NORTHWESTERN UNIVERSITY
MSDS 430: PYTHON FOR DATA SCIENCE

FINAL PROJECT: ANALYSIS ON WINE REVIEWS

By: Chris Festa

Due: June 9th, 2019

INTRODUCTION

Dataset	Wine Reviews 130k wine reviews with variety, location, winery, price, and description
URL	https://www.kaggle.com/zynicide/wine-reviews
GitHub	https://github.com/chrisfesta/wine_reviews_analysis

When I socialize with people I often love to drink wine, however, I am far from an expert. When I buy wine I typically assume that the higher the price of wine often translates to a better taste. Sometimes this is true but it often is false. I could never understand why wine taste good or whether or not I am paying a fair price.

A funny joke I sometimes tell is I went on a date a few years ago and my manager at the time gave me a story to tell about a bottle of wine on the menu. I explained to my date that it is grown from a vila that is next to a volcano in Italy which translates to an excellent grape. She was so impressed that I knew so much about the wine when I only knew what I was told to say, had we had to choose another bottle of wine I would have been in a world of trouble.

Luckily I was able to find a dataset on Kaggle that interprets wine reviews. It comes with a number of attributes to do calculations such as: country, description, designation, points, price, province, province, taster name, title of review, type of grape, and winery.

- There are a number of correlations I can do with this data such as:
- Does the price often correlate to a good tasting wine?
- Does the taster impact the rating of the wine and that makes the rating impartial?
- Taking a deeper look at California. Resarch shows that Chardonnay is the most popular grape planted in California, with close to 95,000 planted acres. Cabernet Sauvignon is a distant second with almost 80,000 acres planted. Merlot and white Zinfandel are also popular
  Reference: https://www.thewinecellarinsider.com/california-wine

My hypothesis will be:

- The vineyard does not play a major role. The country or province is more important for each type of grape
- Price does not correlate to good tasting wine and the ratings will not increase as price increases
- The taster will play an important role and influence the review points of wine
- The research about California wine will be correct and the best wine from California will be Chardonnay, Cabernet Sauvignon, Merlot, and Zifandel

APPROACH

To perform the analysis we will apply several formulas to the price and points of each wine to provide correlations on the country, province, variety, winery, and taster. We do not want data that is outside the norm to misrepresent the result, therefore we will calculate the 5th and 95th percentile to remove wine that is below the 5th and greater than the 95th percentile.

Hypothesis #1: Does the vineyard that make the wine play an important role or does the province or country more important?

The data set provides review points for type of grape (variety) which we can do correlations on average (mean) of each variety broken down by vineyard, province, and country to determine the best wine.

We will look at the 10 most reviewed varieties (type of grape) and create a chart that displays the mean rating of the variety for winery, country, and province to see where the highest rated wine is from.

Hypothesis 2: Does the price often correlate to a good tasting wine?

The data set provides review points and price of wine that we can break down into price ranges to get the mean points for each price range.

We will look at the 20 most reviewed varieties (type of grape), countries, and provinces to determine if the quality of the wine increases or decreases as price increase. Our price range will start at $0 and end at $200 with increments of $10 (i.e. $0 to $10, …, $190 to $200)

Hypothesis 3: Does the taster impact the rating of the wine and that makes the rating impartial?

The dataset provides the taster who reviewed the wine and each wine reviewed contains the country, province, and variety. We will look at the top 10 tasters who reviewed the most wine and get the mean (average) for each taster for the top countries, provinces, and varieties with the most wines reviewed. The result should show if the tasters for each attribute is similar or disparate.

Hypothesis 4: The research about California wine will be correct and the best wine from California will be Chardonnay, Cabernet Sauvignon, Merlot, and Zifandel

The dataset provides the wine reviews for the provinces and we can look at the California province to compare the wine varieties (grapes) to determine which grape is the best. We will get the mean of each grape in California, order the result from best to worst, and determine which grapes are above the mean of all wine in California.

If the hypothesis is true we will prove the Chardonnay, Cabernet Sauvignon, Merlot, and Zifandel are either the highest rated wine by order or above the overall mean across all grapes.

To prevent cheap or expensive wine from distorting the results we will assume the wine must cost between $25 to $75.

CODE APPROACH

A class will be created that we can apply to all the hypothesis' to perform several calculations on price and points. The calculations will compute the following on price and points:

- 95th and 5th percentile
- mean, median, variance, and standard deviation of the dataset that removes values less than the 5th and greater than the 95th percentile
- the average points of wine for a price range with an increment of 10 (0 to 10, 10 to 20, ...)

Variable	Type	Description
attribute	String	Value that is under review (i.e. Pinot Noir)
dataset_size	int	Size of the dataset under review
price_5	float	The 5th percentile of price
price_95	float	The 95th percentile of price
price_mean	float	The mean of price, less percentiles
price_median	float	The median of price, less percentiles
price_variance	float	The variane of price, less percentiles
price_standard_deviation	float	The standard deviation of price, less percentiles
points_5	float	The 5th percentile of points
points_95	float	The 95th percentile of points
points_mean	float	The mean of points, less percentiles
points_median	float	The median of points, less percentiles
points_variance	float	The variance of points, less percentiles
points_standard_deviation	float	The standard deviation of points, less percentiles
price_range_calculations	dictionary	The mean and count of points for each price range between $0 and $200 with increments of $10

MODULES USED IN ANALYSIS

- **Pandas** - library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language. We will use pandas to load in the data into a pandas dataframe and clean the data to normalize the data to be consumable for calculations
- **Numpy** - library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays. We will use Numpy to perform several calculations such as mean, median, variance, standard deviation, and percentiles to assist with correlating the various attributes (country, province, variety, winery, and taster) using the review points and price information for each bottle of wine reviewed to assist with proving or disproving the hypothesis
- **Matplotlib** - a plotting library for the Python programming language that can create graphical representations of data. We will use Matplotlib to extend on what Numpy calculates to visual represent the data to help prove or disprove the hypothesis
- **Random** - a library in Python that is used to generate random numbers. We will use random to select attributes at random in addition to looking at the top 10
- **Operator** - a module in Python that exports a set of efficient functions corresponding to the intrinsic operators of Python. We will use the operator module to sort the dictionaries used in the project
- **Collections** - Collections in Python are containers that are used to store collections of data, for example, list, dict, set, tuple etc. The module improve the functionalities of the built-in collection containers. We will use the collection module to get the count of instances of a value
- **Datetime** - The datetime module supplies classes for manipulating dates and times in both simple and complex ways. We will use the datetime module to capture the current time so we can calculate how long calculations take to compute.

DATA DEFINITIONS

Data Label	Data Type	Data Description
country	String	The country the wine is from
province	String	The province or state that the wine is from
variety	String	The type of grapes used to make the wine (i.e. Pinot Noir)
winery	String	The winery that made the wine
price	Float	The cost for a bottle of wine
points	Int	The number of points WineEnthusiast rated the wine on a scale of 1-100
taster_name	String	The person who rated the wine

Using a Pandas dataframe we load the ***winemag-data-130k-v2.json*** file into a Pandas dataframe. Since the data contains several attributes that are null (NA) we use the pandas method that drops null (NA) values (dropna) so we are only looking at rows that have valid values for each wine that is reviewed.

A good row will be a wine that has the country, province, variety, winery, price, points, and taster with a valid value.

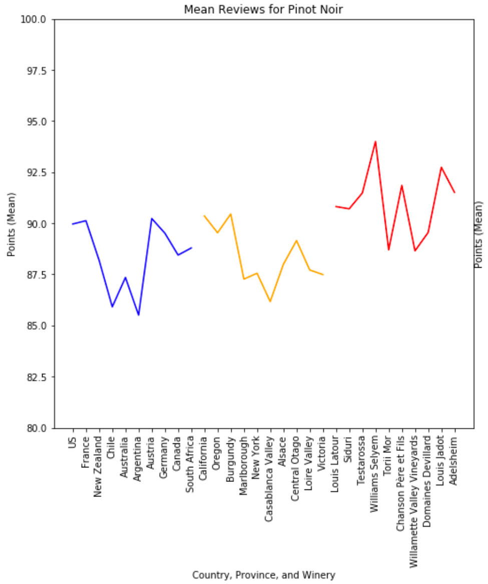
HYPOTHESIS #1

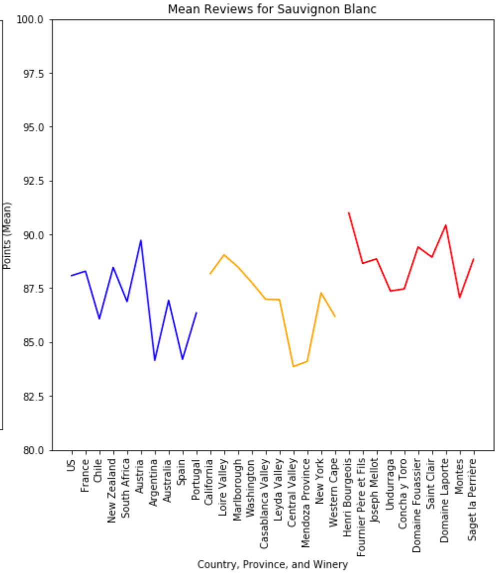
Does the vineyard that make the wine play an important role or does the province or country more important?

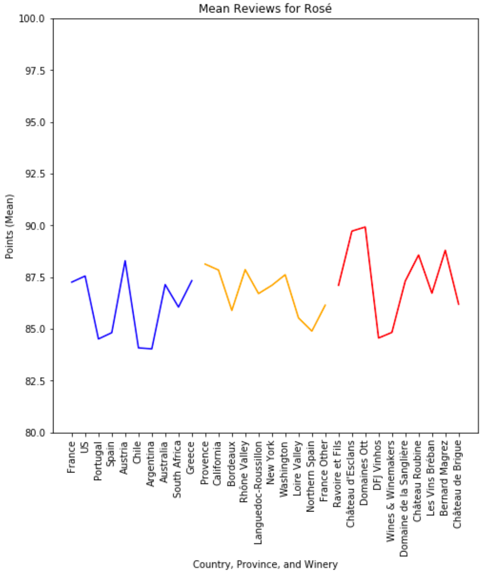
- Countries - represented as a blue line
- Provinces - represented as a orange line
- Wineries - represented as a red line

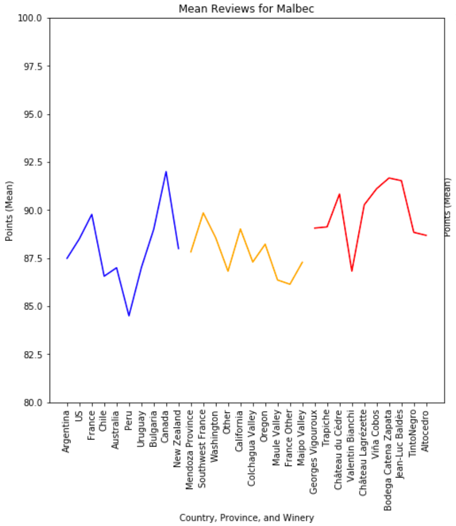
The graphs show that the winery consistently has the highest mean wine reviews (points) compared with the country and province. This would indicate that the vineyard (winery) is the most important aspect to find the best wine for a type of grape (variety).

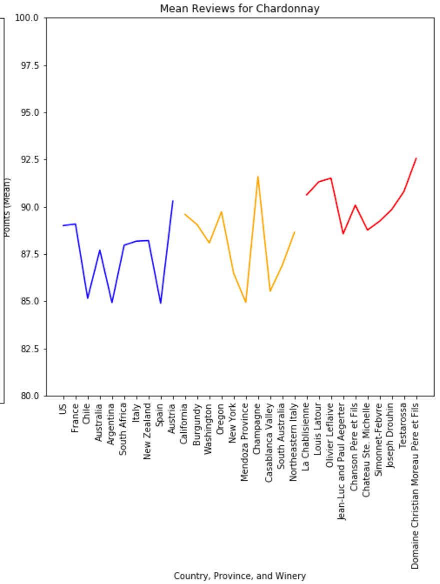
This hypothesis was proven wrong by the analysis, the vineyard plays the most important role for purchasing quality wine for a given type of grape (variety)

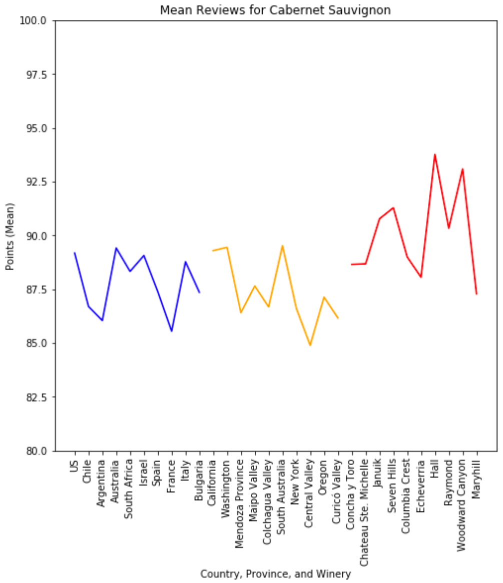

Mean Reviews for Pinot Noir


Mean Reviews for Sauvignon Blanc


Mean Reviews for Rosé


Mean Reviews for Malbec


Mean Reviews for Chardonnay


Mean Reviews for Cabernet Sauvignon

HYPOTHESIS #2

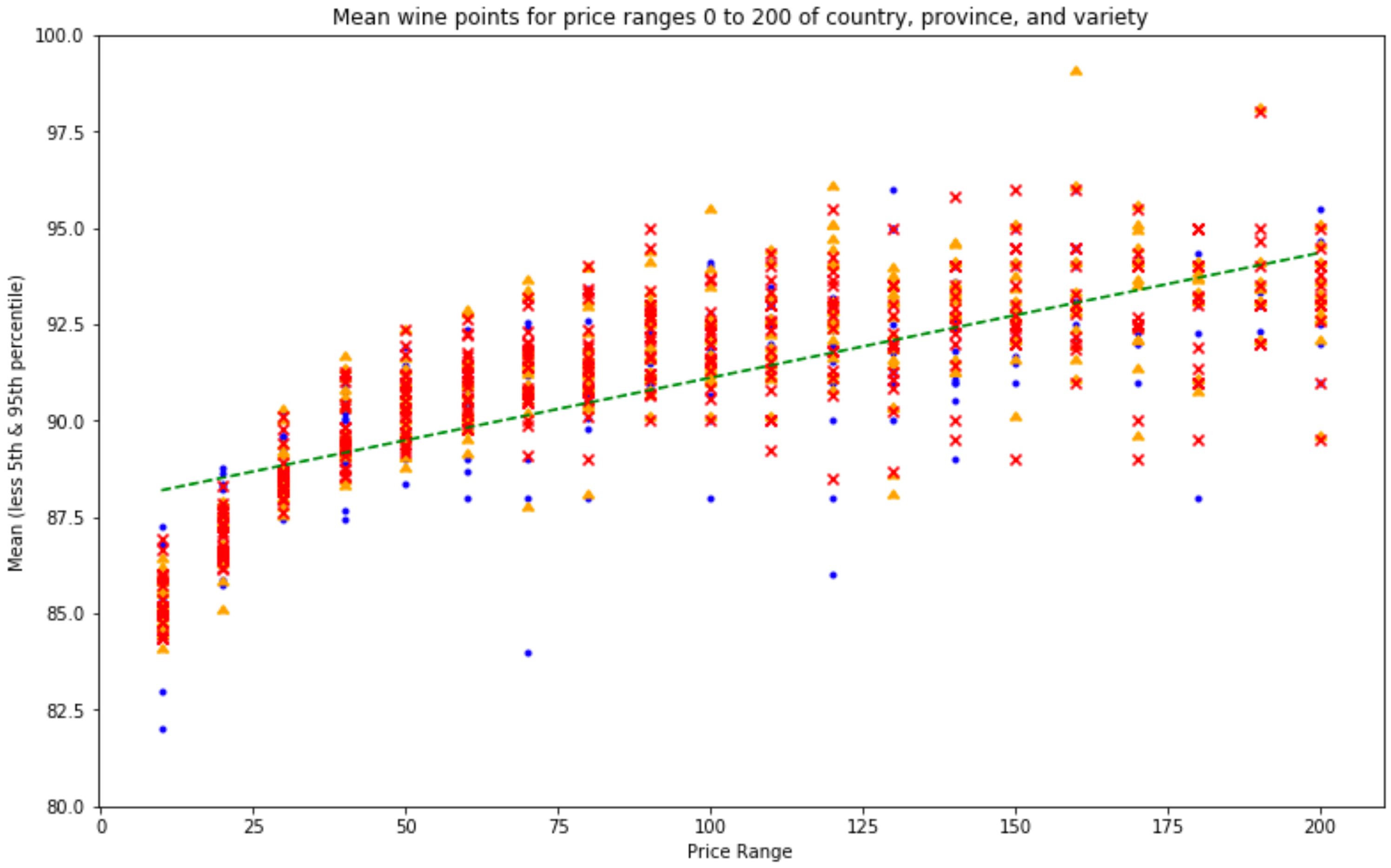
Does the price often correlate to a good tasting wine?

- Countries - represented as blue circles
- Provinces - represented as orange triangles
- Varieties - represented as red 'x'

The graph shows that for countries, provinces, and varieties the average points increases as the price goes up.

The trendline for the scatter plot also represents the quality of the wine increases as price goes up.

This hypothesis was proven wrong by the analysis, as price goes up the quality of wine increases based on the average points for the wine within that range increasing for all attributes (country, province, and variety)



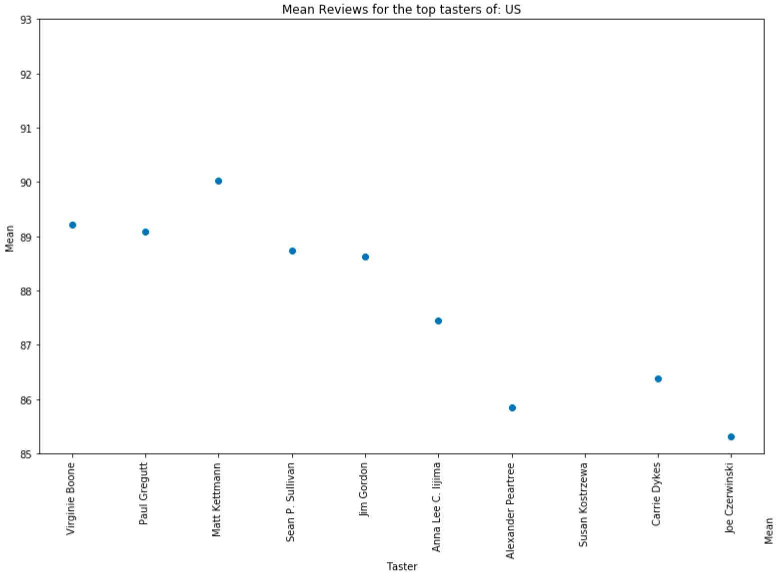
HYPOTHESIS #3

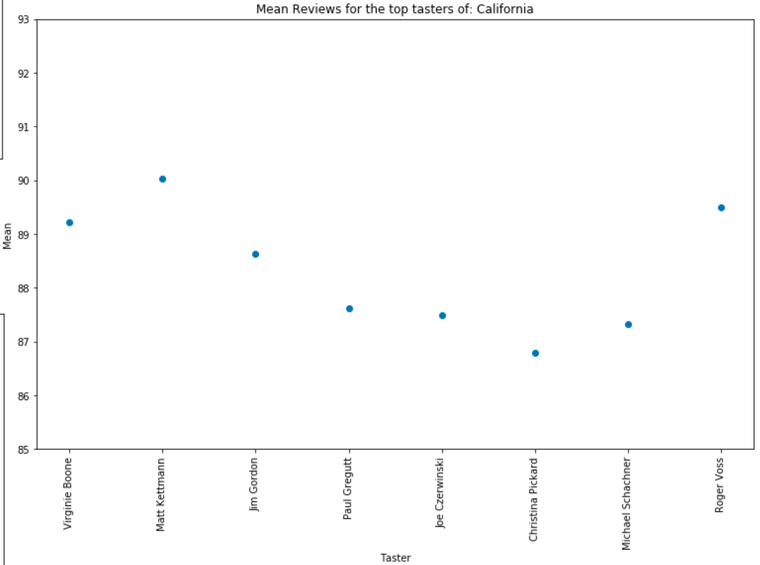
Does the taster impact the rating of the wine and that makes the rating impartial?

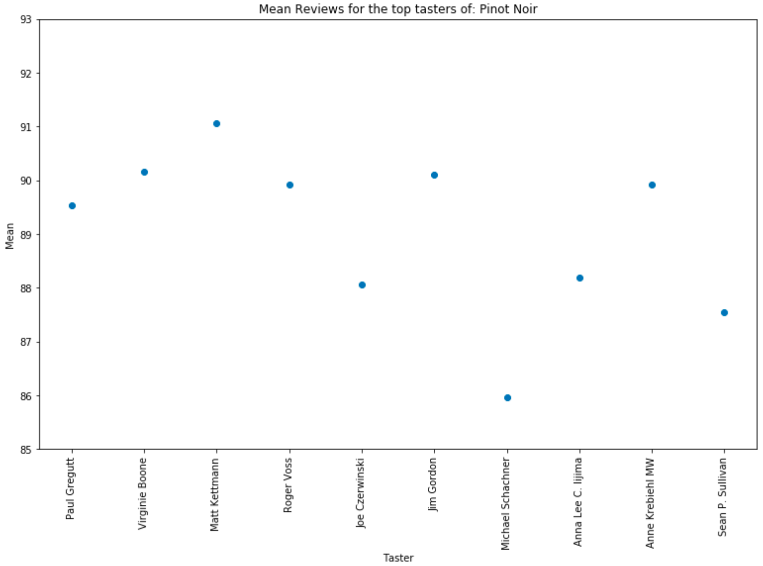
The scatter plot for the variety of 'Pinot Noir', country of 'US', and province of 'California' shows up to the 10 most tasters who reviewed wine and the mean of the points they scored.

The graph shows that the taster is disparate and has a wide range.

This hypothesis was proven correct by the analysis, the taster mean review points is not linear and has a wide range for each taster


Mean Reviews for the top tasters of: US


Mean Reviews for the top tasters of: California


Mean Reviews for the top tasters of: Pinot Noir

HYPOTHESIS #4

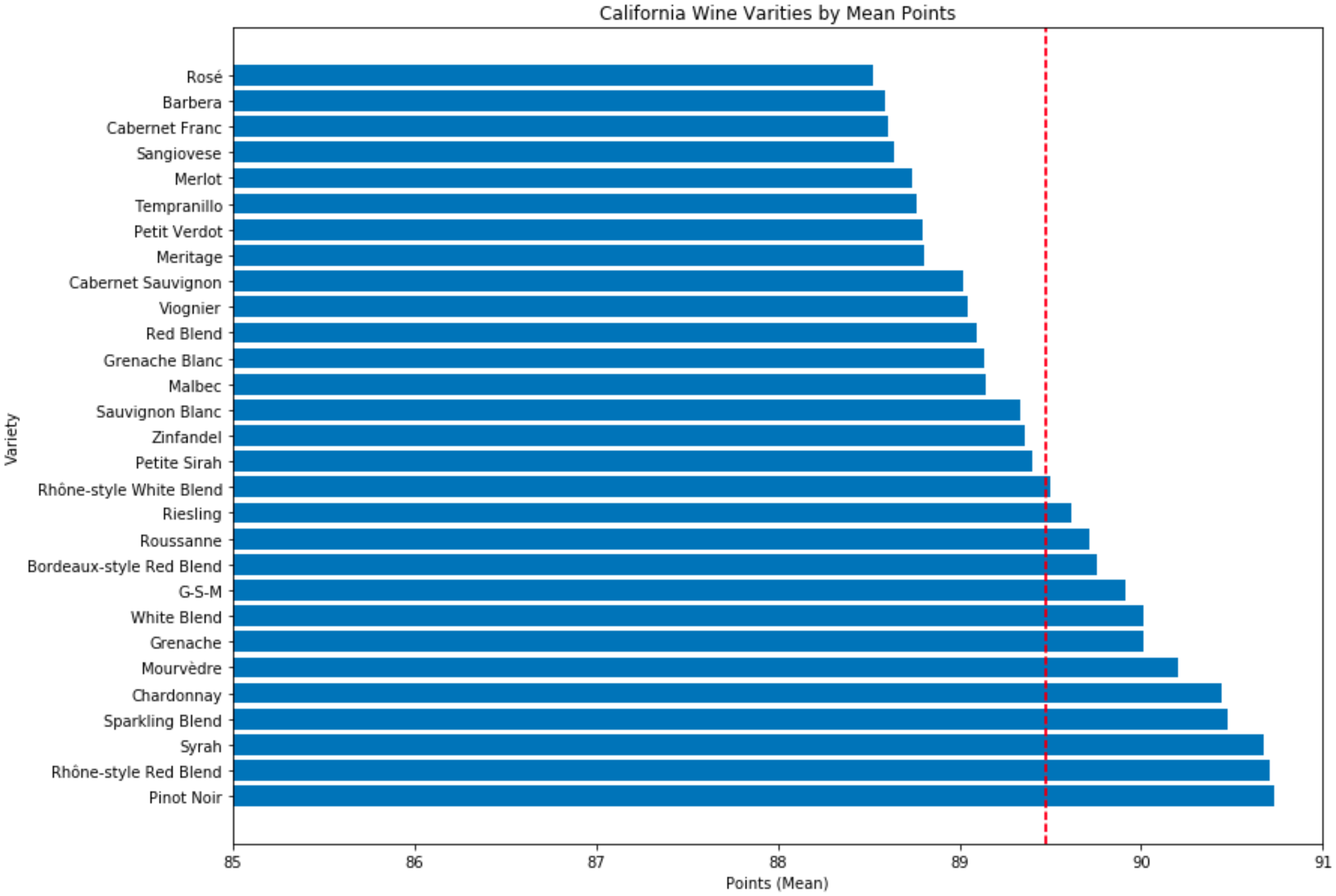
The research about California wine will be correct and the best wine from California will be Chardonnay, Cabernet Sauvignon, Merlot, and Zinfandel

Based on online research we assumed that Chardonnay, Cabernet Sauvignon, Merlot, and Zinfandel are the best wine from California. We set out to prove that by comparing the type of grape and determining if it is one of the highest rated wine or above the overall average where the price is between 25 and 75 dollars.

The results show:

- Chardonnay - above the overall average
- Cabernet Sauvignon - bellow the overall average
- Merlot - bellow the overall average
- Zinfandel - bellow the overall average

Based on the analysis the market research was wrong. While these grapes may be popular they are not the best wine you can purchase from California. The best wine that is above the overall average is: Pinot Noir, Rhône-style Red Blend, Syrah, Sparkling Blend, Chardonnay, Mourvèdre, Grenache, White Blend, G-S-M, Bordeaux-style Red Blend, Roussanne, Riesling, Rhône-style White Blend


California Wine Varities by Mean Points
Variety
Rosé
Barbera
Cabernet Franc
Sangiovese
Merlot
Tempranillo
Petit Verdot
Meritage
Cabernet Sauvignon
Viognier
Red Blend
Grenache Blanc
Malbec
Sauvignon Blanc
Zinfandel
Petite Sirah
Rhône-style White Blend
Riesling
Roussanne
Bordeaux-style Red Blend
G-S-M
White Blend
Grenache
Mourvèdre
Chardonnay
Sparkling Blend
Syrah
Rhône-style Red Blend
Pinot Noir
Points (Mean)