**Final Project**

**Output Summary Report**

**Name**: Rupali S. Metkari

**Date**: 7th April 2021

Output

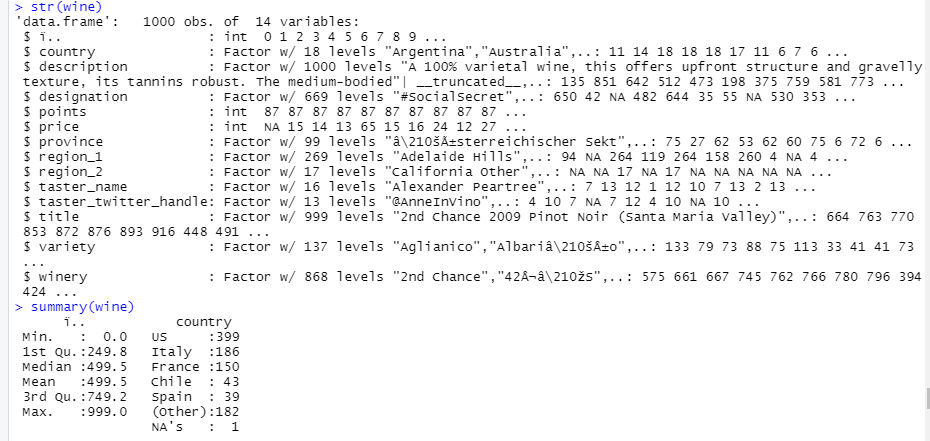


Fig (1)

First, we load Wine\_tasting.csv data. This data contains total 1000 obs. and 14 variables. It contains information on the type of wine, country, region, and winery it is from, the cost, its score, and a little description about the wine.

The next thing I’ll do is a simple summary of the data to see if anything pops out that I want to further analyze.

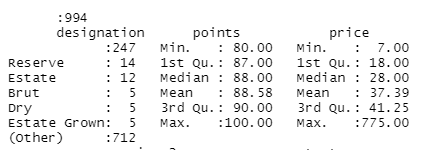


Fig (2)

From the points we can see that both the mean and median are very close (87.89 vs 88.00) meaning there are not many outliers in one direction or the other. However, in price the mean is quite larger than the median (37.39 vs 28.00) and just by seeing the max price (775) we know there will be some outliers of highly priced wines.

**Hypothesis Testing**

Hypothesis testing is used to assess the plausibility of a hypothesis by using sample data. Such data may come from a larger population, or from a data-generating process.

Using points and price variable we did hypothesis tasting. According to given instruction we would to perform two types of t.test . one sample t.test and two sample t.test.

So, we considered question,

1. Which countries wines are costly old or new countries?
2. Which countries wines are tasty based on points?
3. Which country produced the most wine?
4. Which Province was the most reviewed?

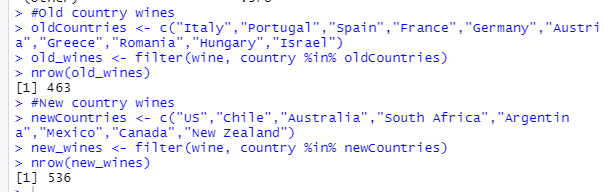


Fig (3)

In Fig (3), we can see create two subsets old countries and new countries from the original dataset. With using three variables country, price, points.

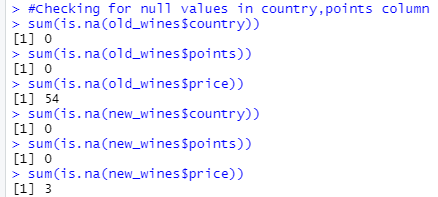


Fig (4)

After that check null values in both subsets. In Fig (3) shows that in price variable null values is exist so next step is these null values replace by median.

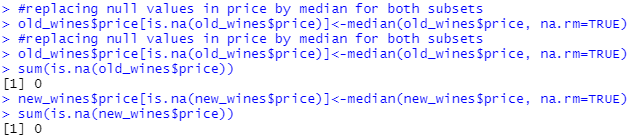


Fig (5)

In Fig (5) shows replace null values by median for both countries. Now all data is clean.

For hypothesis testing and inferential statistics take two questions are:

1. According to points old countries wines are tasty.
2. According to price old countries wines are costly.

**Two sample t.test- based on points:**

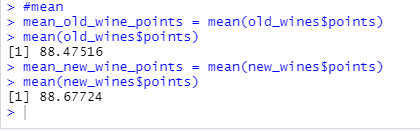


Fig (6)

In Fig (6) we can see that mean of two countries are different. We take significance value is α= 0.05.

Now assume two scenarios:

* H0: Old countries wines are tasty.
* H1: New countries wines are tasty.

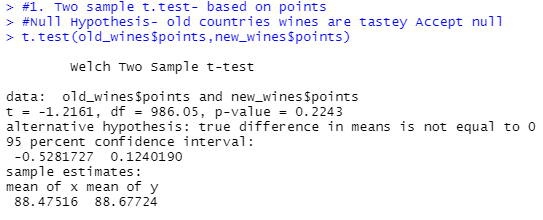


Fig (7)

T- statistics: t = -1.2161, df = 986.05, p-value = 0.2243

for significance value α= 0.05 We accept Null hypothesis as p-value > 0.05 because p value is 0.2243 is greater than significance value. So, it is concluded that “Old countries wines are tasty.”

**Conclusion:** p-value is greater than the significance value, so we reject alternative hypothesis and accept null hypothesis.

**One sample t.test- based on price:**

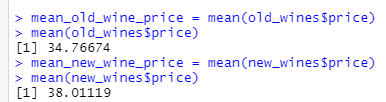


Fig (8)

Old countries mean is 34. 76 and new countries mean is 38.011. Here find two mean. For t.test compare new countries mean to old countries.

Scenario is: Old countries wines are more expensive.

* H0 is Null hypothesis: Old countries wines are more expensive. (μ old = μ new )
* H1 is Alternative hypothesis: New countries wines are more expensive. (μ old ≠ μ new )

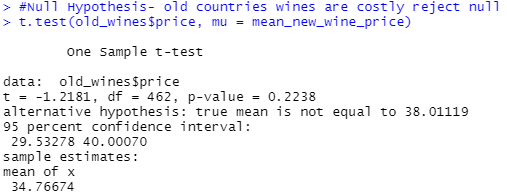


Fig (9)

Here we can see that true mean μ old = 34.76 is not equal to μ new = 38.011, so we reject null hypothesis. And Alternative hypothesis H1 is accepted. Menace old countries wines are not expensive. New countries wines are expensive.

According to me the climates of New country wine regions are often warmer, which tends to result in riper, more alcoholic, full-bodied, and fruit-centered wines. These wines are often made in a more highly extracted and oak-influenced style.

**Linear Regression**

According to Fig (1), many wines are made in US, Italy, France, with 263 of the 399 of US wines coming from California.

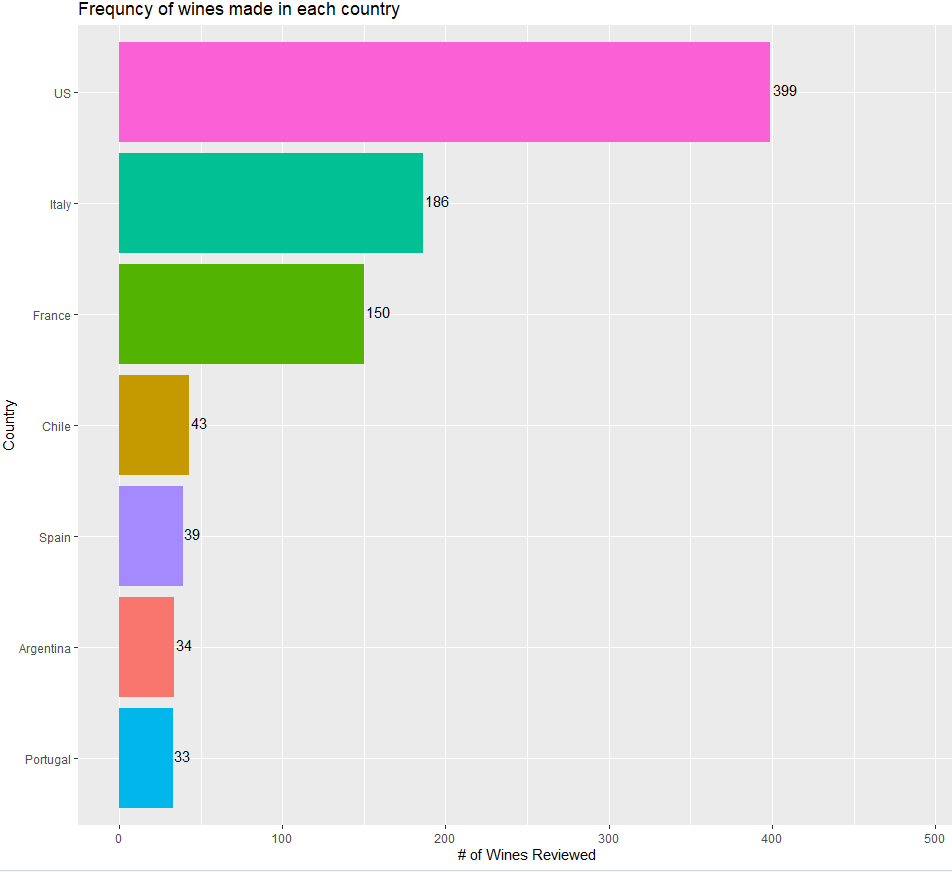


Fig (10)

Fig (10), wines organized the data by country and put together a bar graph showing the number of wines made in each country.

Countries compare in terms of points for each wine. And find is there a difference in quality of wine between each country. Then compare California wines to Italian and French wines. I will visualize this by using a box plot that also includes the mean.

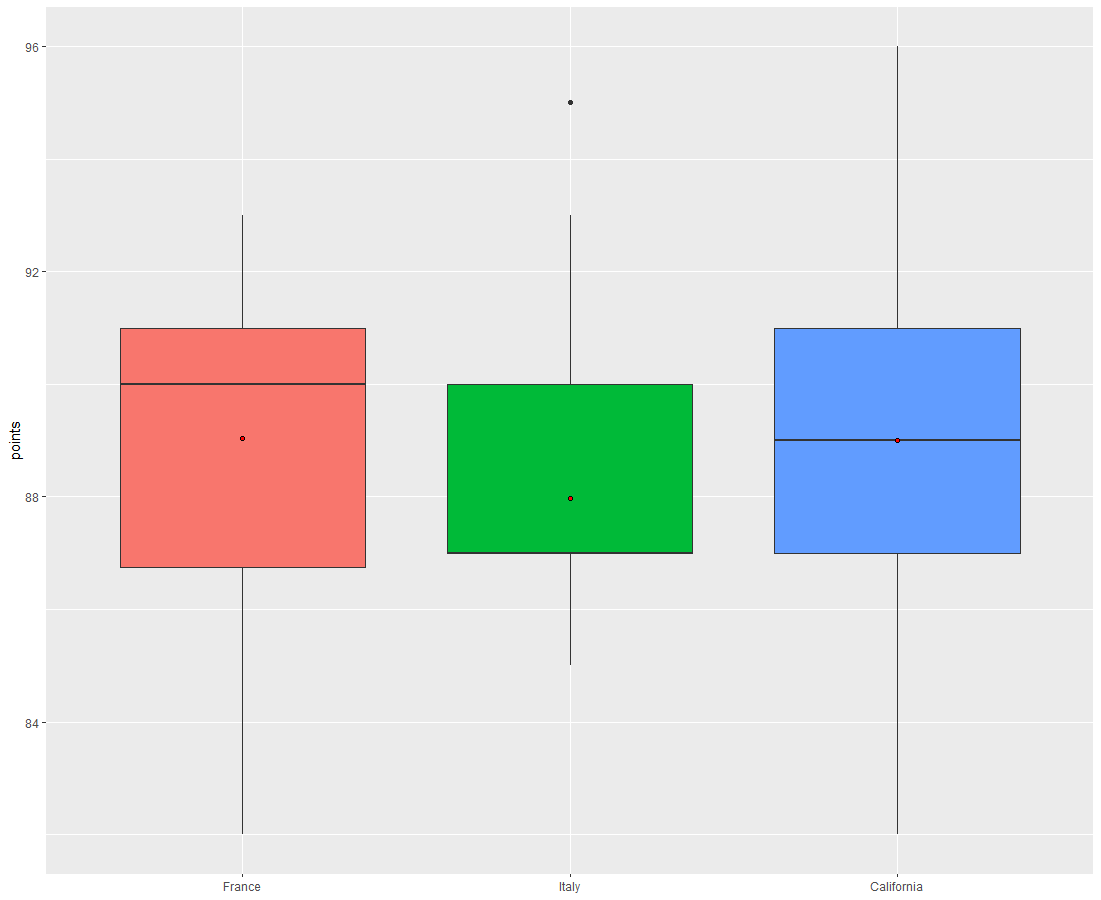


Fig (11)

California has almost the exact same median and mean, both of which are higher than Italy and France, meaning California appears consistently better.

Plot histogram to ensure that each of these groups of data are normally distributed. To check statistically significant difference between the points for the wines in each region.

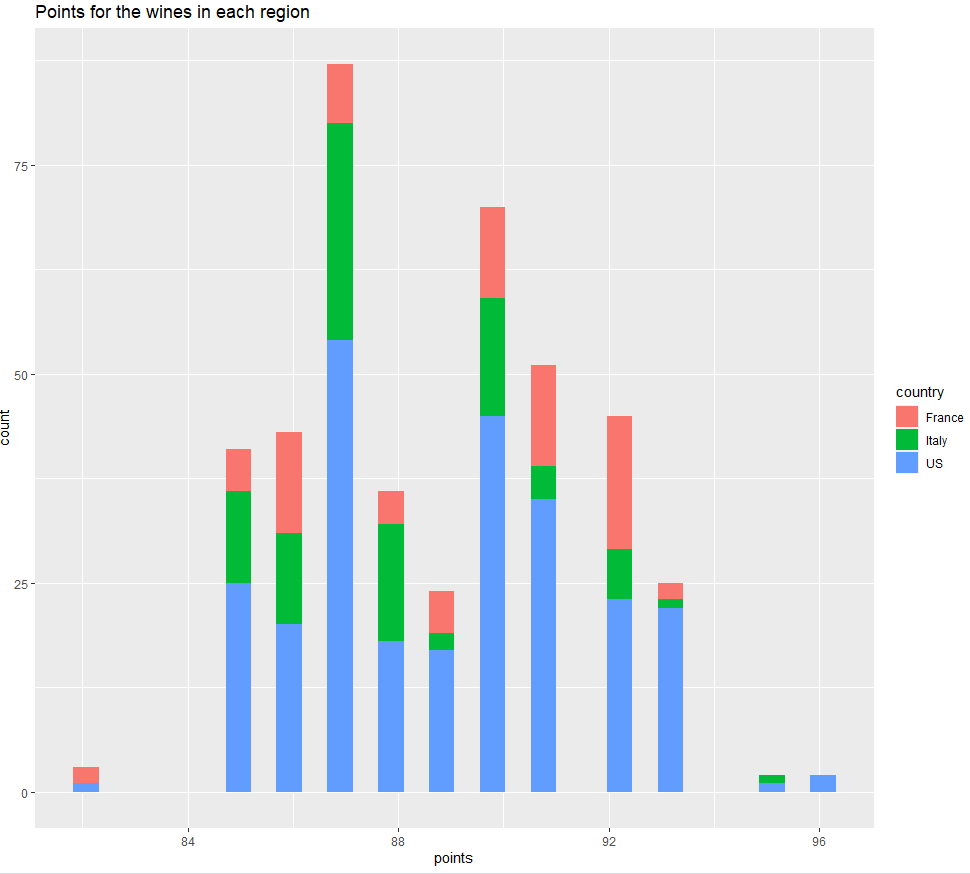


Fig (12)

All three groups normally distributed.

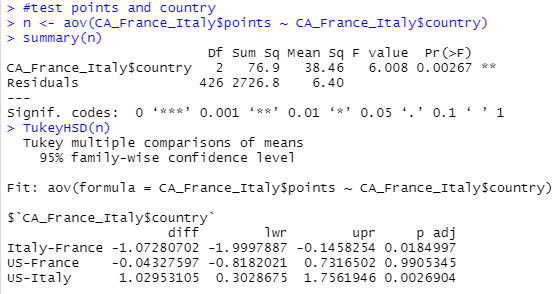


Fig (13)

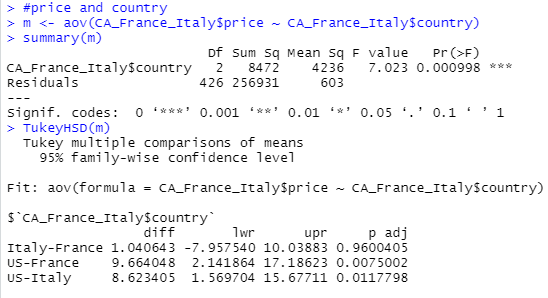


Fig (14)

We can see that there is a statistically significant difference between these three groups (P value is 0.002, which is well below .05), then did Tukey’s test to tell which groups are different from the other. We can see that all three groups are statistically significant different from each athermancy US has the higher rated wine.

From Fig (14) we can conclude that US wine price is higher than the other countries.

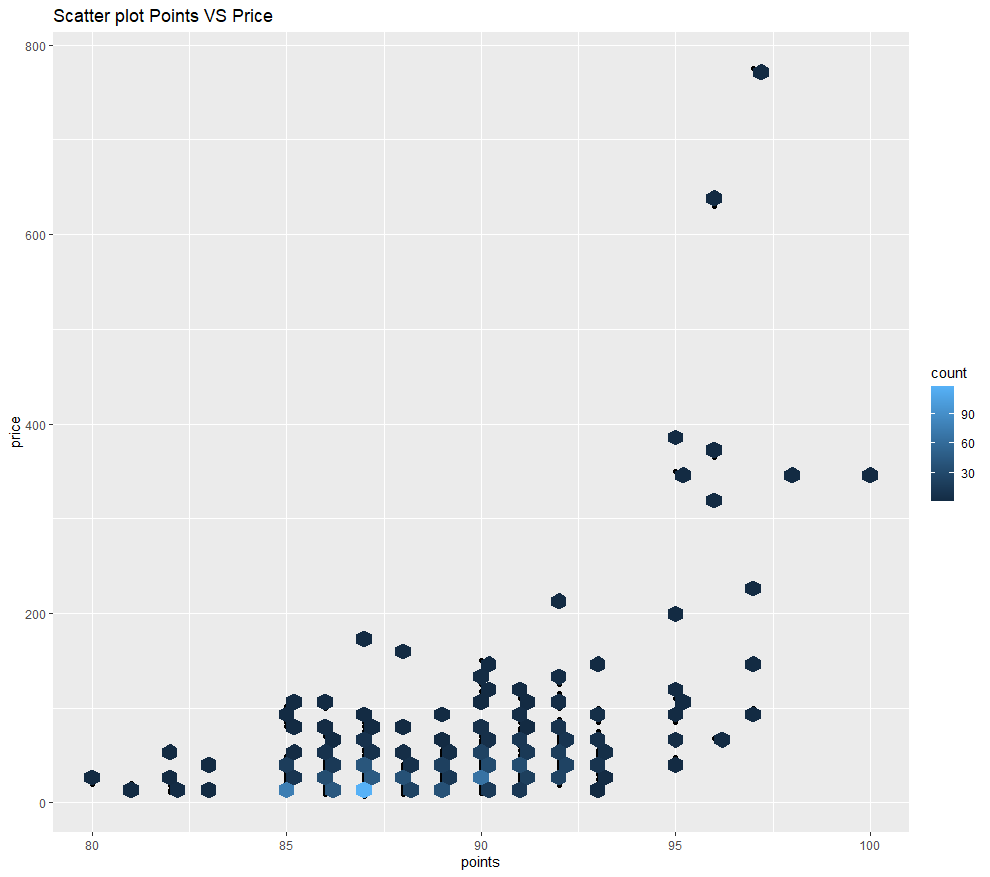


Fig (13)

it does show that there is a bit of a concentration of lower priced wines in the 85-92.5-point range.

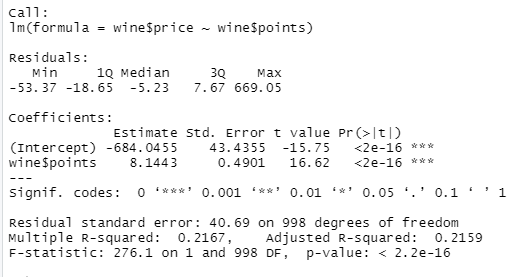


Fig (14)

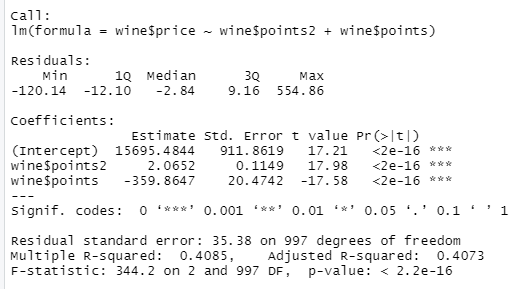


Fig (15)

In Fig (14) linear regression equation would be **y=8.1443x-684.0455**. When wine points below 81, the price would be negative. R squared value at 21.67%, menace variability in price has to do with points. So many factors involved in determining how to price a wine. This could include factors such as the age of the wine, region/country, or the popularity of a winery (a well-known winery may be able to up their prices).

In Fig (15) Once again, we see that the p-value is significant (<2e-16), our R squared value is slightly better (40.55%), but still not optimal. Clearly, we need more information on these wines to create a better equation for predicting the price of a wine.

**Conclusion:**

While the linear regression analysis was a bit of a disappointment, it is by no means the end. Since the points of a wine does not completely translate into the price of a wine, that just means more data is needed. Since Wine Enthusiast focuses on the better wines, it would be smart to explore other data sets of more common wines and data sets that provide more quantitative variables for multiple linear regression (like age of wine).

So, we were able to conclude couple of things. From the hypothesis testing and liner regression. New countries wines are costly. Old countries wines are tasty based on points. US has the highest country to produce wine. US is belonging new countries wine. California was the most reviewed Province.

References:

1. Allan G. Bluman Published by McGraw-Hill Education, 2 Penn Plaza, New York, NY 10121. Copyright © 2018 by McGraw-Hill Education. *Elementary Statistics A Step-by-Step Approach, 10th edition*.
2. T Test in R: One Sample and Paired (with Example), May 2018, Guru99: <https://www.guru99.com/r-t-test-one-sample.html>
3. The Two-sample t-test, July 2019, jmp: <https://www.jmp.com/en_us/statistics-knowledge-portal/t-test/two-sample-t-test.html>
4. Correlation plot in R, Feb 2018, R Coder: [*https://r-coder.com/correlation-plot-r/*](https://r-coder.com/correlation-plot-r/)
5. Complete Introduction to Linear Regression in R,Selva Prabhakaran 2020, ML+:[*https://www.machinelearningplus.com/machine-learning/complete-introduction-linear-regression-r/*](https://www.machinelearningplus.com/machine-learning/complete-introduction-linear-regression-r/)
6. Linear Regression, Selva Prabhakaran 2018, r-statistics*: http://r-statistics.co/Linear-Regression.html*
7. Identifying the Most Important Independent Variables in Regression Models by Jim Frost (May 2020), Statistics: *https://statisticsbyjim.com/regression/identifying-important-independent-variables/*