Who Makes the World's Best Wine?

A statistical analysis of wine reviews

Greg DeVore

August 26th, 2017

Overview

You know the feeling: You're heading to a dinner party and decide to bring a bottle of wine. You stop at the grocery store and find yourself staring at a wall of wines from at least a dozen different countries. Red or white? Which varietal? Which country? The goal of this write-up is to help make that decision a little easier, and help guarantee you impress your friends with your selection. To do that, we'll be analyzing over 150,000 wine reviews from Wine Enthusiast magazine. Each review contains the type of wine, the country and region of origin, the price, and the number of points awarded to the wine (ranging from 80-100, 100 being 'perfect'). When looking at this data, we'll attempt to answer the following questions:

1. Can a country claim to make the best wine in the world in terms of average number of points awarded?
2. How do different countries compare in terms of price and quality? If you're looking for a great wine at a reasonable price, which countries are your best bet?

In a formal sense, in terms of hypothesis testing, our null and alternative hypotheses are as follows: The null hypothesis (H_0) is that there is no difference in terms of points and prices when it comes to different countries. In other words, it makes no difference which country's wine you buy. The alternative hypothesis (H_A), which is what we're trying to prove, is that it does matter where you buy your wine. There is in fact a difference in terms of price and quality depending on the country.

After exploring this data set we will find that, perhaps surprisingly, Austria makes the best wines out of the countries studied according to the average number of points awarded. After that, France and Germany are tied for second place. In terms of price, France makes by far the most expensive wines in terms of average price per bottle. After that, Germany and Italy are tied for second place. Austria, which makes the best wines in terms of points, is in the middle of the pack when it comes to price.

Given the outcomes above, your best best for a bottle of wine in terms of price and quality is Austria. The only catch is that Austria produces mostly white wines. If you're looking for a good red wine, French wines are great in terms of quality, but you're going to pay a premium in terms of price. As an alternative, consider Italy, which has comparable quality but a lower average price.

Exploratory Data Analysis

The wine review data used here is available at https://www.kaggle.com/zynicide/wine-reviews. As stated, there are just over 150,000 reviews. The particular quantities of interest for each review are the country, points awarded, and price. We can start by creating a unique list of countries to see which are represented.

```
country.list <- unique(wine$country)
country.list
```

```
##  [1] "US"                    "Spain"
##  [3] "France"                "Italy"
##  [5] "New Zealand"           "Bulgaria"
##  [7] "Argentina"             "Australia"
##  [9] "Portugal"              "Israel"
## [11] "South Africa"          "Greece"
```

```
## [13] "Chile"                  "Morocco"
## [15] "Romania"                "Germany"
## [17] "Canada"                 "Moldova"
## [19] "Hungary"                "Austria"
## [21] "Croatia"                "Slovenia"
## [23] ""                       "India"
## [25] "Turkey"                 "Macedonia"
## [27] "Lebanon"                "Serbia"
## [29] "Uruguay"                "Switzerland"
## [31] "Albania"                "Bosnia and Herzegovina"
## [33] "Brazil"                 "Cyprus"
## [35] "Lithuania"              "Japan"
## [37] "China"                  "South Korea"
## [39] "Ukraine"                "England"
## [41] "Mexico"                 "Georgia"
## [43] "Montenegro"             "Luxembourg"
## [45] "Slovakia"               "Czech Republic"
## [47] "Egypt"                  "Tunisia"
## [49] "US-France"
```

There are 49 different values for *country*, but note that one of them is blank. Let's take a closer look at those reviews to see if we can fill in the missing values.

```
# Try to identify missing countries
idx <- which(wine$country == "")
wine[idx,'winery',drop = FALSE]
```

```
##           winery
## 1134    Tsililis
## 1441   Büyülübağ
## 68227    Chilcas
## 113017   Chilcas
## 135697   Chilcas
```

A quick search of these wineries tell us the country of origin. We'll use this information to fill in the missing values.

```
# Add missing countries
wine[1134,2] <- 'Greece'
wine[1441,2] <- 'Turkey'
wine[c(68227,113017,135697),2] <- 'Chile'
```

Next, let's look for any reviews that are missing either points or price information.

```
summary(wine$points)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   80.00   86.00   88.00   87.89   90.00  100.00
```

```
summary(wine$price)
```

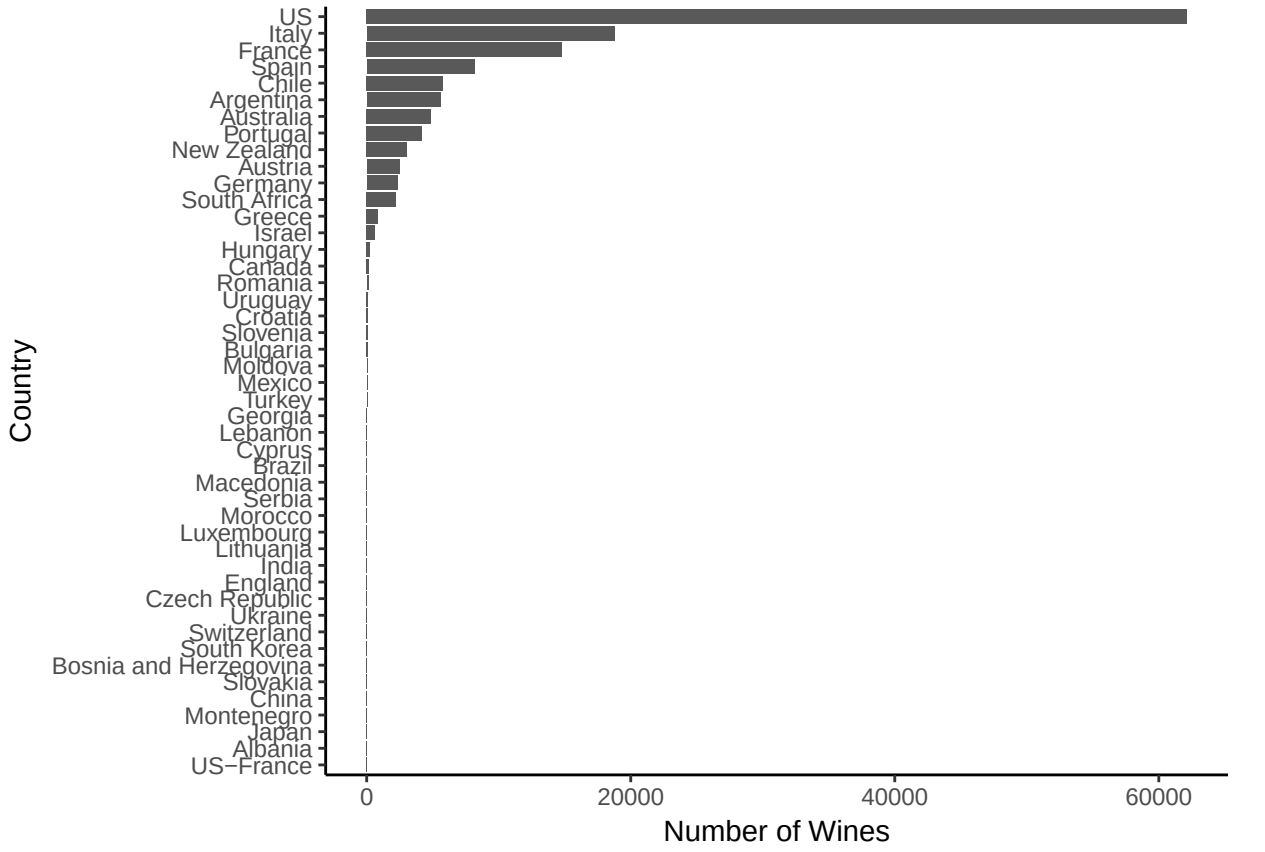
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    4.00   16.00   24.00   33.13   40.00 2300.00   13695
```

It looks like all reviews have a point value assigned, but there are nearly 14,000 wines without price data. Since we need this information, we'll have to remove these wines from the data set.

```
# Remove wines with no price
wine <- filter(wine, !is.na(price))
```

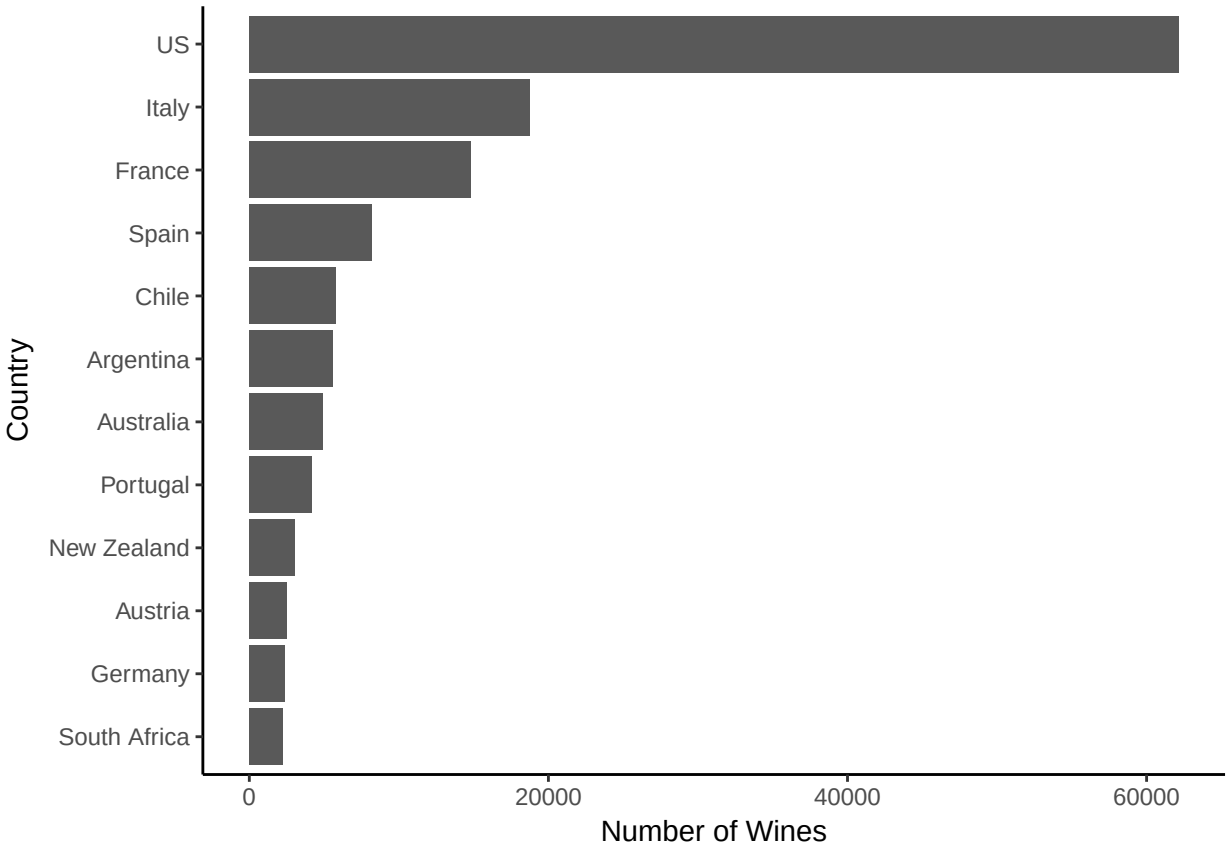
Because there are so many countries represented, we may want to limit our analysis to countries that produce a significant number of wines. To help, let's plot the review counts for each country.

```
# Look at counts per country
wine.count <- wine %>% count(country) %>% arrange(-n)
ggplot(wine.count, aes(x = reorder(country, n), y = n)) + geom_col() +
  xlab('Country') +ylab('Number of Wines') + coord_flip() + theme_classic()
```



Note that most countries have very few wines represented. Since these sample sizes are most likely not statistically meaningful, let's go ahead and remove them from the data set and focus only on countries with a good sample size. Looking at the data, there appears to be a drop off in number of wines after South Africa, so let's remove all countries below it.

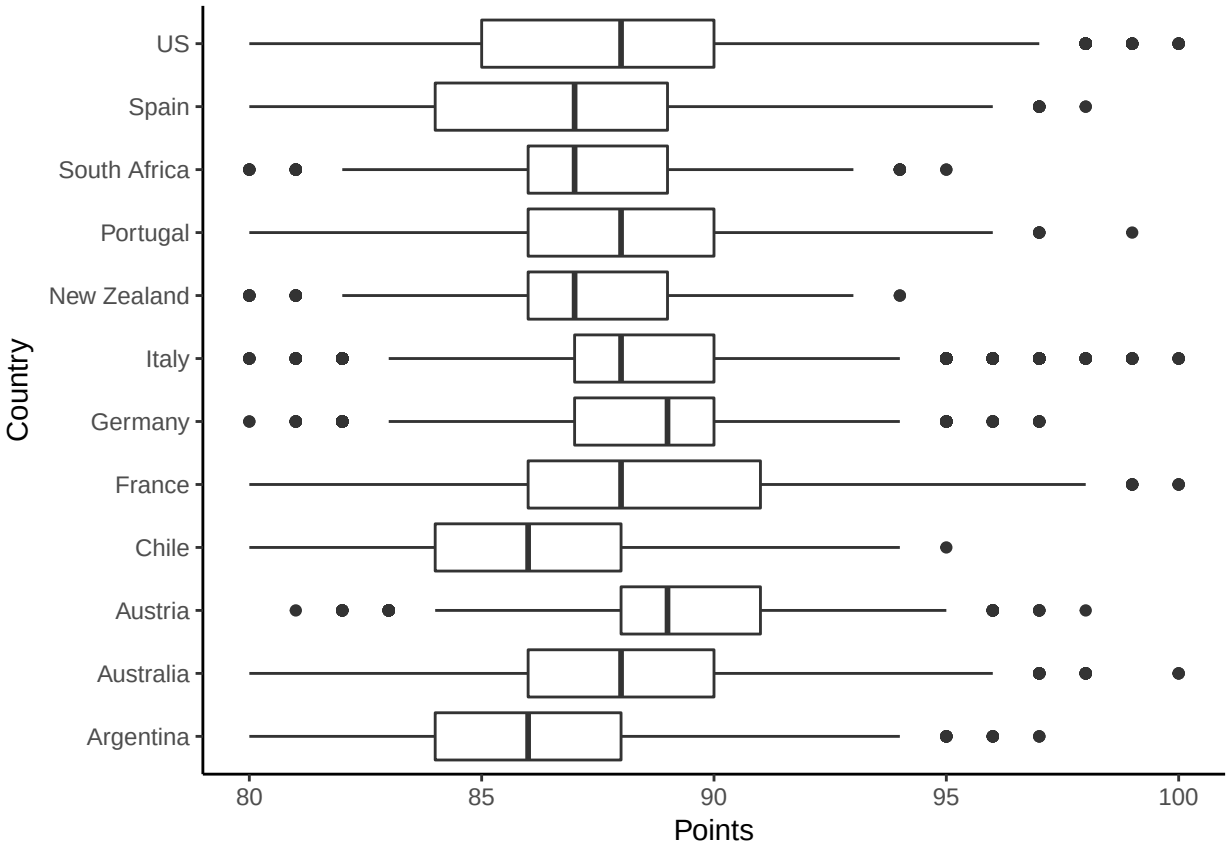
```
# Keep countries with > 1000 wines
keep <- filter(wine.count, n > 2000)
wine.count <- filter(wine.count, country %in% keep$country)
wine <- filter(wine, country %in% keep$country)
# Visualize counts by country
ggplot(wine.count, aes(x = reorder(country, n), y = n)) + geom_col() +
  xlab('Country') +ylab('Number of Wines') + coord_flip() + theme_classic()
```



Now, the remaining 12 countries in our data set have a decent sample size (at least 2,000 wines). Note that the United States has by far the highest number of wines, which shouldn't be surprising since Wine Enthusiast is a US publication.

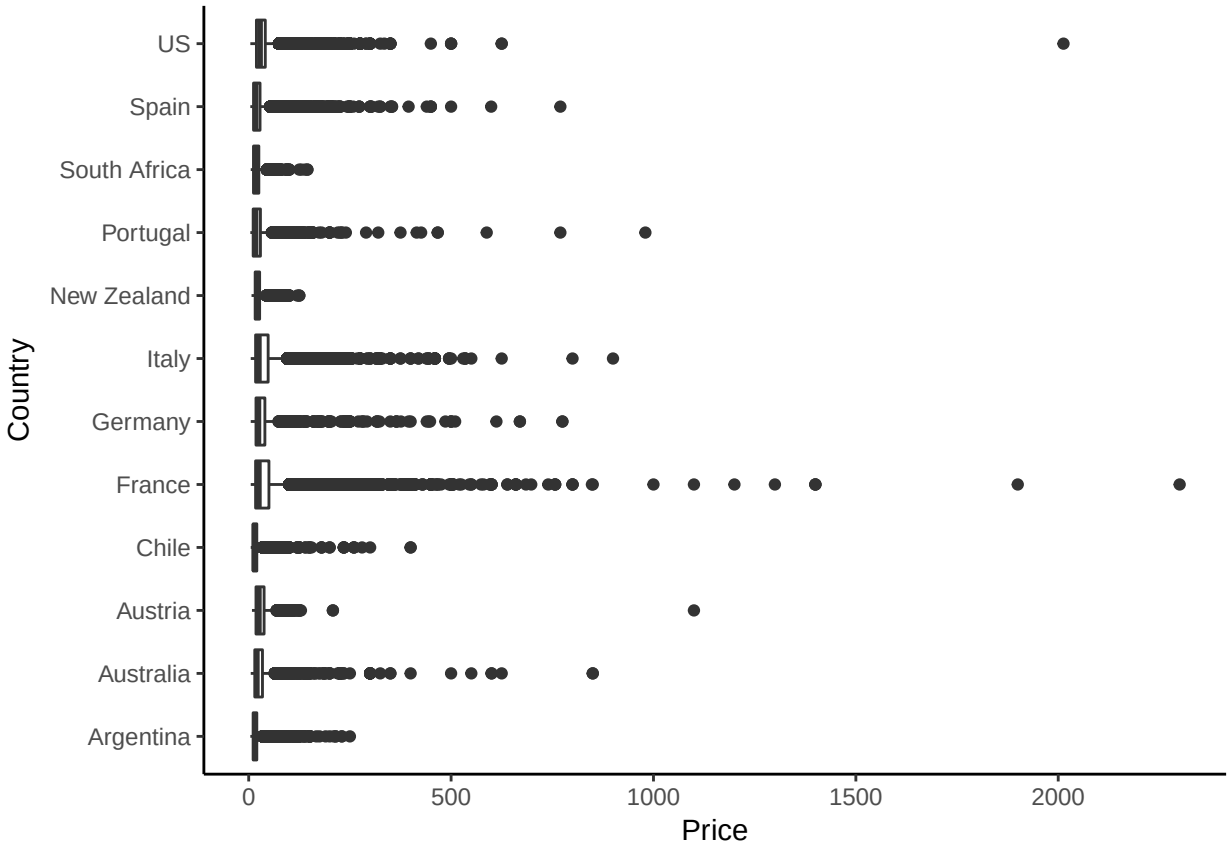
With the final countries selected, let's look at the distribution of points awarded and prices for each.

```
# Look at point distributions
ggplot(wine, aes(country, points)) + geom_boxplot() + xlab('Country') +
  ylab('Points') + coord_flip() + theme_classic()
```



All countries have the majority of their wines scored between 85 and 90 points, with the spreads relatively balanced (suggesting an approximately normal distribution). Note that the US, Italy, France, and Australia have all produced 100 point wines.

```
# Look at price distributions
ggplot(wine, aes(country, price)) + geom_boxplot() + xlab('Country') +
  ylab('Price') + coord_flip() + theme_classic()
```



The distribution of prices for all countries are highly right skewed, with extremely long tails. This makes sense, as most wines tend to be well under 100 dollars, with fewer wines at the higher end of the price range. Note that France appears to have the most expensive wines among all countries.

Let's take a closer look at the distribution of price to see if there are any extreme outliers. It appears that there could be potential outliers in the US, French, and Austrian wines.

```
# Identify extreme price outliers
wine.outliers <- filter(wine, (price > 1500) | (country == 'Austria' & price > 1000))
wine.outliers[, -c(1,3,4,8,9)]
```

```
##   country points price   province                 variety          winery
## 1 Austria     94  1100     Wachau         Grüner Veltliner  Emmerich Knoll
## 2      US     91  2013 California               Chardonnay           Blair
## 3  France     99  2300   Bordeaux Bordeaux-style Red Blend  Château Latour
## 4  France     98  1900   Bordeaux Bordeaux-style Red Blend Château Margaux
```

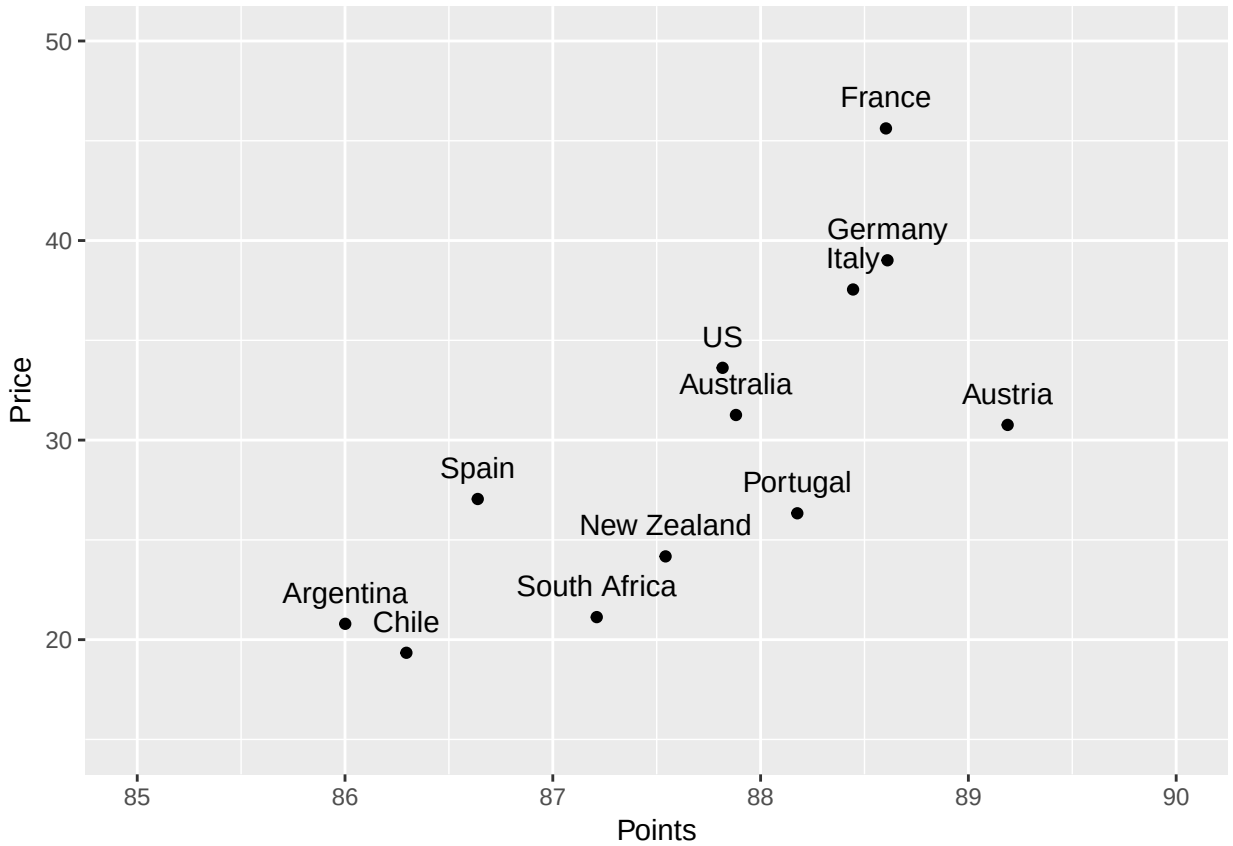
There are four outliers, one from Austria, one from the US, and two from France. The Austrian wine actually retails for just over 100 dollars, so this is evidently a misprint (an extra zero was added). The US wine is from a small winery along the central California coast with typical prices well under 100 dollars. This price point is most likely erroneous, and the value of 2013 indicates that perhaps the year was accidentally entered. The French wines are both from wineries that have been producing wine for at least 500 years, so those prices, although quite high, are most likely legitimate. We'll remove the Austrian and US wines, and leave the French wines in the data set.

```r
# Remove Austrian and US wines only
wine <- filter(wine, !(country == 'US' & price > 2000))
wine <- filter(wine, !(country == 'Austria' & price > 1000))
```

Relationship Between Price and Points

To complement the last two plots, let's take a high level look at price versus points for each country. We can do this by calculating the mean price and points for each country.

```r
# Split data frame by the 12 countries (creates a list of data frames)
wine.country <- split(wine,wine$country)
# Calculate simple means of price and points for each country
country.point.mean <- sapply(wine.country, function(x) mean(x$points))
country.price.mean <- sapply(wine.country, function(x) mean(x$price))
country.means <- data.frame(country = names(wine.country),points = country.point.mean,
  price = country.price.mean)
# Plot quantities
ggplot(country.means, aes(x = points, y= price)) + geom_point() +
  geom_text(aes(label = country), vjust = -1) + xlab('Points') + ylab('Price') +
 xlim(85,90) + ylim(15,50)
```



This plot gives a rough idea of the average points and prices for each country. Note that there is an overall positive correlation between price and points (countries that produce better wines tend to produce more expensive wines). We can't yet say which of these differences are statistically meaningful (we'll get to that later), but this at least gives us a high level view of how the various countries compare.

Next, let's look at whether or not there is a strong correlation between the points awarded to a wine and the price of that wine. We'll split the data by country and compute the individual correlation coefficients. Note that we're applying a log transformation to the prices due to the high amount of skew.

```
as.data.frame(wine %>% group_by(country) %>%
summarize(corr = cor(points,log(price))) %>% arrange(-corr))
```

```
##         country      corr
## 1         Italy 0.7343516
## 2        France 0.7050856
## 3     Australia 0.6868276
## 4      Portugal 0.6398890
## 5     Argentina 0.6194844
## 6  South Africa 0.6170417
## 7       Austria 0.6146309
## 8         Spain 0.6057671
## 9         Chile 0.6040882
## 10      Germany 0.5865599
## 11           US 0.5417158
## 12  New Zealand 0.4552688
```

Note that all countries have a positive correlation coefficient between points and $log(price)$. This means that an increase in points awarded to a wine corresponds to an increase in price, which is somewhat expected. Italy and France show the strongest correlation, while the US and New Zealand show the weakest correlation. For all countries, the correlation coefficient is not too large, ranging from approximately 0.45 to 0.73. This indicates a moderately strong correlation between points and $log(price)$ across all countries.

It is important to acknowledge that the correlation coefficient doesn't tell us how much the price of a wine changes per unit change in points awarded, rather it only tells us about the strength of the relationship between price and points. If we want to investigate the former, we can use regression to model a linear relationship between price and points. We'll do that below and record the slope of the line for each country. As with the correlation coefficient, we're using $log(price)$ instead of price.

```
as.data.frame(wine %>% group_by(country) %>%
summarize(slope = lm(log(price) ~ points)$coefficients[2]) %>% arrange(-slope))
```

```
##         country      slope
## 1         Italy 0.18260267
## 2        France 0.18024399
## 3     Australia 0.15751742
## 4      Portugal 0.15537224
## 5       Germany 0.14060734
## 6  South Africa 0.13582899
## 7         Spain 0.13485290
## 8       Austria 0.12843692
## 9         Chile 0.12069934
## 10    Argentina 0.11825014
## 11           US 0.09198217
## 12  New Zealand 0.08325964
```

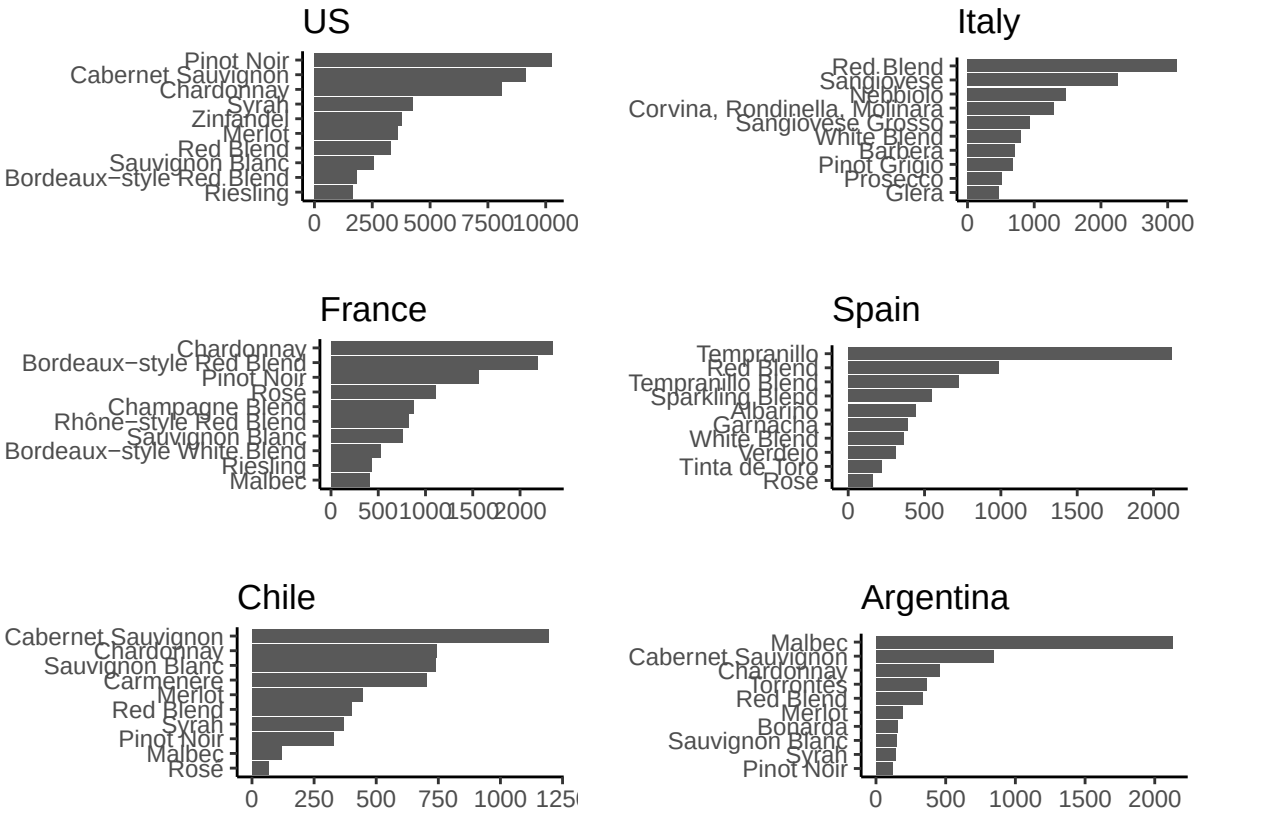
We have to be careful when interpreting the slope, since it is in units of $log(price)$ per point. Using the properties of the logarithm function, we can view the slope as the percent change in price given a unit change in points. For example, the price of an Italian or French wine increases by roughly 18% for each extra point awarded, whereas a wine from New Zealand increases by just over 8%.

Top Varietals By Country

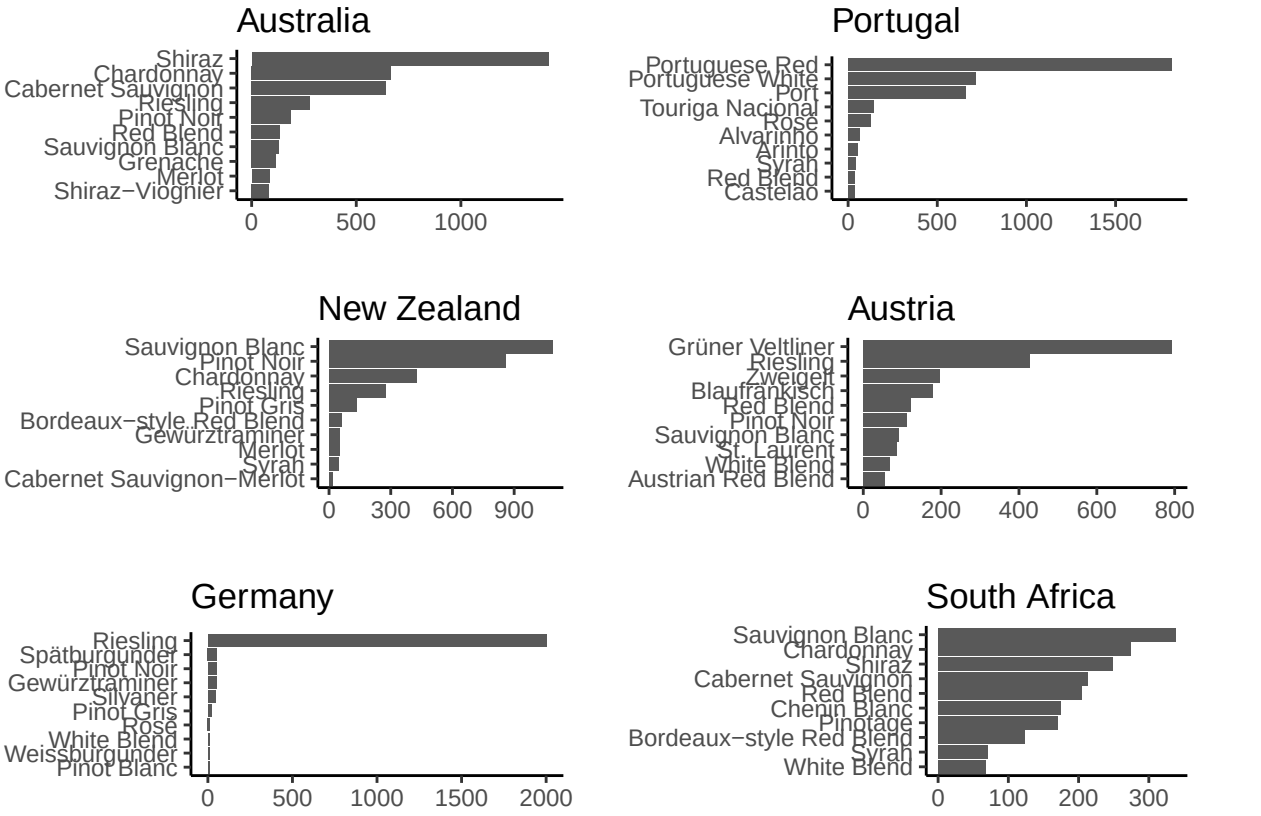
As a final exploratory exercise, let's look at the top 10 varietals produced by each of the 12 countries.

```
# Look at top 10 varietals
wine.var <- wine %>% group_by(country) %>% count(variety) %>% arrange(-n) %>% slice(1:10)

p1 <- ggplot(subset(wine.var, country == 'US'), aes(x = reorder(variety,n), y = n)) +
  geom_col() + xlab('') + ylab('') + ggtitle('US') + coord_flip() + theme_classic()
p2 <- ggplot(subset(wine.var, country == 'Italy'), aes(x = reorder(variety,n), y = n)) +
  geom_col() + xlab('') + ylab('') + ggtitle('Italy') + coord_flip() + theme_classic()
p3 <- ggplot(subset(wine.var, country == 'France'), aes(x = reorder(variety,n), y = n)) +
  geom_col() + xlab('') + ylab('') + ggtitle('France') + coord_flip() + theme_classic()
p4 <- ggplot(subset(wine.var, country == 'Spain'), aes(x = reorder(variety,n), y = n)) +
  geom_col() + xlab('') + ylab('') + ggtitle('Spain') + coord_flip() + theme_classic()
p5 <- ggplot(subset(wine.var, country == 'Chile'), aes(x = reorder(variety,n), y = n)) +
  geom_col() + xlab('') + ylab('') + ggtitle('Chile') + coord_flip() + theme_classic()
p6 <- ggplot(subset(wine.var, country == 'Argentina'),
  aes(x = reorder(variety,n), y = n)) + geom_col() + xlab('') + ylab('') +
  ggtitle('Argentina') + coord_flip() + theme_classic()
grid.arrange(p1, p2, p3, p4, p5, p6, ncol=2)
```



```
p1 <- ggplot(subset(wine.var, country == 'Australia'),
  aes(x = reorder(variety,n), y = n)) + geom_col() + xlab('') + ylab('') +
  ggtitle('Australia') + coord_flip() + theme_classic()
p2 <- ggplot(subset(wine.var, country == 'Portugal'),
  aes(x = reorder(variety,n), y = n)) + geom_col() + xlab('') + ylab('') +
  ggtitle('Portugal') + coord_flip() + theme_classic()
p3 <- ggplot(subset(wine.var, country == 'New Zealand'),
  aes(x = reorder(variety,n), y = n)) + geom_col() + xlab('') + ylab('') +
  ggtitle('New Zealand') + coord_flip() + theme_classic()
p4 <- ggplot(subset(wine.var, country == 'Austria'), aes(x = reorder(variety,n), y = n)) +
  geom_col() + xlab('') + ylab('') + ggtitle('Austria') + coord_flip() + theme_classic()
p5 <- ggplot(subset(wine.var, country == 'Germany'), aes(x = reorder(variety,n), y = n)) +
  geom_col() + xlab('') + ylab('') + ggtitle('Germany') + coord_flip() + theme_classic()
p6 <- ggplot(subset(wine.var, country == 'South Africa'),
  aes(x = reorder(variety,n), y = n)) + geom_col() + xlab('') + ylab('') +
  ggtitle('South Africa') + coord_flip() + theme_classic()
grid.arrange(p1, p2, p3, p4, p5, p6, ncol=2)
```



The main observation from the varietals plots is that certain countries are clearly known for producing a dominant varietal:

- Spanish tempranillo
- Chilean cabernet sauvignon
- Argentinian malbec
- Australian shiraz
- Portuguese reds
- Austrian gruner veltliner
- German riesling

These country/varietal pairings probably look familiar to anyone who enjoys wine. Perhaps the biggest surprise is the domination of German riesling, which is by far the most dominant varietal in terms of fraction of total wines produced for any single country.

Ranking Countries By Points and Price

In the previous section, we explored the data set and got an idea of the distribution of points and prices for each country, the relationship between the two variables for each country, and how the countries compare in terms of both (countries with more expensive wines, higher average points awarded, and vice versa). Now, we're ready to move on to actually ranking the countries and see if there's a clear winner in terms of price and points. To do this, we'll be conducting pairwise comparisons between each of the 12 countries and seeing first if a meaningful difference exists between their means for price and points, and if so, which country comes out ahead. Recall that our null hypothesis (H_0) is that there is no difference in terms of points and prices when it comes to different countries, and our alternative hypothesis (H_A) is that it does matter where you buy your wine.

We're going to start with two classical tests: ANOVA (analysis of variance) and Tukey's HSD. ANOVA will tell us if a significant difference exists between any of our groups, and Tukey's HSD will tell us which groups are actually different. Before we do this however, we need to make sure the tests are appropriate given our data.

ANOVA Power

We need to ensure that our ANOVA test will be powerful enough, in terms of being able to reject the null hypothesis when the alternative hypothesis is true. To do this, we can compute the number of wines required for each group given our desired power level of the test. This depends on the following:

1. Our number of groups, which in this case is 12.
2. The number of wines in each group, which is our unknown.
3. The effect size we're trying to detect (in this case the difference in means), which we'll set to 0.10.
4. The significance level, which is our rate of Type I, or false positive, error probability. We'll set this to 0.01, which means a 1% chance of falsely rejecting H_0. This is important due to our large number of groups.
5. The desired power level (which is 1 minus our Type II, or false negative, error probability.). We'll set this to 0.99, which means we want a 99% percent chance of rejecting H_0 when H_A is true.

We can plug this into a formula and get the required group size given our other data.

```
# We have 12 groups, want to detect effect size of 0.1 (difference in means)
# with a significance level of 0.01 and a power of 99%
pwr.anova.test(k = 12, n = NULL, f = 0.1, sig.level = 0.01, power = 0.99)
```

```
##
##      Balanced one-way analysis of variance power calculation
##
##               k = 12
##               n = 343.4632
##               f = 0.1
##       sig.level = 0.01
##           power = 0.99
##
## NOTE: n is number in each group
```

As stated, our unknown is n, which is the number of wines per country in order to achieve the desired power level. It is just under 350, and the smallest country represented has just over 2,000 samples. This means that our ANOVA test will be powerful enough to detect our desired effect level at our desired significance level.

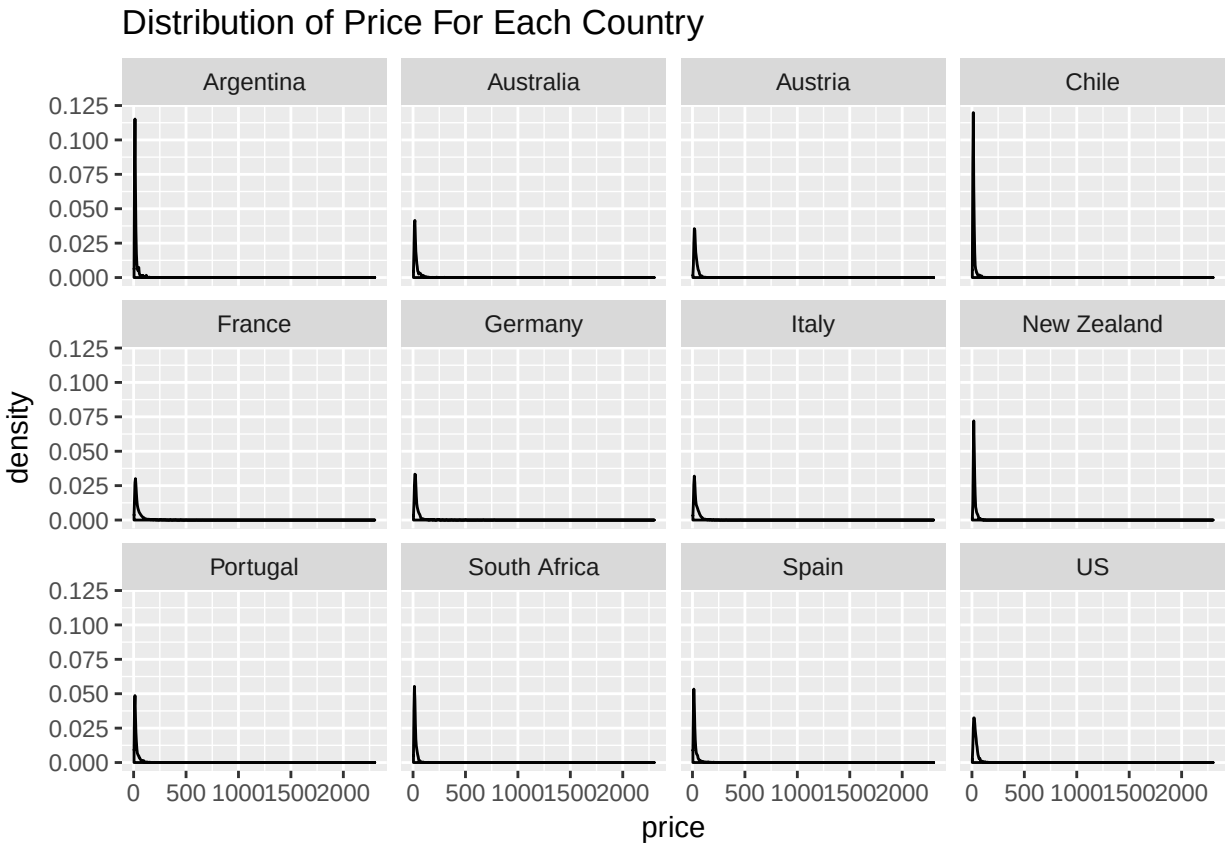
ANOVA Assumptions

Both ANOVA and Tukey's HSD make the following assumptions:

1. The observations are independent, which in our case is true since all of the wines were produced by different countries and selected for review independently by the magazine.
2. The distributions of the residuals are normal. We can ensure this by examining the distributions of our dependent variables (price or points, depending on the test) for each country.
3. The variance within each group should be the same. We can check this by computing the variance of price and points for each country.

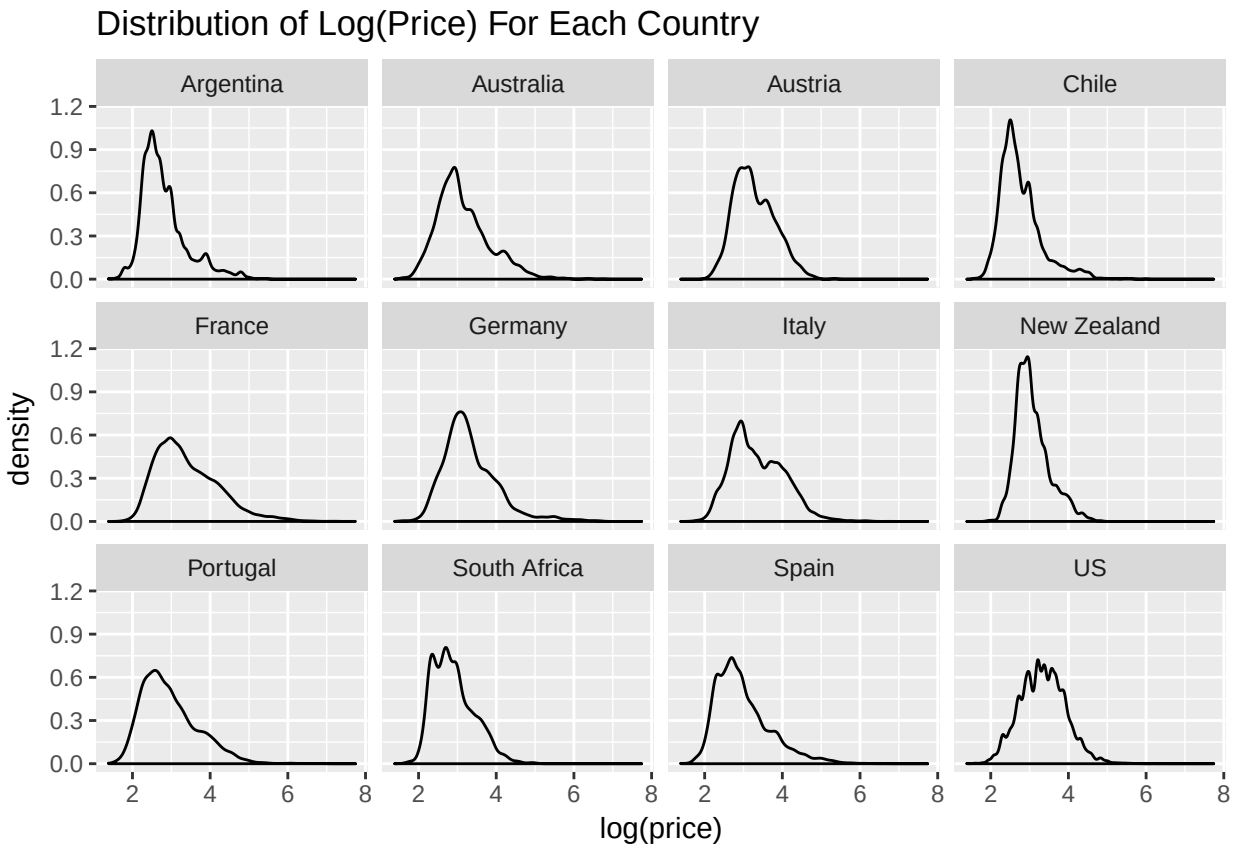
To check (2), we can produce plots of price and points for each country and asses the normality.

```
ggplot(wine, aes(price)) + geom_density() + facet_wrap(~country) +
  ggtitle('Distribution of Price For Each Country')
```



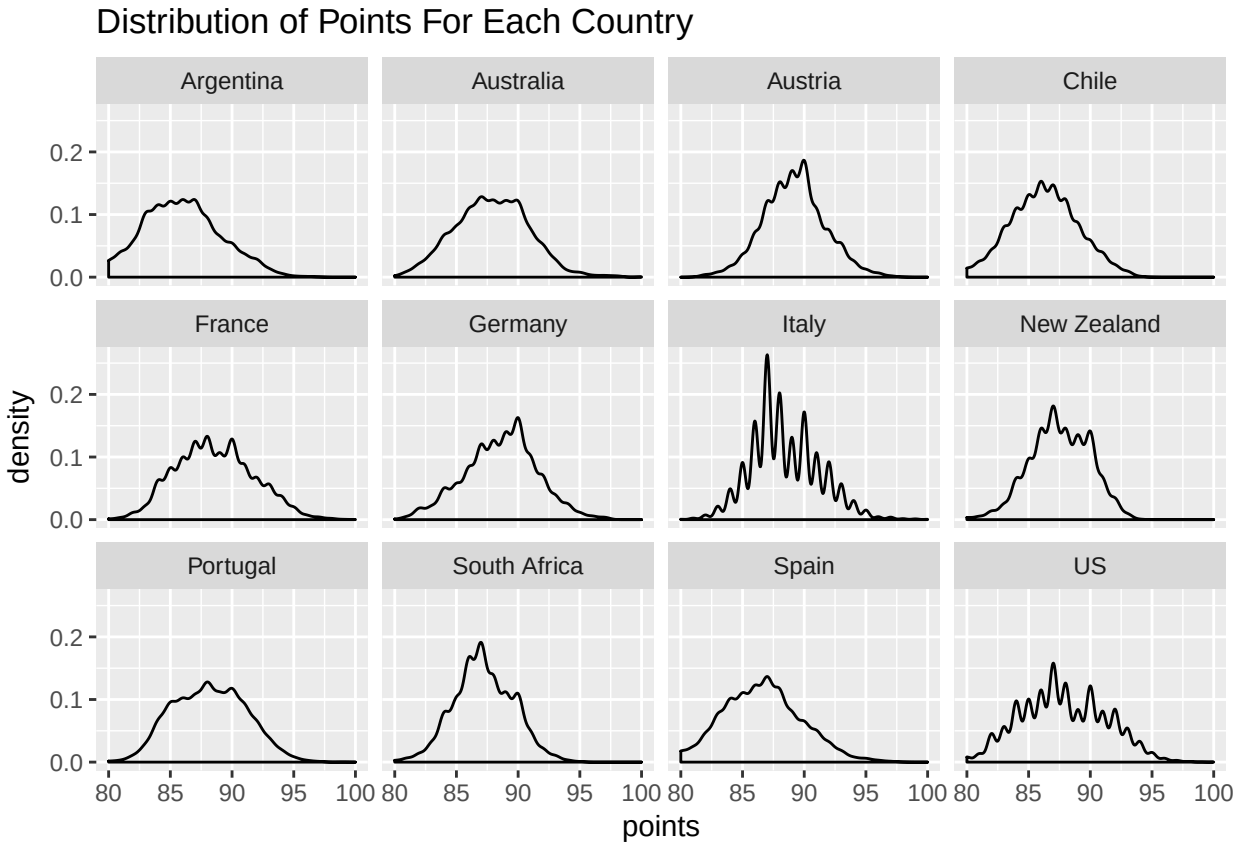
Note that the distribution of price is highly skewed, which is not surprising given that we saw this behavior in the box plots produced during our exploratory data analysis. This violation means our results for ANOVA and Tukey's HSD may not be reliable. To help remedy this, let's try applying a log transform to the price data.

```
ggplot(wine, aes(log(price))) + geom_density() + facet_wrap(~country) +
  ggtitle('Distribution of Log(Price) For Each Country')
```



These distributions, while still skewed, are much closer to normal, so we'll make sure to use $log(price)$ when doing the ANOVA test. Let's move on to the distribution of points.

```
ggplot(wine, aes(points)) + geom_density() + facet_wrap(~country) +
  ggtitle('Distribution of Points For Each Country')
```



These distributions look fairly normal, so we can use the points data as in when doing the ANOVA test.

To check (3), we can compute the standard deviation and variance for points and price within each of the groups. Note that we're using $log(price)$ here because of the skew found in the price data.

```
wine %>% group_by(country) %>% summarize(sd.points = sd(points),sd.price = sd(log(price)),
  var.points = var(points),var.price = var(log(price)))
```

```
## # A tibble: 12 x 5
##         country sd.points  sd.price var.points var.price
##           <chr>     <dbl>     <dbl>      <dbl>     <dbl>
##  1    Argentina  3.093020 0.5904104   9.566773 0.3485845
##  2    Australia  2.981422 0.6837608   8.888874 0.4675289
##  3      Austria  2.487403 0.5197825   6.187174 0.2701738
##  4        Chile  2.708398 0.5411491   7.335417 0.2928424
##  5       France  3.142284 0.8032752   9.873947 0.6452510
##  6      Germany  2.916430 0.6991127   8.505566 0.4887586
##  7        Italy  2.754785 0.6850003   7.588841 0.4692255
##  8  New Zealand  2.402068 0.4392906   5.769928 0.1929762
##  9     Portugal  2.929330 0.7112744   8.580977 0.5059112
## 10 South Africa  2.394368 0.5270707   5.733000 0.2778035
## 11        Spain  3.128904 0.6965413   9.790043 0.4851698
## 12           US  3.410177 0.5790408  11.629310 0.3352882
```

Note that the variance for both point and price differ significantly between countries. This violation means our results for ANOVA and Tukey's HSD may not be reliable. Because our group sizes are different (the number of wines range from 2,000 to 60,000 per country), a rule of thumb is the largest value of standard deviation should be no more than twice the smallest value. Our data comes close to violating this (0.8/0.43 ~= 1.9), so we should proceed with caution when using the ANOVA test.

ANOVA Comparison of Wine Points

Now that we've looked at the requirements for ANOVA, and saw that our data satisfies most of them, we're ready to proceed (with caution) with ANOVA and Tukey's HSD. As stated, we're going to be looking at pairwise comparisons between the countries in our data set. Recall that we have 12 groups, so that means we have

```
choose(12,2)
```

```
## [1] 66
```

pairwise comparisons in our test. Let's proceed and create an ANOVA model for points awarded, using a 99% confidence level.

```
# ANOVA test for points
df_aov_points <- aov(points ~ country, data = wine)
summary(df_aov_points)
```

```
##                 Df  Sum Sq Mean Sq F value Pr(>F)
## country         11   67421    6129   618.6 <2e-16 ***
## Residuals   134417 1331906      10
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
print(df_aov_points)
```

```
## Call:
##    aov(formula = points ~ country, data = wine)
##
## Terms:
##                   country Residuals
## Sum of Squares    67420.9 1331905.9
## Deg. of Freedom        11    134417
##
## Residual standard error: 3.147818
## Estimated effects may be unbalanced
```

Note that the p-value from our ANOVA model is essentially zero, which means that a significant difference in points exists between at least two countries, which isn't surprising given our summary charts created earlier. Let's run Tukey's HSD using our ANOVA model to see which countries are different.

```
tukey_anova_points = TukeyHSD(df_aov_points, conf.level = 0.99)
tukey_anova_points
```

```
##   Tukey multiple comparisons of means
##     99% family-wise confidence level
##
## Fit: aov(formula = points ~ country, data = wine)
##
## $country
##                                 diff         lwr          upr     p adj
## Australia-Argentina       1.880413614  1.64987353  2.110953696 0.0000000
## Austria-Argentina         3.187886594  2.90384232  3.471930865 0.0000000
## Chile-Argentina           0.294297893  0.07327387  0.515321921 0.0000406
## France-Argentina          2.602240248  2.41732052  2.787159971 0.0000000
## Germany-Argentina         2.609492759  2.31984737  2.899138148 0.0000000
## Italy-Argentina           2.443932467  2.26449229  2.623372649 0.0000000
## New Zealand-Argentina     1.541597088  1.27705694  1.806137233 0.0000000
```

```
## Portugal-Argentina        2.175650216  1.93477703  2.416523407 0.0000000
## South Africa-Argentina    1.210369976  0.91575251  1.504987446 0.0000000
## Spain-Argentina           0.637651569  0.43317858  0.842124561 0.0000000
## US-Argentina              1.816075005  1.65161021  1.980539796 0.0000000
## Austria-Australia         1.307472980  1.01730868  1.597637284 0.0000000
## Chile-Australia          -1.586115721 -1.81495146 -1.357279981 0.0000000
## France-Australia          0.721826634  0.52763735  0.916015921 0.0000000
## Germany-Australia         0.729079145  0.43342965  1.024728644 0.0000000
## Italy-Australia           0.563518853  0.37454014  0.752497566 0.0000000
## New Zealand-Australia    -0.338816526 -0.60991736 -0.067715690 0.0001851
## Portugal-Australia        0.295236602  0.04717599  0.543297218 0.0005249
## South Africa-Australia   -0.670043638 -0.97056590 -0.369521379 0.0000000
## Spain-Australia          -1.242762046 -1.45565493 -1.029869162 0.0000000
## US-Australia             -0.064338609 -0.23916093  0.110483708 0.9680526
## Chile-Austria            -2.893588701 -3.17625142 -2.610925979 0.0000000
## France-Austria           -0.585646346 -0.84107128 -0.330221410 0.0000000
## Germany-Austria          -0.578393835 -0.91742340 -0.239364271 0.0000000
## Italy-Austria            -0.743954127 -0.99544045 -0.492467801 0.0000000
## New Zealand-Austria      -1.646289506 -1.96413866 -1.328440357 0.0000000
## Portugal-Austria         -1.012236378 -1.31067647 -0.713796287 0.0000000
## South Africa-Austria     -1.977516618 -2.32080374 -1.634229498 0.0000000
## Spain-Austria            -2.550235026 -2.82015316 -2.280316893 0.0000000
## US-Austria               -1.371811589 -1.61284117 -1.130782004 0.0000000
## France-Chile              2.307942355  2.12515185  2.490732864 0.0000000
## Germany-Chile             2.315194866  2.02690418  2.603485548 0.0000000
## Italy-Chile               2.149634574  1.97238942  2.326879729 0.0000000
## New Zealand-Chile         1.247299195  0.98424301  1.510355376 0.0000000
## Portugal-Chile            1.881352323  1.64210985  2.120594795 0.0000000
## South Africa-Chile        0.916072083  0.62278635  1.209357813 0.0000000
## Spain-Chile               0.343353675  0.14080425  0.545903104 0.0000000
## US-Chile                  1.521777112  1.35971005  1.683844177 0.0000000
## Germany-France            0.007252511 -0.25438694  0.268891959 1.0000000
## Italy-France             -0.158307781 -0.28776638 -0.028849183 0.0002985
## New Zealand-France       -1.060643160 -1.29418596 -0.827100364 0.0000000
## Portugal-France          -0.426590032 -0.63294081 -0.220239254 0.0000000
## South Africa-France      -1.391870272 -1.65900358 -1.124736966 0.0000000
## Spain-France             -1.964588680 -2.12697696 -1.802200397 0.0000000
## US-France                -0.786165243 -0.89391225 -0.678418238 0.0000000
## Italy-Germany            -0.165560292 -0.42335609  0.092235510 0.4042351
## New Zealand-Germany      -1.067895671 -1.39076002 -0.745031327 0.0000000
## Portugal-Germany         -0.433842543 -0.73761843 -0.130066654 0.0000060
## South Africa-Germany     -1.399122783 -1.74705862 -1.051186941 0.0000000
## Spain-Germany            -1.971841191 -2.24764747 -1.696034915 0.0000000
## US-Germany               -0.793417754 -1.04102342 -0.545812091 0.0000000
## New Zealand-Italy        -0.902335379 -1.13156389 -0.673106863 0.0000000
## Portugal-Italy           -0.268282250 -0.46973725 -0.066827247 0.0000405
## South Africa-Italy       -1.233562490 -1.49693234 -0.970192641 0.0000000
## Spain-Italy              -1.806280898 -1.96240082 -1.650160979 0.0000000
## US-Italy                 -0.627857462 -0.72590282 -0.529812105 0.0000000
## Portugal-New Zealand      0.634053128  0.35411237  0.913993886 0.0000000
## South Africa-New Zealand -0.331227112 -0.65855934 -0.003894885 0.0084655
## Spain-New Zealand        -0.903945520 -1.15325685 -0.654634192 0.0000000
## US-New Zealand            0.274477917  0.05677261  0.492183225 0.0001518
## South Africa-Portugal    -0.965280240 -1.27380057 -0.656759912 0.0000000
```

```
## Spain-Portugal           -1.537998648 -1.76204004 -1.313957258 0.0000000
## US-Portugal              -0.359575211 -0.54781444 -0.171335983 0.0000000
## Spain-South Africa       -0.572718408 -0.85374172 -0.291695093 0.0000000
## US-South Africa           0.605705029  0.35230107  0.859108987 0.0000000
## US-Spain                  1.178423436  1.03977526  1.317071616 0.0000000
```

There's a lot of data here, given that we have 66 pairwise comparisons. Note that nearly all of the pairs have a significant difference at a 99% confidence level (p-value less than 0.01). The following countries were too close (the confidence interval for the difference in means included zero) to declare a significant difference in terms of points:

- US versus Australia
- Germany versus France
- Italy versus Germany

ANOVA Comparison of Wine Prices

Having looked at points, let's create an ANOVA model for $log(price)$, again using a 99% confidence level.

```
# ANOVA test for log(price)
df_aov_price <- aov(log(price) ~ country, data = wine)
summary(df_aov_price)
```

```
##                 Df Sum Sq Mean Sq F value Pr(>F)
## country         11   4946   449.6    1115 <2e-16 ***
## Residuals   134417  54212     0.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
print(df_aov_price)
```

```
## Call:
##    aov(formula = log(price) ~ country, data = wine)
##
## Terms:
##                   country Residuals
## Sum of Squares   4945.79  54211.60
## Deg. of Freedom       11    134417
##
## Residual standard error: 0.6350662
## Estimated effects may be unbalanced
```

Once again, the p-value from our ANOVA model is essentially zero, which means that a significant difference in price exists between at least two countries, which again isn't surprising given our summary charts created earlier. Let's run Tukey's HSD using our ANOVA model to see which countries are different.

```
tukey_anova_price = TukeyHSD(df_aov_price, conf.level = 0.99)
tukey_anova_price
```

```
##   Tukey multiple comparisons of means
##     99% family-wise confidence level
##
## Fit: aov(formula = log(price) ~ country, data = wine)
##
## $country
##                                diff          lwr           upr     p adj
## Australia-Argentina       0.33875352  0.292242511  0.3852645372 0.0000000
```

```
## Austria-Argentina         0.47344941  0.416144029  0.5307547872 0.0000000
## Chile-Argentina          -0.04321503 -0.087806198  0.0013761371 0.0152248
## France-Argentina          0.59902604  0.561718841  0.6363332295 0.0000000
## Germany-Argentina         0.51368707  0.455251673  0.5721224608 0.0000000
## Italy-Argentina           0.55180332  0.515601611  0.5880050260 0.0000000
## New Zealand-Argentina     0.26738368  0.214013217  0.3207541366 0.0000000
## Portugal-Argentina        0.14483769  0.096241997  0.1934333921 0.0000000
## South Africa-Argentina    0.08666231  0.027223813  0.1461008155 0.0000032
## Spain-Argentina           0.17241299  0.131160962  0.2136650190 0.0000000
## US-Argentina              0.53182091  0.498640458  0.5650013609 0.0000000
## Austria-Australia         0.13469588  0.076155800  0.1932359677 0.0000000
## Chile-Australia          -0.38196855 -0.428135720 -0.3358013892 0.0000000
## France-Australia          0.26027251  0.221095200  0.2994498218 0.0000000
## Germany-Australia         0.17493354  0.115286832  0.2345802543 0.0000000
## Italy-Australia           0.21304979  0.174923706  0.2511758819 0.0000000
## New Zealand-Australia    -0.07136985 -0.126063913 -0.0166757809 0.0000673
## Portugal-Australia       -0.19391583 -0.243961576 -0.1438700830 0.0000000
## South Africa-Australia   -0.25209121 -0.312720991 -0.1914614287 0.0000000
## Spain-Australia          -0.16634053 -0.209291259 -0.1233898085 0.0000000
## US-Australia              0.19306739  0.157797323  0.2283374475 0.0000000
## Chile-Austria            -0.51666444 -0.573691093 -0.4596377841 0.0000000
## France-Austria            0.12557663  0.074045143  0.1771081110 0.0000000
## Germany-Austria           0.04023766 -0.028160896  0.1086362140 0.5491298
## Italy-Austria             0.07835391  0.027617033  0.1290907876 0.0000005
## New Zealand-Austria      -0.20606573 -0.270191179 -0.1419402834 0.0000000
## Portugal-Austria         -0.32861171 -0.388821421 -0.2684020058 0.0000000
## South Africa-Austria     -0.38678709 -0.456044602 -0.3175295851 0.0000000
## Spain-Austria            -0.30103642 -0.355491876 -0.2465809592 0.0000000
## US-Austria                0.05837150  0.009744251  0.1069987513 0.0004408
## France-Chile              0.64224107  0.605363436  0.6791186952 0.0000000
## Germany-Chile             0.55690210  0.498740013  0.6150641819 0.0000000
## Italy-Chile               0.59501835  0.559259483  0.6307772141 0.0000000
## New Zealand-Chile         0.31059871  0.257527635  0.3636697803 0.0000000
## Portugal-Chile            0.18805272  0.139786022  0.2363194281 0.0000000
## South Africa-Chile        0.12987734  0.070707519  0.1890471702 0.0000000
## Spain-Chile               0.21562802  0.174764068  0.2564919742 0.0000000
## US-Chile                  0.57503594  0.542339224  0.6077326549 0.0000000
## Germany-France           -0.08533897 -0.138124218 -0.0325537182 0.0000001
## Italy-France             -0.04722272 -0.073340737 -0.0211046964 0.0000000
## New Zealand-France       -0.33164236 -0.378759163 -0.2845255532 0.0000000
## Portugal-France          -0.45418834 -0.495819209 -0.4125574725 0.0000000
## South Africa-France      -0.51236372 -0.566257346 -0.4584700961 0.0000000
## Spain-France             -0.42661304 -0.459374565 -0.3938515244 0.0000000
## US-France                -0.06720513 -0.088942875 -0.0454673771 0.0000000
## Italy-Germany             0.03811625 -0.013893550  0.0901260529 0.2059319
## New Zealand-Germany      -0.24630339 -0.311440644 -0.1811661366 0.0000000
## Portugal-Germany         -0.36884937 -0.430135567 -0.3075631779 0.0000000
## South Africa-Germany     -0.42702475 -0.497220132 -0.3568293736 0.0000000
## Spain-Germany            -0.34127408 -0.396917456 -0.2856306970 0.0000000
## US-Germany                0.01813384 -0.031820119  0.0680878030 0.9711610
## New Zealand-Italy        -0.28441964 -0.330666049 -0.2381732341 0.0000000
## Portugal-Italy           -0.40696562 -0.447608779 -0.3663224688 0.0000000
## South Africa-Italy       -0.46514100 -0.518275359 -0.4120066493 0.0000000
## Spain-Italy              -0.37939033 -0.410887218 -0.3478934378 0.0000000
```

```
## US-Italy                 -0.01998241 -0.039762869 -0.0002019491 0.0086694
## Portugal-New Zealand     -0.12254598 -0.179023486 -0.0660684793 0.0000000
## South Africa-New Zealand -0.18072136 -0.246760003 -0.1146827227 0.0000000
## Spain-New Zealand        -0.09497069 -0.145268762 -0.0446726109 0.0000000
## US-New Zealand            0.26443723  0.220515610  0.3083588547 0.0000000
## South Africa-Portugal    -0.05817538 -0.120418756  0.0040679958 0.0238616
## Spain-Portugal            0.02757530 -0.017624618  0.0727752106 0.4894189
## US-Portugal               0.38698321  0.349006317  0.4249601129 0.0000000
## Spain-South Africa        0.08575068  0.029054769  0.1424465834 0.0000010
## US-South Africa           0.44515860  0.394034839  0.4962823507 0.0000000
## US-Spain                  0.35940792  0.331435918  0.3873799192 0.0000000
```

Note that as with points, nearly all of the pairs have a significant difference at a 99% confidence level. The following countries were too close (the confidence interval for the difference in means included zero) to declare a significant difference in terms of price:

- Chile versus Argentina
- Germany versus Austria
- Italy versus Germany
- US versus Germany
- South Africa versus Portugal
- Spain versus Portugal

An Alternative Approach (Bootstrap)

ANOVA and Tukey's HSD show that nearly all countries have a significant difference in terms of points and price, but recall that we may have violated some of the assumptions required for the results to be considered valid. As an alternative, let's use bootstrap methods to compute the difference in means for each country and compare the results. Bootstrap methods make no assumptions about normality or variance, and will let us create an artificial population by repeatedly sampling our wine data with replacement to create 'new' samples of the same data. Our algorithm will be:

1. Iterate over all 66 country pairs.
2. For each pair of countries, sample both populations with replacement to create new samples with the same number of wines as the original population. Compute the mean (for points or price) for both countries and subtract to get a difference in means.
3. Repeat (2) 10,000 times.
4. Compute the confidence interval for the distribution of the difference in means.
5. If the 99% confidence interval excludes zero, declare one of the countries the 'winner' of the match-up. If the interval includes zero, neither country can be declared the winner.
6. Summarize results for all 66 comparisons and see which countries came out on top.

Once again, we'll be using a 99% confidence level because of the large number of groups being tested. As a first step, let's compute all country pairs.

```
# Confidence interval
p = 0.01
# Create pairwise comparisons
country.pairs <- combn(keep$country,2)
country1 <- country.pairs[1,]
country2 <- country.pairs[2,]
```

Bootstrap Comparison of Wine Points

The code below computes and compares the bootstrapped difference in point means between each country. Afterwards, a data frame is created to summarize the results. There are columns for the countries being compared, the confidence interval for their difference in means, and a winner, if the confidence interval excludes zero.

```
# Rank countries by points
ci.lower <- vector(length = ncol(country.pairs))
ci.upper <- vector(length = ncol(country.pairs))
results <- vector(length = ncol(country.pairs))
for (i in seq(1,ncol(country.pairs))) {
  p1 <- wine.country[[country1[i]]]$points
  p2 <- wine.country[[country2[i]]]$points
  mean.boot.p1p2 = two.boot(p1, p2, mean, R = 10000)
  ci <- quantile(mean.boot.p1p2$t, probs = c(p/2, 1-p/2))
  if (0 > ci[1] && 0 < ci[2]) {
    result <- 'NA'
  } else if (ci[1] > 0 && ci[2] > 0) {
    result <- country1[i]
  } else {
    result <- country2[i]
  }
  ci.lower[i] <- ci[1]
  ci.upper[i] <- ci[2]
  results[i] <- result
}
country.points.ranking <- data.frame(country1 = country1, country2 = country2,
  ciL = ci.lower, ciU = ci.upper, result = results)
```

```
country.points.ranking
```

```
##       country1     country2        ciL          ciU       result
## 1           US        Italy -0.6898858 -0.564443994        Italy
## 2           US       France -0.8615594 -0.709142081       France
## 3           US        Spain  1.0814123  1.274425017           US
## 4           US        Chile  1.4234008  1.616973783           US
## 5           US    Argentina  1.7019033  1.927795280           US
## 6           US    Australia -0.1771581  0.050317913         <NA>
## 7           US     Portugal -0.4807559 -0.237445414     Portugal
## 8           US  New Zealand  0.1530841  0.393234017           US
## 9           US      Austria -1.5084472 -1.239028768      Austria
## 10          US      Germany -0.9491309 -0.633780190      Germany
## 11          US South Africa  0.4736766  0.742481652           US
## 12      Italy       France -0.2409121 -0.075496852       France
## 13      Italy        Spain  1.7013089  1.906102927        Italy
## 14      Italy        Chile  2.0462927  2.255747518        Italy
## 15      Italy    Argentina  2.3266176  2.560911893        Italy
## 16      Italy    Australia  0.4418257  0.681336358        Italy
## 17      Italy     Portugal  0.1414168  0.398878191        Italy
## 18      Italy  New Zealand  0.7795032  1.025796760        Italy
## 19      Italy      Austria -0.8818239 -0.603079413      Austria
## 20      Italy      Germany -0.3318254 -0.004877526      Germany
## 21      Italy South Africa  1.0922705  1.372381419        Italy
## 22     France        Spain  1.8535794  2.078702004       France
```

```
## 23      France        Chile  2.1955415  2.420319011       France
## 24      France    Argentina  2.4793521  2.730331809       France
## 25      France    Australia  0.5926963  0.851666101       France
## 26      France     Portugal  0.2955535  0.559685023       France
## 27      France  New Zealand  0.9291193  1.192755765       France
## 28      France      Austria -0.7282306 -0.442754490      Austria
## 29      France      Germany -0.1700585  0.157815158         <NA>
## 30      France South Africa  1.2502868  1.534716442       France
## 31       Spain        Chile  0.2223397  0.473086078        Spain
## 32       Spain    Argentina  0.5010915  0.775888260        Spain
## 33       Spain    Australia -1.3878445 -1.104020693    Australia
## 34       Spain     Portugal -1.6882004 -1.392884015     Portugal
## 35       Spain  New Zealand -1.0454813 -0.762219387  New Zealand
## 36       Spain      Austria -2.7086408 -2.394734715      Austria
## 37       Spain      Germany -2.1515586 -1.791191624      Germany
## 38       Spain South Africa -0.7229063 -0.416881346 South Africa
## 39       Chile    Argentina  0.1510611  0.430750214        Chile
## 40       Chile    Australia -1.7284159 -1.447443505    Australia
## 41       Chile     Portugal -2.0271496 -1.733382870     Portugal
## 42       Chile  New Zealand -1.3915146 -1.105046363  New Zealand
## 43       Chile      Austria -3.0498725 -2.732650556      Austria
## 44       Chile      Germany -2.4969279 -2.133603167      Germany
## 45       Chile South Africa -1.0761942 -0.758123896 South Africa
## 46   Argentina    Australia -2.0307680 -1.732341204    Australia
## 47   Argentina     Portugal -2.3301617 -2.021329869     Portugal
## 48   Argentina  New Zealand -1.6971816 -1.393735692  New Zealand
## 49   Argentina      Austria -3.3571415 -3.022428150      Austria
## 50   Argentina      Germany -2.7985445 -2.418501187      Germany
## 51   Argentina South Africa -1.3791751 -1.040191744 South Africa
## 52   Australia     Portugal -0.4558028 -0.134266753     Portugal
## 53   Australia  New Zealand  0.1823277  0.496958103    Australia
## 54   Australia      Austria -1.4747558 -1.138152422      Austria
## 55   Australia      Germany -0.9294690 -0.541570615      Germany
## 56   Australia South Africa  0.5044135  0.835309871    Australia
## 57    Portugal  New Zealand  0.4739230  0.789660721     Portugal
## 58    Portugal      Austria -1.1856267 -0.831821056      Austria
## 59    Portugal      Germany -0.6275222 -0.238584453      Germany
## 60    Portugal South Africa  0.7886251  1.142434357     Portugal
## 61 New Zealand      Austria -1.8155642 -1.474762349      Austria
## 62 New Zealand      Germany -1.2600722 -0.879938390      Germany
## 63 New Zealand South Africa  0.1629521  0.507111207  New Zealand
## 64     Austria      Germany  0.3771504  0.777944302      Austria
## 65     Austria South Africa  1.7940976  2.161584098      Austria
## 66     Germany South Africa  1.1981829  1.604860847      Germany
```

Note that in nearly every case, the difference in point means was significant enough to declare a winner. As with ANOVA, in a few cases the 99% confidence interval includes zero, so we can't say that a statistically significant difference exists between points in these cases:

- US versus Australia
- Germany versus France

Note that these pairs were also too close to call using ANOVA, however the pair of Italy versus Germany was not included. The bootstrapped confidence interval just excludes zero ($[-0.33, -0.004]$), and with ANOVA it just included zero ($[-0.42, 0.09]$). Perhaps the slight violations of the assumptions required for ANOVA led

to the difference. Recall that the distribution of points for each country was nearly normal, which is why the results mostly agree between the two methods. Let's tally up the winners and have a look at the overall results.

```
points.ranked <- as.data.frame(country.points.ranking %>% group_by(result) %>%
  summarise(wins = length(result)) %>% arrange(-wins))
points.ranked
```

```
##          result wins
## 1       Austria   11
## 2        France    9
## 3       Germany    9
## 4         Italy    8
## 5      Portugal    7
## 6     Australia    5
## 7            US    5
## 8   New Zealand    4
## 9  South Africa    3
## 10        Spain    2
## 11         <NA>    2
## 12        Chile    1
```

Austria is the clear winner in terms of average points awarded, as it beat all 11 other countries with a statistically significant difference. France and Germany are tied for second place, as they beat 9 out of their 11 competitors. Note that Argentina is not included, as it did not beat any of the other countries.

Bootstrap Comparison of Wine Prices

Next, we'll use the same bootstrap approach on the price of wines. The code used is identical, except now the difference in means for price is being computed, rather than points. The code is omitted for brevity.

```
country.price.ranking
```

```
##       country1     country2         ciL         ciU       result
## 1           US        Italy  -4.6737952  -3.1812425        Italy
## 2           US       France -13.5203840 -10.5674575       France
## 3           US        Spain   5.5771010   7.5110615           US
## 4           US        Chile  13.5329341  14.9824932           US
## 5           US    Argentina  12.0817674  13.5369547           US
## 6           US    Australia   0.8850967   3.7615499           US
## 7           US     Portugal   5.7699523   8.5840570           US
## 8           US  New Zealand   8.7614469  10.1257855           US
## 9           US      Austria   1.8232033   3.8396012           US
## 10          US      Germany  -8.5723508  -2.5637818      Germany
## 11          US South Africa  11.6786359  13.2617875           US
## 12       Italy       France  -9.7772516  -6.4507322       France
## 13       Italy        Spain   9.2629402  11.7056783        Italy
## 14       Italy        Chile  17.2545016  19.1733209        Italy
## 15       Italy    Argentina  15.7869874  17.7519496        Italy
## 16       Italy    Australia   4.6198185   7.8240961        Italy
## 17       Italy     Portugal   9.5807263  12.7416866        Italy
## 18       Italy  New Zealand  12.4347353  14.3312924        Italy
## 19       Italy      Austria   5.5912511   7.9937062        Italy
## 20       Italy      Germany  -4.7500699   1.5313586         <NA>
## 21       Italy South Africa  15.3655166  17.4671749        Italy
```

```
## 22      France        Spain  16.8044643  20.3826341       France
## 23      France        Chile  24.7099682  27.9412776       France
## 24      France    Argentina  23.2145678  26.4873176       France
## 25      France    Australia  12.3598571  16.3790042       France
## 26      France     Portugal  17.2486531  21.3471009       France
## 27      France  New Zealand  19.8848804  23.1215641       France
## 28      France      Austria  13.1077133  16.6453447       France
## 29      France      Germany   3.1863284   9.8584172       France
## 30      France South Africa  22.8566826  26.1397943       France
## 31       Spain        Chile   6.5586665   8.8901816        Spain
## 32       Spain    Argentina   5.0868044   7.4248819        Spain
## 33       Spain    Australia  -5.9994621  -2.5468574    Australia
## 34       Spain     Portugal  -1.0280683   2.3469786         <NA>
## 35       Spain  New Zealand   1.7478456   4.0276512        Spain
## 36       Spain      Austria  -5.0787389  -2.2939771      Austria
## 37       Spain      Germany -15.2489645  -8.9648331      Germany
## 38       Spain South Africa   4.6739849   7.1815359        Spain
## 39       Chile    Argentina  -2.4097353  -0.4695669    Argentina
## 40       Chile    Australia -13.5474138 -10.3916621    Australia
## 41       Chile     Portugal  -8.6005632  -5.4989504     Portugal
## 42       Chile  New Zealand  -5.7663331  -3.9257300  New Zealand
## 43       Chile      Austria -12.6270088 -10.2466663      Austria
## 44       Chile      Germany -22.8669178 -16.7619326      Germany
## 45       Chile South Africa  -2.8041946  -0.7828106 South Africa
## 46   Argentina    Australia -12.1262067  -8.9426010    Australia
## 47   Argentina     Portugal  -7.1130592  -4.0683005     Portugal
## 48   Argentina  New Zealand  -4.2999458  -2.4432559  New Zealand
## 49   Argentina      Austria -11.1641951  -8.7976098      Austria
## 50   Argentina      Germany -21.4951719 -15.2338367      Germany
## 51   Argentina South Africa  -1.3353706   0.6911458         <NA>
## 52   Australia     Portugal   2.9634359   6.8633064    Australia
## 53   Australia  New Zealand   5.5952066   8.7041215    Australia
## 54   Australia      Austria  -1.2135389   2.2612768         <NA>
## 55   Australia      Germany -11.2826246  -4.4676951      Germany
## 56   Australia South Africa   8.5056147  11.8124017    Australia
## 57    Portugal  New Zealand   0.6807509   3.7431918     Portugal
## 58    Portugal      Austria  -6.0937487  -2.6597112      Austria
## 59    Portugal      Germany -16.0947413  -9.3713385      Germany
## 60    Portugal South Africa   3.6622162   6.8336031     Portugal
## 61 New Zealand      Austria  -7.7566829  -5.4286434      Austria
## 62 New Zealand      Germany -18.1516889 -11.9873901      Germany
## 63 New Zealand South Africa   2.0673288   4.0420808  New Zealand
## 64     Austria      Germany -11.5247589  -5.2196166      Germany
## 65     Austria South Africa   8.3595997  10.8731914      Austria
## 66     Germany South Africa  14.9719578  21.1100924      Germany
```

Again, in nearly every case the difference in price means was significant enough to declare a winner. As with ANOVA, in a few cases the 99% confidence interval includes zero, so we can't say that a statistically significant difference exists between price in these cases:

- Italy versus Germany
- Spain versus Portugal
- Argentina versus South Africa
- Austria versus Australia

Note that the first two pairs were also too close to call using ANOVA, however the rest of the pairs listed in the ANOVA section are missing. In their place, two new pairs were included. Recall that the distribution of price for each country was still fairly skewed, even after a log transformation was applied. This may explain the larger difference in bootstrapped means of price when compared to points, where the results mostly agreed between the two methods. As before, let's tally up the winners and have a look at the overall results.

```
price.ranked <- as.data.frame(country.price.ranking %>% group_by(result) %>%
  summarise(wins = length(result)) %>% arrange(-wins))
price.ranked
```

```
##          result wins
## 1        France   11
## 2       Germany    9
## 3         Italy    9
## 4            US    8
## 5     Australia    6
## 6       Austria    6
## 7      Portugal    4
## 8         Spain    4
## 9          <NA>    4
## 10  New Zealand    3
## 11    Argentina    1
## 12 South Africa    1
```

France is the clear winner (or loser, if you're looking for a less expensive bottle) in terms of price, as it beat all 11 other countries with a statistically significant difference. Germany and Italy are tied for second place, as they beat 9 out of their 11 competitors. Note that Chile is not included, as it makes the least expensive wines and did not beat any of the other countries.

Summary and Conclusion

Our goal was to help reduce the anxiety one often feels when trying to pick out a bottle of wine given the numerous choices available. To do this, we applied statistical methods to over 150,000 wine reviews in an attempt to answer the following questions:

1. Can a country claim to make the best wine in the world in terms of average number of points awarded?
2. How do different countries compare in terms of price and quality? If you're looking for a great wine at a reasonable price, which countries are your best bet?

Recall that our null hypothesis (H_0) was that there is no difference in quality and price when it comes to wines produced by different countries, and our alternative hypothesis (H_A) was that it does matter from which country you buy your wine, there is in fact a significant difference.

Using both classical (ANOVA and Tukey's HSD) and bootstrap methods, we found that there was a statistically significant difference between wines produced by different countries in terms of both points and price. The results mostly agreed between the two methods, with a few differences. The distribution of prices for each country were still skewed, even after applying a log transformation, so the results from the classical methods may not have been reliable. The bootstrap approach provided better, more reliable results, as there are no assumptions about the normality or variance of the data.

After running our tests, we found that Austria makes the best wines out of the countries studied according to the average number of points awarded. After that, France and Germany are tied for second place. In terms of price, France makes by far the most expensive wines in terms of average price for a bottle. After that, Germany and Italy are tied for second place. Austria, which makes the best wines in terms of points, is in the middle of the pack when it comes to price.

Given the outcomes above, your best best for a bottle of wine in terms of price and quality is Austria. The only catch is that Austria produces mostly white wines. If you're looking for a good red wine, French wines are great in terms of quality, but you're going to pay a premium in terms of price. As an alternative, consider Italy, which has comparable quality but a lower average price.

With these results, it is important to consider possible sources of bias or error:

1. The year each wine was bottled was not included in the data. Older wines tend to increase in price due to the rarity of the wine and the price of storing it for a number of years. Certain countries with a longer history of production (such as those in Europe) may have a disproportionate number of older wines, which could skew the average price higher when compared to a younger country with more new wines.
2. The US had three times as many wines as the next largest country, and almost 30 times as many as the smallest country. This may have driven the average points awarded down, as the fact that so many more wines were available could lead to a dilution in terms of quality. A country with fewer wines may have only made their 'best' wines available for review, and smaller wineries that didn't produce as many quality wines may have been omitted. This would tend to drive the quality of wines for smaller countries upward.