STA 3024

Project Phase A

Eric Fernandez

Motivation

Wine making is a lengthy process that involves several factors such as environmental conditions, chemical properties of materials used, type of grape and others. It takes years of expertise to know how to produce and determine what makes a good wine and, even after a wine is produced, the flavor is constantly changing over time. Quality of wines is usually determined by sommeliers. Based on the information provided by the wine maker and the reviews by sommeliers, is there a way to meet the requirements of what makes a good wine and, by doing so, produce better wines of different varieties? Over the semester, I hope to find answers to some of the questions below:

1)Is there a way to determine what makes a great wine based on specific descriptors?

2)If so, is there any relationship between good quality and price? Can you predict how much a wine will cost based on a review by a sommelier?

3)Can we identify regions that produce better wine than others?

4)If there are regions that produce better wine than others, are there special conditions in these regions that help produce these results?

5)If there are special conditions in these regions, are these conditions replicable in regions that do not produce wine that could potentially produce similar quality wine?

By addressing these questions, I look to understand what are the main factors that determine the quality in wine.

Data Description

The dataset I am using was collected by Zack Thoutt scraping the website WineEnthusiast during the week of June 15th, 2017. The code used to scrap the data can be found here.

The columns describe the following attributes for every data point:

- *Points*: the number of points WineEnthusiast rated the wine on a scale of 1-100. WineEnthusiasts only post reviews for wines that score >=80.
- *Variety*: the type of grapes used to make the wine (ie Pinot Noir)
- *Description*: a few sentences from a sommelier describing the wine's taste, smell, look, feel, etc.
- *Country*: the country that the wine is from
- *Province*: the province or state that the wine is from
- *Region 1*: the wine growing area in a province or state (ie Napa)
- *Region 2*: sometimes there are more specific regions specified within a wine growing area (ie Rutherford inside the Napa Valley), but this value can sometimes be blank
- *Winery*: the winery that made the wine
- *Designation*: the vineyard within the winery where the grapes that made the wine are from
- *Price*: the cost for a bottle of the wine

These descriptions were created by the Zack Thoutt. The dataset was downloaded as a csv file from Kaggle(https://www.kaggle.com/zynicide/wine-reviews). By having the price, location and description from sommeliers, I will be able to answer questions 1 through 3. 4 will require more digging into the processes and conditions.

SAS Implementation

Some of the issues I encountered in this dataset were missing values for region fields and incorrect formatting when importing. Incorrect formatting was due to the fact that several lines were split into two or more in the original dataset file which caused SAS to have errors when reading the csv file. By concatenating the lines that were separated, this error was solved. Region data is not relevant to questions 1 and 2 but questions 3 and 4 need region fields so wines with no region data will not be considered for these particular questions.

Code:

```
/* Eric Fernandez Project-Phase A*/
/* I certify that the SAS code given is my original and exclusive work*/


/* To read the file:
 Create a new folder.
 Upload 'winemag-data_first150k.csv' to the folder
 Right click on 'winemag-data_first150k.csv' and select Properties
 Copy the path name and paste to the filename statement below
 Add a slash and the file name to the end of the path
*/
FILENAME CSV "~/datasets/winemag-data_first150k.csv" TERMSTR=LF;
/** Import the CSV file.  **/
PROC IMPORT DATAFILE=CSV
                    OUT=WineReviews
                    DBMS=CSV
                    REPLACE;
RUN;
/*Print out the first 20 reviews out of 150,000 reviews*/
proc print data=WineReviews(obs=20);
run;
```

Output:

Obs	number	country	description	designation	points	price	province	region_1	region_2	variety	winery
1	0	US	This tremendous 100% varietal wine hails from Oakville and was aged over three years in	Martha's Vineyard	96	235	California	Napa Valley	Napa	Cabernet Sauvignon	Heitz

Obs	number	country	description	designation	points	price	province	region_1	region_2	variety	winery
			oak. Juicy red-cherry fruit and a compelling hint of caramel greet the palate, framed by elegant, fine tannins and a subtle minty tone in the background. Balanced and rewarding from start to finish, it has years ahead of it to develop further nuance. Enjoy 2022–2030.								
2	1	Spain	Ripe aromas of fig, blackberry and cassis are softened and sweetened by a slathering of oaky chocolate and vanilla. This is full,	Carodorum Selección Especial Reserva	96	110	Northern Spain	Toro		Tinta de Toro	Bodega Carmen Rodríguez

Obs	number	country	description	designation	points	price	province	region_1	region_2	variety	winery
			layered, intense and cushioned on the palate, with rich flavors of chocolaty black fruits and baking spices. A toasty, everlasting finish is heady but ideally balanced. Drink through 2023.								
3	2	US	Mac Watson honors the memory of a wine once made by his mother in this tremendously delicious, balanced and complex botrytised white. Dark gold in color, it layers toasted hazelnut, pear compote and orange	Special Selected Late Harvest	96	90	California	Knights Valley	Sonoma	Sauvignon Blanc	Macauley

Obs	number	country	description	designation	points	price	province	region_1	region_2	variety	winery
			peel flavors, reveling in the succulence of its 122 g/L of residual sugar.								
4	3	US	This spent 20 months in 30% new French oak, and incorporates fruit from Ponzi's Aurora, Abetina and Madrona vineyards, among others. Aromatic, dense and toasty, it deftly blends aromas and flavors of toast, cigar box, blackberry, black cherry, coffee and graphite. Tannins are polished to a fine sheen, and frame a finish	Reserve	96	65	Oregon	Willamette Valley	Willamette Valley	Pinot Noir	Ponzi

Obs	number	country	description	designation	points	price	province	region_1	region_2	variety	winery
			loaded with dark chocolate and espresso. Drink now through 2032.								
5	4	France	This is the top wine from La Bégude, named after the highest point in the vineyard at 1200 feet. It has structure, density and considerable acidity that is still calming down. With 18 months in wood, the wine has developing an extra richness and concentration. Produced by the Tari family, formerly of Château Giscours in Margaux, it is a wine made for	La Brûlade	95	66	Provence	Bandol		Provence red blend	Domaine de la Bégude

Obs	number	country	description	designation	points	price	province	region_1	region_2	variety	winery
			aging. Drink from 2020.								
6	5	Spain	Deep, dense and pure from the opening bell, this Toro is a winner. Aromas of dark ripe black fruits are cool and moderately oaked. This feels massive on the palate but sensationally balanced. Flavors of blackberry, coffee, mocha and toasty oak finish spicy, smooth and heady. Drink this exemplary Toro through 2023.	Numanthia	95	73	Northern Spain	Toro		Tinta de Toro	Numanthia
7	6	Spain	Slightly gritty black-fruit aromas include a sweet note of pastry	San Román	95	65	Northern Spain	Toro		Tinta de Toro	Maurodos

Obs	number	country	description	designation	points	price	province	region_1	region_2	variety	winery
			along with a hint of prune. Wall-to-wall saturation ensures that all corners of one's mouth are covered. Flavors of blackberry, mocha and chocolate are highly impressive and expressive, while this settles nicely on a long finish. Drink now through 2024.								
8	7	Spain	Lush cedary black-fruit aromas are luxe and offer notes of marzipan and vanilla. This bruiser is massive and tannic on the palate, but still lush and friendly. Chocolate	Carodorum Único Crianza	95	110	Northern Spain	Toro		Tinta de Toro	Bodega Carmen Rodríguez

Obs	number	country	description	designation	points	price	province	region_1	region_2	variety	winery
			is a key flavor, while baked berry and cassis flavors are hardly wallflowers. On the finish, this is tannic and deep as a sea trench. Drink this saturated black-colored Toro through 2023.								
9	8	US	This re-named vineyard was formerly bottled as deLancellotti. You'll find striking minerality underscoring chunky black fruits. Accents of citrus and graphite comingle, with exceptional midpalate concentration. This is a wine to	Silice	95	65	Oregon	Chehalem Mountains	Willamette Valley	Pinot Noir	Bergström

Obs	number	country	description	designation	points	price	province	region_1	region_2	variety	winery
			cellar, though it is already quite enjoyable. Drink now through 2030.								
10	9	US	The producer sources from two blocks of the vineyard for this wine—one at a high elevation, which contributes bright acidity. Crunchy cranberry, pomegranate and orange peel flavors surround silky, succulent layers of texture that present as fleshy fruit. That delicately lush flavor has considerable length.	Gap's Crown Vineyard	95	60	California	Sonoma Coast	Sonoma	Pinot Noir	Blue Farm

Obs	number	country	description	designation	points	price	province	region_1	region_2	variety	winery
11	10	Italy	Elegance, complexity and structure come together in this drop-dead gorgeous winethat ranks among Italy's greatest whites. It opens with sublime yellow spring flower, aromatic herb and orchard fruit scents. The creamy, delicious palate seamlessly combines juicy white peach, ripe pear and citrus flavors while white almond and savory mineral notes grace the lingering finish.	Ronco della Chiesa	95	80	Northeastern Italy	Collio		Friulano	Borgo del Tiglio

Obs	number	country	description	designation	points	price	province	region_1	region_2	variety	winery
12	11	US	From 18-year-old vines, this supple well-balanced effort blends flavors of mocha, cherry, vanilla and breakfast tea. Superbly integrated and delicious even at this early stage, this wine seems destined for a long and savory cellar life. Drink now through 2028.	Estate Vineyard Wadensvil Block	95	48	Oregon	Ribbon Ridge	Willamette Valley	Pinot Noir	Patricia Green Cellars
13	12	US	A standout even in this terrific lineup of 2015 releases from Patricia Green, the Weber opens with a burst of cola and tobacco scents and accents. It continues,	Weber Vineyard	95	48	Oregon	Dundee Hills	Willamette Valley	Pinot Noir	Patricia Green Cellars

Obs	number	country	description	designation	points	price	province	region_1	region_2	variety	winery
			subtle and detailed, with flavors of oranges, vanilla, tea and milk chocolate discreetly threaded through ripe blackberry fruit.								
14	13	France	This wine is in peak condition. The tannins and the secondary flavors dominate this ripe leather-textured wine. The fruit is all there as well: dried berries and hints of black-plum skins. It is a major wine right at the point of drinking with both the mature flavors and the fruit in the right balance.	Château Montus Prestige	95	90	Southwest France	Madiran		Tannat	Vignobles Brumont

Obs	number	country	description	designation	points	price	province	region_1	region_2	variety	winery
15	14	US	With its sophisticated mix of mineral, acid and tart fruits, this seductive effort pleases from start to finish. Supple and dense, it's got strawberry, blueberry, plum and black cherry, a touch of chocolate, and that underlying streak of mineral. All these elements are in good proportion and finish with an appealing silky texture. It's delicious already, but give it another decade for full enjoyment. Drink now through 2028.	Grace Vineyard	95	185	Oregon	Dundee Hills	Willamette Valley	Pinot Noir	Domaine Serene

Obs	number	country	description	designation	points	price	province	region_1	region_2	variety	winery
16	15	US	First made in 2006, this succulent luscious Chardonnay is all about minerality. It's got a rich core of butterscotch and the seemingly endless layers of subtle flavors that biodynamic farming can bring. It spends 18 months on the lees prior to bottling. Drink now through 2028.	Sigrid	95	90	Oregon	Willamette Valley	Willamette Valley	Chardonnay	Bergström
17	16	US	This blockbuster, powerhouse of a wine suggests blueberry pie and chocolate as it opens in the glass. On the palate, it's smooth and seductively	Rainin Vineyard	95	325	California	Diamond Mountain District	Napa	Cabernet Sauvignon	Hall

Obs	number	country	description	designation	points	price	province	region_1	region_2	variety	winery
			silky, offering complex cedar, peppercorn and peppery oak seasonings amidst its dense richness. It finishes with finesse and spice.								
18	17	Spain	Nicely oaked blackberry, licorice, vanilla and charred aromas are smooth and sultry. This is an outstanding wine from an excellent year. Forward barrel-spice and mocha flavors adorn core blackberry and raspberry fruit, while this runs long and tastes vaguely chocolaty on the	6 Años Reserva Premium	95	80	Northern Spain	Ribera del Duero		Tempranillo	Valduero

Obs	number	country	description	designation	points	price	province	region_1	region_2	variety	winery
			velvety finish. Enjoy this top-notch Tempranillo through 2030.								
19	18	France	Coming from a seven-acre vineyard named after the dovecote on the property, this is a magnificent wine. Powered by both fruit tannins and the 28 months of new wood aging, it is darkly rich and with great concentration. As a sign of its pedigree, there is also elegance here, a restraint which is new to this wine. That makes it a wine for long-term aging.	Le Pigeonnier	95	290	Southwest France	Cahors		Malbec	Château Lagrézette

Obs	number	country	description	designation	points	price	province	region_1	region_2	variety	winery
			Drink from 2022.								
20	19	US	This fresh and lively medium-bodied wine is beautifully crafted, with cherry blossom aromas and tangy acidity. Layered and seductive, it offers a crisp mix of orange peel, cherry, pomegranate and baking spice flavors that are ready for the table or the cellar.	Gap's Crown Vineyard	95	75	California	Sonoma Coast	Sonoma	Pinot Noir	Gary Farrell

Future Direction

For the next steps, in order to analyze and learn about the dataset more I planned to:

- Find models to predict the variables I am looking for and compared them for this use case.
- Find better ways to visualize the data.

- Research about processes used to create wine and conditions of regions with good wine quality in this dataset.
- Create a dictionary of most common words used by sommeliers.
- Find regions that have similar conditions to the ones used in the dataset that currently do not produce wine to check if there are new regions that could potentially produce similar kind of wines.

We, the project team members, certify that the percentage of the effort listed by each of our names below is an accurate account of the original effort contributed by each team member in the producing of this project and report:

Name (Printed)	Percent of Total Effort	Statistics Major?
Eric Fernandez	100	No

Project Phase B

Introduction

My main motivation for this project is to find if there are significant relationships between countries and quality of wine, varieties and price, and quality and price. This analysis can help produce better wines, predict numerically how good it would be and decide which factors are most important when producing a high quality wine.

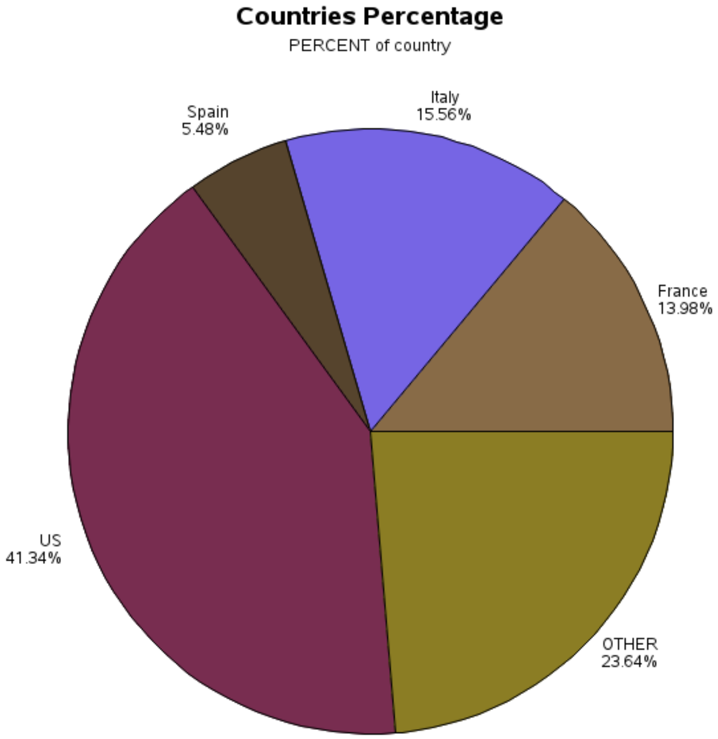
The variables of interest used for this phase are:

- *Points*: the number of points WineEnthusiast rated the wine on a scale of 1-100. WineEnthusiasts only post reviews for wines that score >=80.
- *Country*: the country that the wine is from
- *Price*: the cost for a bottle of the wine
- *Variety*: the type of grapes used to make the wine (ie Pinot Noir)

In this phase, I will explore the wine review dataset by using graphical displays and numerical summaries to find if there is a relationship between price and quality and how each country category compares to each other.

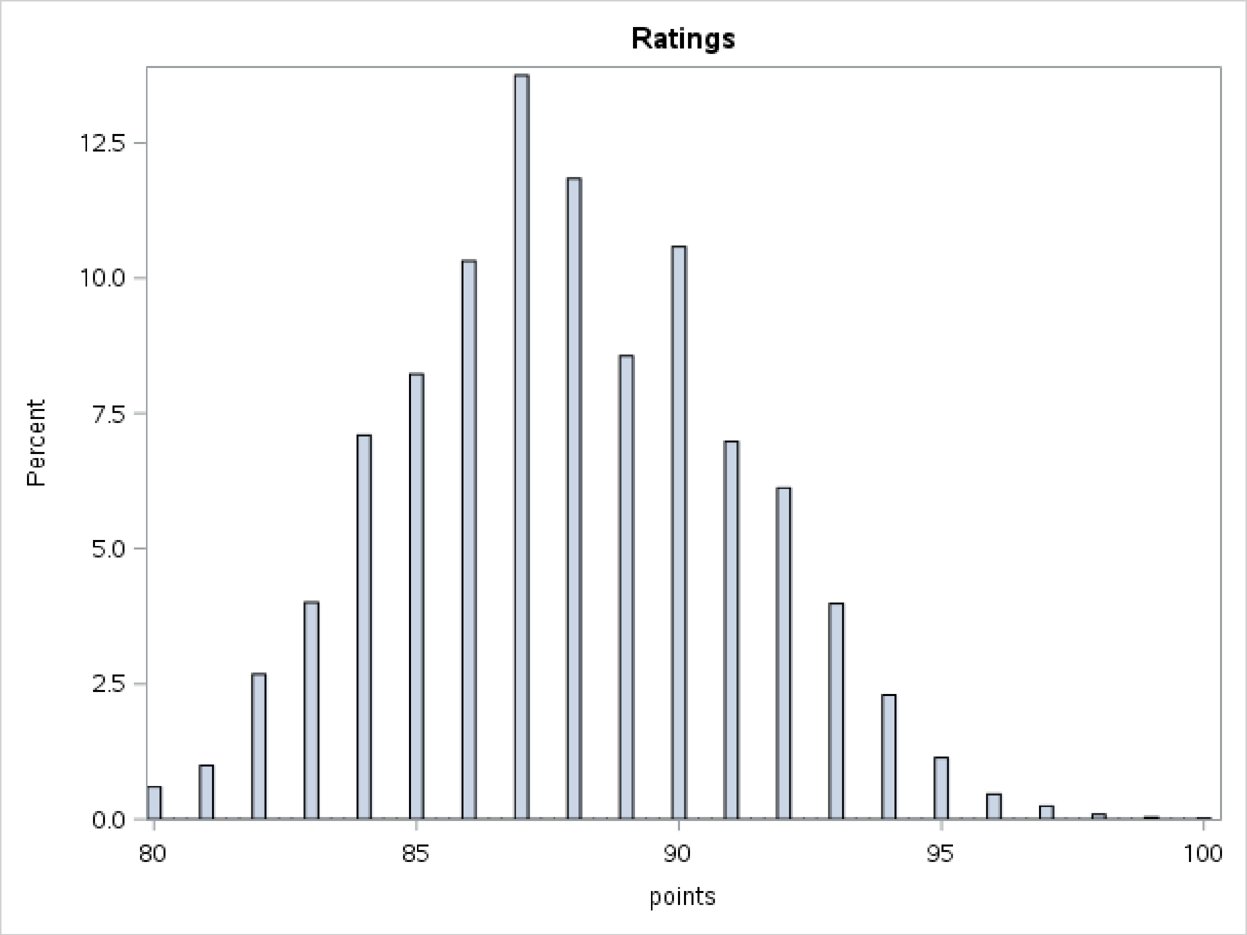
Data Exploration via Graphical Display

a)



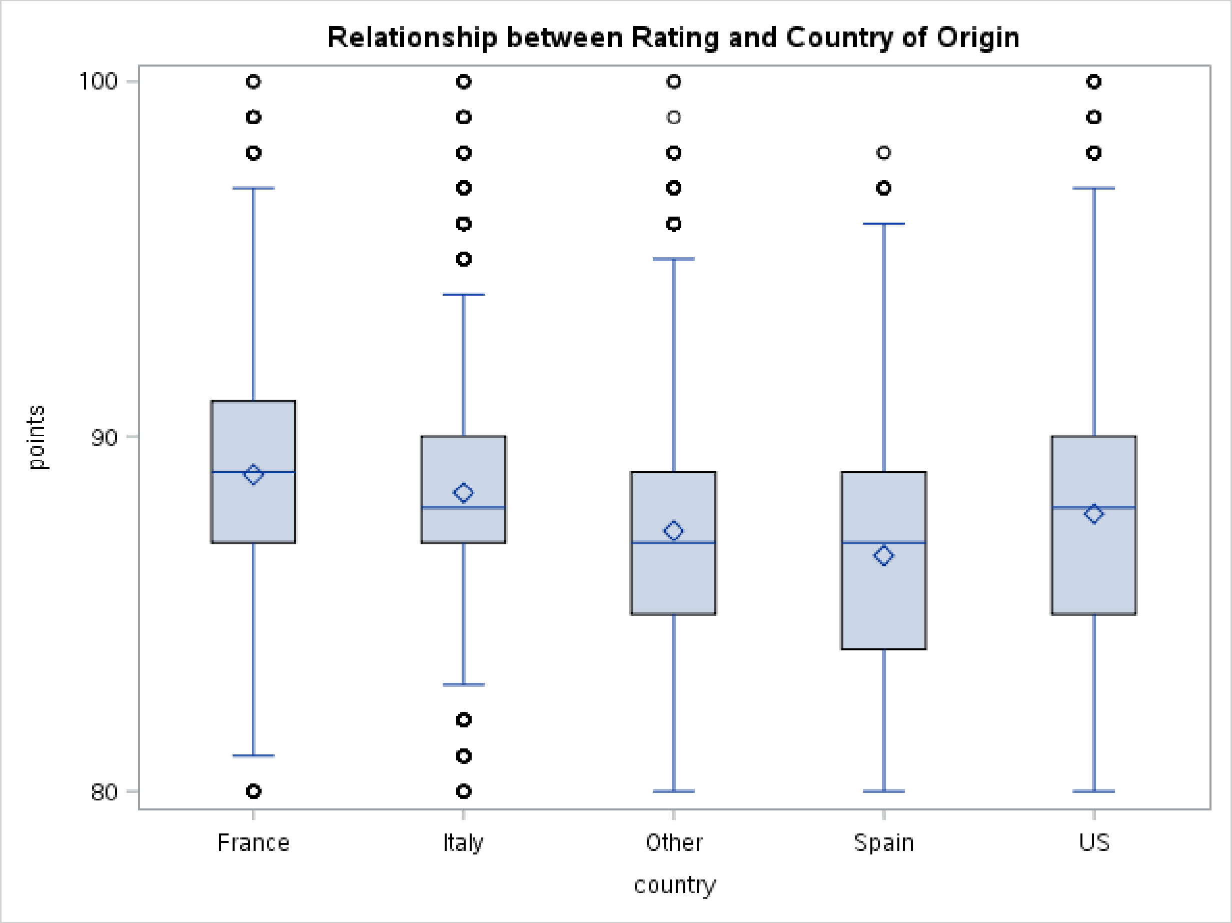
The majority of the wines reviewed are from the US(41.34%). There are similar amounts of wines reviews of wines coming from France and Italy. The “Other Countries” category consists of a of 47 countries. This group includes: Albania, Argentina, Austria, Australia, Bosnia, Brazil, Bulgaria, Canada, Chile, China, Croatia, Cyprus, Czech Republic, Egypt, England, Georgia, Germany, Greece, Hungary, India, Israel, Japan, Lebanon, Lithuania, Luxembourg, Macedonia, Moldovia, Montenegro, Morocco, New Zealand, Portugal, Romania, Serbia, Slovakia, Slovenia, South Africa, Switzerland, Tunisia, Turkey, Ukraine and Uruguay.

b)



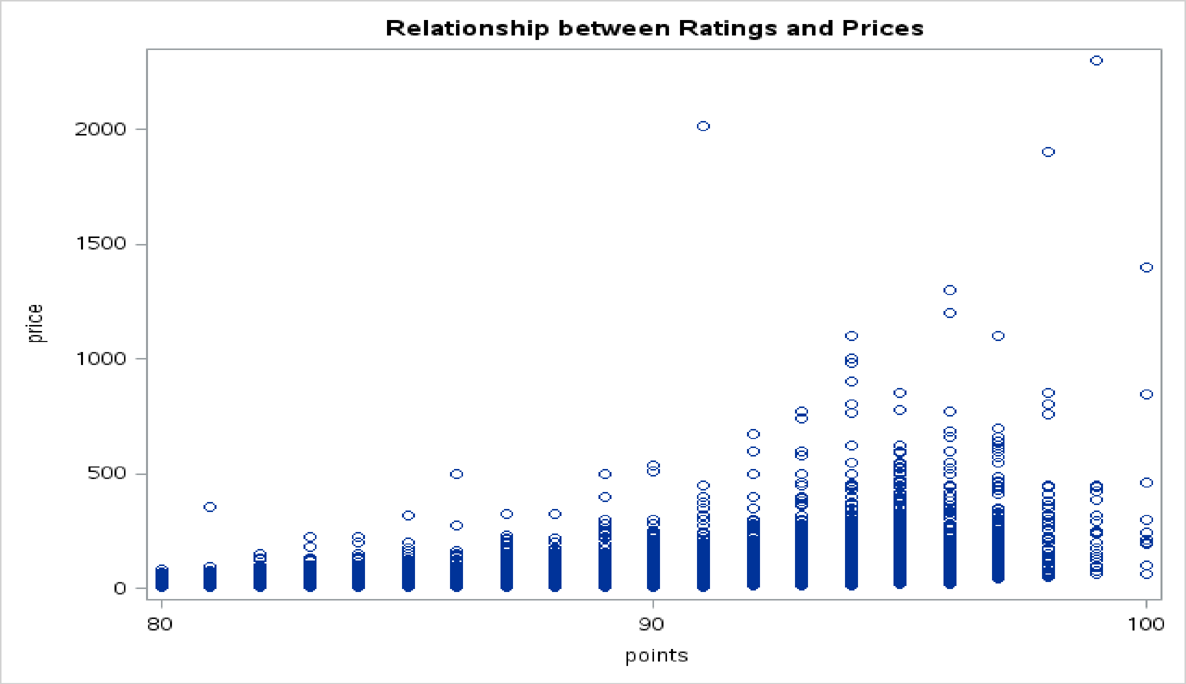
The ratings follow slightly a bell shaped right-skewed distribution. The center is at 87. Most of the wine ratings are above the median.

d)



This dataset contains outliers in every country category. Spain, U.S. and “Other Countries” have outliers above the third quartile while countries like Italy and France have outliers below and above the first and third quartile respectively. The highest median is from France, around 89 points while the “Other Countries” category and Spain have the lowest medians.

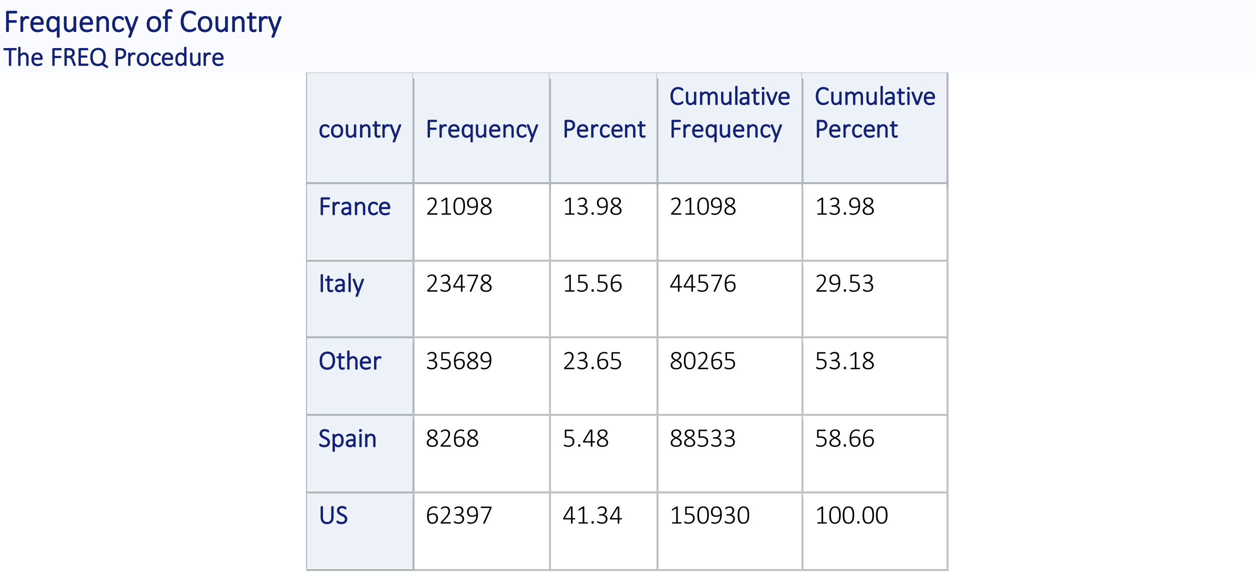
e)



The density of points displayed in the graph suggests that there is a high amount of wines priced below $500 dollars. Wines with ratings above 90 tend to be more expensive with one outlier reaching above $2,000 dollars.

Data Exploration via Numerical Summaries

a)



Frequency of Country

The FREQ Procedure

country	Frequency	Percent	Cumulative Frequency	Cumulative Percent
France	21098	13.98	21098	13.98
Italy	23478	15.56	44576	29.53
Other	35689	23.65	80265	53.18
Spain	8268	5.48	88533	58.66
US	62397	41.34	150930	100.00

This table summarizes numerically the graphical representation of the countries in the pie chart of in the data exploration part showing once again that majority of wines reviewed in this dataset come from the U.S.

b)

Descriptive Analysis of Prices of Wine

The MEANS Procedure

Analysis Variable : price				
N	Mean	Std Dev	Minimum	Maximum
137235	33.1314825	36.3225362	4.0000000	2300.00

Wines in this dataset can reach the price of $2,300.00 dollars and can be as low as $4.00 dollars. The mean of the dataset being $33.13 dollars. The standard deviation is large because the min and max are far apart due to outliers.

d)A descriptive analysis of prices of wine per variety is included in Table 1 of the appendix. This analysis shows that the mean, minimum, maximum value vary largely between wine varieties. Not all wine varieties are included in Table 1 however, it shows a representation of the sporadic changes in price.

e)

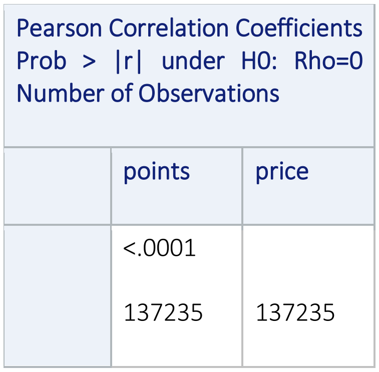
Relationship between Rating and Price

The CORR Procedure

2 Variables:	points price

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
points	150930	87.88842	3.22239	13264999	80.00000	100.00000
price	137235	33.13148	36.32254	4546799	4.00000	2300

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations		
	points	price
points	1.00000 150930	0.45986 <.0001 137235
price	0.45986	1.00000



Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	points	price
	<.0001 137235	137235

The correlation coefficient obtained is 0.45986 suggesting that there is a moderately positive correlation between price and quality points.

Conclusion

The graphical analysis suggests that the majority of the wines reviewed in this dataset are from the U.S.(41.34 percent) with the "Other Countries" category having the second highest percentage(23.64%), followed by France(13.98 percent) and Italy(15.56 percent) with similar percentages and finally Spain(5.48) with the lowest percentage. We can also see that the median for the quality of wines in this dataset ranging from 80 to 100 is 87. The results suggest that there is a moderately positive correlation between price and quality with some outliers surpassing $2,000. For the next phase of the project, I will run an ANOVA test on the quality points variable in order to observe if there exists a statistical significance in the difference between means that can determine if there is a country that produces better wine on average. From there, I would like to analyze the quality of wine based on the regions of the best country/countries and check whether there are special conditions these wines are prepared.

We, the project team members, certify that the percentage of the effort listed by each of our names below is an accurate account of the original effort contributed by each team member in the producing of this project and report:

Name (Printed)	Percent of Total Effort	Statistics Major?
Eric Fernandez	100	No

Appendix

Table1

Descriptive analysis of Prices of Wine per Variety

The MEANS Procedure

Analysis Variable : price

variety	N Obs	N	Mean	Std Dev	Minimum	Maximum
Agiorgitiko	120	117	19.2991453	10.0243367	8.0000000	65.0000000
Aglianico	317	259	33.1698842	19.0465083	6.0000000	130.0000000
Aidani	1	1	27.0000000	.	27.0000000	27.0000000
Airen	6	6	8.8333333	0.7527727	8.0000000	10.0000000
Albana	17	15	33.9333333	19.2816221	8.0000000	66.0000000
Albariño	537	530	19.9924528	7.6472279	10.0000000	110.0000000
Albarossa	1	1	40.0000000	.	40.0000000	40.0000000
Albarín	1	1	15.0000000	.	15.0000000	15.0000000
Aleatico	11	10	37.9000000	7.4304180	30.0000000	50.0000000
Alfrocheiro	18	18	24.0000000	11.9114379	11.0000000	40.0000000
Alicante	10	10	24.3000000	3.8600518	15.0000000	30.0000000
Alicante Bouschet	42	39	29.7179487	33.4474834	7.0000000	150.0000000
Aligoté	30	30	17.8333333	4.8358099	11.0000000	28.0000000
Alsace white blend	52	51	33.6470588	23.0649722	10.0000000	98.0000000
Altesse	1	1	18.0000000	.	18.0000000	18.0000000
Alvarelhão	2	2	18.0000000	0	18.0000000	18.0000000
Alvarinho	77	63	16.3492063	5.8672374	11.0000000	45.0000000

Analysis Variable : price

variety	N Obs	N	Mean	Std Dev	Minimum	Maximum
Alvarinho-Chardonn	3	2	10.0000000	1.4142136	9.0000000	11.0000000
Angevine	5	5	12.4000000	0.8944272	12.0000000	14.0000000
Ansonica	4	1	18.0000000	.	18.0000000	18.0000000
Antão Vaz	16	15	23.4666667	5.3966480	13.0000000	30.0000000
Apple	6	6	31.0000000	4.7328638	25.0000000	35.0000000
Aragonez	9	8	24.1250000	14.2070154	10.0000000	45.0000000
Aragonês	15	13	30.5384615	21.9340503	8.0000000	70.0000000
Argaman	3	3	36.6666667	1.1547005	36.0000000	38.0000000
Arinto	72	54	16.1851852	6.9555844	7.0000000	40.0000000
Arneis	64	63	19.2857143	5.4252355	14.0000000	50.0000000
Asprinio	1	0	.	.	.	.
Assyrtico	67	67	23.3432836	6.4703338	13.0000000	40.0000000
Assyrtiko	8	8	21.5000000	4.8403070	17.0000000	30.0000000
Athiri	2	2	18.0000000	0	18.0000000	18.0000000
Austrian Red Blend	67	55	37.7636364	18.9882650	15.0000000	115.0000000
Austrian white ble	47	36	28.3888889	18.7102383	15.0000000	110.0000000
Auxerrois	17	14	24.6428571	4.4133912	16.0000000	32.0000000
Avesso	3	3	14.6666667	1.5275252	13.0000000	16.0000000
Azal	1	1	13.0000000	.	13.0000000	13.0000000
Baco Noir	9	9	24.2222222	4.2946996	18.0000000	30.0000000

Analysis Variable : price

variety	N Obs	N	Mean	Std Dev	Minimum	Maximum
Baga	22	16	31.6250000	22.4703508	9.0000000	70.0000000
Baga-Touriga Nacio	1	1	20.0000000	.	20.0000000	20.0000000
Barbera	1365	967	25.9017580	14.2923363	9.0000000	163.0000000
Bastardo	7	7	30.5714286	0.9759001	30.0000000	32.0000000
Bical	13	9	15.2222222	7.7585079	9.0000000	28.0000000
Black Monukka	4	4	25.0000000	0	25.0000000	25.0000000
Black Muscat	13	13	25.9230769	9.1965713	10.0000000	38.0000000
Blatina	3	3	12.6666667	0.5773503	12.0000000	13.0000000
Blauburgunder	1	1	19.0000000	.	19.0000000	19.0000000
Blauer Portugieser	7	7	15.4285714	1.1338934	14.0000000	17.0000000
Blaufränkisch	227	191	29.0261780	16.8136644	9.0000000	129.0000000
Bobal	16	16	14.6875000	9.0753788	6.0000000	46.0000000
Bombino Bianco	1	1	30.0000000	.	30.0000000	30.0000000
Bonarda	152	152	15.0460526	5.3960236	9.0000000	38.0000000
Bordeaux-style Red	7347	4545	49.1634763	72.6755850	7.0000000	2300.00
Bordeaux-style Whi	1261	580	36.7206897	91.3422907	8.0000000	1000.00
Bovale	7	4	37.5000000	8.6602540	30.0000000	45.0000000
Boğazkere	3	3	25.0000000	6.9282032	21.0000000	33.0000000
Brachetto	25	24	18.2916667	4.0698164	11.0000000	27.0000000
Braucol	3	3	27.0000000	16.7032931	12.0000000	45.0000000

Analysis Variable : price

variety	N Obs	N	Mean	Std Dev	Minimum	Maximum
Bual	4	3	34.0000000	2.0000000	32.0000000	36.0000000
Bukettraube	2	2	18.0000000	0	18.0000000	18.0000000
Cabernet	20	18	20.2222222	9.0719708	11.0000000	45.0000000
Cabernet Blend	305	301	61.0000000	59.6369572	8.0000000	500.0000000
Cabernet Franc	1363	1310	32.8152672	20.5014169	9.0000000	180.0000000
Cabernet Franc-Cab	3	3	34.0000000	6.9282032	26.0000000	38.0000000
Cabernet Franc-Car	6	6	18.5000000	17.9080987	10.0000000	55.0000000
Cabernet Franc-Mal	1	1	22.0000000	.	22.0000000	22.0000000
Cabernet Franc-Mer	10	9	45.5555556	17.6501495	28.0000000	80.0000000
Cabernet Franc-Tem	2	2	18.0000000	0	18.0000000	18.0000000
Cabernet Merlot	52	48	23.2083333	18.1412406	8.0000000	70.0000000
Cabernet Moravia	1	1	18.0000000	.	18.0000000	18.0000000
Cabernet Pfeffer	1	1	25.0000000	.	25.0000000	25.0000000
Cabernet Sauvignon	13470	13322	41.4960967	34.9645721	4.0000000	625.0000000
Cabernet-Shiraz	1	1	150.0000000	.	150.0000000	150.0000000
Cabernet-Syrah	12	12	26.0000000	7.9772404	16.0000000	40.0000000
Cannonau	43	35	35.2285714	22.3371041	15.0000000	91.0000000
Caprettone	1	1	19.0000000	.	19.0000000	19.0000000
Carignan	74	74	40.8378378	88.5230451	14.0000000	770.0000000
Carignan-Grenache	7	7	33.7142857	16.0801564	20.0000000	65.0000000

Analysis Variable : price

variety	N Obs	N	Mean	Std Dev	Minimum	Maximum
Carignan-Syrah	1	1	80.0000000	.	80.0000000	80.0000000
Carignane	26	25	25.1600000	6.9382995	11.0000000	42.0000000
Carignano	66	58	38.9482759	21.2688904	11.0000000	91.0000000
Carineña	1	1	8.0000000	.	8.0000000	8.0000000
Cariñena-Garnacha	3	3	31.0000000	0	31.0000000	31.0000000
Carmenère	761	746	21.3270777	24.4216373	6.0000000	235.0000000
Carmenère-Caberne	22	20	16.0500000	2.3277502	13.0000000	20.0000000
Carmenère-Syrah	10	10	16.4000000	10.8852602	10.0000000	37.0000000
Carnelian	1	1	14.0000000	.	14.0000000	14.0000000
Carricante	23	22	44.5454545	37.3627358	21.0000000	195.0000000
Casavecchia	6	6	42.3333333	13.4709564	25.0000000	55.0000000
Castelão	37	37	10.8918919	2.5252479	7.0000000	17.0000000
Catalanesca	1	1	19.0000000	.	19.0000000	19.0000000
Catarratto	31	27	18.2962963	5.0825101	12.0000000	30.0000000
Cayuga	3	3	20.3333333	2.3094011	19.0000000	23.0000000
Cerceal	3	3	43.3333333	11.5470054	30.0000000	50.0000000
Cesanese d'Affile	18	9	22.0000000	7.5828754	16.0000000	35.0000000
Chambourcin	16	16	19.0000000	5.6920998	10.0000000	26.0000000
Champagne Blend	1238	1003	78.6271186	74.9159778	7.0000000	505.0000000
Charbono	40	40	31.3500000	6.1458762	16.0000000	40.0000000

Analysis Variable : price

variety	N Obs	N	Mean	Std Dev	Minimum	Maximum
Chardonel	1	1	11.0000000	.	11.0000000	11.0000000
Chardonelle	1	1	30.0000000	.	30.0000000	30.0000000
Chardonnay	14482	13775	32.2471869	45.1487992	4.0000000	2013.00
Chardonnay Weissbu	3	3	25.0000000	0	25.0000000	25.0000000

SAS Code

```
/* Eric Fernandez Project-Phase B*/
/* I certify that the SAS code given is my original and exclusive work*/

/* To read the file:
 Create a  new folder.
 Upload 'winemag-data_first150k.csv' to the folder
 Right click on 'winemag-data_first150k.csv' and select Properties
 Copy the path name and paste to the filename statement below
 Add a slash and the file name to the end of the path
*/
FILENAME CSV "~/datasets/winemag-data_first150k.csv" TERMSTR=LF;


/** Import the CSV file.  **/
PROC IMPORT DATAFILE=CSV
                    OUT=WineReviews
                    DBMS=CSV
                    REPLACE;
RUN;

/* Section 2 */

/* 2(a) */
/* Single Categorical Variable: Country of Origin */
Proc gchart data=WineReviews;  /* general bar charting proc */
    pie country/type=percent; /* pie chart */
   title 'Countries Percentage' ;
Run;

/* 2(b) */
/* Single Quantitative Variable: Rating Points */
Proc sgplot data=WineReviews;
        histogram points;
        title 'Ratings';
Run;
title ;

/* 2(d) */
/* Created a new dataset with other countries that are not France, US,
  Spain or Italy merged into one group of countries called Other */
Data WineReviewsB;
        Set WineReviews;
        /* If countries are not Spain, US, Italy or France then change to Other*/
        if Country not in ('Spain' 'US' 'Italy' 'France') then country = 'Other';
Run;

/* Relationship between Quantitative and Categorial Response:
  Quantitative: Rating Points Categorical: Country of Origin*/
Proc sgplot data=WineReviewsB;
        vbox points / /* This is the quantitative variable for the y-axis */
        category = country; /* This is the categorical variable */
        title 'Relationship between Rating and Country of Origin';
Run;
title ;
```

```
/* 2(e) */
/* Relationship between Quantitative Variables:
  Quantitative Variables: x=Points y= Price*/
Proc sgplot data=WineReviews;
        scatter x=Points  y=Price; /* Quantitative Variables */
        title 'Relationship between Ratings and Prices';
Run;
title ;
/* Section 3 */

/* 3(a) */
/* Single Categorical Variable: Variety of Wine */
/* Counting varieties using proc freq */
Proc freq data=WineReviewsB;
        tables country; /* count the number of each type of variety */
        title 'Frequency of Country';
Run;
title ;

/* 3(b) */
/* Single Quantitative Variable: Price of Wines */
Proc means data=WineReviews;
        var Price;
        title 'Descriptive analysis of Prices of Wine';
Run;
title ;

/* 3(d) */
/* Relationship between Price and Variety  */
Proc means data=WineReviews;
   var price;
        class Variety;
        title 'Descriptive analysis of Prices of Wine per Variety';
Run;
title ;

/* 3(e) */
/* Relationship between points and price */
Proc corr data=WineReviews;
        var points price;
        title 'Relationship between Rating and Price';
Run;
title ;
```