and see if specific regions are preferred by reviewers.

Data Description

This project uses data from a 2017 Kaggle set of wine reviews from Wine Enthusiast Magazine (https://www.kaggle.com/zynicide/wine-reviews#winemag-data-130k-v2.csv). The original set contains 129,970 rows. This data set was too large to import to Apex Oracle however and needed to be shrunk somewhat. To reduce the set into a more manageable size, we chose to use 2000 of the reviews and eliminated the 2nd region column. We eliminated the 2nd region column because most records didn't have a 2nd region so it was not relevant to our analysis. Table 1 shows a description of the data fields.

Table 1:

Field	Туре	Description	
Country	Text	Country the wine is from.	
Description	Text	Wine taste description.	
Designation	Text	Vineyard within the winery where the grapes that made the wine are from.	
Points	Numeric	Number of points the WineEnthusiast rated the wine on a scale of 1-100.	
Price	Numeric	Cost for a bottle of wine.	
Province	Text	Province or state that wine is from.	
Region	Text	Wine growing area in a province or state.	
TasterFirstName	Text	First name of taster.	
TasterMiddleName	Text	Middle name of taster.	
TasterLastName	Text	Last name of taster.	
Handle	Text	Twitter handle of taster that uniquely identifies each taster.	
Title	Text	Title of the wine.	
Variety	Text	Types of grapes used to make the wine.	
WineID	Numeric	Unique ID for each wine.	
Winery	Text	Winery that made the wine.	
Winery	Numeric	Unique ID for each winery.	

The primary entity in this database is WINE, which is identified by Title. The raw data is relatively flat, with no multivalued data but some composite data identified in Title. A solution is to add a Wine ID and break apart Title into (Brand, Year). In our ERD there are several optional tabs that that are not representative of the data, all locations have wineries in this data, all wineries have bottled wine, but

not all wine has been reviewed by a reviewer. Some rows are missing data and will be dropped. Figure 1 displays the ERD model for our data.

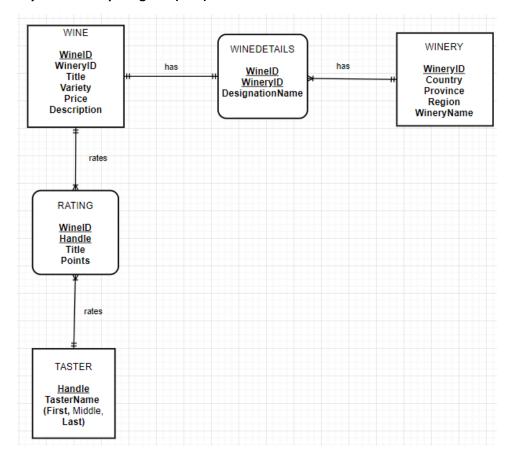
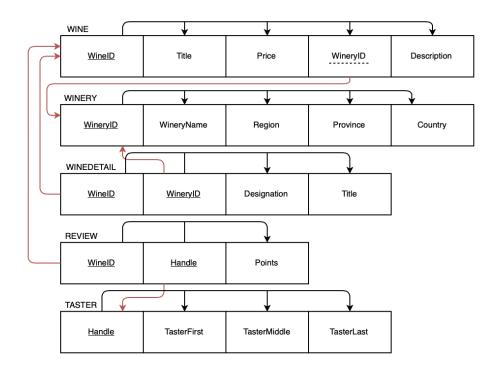


Figure 1: Entity Relationship Diagram (ERD)

Based on the ERD, we normalized the data and created a relational schema with 5 tables. Figure 2 displays the graphical relational schema for the wine data. We have 3 unique ID representing our data which are WineID, WineryID, and Handle.

Figure 2. Graphical Relational Schema



Database Implementation

To implement the database in APEX, we wrote a CREATE TABLE command for each table in the relational schema. Tables WINERY, WINE, and TASTER were entered first due to WINEDETAIL and REVIEW using data from the first three primary tables. Instead of using insert queries, data was uploaded directly to Apex Oracle through the uploading data tool.

The table WINERY is the winery where each wine is produced at and its location.

WINERY

CREATE TABLE WINERY (

WineryID NUMBER(7,0) NOT NULL,

Country VARCHAR2(20) NOT NULL,

Province VARCHAR2(50) NOT NULL,

Region VARCHAR2(100) NOT NULL,

WineryName VARCHAR2(100) NOT NULL,

CONSTRAINT WINERY_PK PRIMARY KEY (WineryID));

INSERT INTO WINERY VALUES (2003000, 'France', 'Alsace', 'Alsace', 'Test Winery')

The table WINE is to show the information on what wines were made where, cost, variety, title, and their description information.

<u>WINE</u>

```
CREATE TABLE WINE (
WineryID NUMBER(7,0) NOT NULL,
Title VARCHAR2(500) NOT NULL,
Variety VARCHAR2(50) NOT NULL,
Price NUMBER(3, 0) NOT NULL,
Description VARCHAR(700) NOT NULL,
WineID NUMBER(7,0) NOT NULL,
CONSTRAINT WINE_PK PRIMARY KEY (WineID),
CONSTRAINT WINE_FK FOREIGN KEY (WineryID) REFERENCES WINERY (WineryID));
INSERT INTO WINE VALUES (2003000, 'Test Wine', 'Riesling', 9, 'Test wine description', 3004000)
Taster table has information on the wine reviewer including their name and twitter handle as a primary
```

TASTER

key.

CREATE TABLE TASTER (

Handle VARCHAR2(50) NOT NULL,
TasterFirstName VARCHAR2(50) NOT NULL,
TasterMiddleName VARCHAR2(10),
TasterLastName VARCHAR2(50) NOT NULL,
CONSTRAINT TASTER_PK PRIMARY KEY (Handle));

INSERT INTO TASTER (Handle, TasterFirstName, TasterLastName) VALUES ('@TestTaster', 'Test', 'Taster')

WINEDETAIL is in the schema to separate wines from each other that share a designation, this is due to some designations having the same name even though they are at different wineries. The purpose was to retain 3rd normal form.

WINEDETAIL

```
CREATE TABLE WINEDETAIL (
WineryID NUMBER(7,0) NOT NULL,
WineID NUMBER(7,0) NOT NULL,
DesignationName VARCHAR2(300) NOT NULL,
CONSTRAINT DETAIL_PK PRIMARY KEY (WineID, WineryID));
INSERT INTO WINEDETAIL VALUES (2003000, 3004000, 'Test Designation')
```

Finally, REVIEW combines the taster table and the WINE table to allow us to query the scores that each reviewer gave each wine.

REVIEW

CREATE TABLE RATING (

Title VARCHAR2(500) NOT NULL,
Handle VARCHAR2(50) NOT NULL,
Points NUMBER(3,0) NOT NULL,
WineID NUMBER(7,0) NOT NULL,
CONSTRAINT RATING_PK PRIMARY KEY (WineID, Handle));

INSERT INTO RATING VALUES ('Test Wine', '@TestTaster', 83, 3004000)

Proposed Analysis

Our analysis is to help new wine connoisseurs in college learn basic information on what experienced wine tasters look for in wine, where better quality wines come from, and if it is reasonable that a student on a budget can afford particular wine.

Question 1: Wine Tasters

What are the names of the 5 most experienced wine tasters and the count of wines they have tasted?

To answer this question, we composed a simple query that counts the ratings for each wine taster, ranked it by a descending order and fetched the first 5 rows to show us the top 5 most experienced wine tasters.

SELECT Distinct(TasterFirstName | | ' ' | | TasterLastName) AS Name, COUNT(RATING.Handle) AS Count FROM RATING JOIN TASTER ON RATING.Handle = TASTER.Handle

GROUP BY TasterFirstName, TasterLastName
Order BY Count DESC
FETCH FIRST 5 ROWS ONLY;

The results from this query are shown in Figure 3 below. Out of all the wine tasters, Roger had the highest amount wines tasted with 331. The next most experienced was Michael with 330, followed by Virginie Boone with 273, Kerin with 268 and Paul with 248 wines tasted.

NAME	COUNT
Roger Voss	331
Michael Schachner	330
Virginie Boone	273
Kerin O,ÄôKeefe	268
Paul Gregutt	248

Figure 3: Top 5 Wine Tasters

Question 2: Ratings Based on Geography

Which geographic locations (countries) tend to have higher wine ratings? Do these areas tend to have more wineries?

We composed a subquery to answer this question. The inside query determines the top 5 countries based on their average rating. The outside query uses the countries from the inside query to count the number of wineries in that country.

The results from this query are shown below in Figure 4. The graph showing these results can be found in Figure 14.

COUNTRY	COUNT
US	932
France	377
Italy	281
Australia	70
Canada	2

Figure 4: Top 5 Countries with Higher Wine Rating

Question 3: Highest Wine Prices Based on Province

Which provinces have the highest wine prices?

To find out which provinces have the highest wine prices, we created a simple query. This query finds the max price of each province and only shows the provinces with the 10 highest prices.

SELECT Province, Max(Price) AS Max_Price
FROM WINERY JOIN WINE ON WINERY.WineryID = WINE.WineryID
GROUP BY Province
ORDER BY Max(Price) DESC
FETCH FIRST 10 ROWS ONLY;

PROVINCE	MAXPRICE		
Burgundy	630		
Tuscany	400		
Victoria	350		
Northern Spain	300		
California	260		
Piedmont	225		
South Australia	225		
Mendoza Province	215		
Rhône Valley	180		
Bordeaux	170		

Figure 5: Top 10 Province with Highest Wine Price

When comparing wine price by province we find the specific locations that wine will be sourced from quality input and processes. Burgundy produces the highest priced wines, so it seems unwise to buy wine from them as a beginner.

Question 4: Comparing Price to Rating

Is the cost of a bottle directly related to the rating it receives?

In we use a join query to find if there is a correlation between high wine ratings and wine price.

SELECT TO_CHAR(Price, '\$999,999') AS Price, ROUND(AVG(Points),1) AS AveragePoints, WINE.TITLE

FROM WINE JOIN RATING ON WINE.WineID = RATING.WineID

GROUP BY Price, Wine. Title

ORDER BY Price DESC;

PRICE	AVERAGEPOINTS	TITLE
\$630	96	Louis Latour 2014 Le Montrachet (Montrachet)
\$400	90	Tua Rita 2013 Redigaffi Merlot (Toscana)
\$350	95	Ch√¢teau de la Tour 2013 Vieilles Vignes (Clos de Vougeot)
\$350	100	Chambers Rosewood Vineyards NV Rare Muscat (Rutherglen)
\$350	98	Chambers Rosewood Vineyards NV Rare Muscadelle (Rutherglen)
\$300	95	Artadi 2011 Viña El Pison (Rioja)
\$260	94	Domaine Leflaive 2009 Les Pucelles Premier Cru (Puligny-Montrachet)
\$260	92	Stag's Leap Wine Cellars 2013 Cask 23 Estate Cabernet Sauvignon (Napa Valley)
\$225	95	Giacomo Conterno 2011 Francia (Barolo)
\$225	97	Torbreck 2012 RunRig Shiraz-Viognier (Barossa)

Figure 6: Relationship between Price and Average Rating

Figure 6 shows a cluster of similarly priced wines receiving scores above 90%, though we are descending our group by clause by price, there is a non-linear connection in points. Either

way, it is unlikely that wines at the highest price ranges will be noticeably better after the \$200 mark.

Question 5: Top 5 Varieties by Average Rating

What are the top 5 varieties of wine by their average rating? Are these wines typically red or white?

To find the 5 varieties with the highest average rating, we created a simple query. This query pulls the points from each rating and groups that by variety to get the average rating by variety.

SELECT Variety, ROUND(AVG(Points), 1) AS AveragePoints
FROM WINE JOIN RATING ON WINE.WineID = RATING.WineID
GROUP BY Variety
ORDER BY AVG(Points) DESC
FETCH FIRST 5 ROWS ONLY;

Figure 7 shows the top 5 varieties with their average rating.

VARIETY	AVERAGEPOINTS		
Muscadelle	98		
Muscat	95.7		
Shiraz-Viognier	95		
Cabernet Sauvignon-Syrah	94.5		
Cabernet Sauvignon-Malbec	94		

Figure 7: Top 5 Varieties by Average Rating

Figure 7 shows that there is a preference for white wines over red wines in the top percentages of wine ratings. This suggests that if a new wine taster wants to start with typically higher rated wines, they should begin with one of these five varieties.

Question 6: College-Friendly Wines

What wines are college friendly (price less then 20)?

In this query, there is a TO_CHAR in order to specify that our "price" column should be represented in dollars. There is also a case clause to simplify the answer if a wine is college friendly (when <\$20)

SELECT title, TO_CHAR(price,'\$999,999') as price, country, CASE
WHEN price < 20 THEN 'College Friendly'
ELSE 'Not College Friendly'
END as Friendly

FROM winery JOIN wine ON WINERY.WineryID = WINE.WineryID ORDER BY Friendly, Price;

TITLE	PRICE	COUNTRY	FRIENDLY
Felix Solis 2013 Flirty Bird Syrah (Vino de la Tierra de Castilla)	\$4	Spain	College Friendly
Don Cristobal 1492 2010 1492 Red (Mendoza)	\$7	Argentina	College Friendly
Dominio de Eguren 2011 Protocolo Rosè (Vino de la Tierra de Castilla)	\$7	Spain	College Friendly
Dominio de Eguren 2010 Protocolo Blanco White (Vino de la Tierra de Castilla)	\$7	Spain	College Friendly
Viña Albali 2012 Smooth Red Tempranillo (Valdepeñas)	\$8	Spain	College Friendly
Marqu√@s de C√°ceres 2006 Blanco Fermentado en Barrica White (Rioja)	\$8	Spain	College Friendly
Vieil Armand 2013 Tradition Pinot Blanc (Alsace)	\$8	France	College Friendly
Bodegas Fontana 2010 Mesta Tempranillo (Ucl√@s)	\$8	Spain	College Friendly
Vicente Gandia 2010 Organic Cabernet Sauvignon (Spain)	\$8	Spain	College Friendly

Figure 8: College-Friendly Wines

The data shown in figure 8 shows that there are many wines within the price range of a college student and there are several countries to choose from. Meaning that it is viable that a college student can have a varied palate without breaking the bank,

Web Design

https://apex.oracle.com/pls/apex/f?p=117109:1:708219334313961

The application design has several features of note; there is a home page that displays a several bottles of wine and a navigation bar on the side of the page. The navigation bar has 3 icons; a house for the home screen, three bars for the data tables, and 9 squares for our queries for differentiation purposes. There is an icon in the top left corner to replace the White lettering that said, "wine reviews" because it did not fit the color scheme. We also elected to create a custom color scheme based on the picture of wine bottles on our home page, giving the application a look that is both more mature and softer on the eyes than the blue and white default coloring.

Home Page

The home page contains a short introduction of our project, a navigation menu, and an image of wines (https://www.self.com/gallery/amazing-red-wines-under-20-dollars). Figure 9 shows the screenshot of the home page.

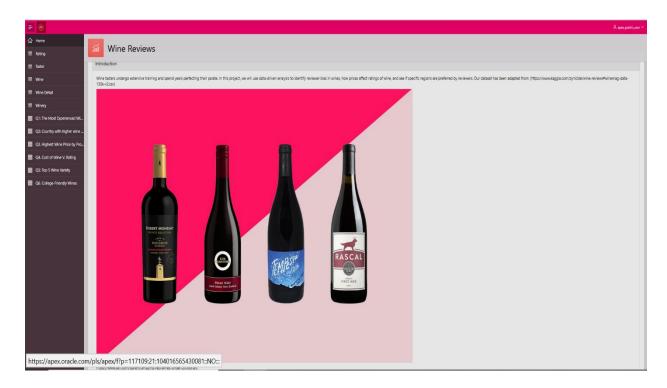


Figure 9: Home Page

Tables

We created an interactive report for each table in our data set. Users can easily search and filter results. (Figure 10 -Figure 14)

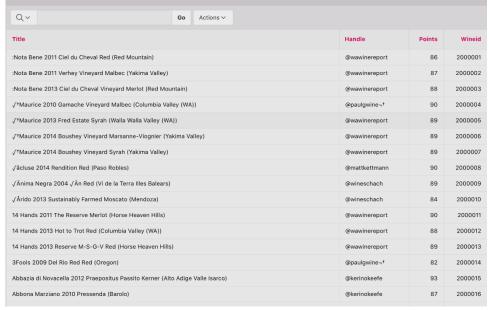


Figure 10: Rating Table

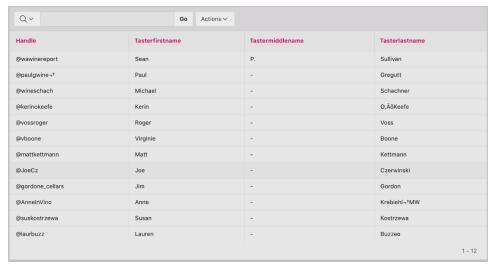


Figure 11: Taster Table

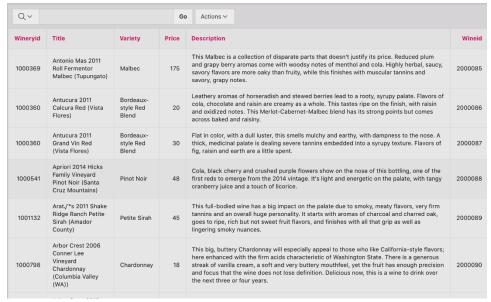


Figure 12: Wine Table

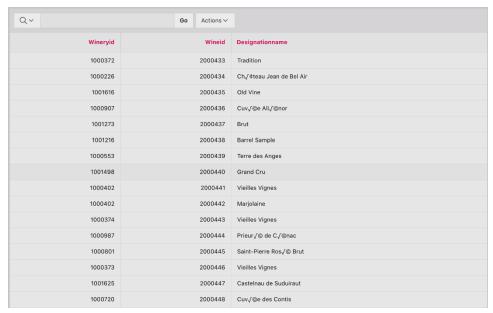


Figure 13: Wine Detail Table

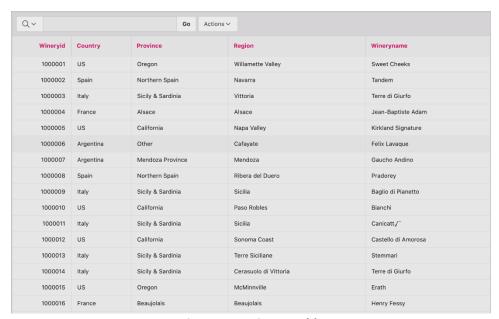


Figure 14: Winery Table

Queries

Query 1: We want to figure the most experienced wine taster, by creating a report, users can easily see the top 5 most experienced tasers and how many ratings they have done. (Figure 15)

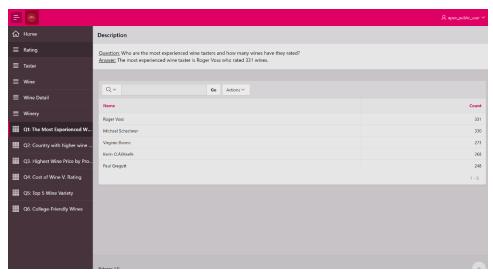


Figure 15: Q1 - Top 5 Wine Tasters

Query 2: Our second question looks at the five countries with the highest average ratings to see how many wineries are in those countries. We chose to use a bar chart here with the intent to show graphically how drastically different each country is when it comes to wine scores, with some nearly quadrupling the prior bar's ratings. (Figure 16).

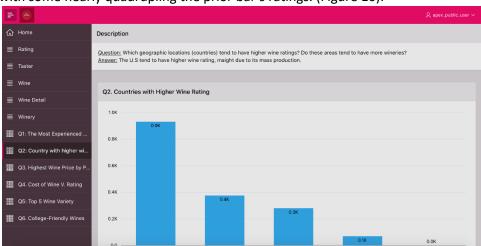


Figure 16: Q2 – Number of Wineries in Countries with Higher Wine Ratings

Query 3: The next question asks about the wine prices in province in each region and find the top 10 provinces with the higher wine prices. We chose to use a bar chart to visualize our results and a table on the side to easily lookup the information. (Figure 17)



Figure 17: Q3 – Highest Wine Price by Province

<u>Query 4:</u> The 4th query is a simple join query that uses the average function on the points column for particular wine. We grouped the scores by descending wine price and compared several prices from the top, middle, and bottom rows of the query. (Figure 18)

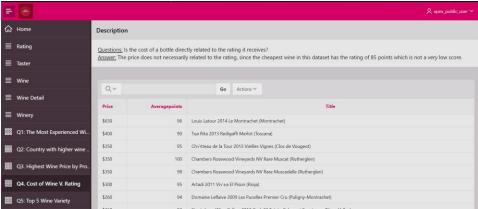


Figure 18: Q4 - Average Points and Price by Wine

Query 5: The 5th query shows the top five wine varieties based on their highest average points in table form as seen in Figure 19. The purpose was to find if red or white wines scored the highest in ratings. The returns show that the top two wines are white but the second three are reds. This shows that both colors of wine receive high ratings on average. (Figure 19)

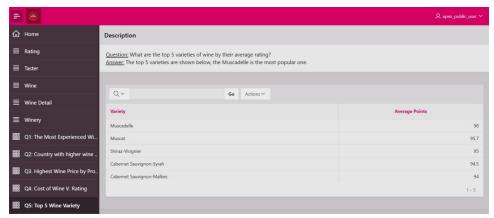


Figure 19: Q5 - Top Five Wine Varieties

Query 6: In the 6th query we wanted to show the most affordable wines for college kids. We concluded that a price of under \$20 is reasonable for college students, so we showed all the wines that are priced under \$20 and named them "College Friendly." (Figure 20)

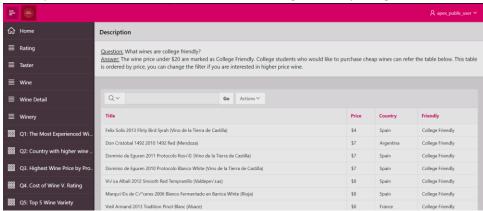


Figure 20: Q6 – The Most College Friendly Wines