Analysis of Wine Tasting

Jillien Chu

Northeastern University

Author Note

Jillien Chu, Department of Analytics, Northeastern University.

Correspondence concerning this article should be addressed to Jillien Chu, Department of Analytics, Northeastern University, Boston, MA, 02115. Contact: chu.jil@northeastern.edu

Table of Contents

Introduction

The final project is associated with the course ALY 6010: Probability Theory and Introductory Statistics. Various findings have been found from the wine tasting dataset, provided by the instructor Prof. Tom Breur, with focus on the country, points, price, province, title, variety, winery and direction. This final project report is the study of the relationships between different variables like points, price, direction and consists of descriptive, exploratory data analysis and the interpretations of hypothesis testing which has been conducted on the wine tasting dataset. The aim of the project is to analyze if wine prices are based on the wine points, to identify if the price of east world wines and west world wines are equal and to identify if the rating of the east world wines and west world wines are equal.

Variables Of The Dataset

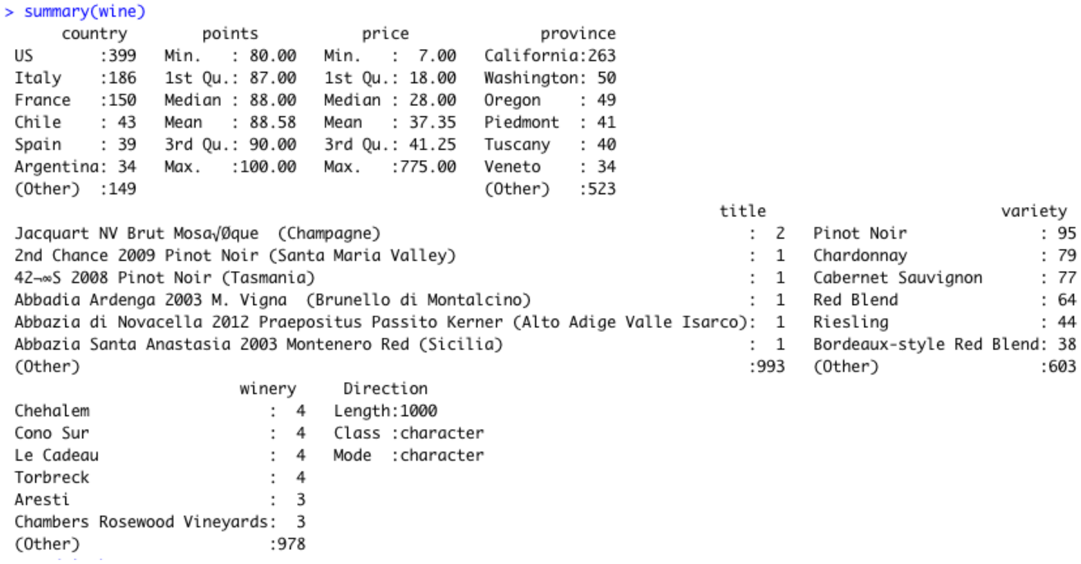
The wine tasting dataset had a total 1000 records with 13 fields but 1000 records with 8 fields were used for this analysis. The variables of interest for this dataset are:

- Country – Wine produced in the country
- Points – Wine score or rating of the wines
- Price – Price of the wine
- Province – Province where the wine is being produced
- Title – Name of the wine
- Variety – It displays the kind of wine
- Winery – The wine company or winery that produces the wine
- Direction – The wineries located in which side of the world i.e., East World or West World

Summary of the Variables

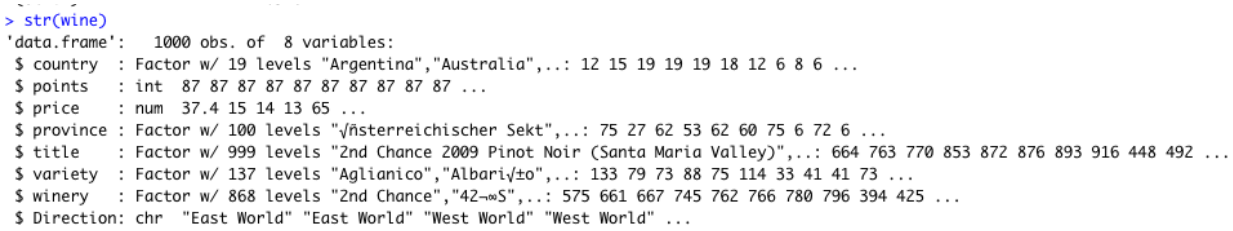
The wine tasting dataset for the project was provided by the instructor. The dataset was divided into two categories i.e., East World and West World where West World consisted of USA, Mexico, Chile, Argentina and Canada and the East world consisted of the rest of the countries. Using these quantitative variables, I conducted the hypothesis test to find out a few questions for this dataset. The following are the descriptive characteristics of the dataset:

Table 1: Summary of the dataset



```
> summary(wine)
      country        points           price             province
 US       :399   Min.   : 80.00   Min.   :  7.00   California:263
 Italy    :186   1st Qu.: 87.00   1st Qu.: 18.00   Washington: 50
 France   :150   Median : 88.00   Median : 28.00   Oregon    : 49
 Chile    : 43   Mean   : 88.58   Mean   : 37.35   Piedmont  : 41
 Spain    : 39   3rd Qu.: 90.00   3rd Qu.: 41.25   Tuscany   : 40
 Argentina: 34   Max.   :100.00   Max.   :775.00   Veneto    : 34
 (Other)  :149                                     (Other)   :523
                                                                  title                          variety
 Jacquart NV Brut Mosa√Øque  (Champagne)                              :  2   Pinot Noir              : 95
 2nd Chance 2009 Pinot Noir (Santa Maria Valley)                      :  1   Chardonnay              : 79
 42¬∞S 2008 Pinot Noir (Tasmania)                                     :  1   Cabernet Sauvignon      : 77
 Abbadia Ardenga 2003 M. Vigna  (Brunello di Montalcino)              :  1   Red Blend               : 64
 Abbazia di Novacella 2012 Praepositus Passito Kerner (Alto Adige Valle Isarco):  1   Riesling                : 44
 Abbazia Santa Anastasia 2003 Montenero Red (Sicilia)                 :  1   Bordeaux-style Red Blend: 38
 (Other)                                                              :993   (Other)                 :603
                        winery      Direction
 Chehalem                  :  4   Length:1000
 Cono Sur                  :  4   Class :character
 Le Cadeau                 :  4   Mode  :character
 Torbreck                  :  4
 Aresti                    :  3
 Chambers Rosewood Vineyards:  3
 (Other)                   :978
```

Table 2: Structure of the dataset



```
> str(wine)
'data.frame':	1000 obs. of  8 variables:
 $ country  : Factor w/ 19 levels "Argentina","Australia",..: 12 15 19 19 19 18 12 6 8 6 ...
 $ points   : int  87 87 87 87 87 87 87 87 87 87 ...
 $ price    : num  37.4 15 14 13 65 ...
 $ province : Factor w/ 100 levels "√ñsterreichischer Sekt",..: 75 27 62 53 62 60 75 6 72 6 ...
 $ title    : Factor w/ 999 levels "2nd Chance 2009 Pinot Noir (Santa Maria Valley)",..: 664 763 770 853 872 876 893 916 448 492 ...
 $ variety  : Factor w/ 137 levels "Aglianico","Albari√±o",..: 133 79 73 88 75 114 33 41 41 73 ...
 $ winery   : Factor w/ 868 levels "2nd Chance","42¬∞S",..: 575 661 667 745 762 766 780 796 394 425 ...
 $ Direction: chr  "East World" "East World" "West World" "West World" ...
```

From the outputs above, we infer that the analysis is conducted for 1000 observations of 8 fields in total for this dataset. The mean price of the wine is 37.35, the mean rating is 88.58 and the greatest number of wines were produced in US followed by Italy. Also, there are multiple kinds of wines, but Pinot Noir is produced the most followed by Chardonnay.

Questions exploration:

1. Does points have a relation with wine pricing?
2. Does direction have a relation with wine pricing?
3. Does direction have a relation with wine rating?
4. Does the mean price differ for wines produced in east world countries and the west world countries?
5. Does the mean rating differ for wines produced in east world countries and the west world countries?
6. Are the mean average points of wine equal to 85?

Since the dataset is regarding wines, initially the east world countries i.e., France, Italy, Spain etc. were considered the top producing wine regions until the west world countries i.e., United States of America, Chile, etc. began to grow and get popular rapidly as they started producing world class wines that were of similar level as the other famous east world country wines. As such, I wanted to analyze if the price and ratings of the east world wines were equal to the west world wines and if the rating of wine plays a role in the pricing of wines.

To analyze and explore the questions, I conducted the following hypothesis tests:

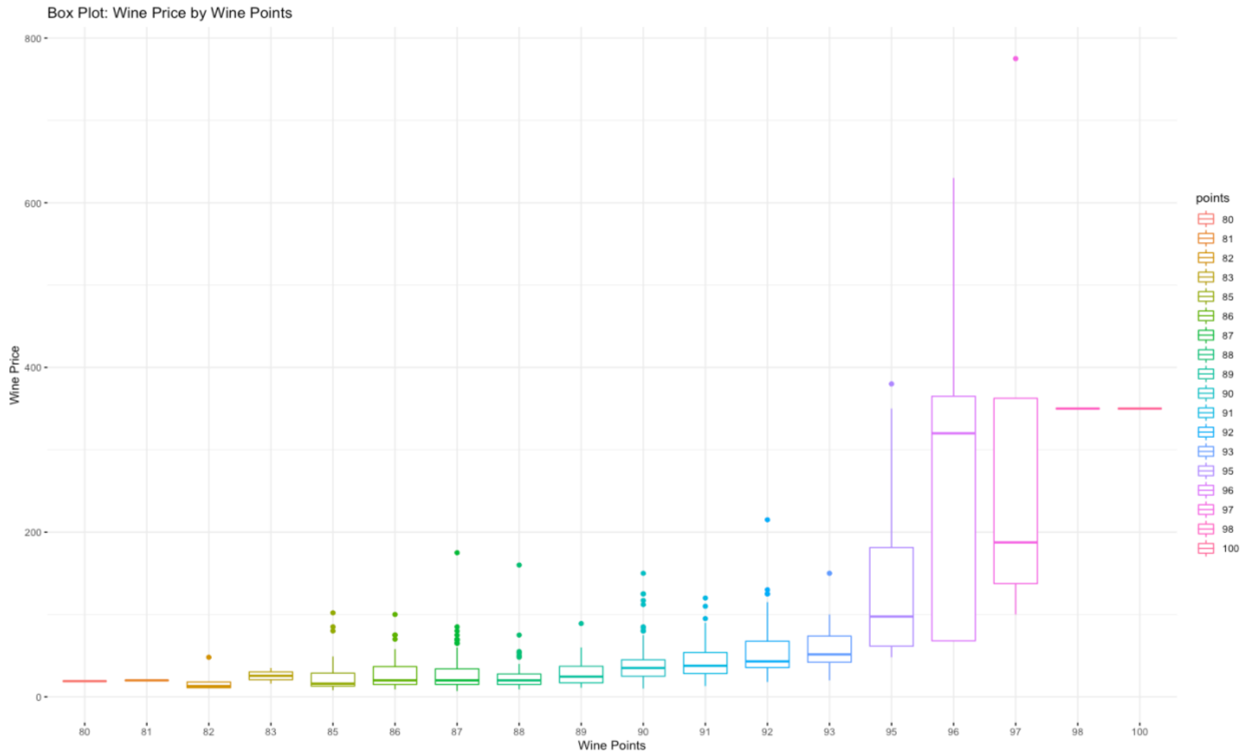
1. Two sample test for analyzing the relation between points and wine prices

Hypothesis Test:

Null Hypothesis: There is no significant linear correlation between price and points in the population

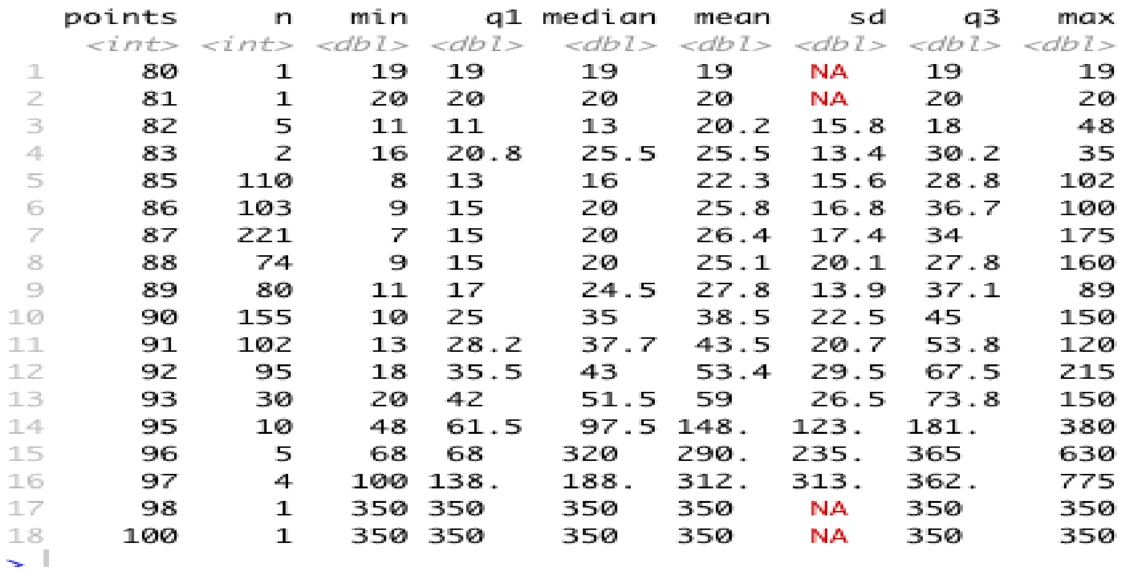
Alternative Hypothesis: There is significant linear correlation between price and points in the population

Plot 1: Box Plot for comparing the price of wines with the wine rating



This boxplot compares the price of wines with the wine points. From the plot, we observe that as the wine points increase, the wine prices increase as well.

Table 3: Summary of points and price of wines



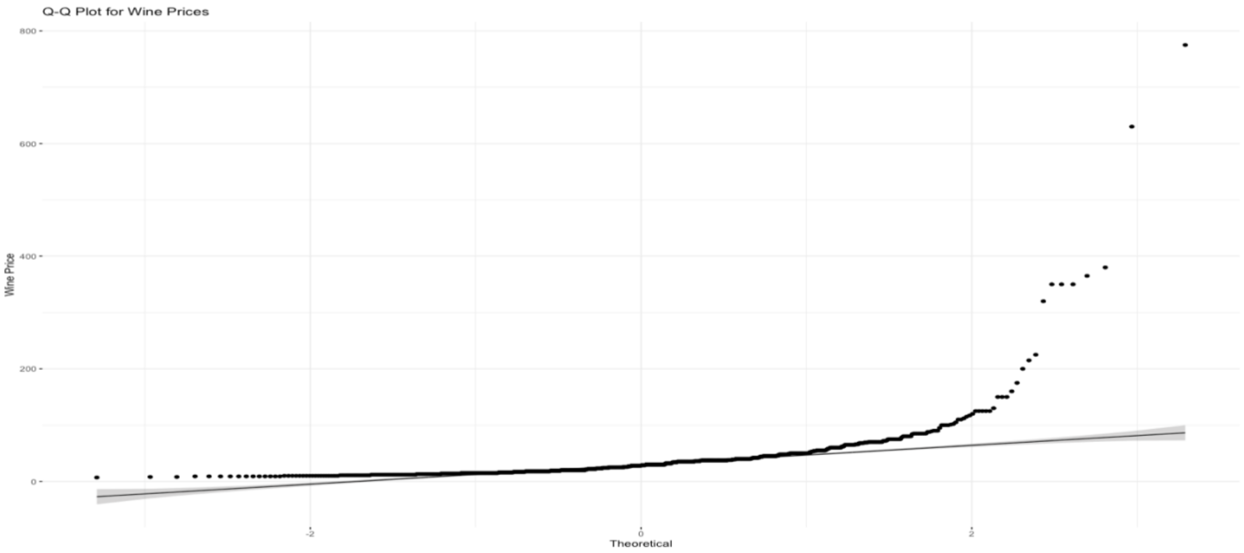
```
   points     n   min    q1 median  mean    sd    q3   max
    <int> <int> <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl>
 1     80     1    19  19     19    19    NA    19      19
 2     81     1    20  20     20    20    NA    20      20
 3     82     5    11  11     13    20.2  15.8  18      48
 4     83     2    16  20.8   25.5  25.5  13.4  30.2    35
 5     85   110     8  13     16    22.3  15.6  28.8   102
 6     86   103     9  15     20    25.8  16.8  36.7   100
 7     87   221     7  15     20    26.4  17.4  34     175
 8     88    74     9  15     20    25.1  20.1  27.8   160
 9     89    80    11  17     24.5  27.8  13.9  37.1    89
10     90   155    10  25     35    38.5  22.5  45     150
11     91   102    13  28.2   37.7  43.5  20.7  53.8   120
12     92    95    18  35.5   43    53.4  29.5  67.5   215
13     93    30    20  42     51.5  59    26.5  73.8   150
14     95    10    48  61.5   97.5 148.  123.  181.    380
15     96     5    68  68    320   290.  235.  365     630
16     97     4   100 138.   188.  312.  313.  362.    775
17     98     1   350 350    350   350    NA   350     350
18    100     1   350 350    350   350    NA   350     350
>
```

From the table above, we observe that as the wine points increase, the mean price of wines mostly increases as well. To find out if points and price have a relation, I conducted the following tests which can be found below.

Preliminary Test:

I conducted the preliminary test to check if the data met the conditions for test assumptions. The entire data was taken for the two variables i.e., for price and points. To find out the relation between the two variables, I checked for the data distribution i.e., if it is normally distributed or not, conducted the Shapiro Wilk test and correlation.

Plot 2: Q-Q Plot to check for normal distribution

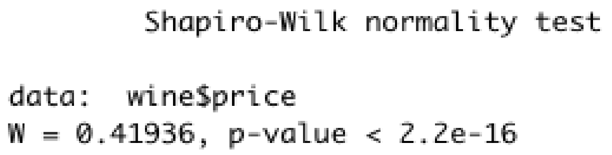


The Q-Q Plot was used for a visual check for data normality. From the plot, it can be inferred that the data is mostly not normally distributed. To be sure, I conducted the Shapiro-Wilk test.

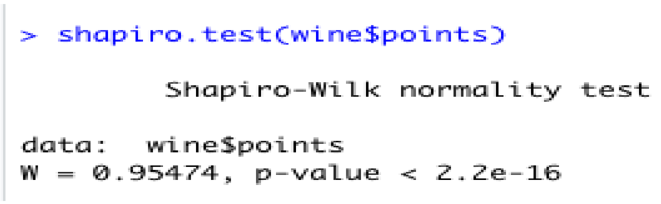
Shapiro-Wilk Test:

Null Hypothesis: The data is normally distributed

Alternative Hypothesis: The data is not normally distributed



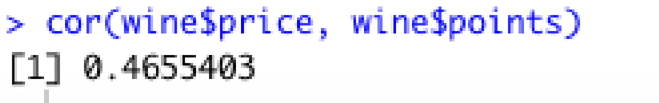
```
	Shapiro-Wilk normality test

data:  wine$price
W = 0.41936, p-value < 2.2e-16
```



```
> shapiro.test(wine$points)

	Shapiro-Wilk normality test

data:  wine$points
W = 0.95474, p-value < 2.2e-16
```

From this, we observe that the two p-values are lesser than the significance level of 0.05 which implies that the distribution of the data is different from the normal distribution. Thus, it is not normally distributed.

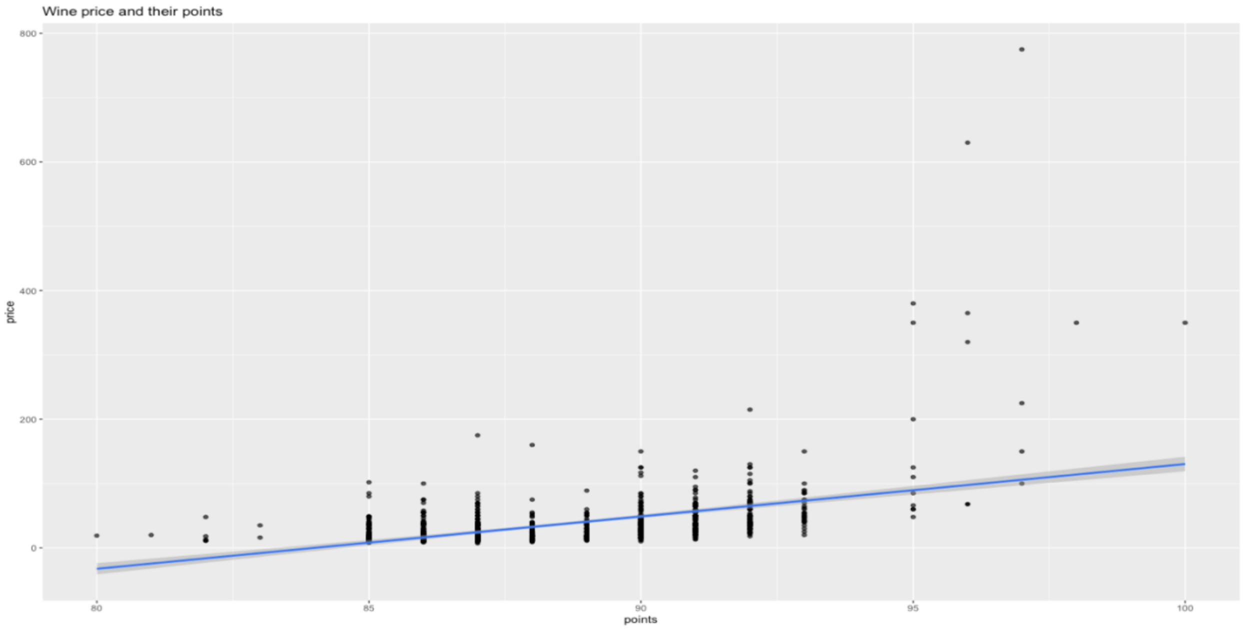
Correlation:



```
> cor(wine$price, wine$points)
[1] 0.4655403
```

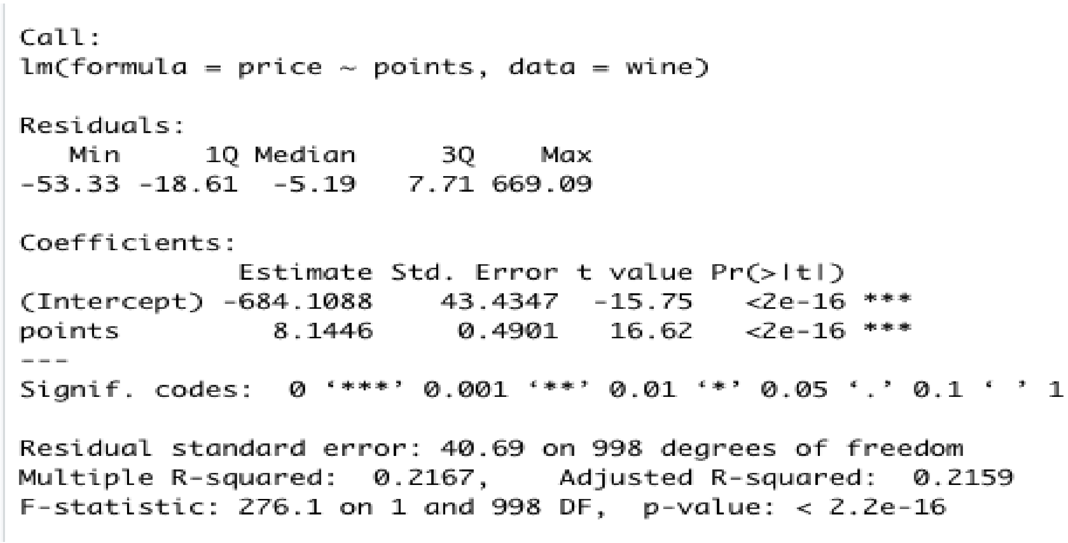
From this correlation, we infer that there is a moderate positive linear relationship between price and points of wine.

Plot 3: Scatterplot for wine points and wine price



This plot represents the wine price and the wine points. From this, we observe that price and points have a moderate, positive linear relationship. Thus, it can be inferred that when the wine points increase, the wine price increases too. Also, there is greater number of wines with the wine points ranging between 85 and 93.

Regression:



```
Call:
lm(formula = price ~ points, data = wine)

Residuals:
   Min     1Q Median     3Q    Max 
-53.33 -18.61  -5.19   7.71 669.09 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -684.1088    43.4347  -15.75   <2e-16 ***
points         8.1446     0.4901   16.62   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 40.69 on 998 degrees of freedom
Multiple R-squared:  0.2167,	Adjusted R-squared:  0.2159 
F-statistic: 276.1 on 1 and 998 DF,  p-value: < 2.2e-16
```

From this, we find that the p-value (2.2e-16) is smaller than the significance level of 0.05, thus we reject the null hypothesis and can conclude that there is a linear correlation between points and price in the population.

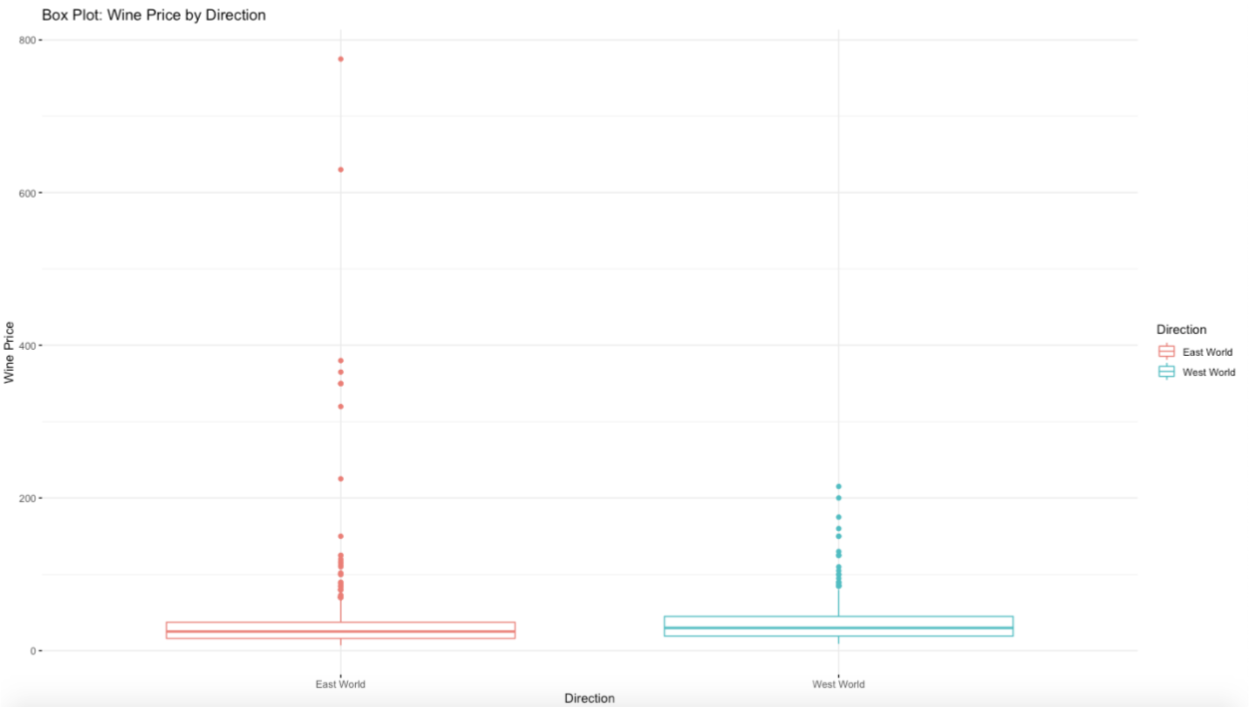
2. Two sample test for analyzing the relation between direction and wine prices

Hypothesis Test:

Null Hypothesis: There is no significant linear correlation between price and direction in the population

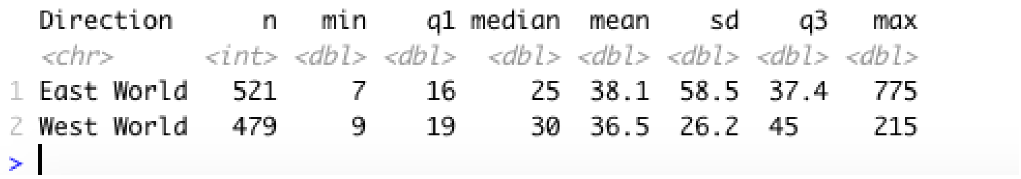
Alternative Hypothesis: There is significant linear correlation between price and direction in the population

Plot 4: Box Plot for comparing the price of wines with the direction (east world or west world)


Box Plot: Wine Price by Direction

This boxplot represents the wine prices with the direction where the wine is produced i.e., east world wineries and west world wineries. From this plot, we can observe that there is a slight difference in the two averages.

Table 4: Summary of price of wines and the direction (east world and west world)



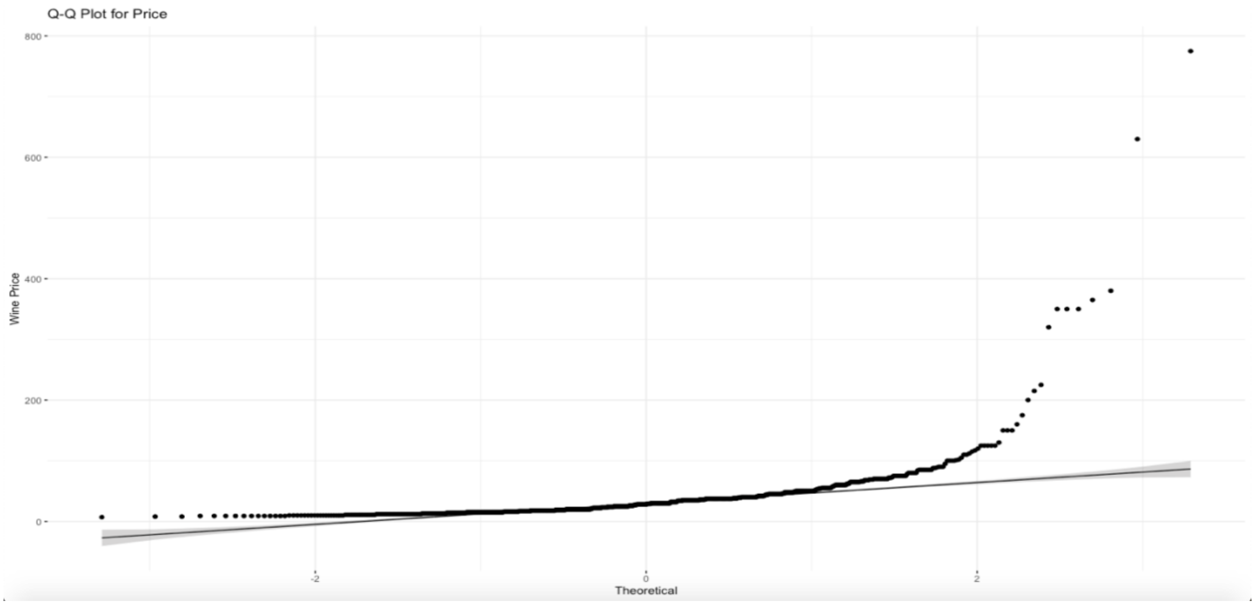
```
  Direction       n   min    q1 median  mean    sd    q3   max
  <chr>       <int> <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl>
1 East World    521     7    16     25  38.1  58.5  37.4   775
2 West World    479     9    19     30  36.5  26.2  45     215
> 
```

From the table above, we can observe that the mean price of east and west world wines varies. The mean of the east world wines is 38.1 whereas the mean of the west world wines is 36.5. From this, we can infer that the east world wines are slightly more expensive than the wines produced by the west world countries.

Preliminary Test:

I conducted the preliminary test to check if the data met the conditions for test assumptions. The entire data was taken for the two variables i.e., for price and direction (east world and west world). To find out the relation between the two variables, I checked for the data distribution i.e., if it is normally distributed or not, conducted the Shapiro Wilk test and its correlation.

Plot 5: Q-Q Plot to check for normal distribution

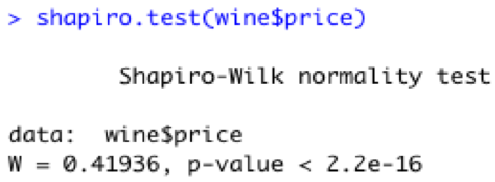


The Q-Q Plot was used for a visual check for data normality. From the plot, it can be inferred that the data is mostly not normally distributed. To be sure, I conducted the Shapiro-Wilk test.

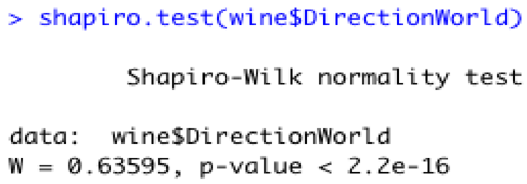
Shapiro-Wilk Test:

Null Hypothesis: The data is normally distributed

Alternative Hypothesis: The data is not normally distributed



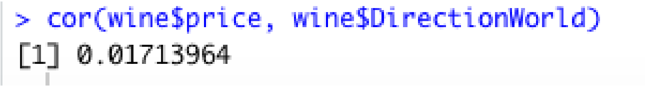
```
> shapiro.test(wine$price)

        Shapiro-Wilk normality test

data:  wine$price
W = 0.41936, p-value < 2.2e-16
```



```
> shapiro.test(wine$DirectionWorld)

        Shapiro-Wilk normality test

data:  wine$DirectionWorld
W = 0.63595, p-value < 2.2e-16
```

From this, we observe that the two p-values are lesser than the significance level of 0.05 which implies that the distribution of the data is different from the normal distribution.

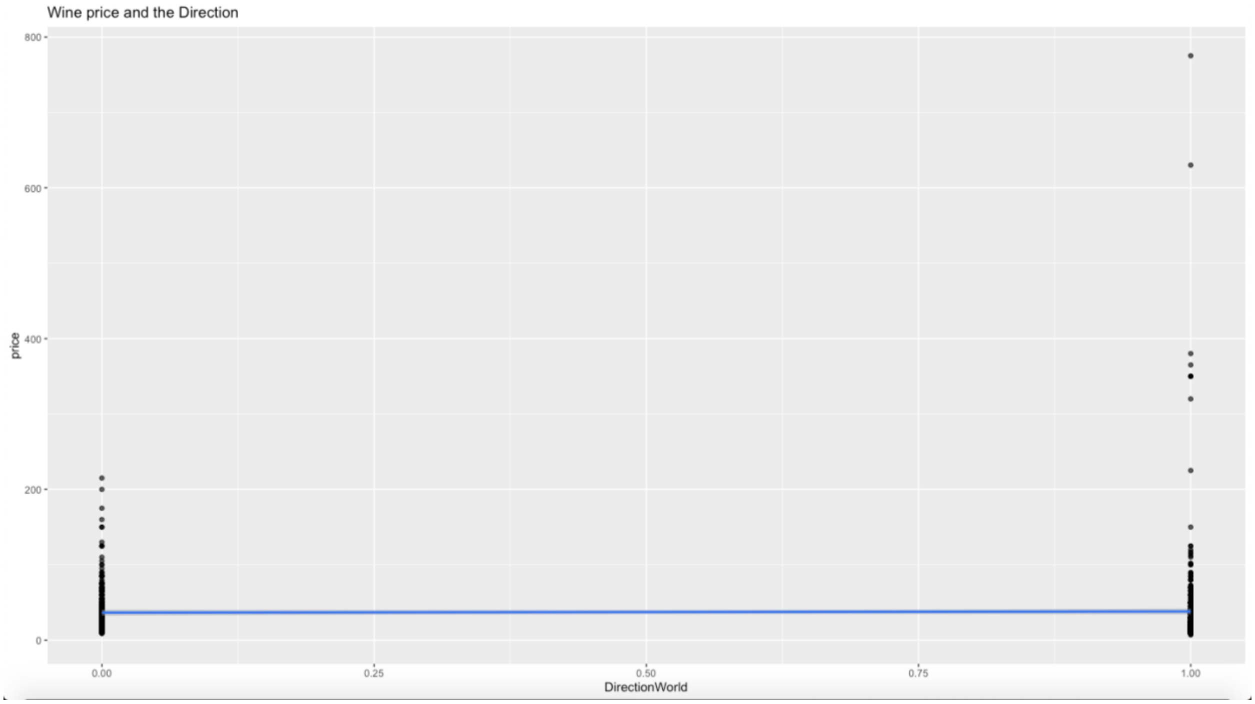
Correlation:



```
> cor(wine$price, wine$DirectionWorld)
[1] 0.01713964
```

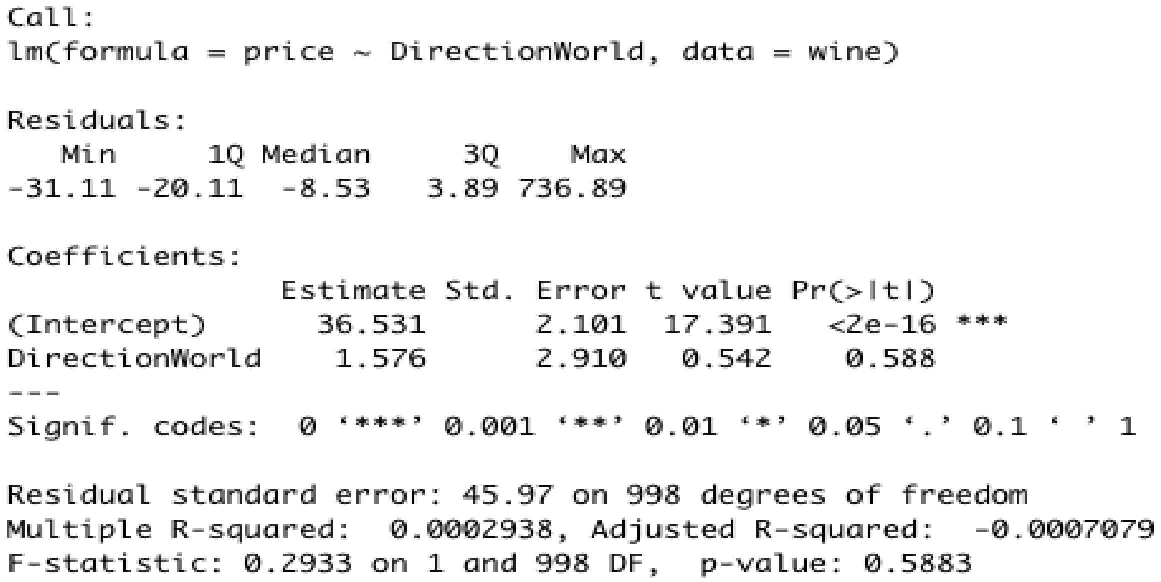
From this correlation, we infer that there is a negligible correlation between price of wine and the direction (east wine and west wine).

Plot 6: Scatterplot for wine points and wine price



This plot represents the wine price and the direction of the winery (east world or west world). From this, we observe that the wine price and direction have a negligible correlation relationship.

Regression:



```
Call:
lm(formula = price ~ DirectionWorld, data = wine)

Residuals:
   Min     1Q Median     3Q    Max 
-31.11 -20.11  -8.53   3.89 736.89 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)      36.531      2.101  17.391   <2e-16 ***
DirectionWorld    1.576      2.910   0.542    0.588    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 45.97 on 998 degrees of freedom
Multiple R-squared:  0.0002938,	Adjusted R-squared:  -0.0007079 
F-statistic: 0.2933 on 1 and 998 DF,  p-value: 0.5883
```

From this, we find that the p-value (0.5883) is greater than the significance level of 0.05, thus we fail to reject the null hypothesis and can conclude that there is not enough evidence to prove that there is a linear correlation between direction and price in the population.

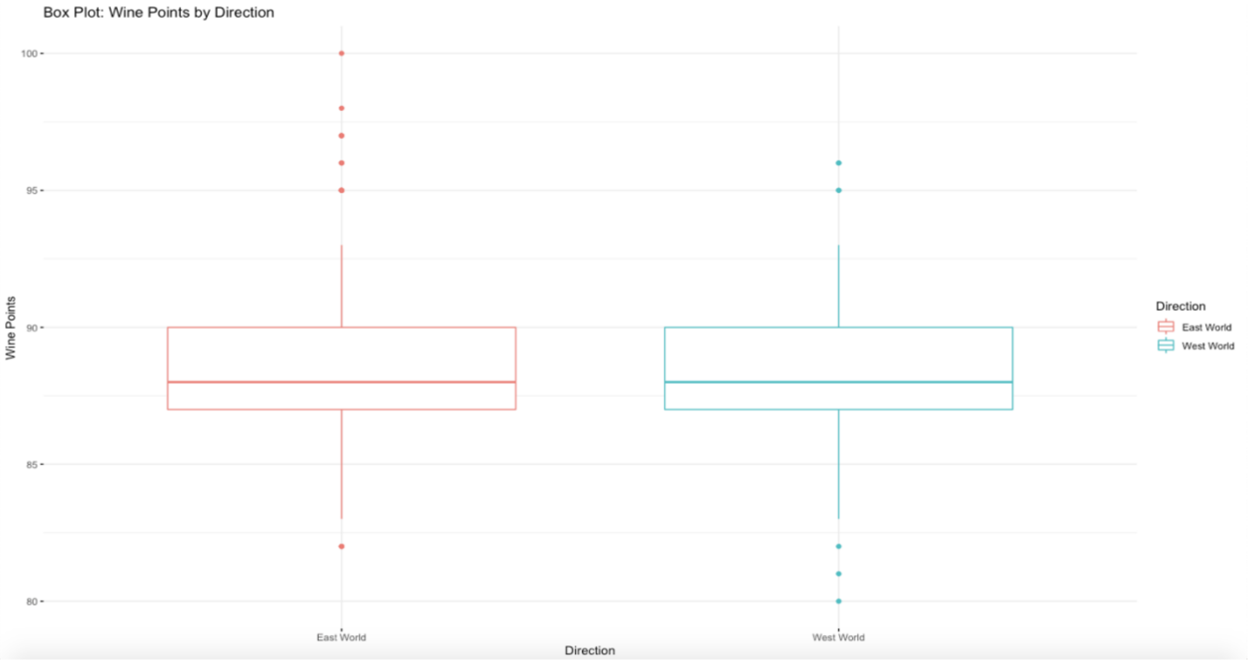
3. Two sample test for analyzing if there is a relation between direction and wine points

Hypothesis Test:

Null Hypothesis: There is no significant linear correlation between points and direction in the population

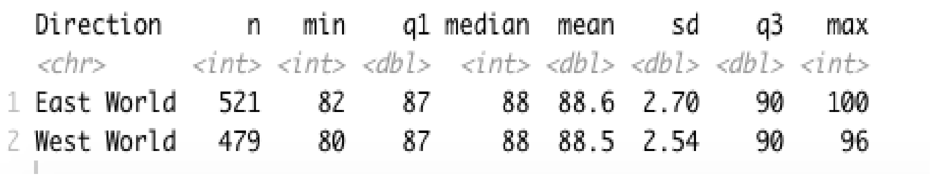
Alternative Hypothesis: There is significant linear correlation between points and direction in the population

Plot7: Box Plot for comparing the price of wines with the direction (east world or west world)



This boxplot represents the wine points or rating with the direction where the wine is produced i.e., east world wineries and west world wineries. From this plot, we can observe that there is not much difference in the two averages.

Table 5: Summary of points of wines and the direction (east world and west world)

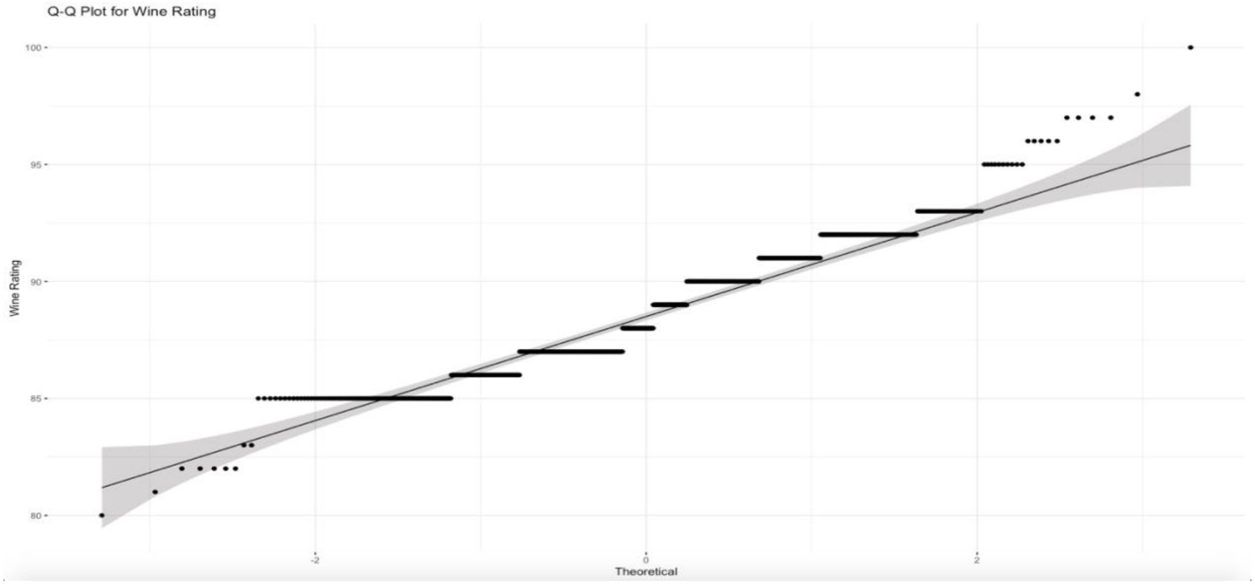


```
  Direction       n   min    q1 median  mean    sd    q3   max
  <chr>       <int> <int> <dbl>  <int> <dbl> <dbl> <dbl> <int>
1 East World    521    82    87     88  88.6  2.70    90   100
2 West World    479    80    87     88  88.5  2.54    90    96
```

The table consists of 1000 observations. From this, we can observe that the East World were 521 in total with minimum rating of 82 and maximum rating of 100 and a mean of 88.6 whereas the

West World consisted of 479 observations in total with minimum rating of 80, maximum rating of 96 and a mean of 88.5. Comparing the two, we can say that the minimum rating of west world wine is lesser than the east world rating and the maximum rating of east world wine is greater than the west world rating. Also, the average rating of wines produced in the east world is almost equal to the average rating of wines produced in the west world. Thus, we can say that the average ratings of the east world wines and west world wines are almost equal.

Plot 8: Q-Q Plot for Wine Rating

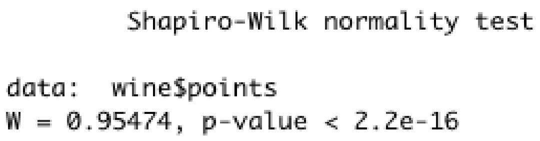


The Q-Q Plot was used for a visual check for data normality. From the plot, it can be inferred that the data is mostly not normally distributed. To be sure, I conducted the Shapiro-Wilk test.

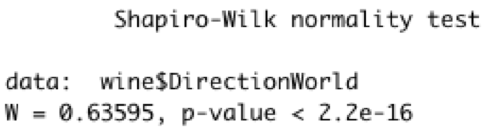
Shapiro-Wilk Test:

Null Hypothesis: The data is normally distributed

Alternative Hypothesis: The data is not normally distributed



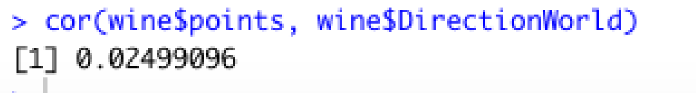
```
        Shapiro-Wilk normality test

data:  wine$points
W = 0.95474, p-value < 2.2e-16
```



```
        Shapiro-Wilk normality test

data:  wine$DirectionWorld
W = 0.63595, p-value < 2.2e-16
```

From this, we observe that the two p-values are lesser than the significance level of 0.05 which implies that the distribution of the data is different from the normal distribution.

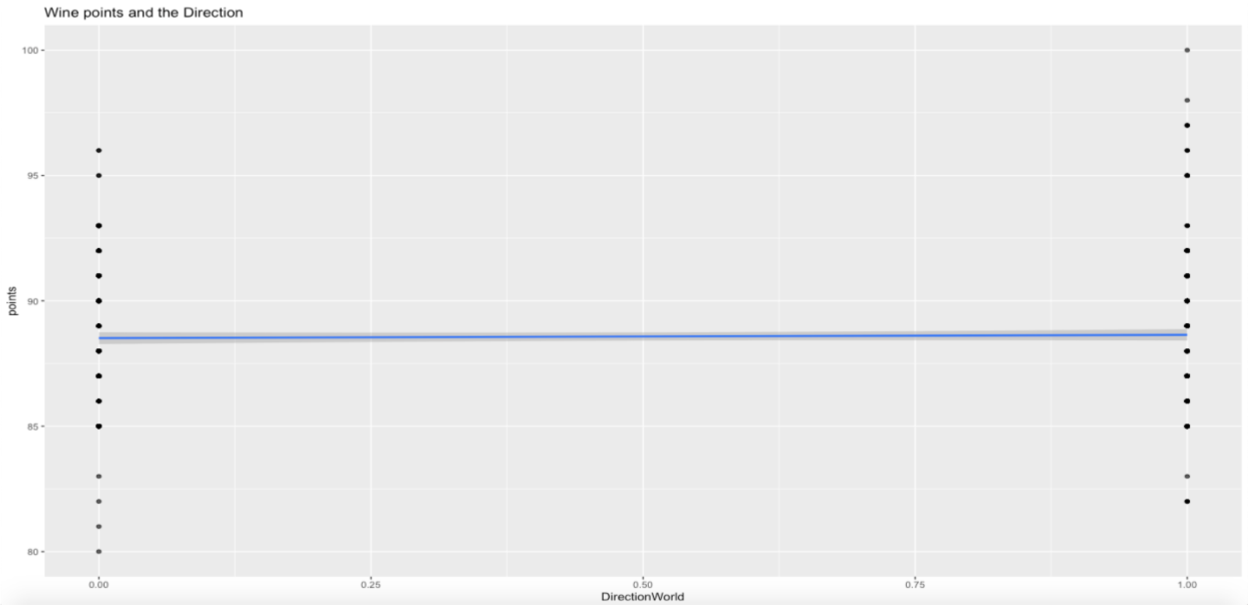
Correlation:



```
> cor(wine$points, wine$DirectionWorld)
[1] 0.02499096
```

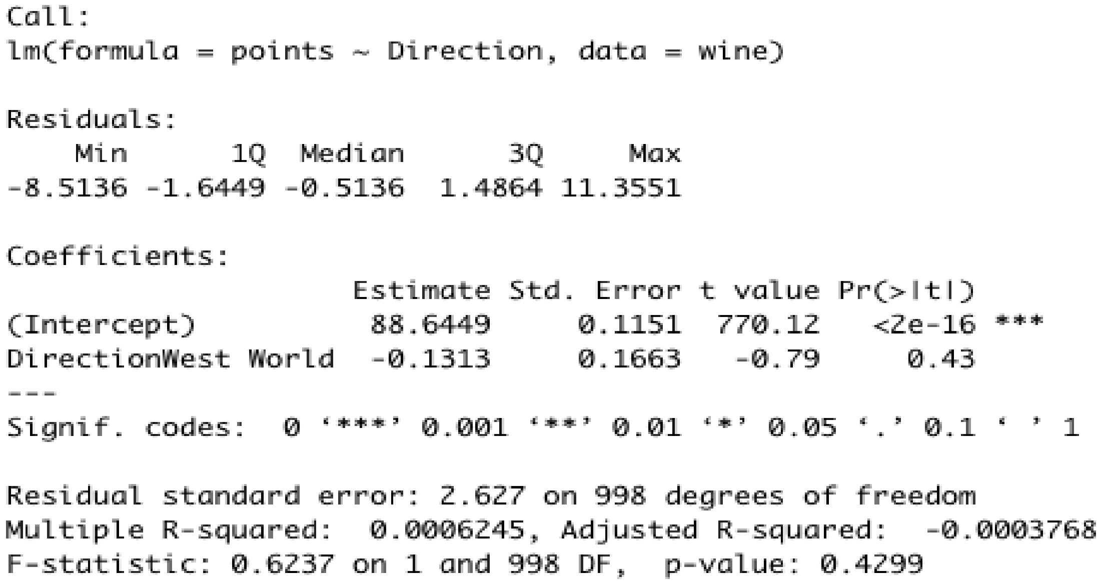
From this correlation, we infer that there is a negligible correlation between wine points and the direction (east wine and west wine).

Plot 9: Scatterplot for wine points and wine price



This plot represents the wine points and the direction of the winery (east world or west world). From this, we observe that the wine price and direction have a negligible correlation relationship.

Regression:



```
Call:
lm(formula = points ~ Direction, data = wine)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.5136 -1.6449 -0.5136  1.4864 11.3551 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)          88.6449     0.1151  770.12   <2e-16 ***
DirectionWest World  -0.1313     0.1663   -0.79     0.43    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.627 on 998 degrees of freedom
Multiple R-squared:  0.0006245, Adjusted R-squared:  -0.0003768 
F-statistic: 0.6237 on 1 and 998 DF,  p-value: 0.4299
```

From this, we find that the p-value (0.4299) is greater than the significance level of 0.05, thus we fail to reject the null hypothesis and can conclude that there is not enough evidence to prove that there is a linear correlation between direction and points in the population.

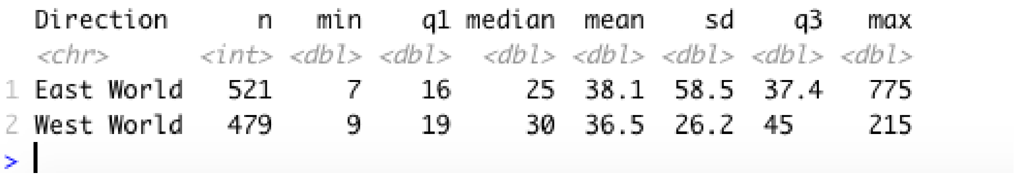
4. Does the mean price differ for wines produced in east world countries and the west world countries?

Hypothesis Testing

Null Hypothesis: No difference in the wine price of east world and west world wines

Alternative Hypothesis: There is difference in the wine price of east world and west world wines

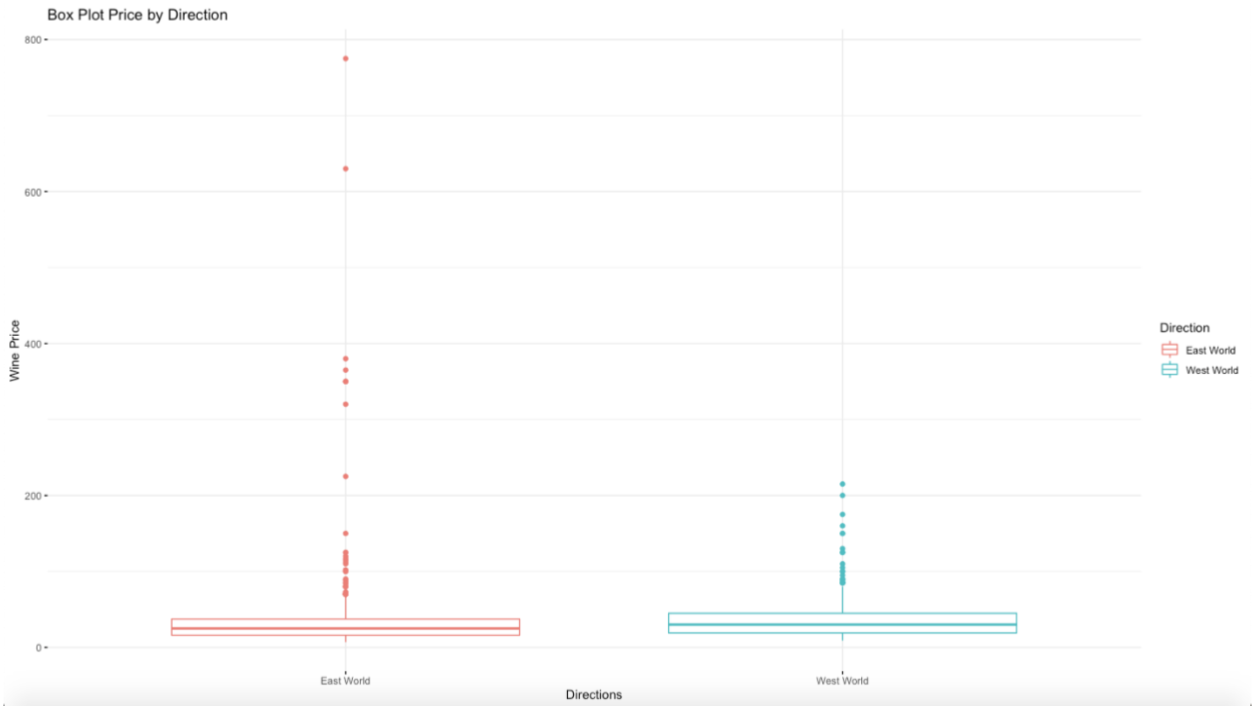
Table 6: Summary of the wine price with the direction (East or West World)



```
  Direction       n   min    q1 median  mean    sd    q3   max
  <chr>       <int> <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl>
1 East World    521     7    16     25  38.1  58.5  37.4   775
2 West World    479     9    19     30  36.5  26.2  45     215
> 
```

This table consists of the entire dataset's summary i.e., for 1000 observations. From this we can infer that the East World were 521 observations in total with minimum wine price of 7, the maximum price of wine as 775 and the average price was 38.1, whereas the West World consisted of 479 observations in total with minimum wine price of 9, the maximum wine price of 215 and the average price was 36.5. Comparing the two, it's easy to say that average mean price of the wine produced in east world is greater than the wine produced in west world. Also, the minimum price of east world wine is lesser than the west world price of wine and the maximum price of east world wine is greater than the west world price.

Plot 10: Box Plot for comparing the price of the wines produced in different worlds (directions)

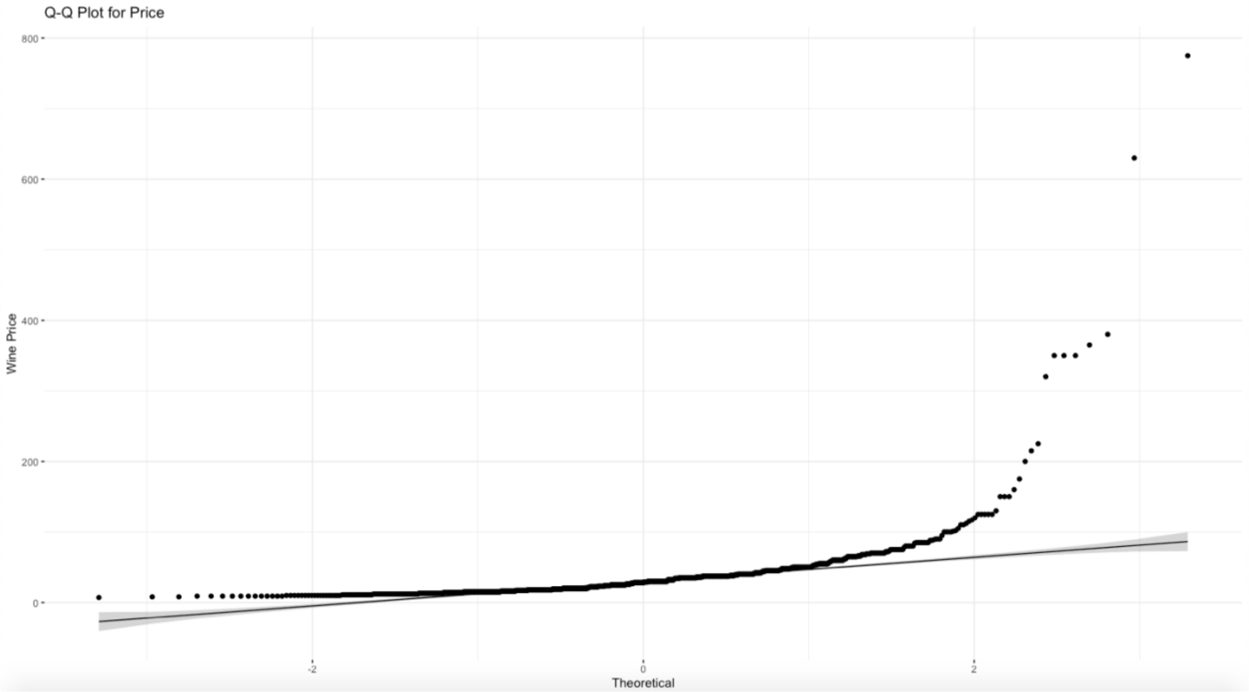


This boxplot compares the price of wines produced in east world wineries and west world wineries. From this plot, we can observe that there is a slight difference in the two averages. Also, both east world and west world wineries have outliers.

Preliminary Test:

I conducted the preliminary test to check if the data met the conditions for test assumption and to determine which test to conduct. Since the groups are independent, I conducted the two-sample unpaired test. To figure out if the data is normally distributed, I conducted the Shapiro Wilk test along with the variances.

Plot 11: Q-Q Plot to check normal distribution

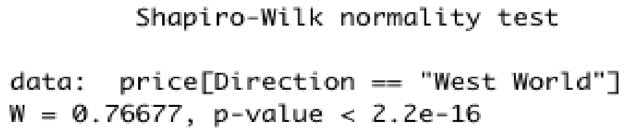


The Q-Q Plot was used for a visual check for data normality. From the plot, it can be inferred that the data is mostly not normally distributed. To be sure, we will be conducting the Shapiro-Wilk test.

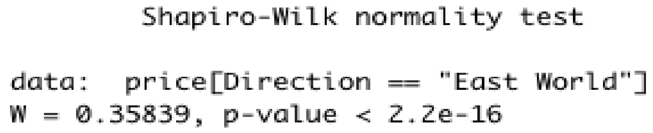
Shapiro-Wilk Test:

Null Hypothesis: The data is normally distributed

Alternative Hypothesis: The data is not normally distributed



```
        Shapiro-Wilk normality test

data:  price[Direction == "West World"]
W = 0.76677, p-value < 2.2e-16
```

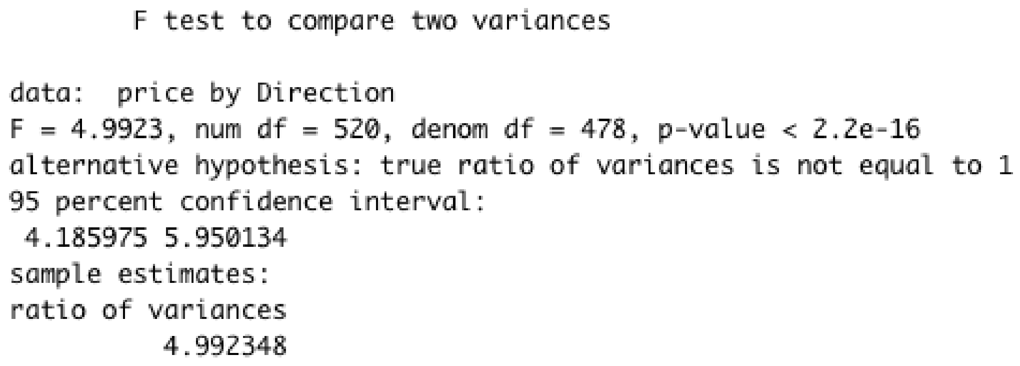


```
        Shapiro-Wilk normality test

data:  price[Direction == "East World"]
W = 0.35839, p-value < 2.2e-16
```

From this, we observe that the two p-values are lesser than the significance level of 0.05 implying that the distribution of the data is different from the normal distribution. Thus, it is not normally distributed and as such, we will use the Wilcoxon test.

Variance Test

The F-test is used to test for the homogeneity in the group's variances. From the test, we find that the p-value is lesser than the significance level of 0.05. Thus, there is a significant difference between the variances and a Wilcoxon test will be conducted.

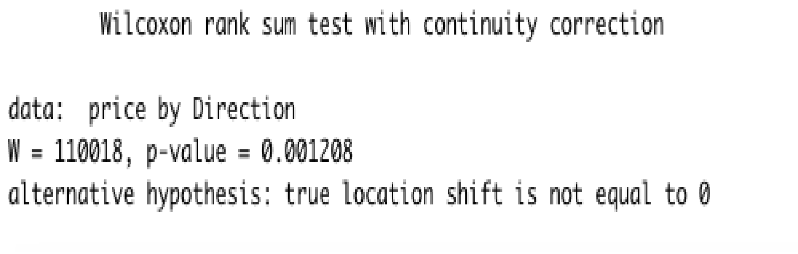


```
        F test to compare two variances

data:  price by Direction
F = 4.9923, num df = 520, denom df = 478, p-value < 2.2e-16
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 4.185975 5.950134
sample estimates:
ratio of variances 
          4.992348 
```

Computing

A confidence interval of 95% and α = 0.05 for our test. According to the rule, we reject the null hypothesis if the p-value is less than the 0.05.

Result:



```
        Wilcoxon rank sum test with continuity correction

data:  price by Direction
W = 110018, p-value = 0.001208
alternative hypothesis: true location shift is not equal to 0
```

From the above, we gather that the p-value of 0.001208 is lesser than the significance level or alpha of 0.05. Thus, we reject the null hypothesis and conclude that the price of wines produced in the east world and west world are significantly different.

Interpretation of the results:

From this result, we can conclude that since the p-value is less than 0.05, we reject the null hypothesis and conclude that the mean price was significantly different between east world and west world.

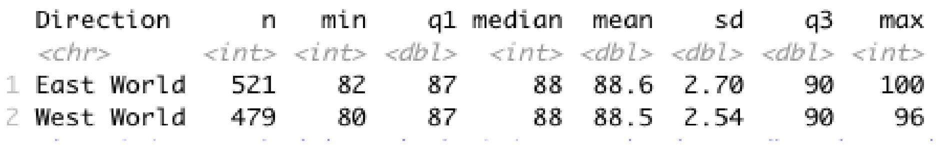
5. Does the mean rating differ for wines produced in east world countries and the west world countries?

Hypothesis Test

Null Hypothesis: No difference in the wine rating of east world and west world wines

Alternative Hypothesis: There is difference in the wine rating of east world and west world wines

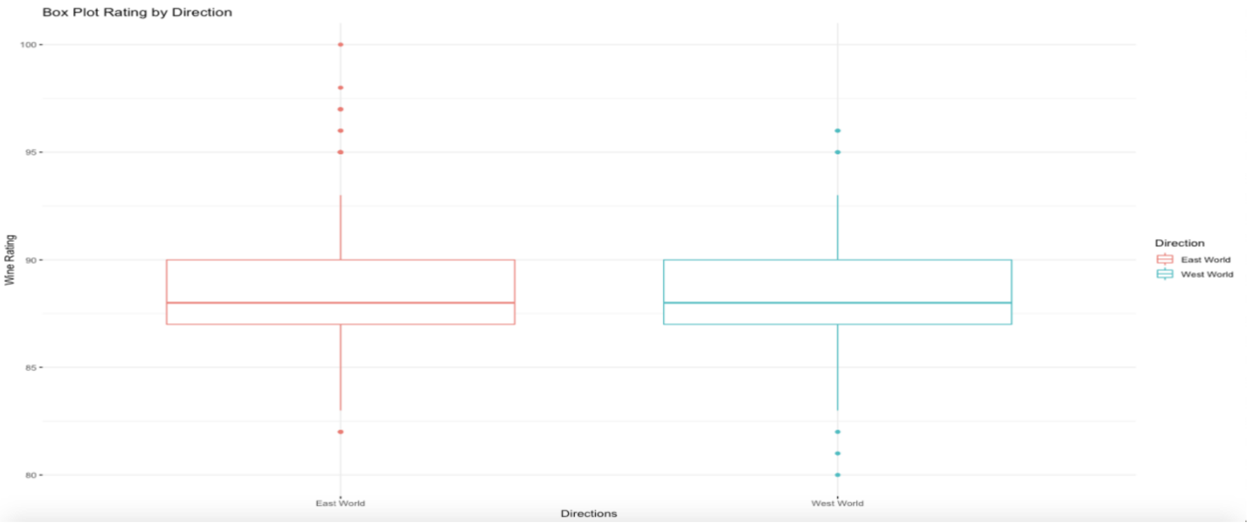
Table 7: Summary of wine ratings with the direction (East or West World)



```
  Direction        n   min    q1 median  mean    sd    q3   max
  <chr>        <int> <int> <dbl>  <int> <dbl> <dbl> <dbl> <int>
1 East World     521    82    87     88  88.6  2.70    90   100
2 West World     479    80    87     88  88.5  2.54    90    96
```

This table consists of the summary for 1000 observations. From this we can infer that the East World were 521 in total with minimum rating of 82 and maximum rating of 100 and a mean of 88.6 whereas the West World consisted of 479 observations in total with minimum rating of 80, maximum rating of 96 and a mean of 88.5. Comparing the two, we can say that the minimum rating of east world wine is greater than the west world rating and the maximum rating of east world wine is greater than the west world rating. Also, the average rating of wines produced in the east world is equal to the average rating of wines produced in the west world.

Plot 12: Box Plot for comparing the rating or points of the wines produced in the two different worlds (directions)

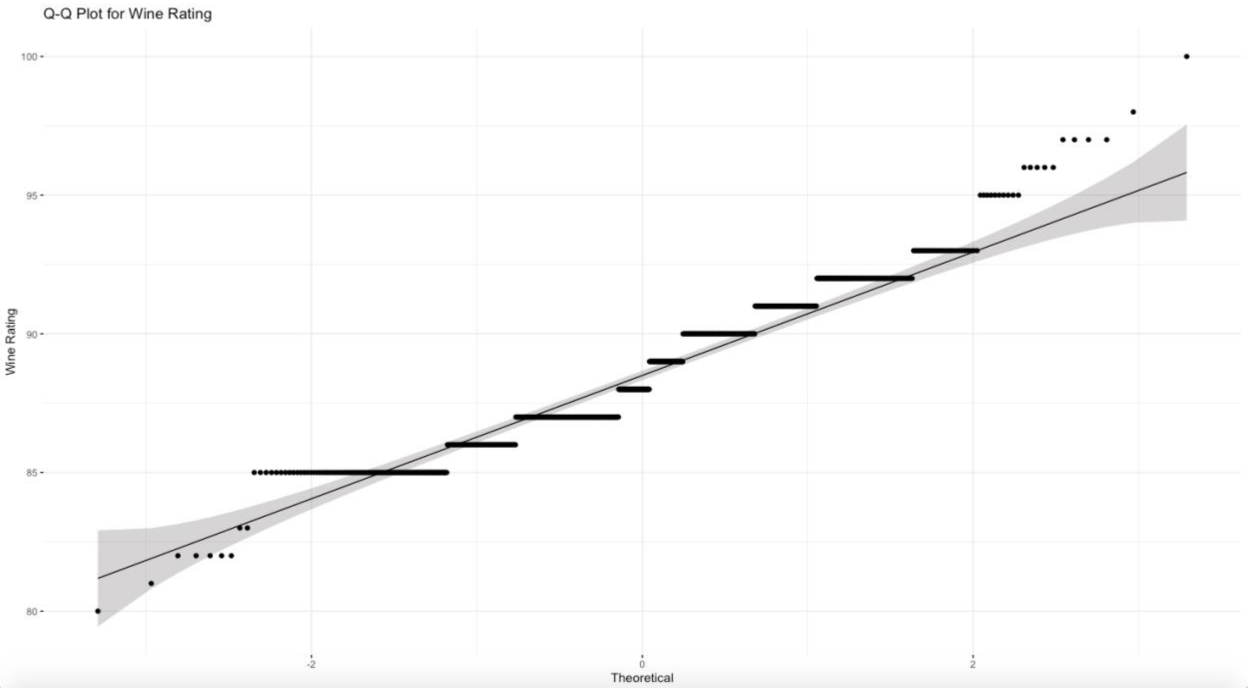


This boxplot compares the ratings of wines produced in east world wineries and west world wineries. From this plot, we can observe that there is not much difference in the two averages. Also, both east world and west world wineries have outliers.

Preliminary Test:

I conducted the preliminary test to check if the data met the conditions for test assumption and to determine which test to conduct. Since the groups are independent, I conducted the two-sample unpaired test. To know if the data is normally distributed, I conducted the Shapiro Wilk test along with the variances.

Plot 13: Q-Q Plot to check normal distribution

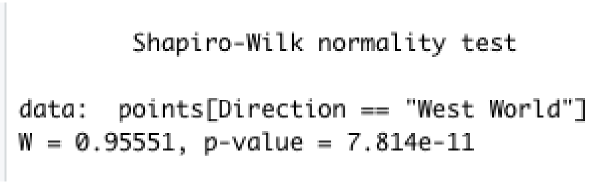


The Q-Q Plot was used for a visual check for data normality. From the plot, it can be inferred that the data is most likely not normally distributed. To be sure, we will be conducting the Shapiro-Wilk test.

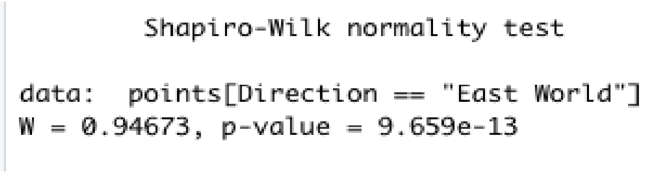
Shapiro-Wilk Test:

Null Hypothesis: The data is normally distributed

Alternative Hypothesis: The data is not normally distributed



```
        Shapiro-Wilk normality test

data:  points[Direction == "West World"]
W = 0.95551, p-value = 7.814e-11
```

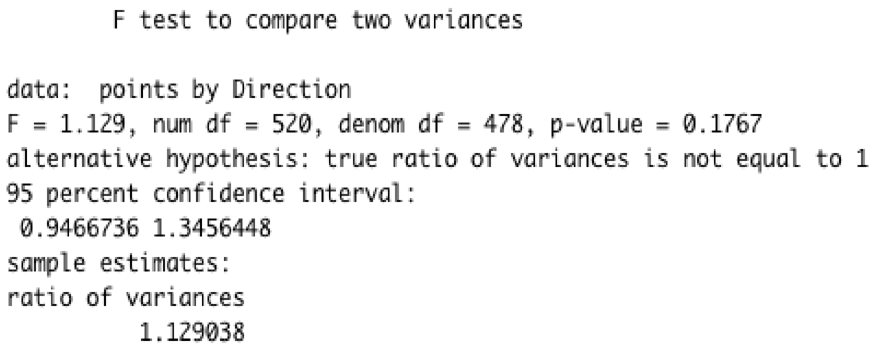


```
        Shapiro-Wilk normality test

data:  points[Direction == "East World"]
W = 0.94673, p-value = 9.659e-13
```

From this, we observe that the p-value for east world and west world is lesser than the significance level of 0.05. As such, it is not normally distributed. Since, the dataset is exhibiting a non-normal distribution, I have chosen the non-parametric two sample Wilcoxon test.

Variance Test

The F-test is used to test for the homogeneity in the group's variances. From the test, we find that the p-value is greater than the significance level of 0.05. Thus, there is no significant difference between the variances and a Wilcoxon test will be conducted.

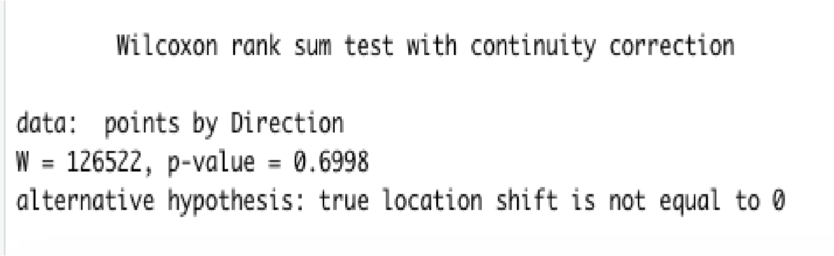


```
        F test to compare two variances

data:  points by Direction
F = 1.129, num df = 520, denom df = 478, p-value = 0.1767
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.9466736 1.3456448
sample estimates:
ratio of variances
          1.129038
```

Computing

A confidence interval of 95% and α = 0.05 for the test. According to the rule, we reject the null hypothesis if the p-value is less than the 0.05.

Result:



```
        Wilcoxon rank sum test with continuity correction

data:  points by Direction
W = 126522, p-value = 0.6998
alternative hypothesis: true location shift is not equal to 0
```

From the above, we observe that the p-value of 0.6998 is greater than the significance level or alpha of 0.05. Thus, we fail to reject the null hypothesis and conclude that the rating of wines produced in the east world and west world are not significantly different.

Interpretation of the results:

From this result, we can conclude that since the p-value was greater than 0.05, we failed to reject the null hypothesis which indicates that the data used didn’t have enough evidence to conclude that the mean rating was significantly different between east world and west world.

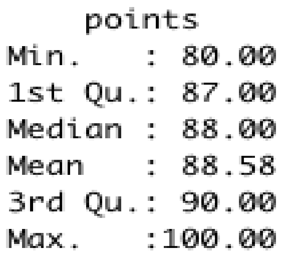
6. Are the mean average points of wine equal to 85?

Hypothesis Test

Null Hypothesis: The mean average points of wine is equal to 85

Alternative Hypothesis: The mean average points of wine is not equal to 85

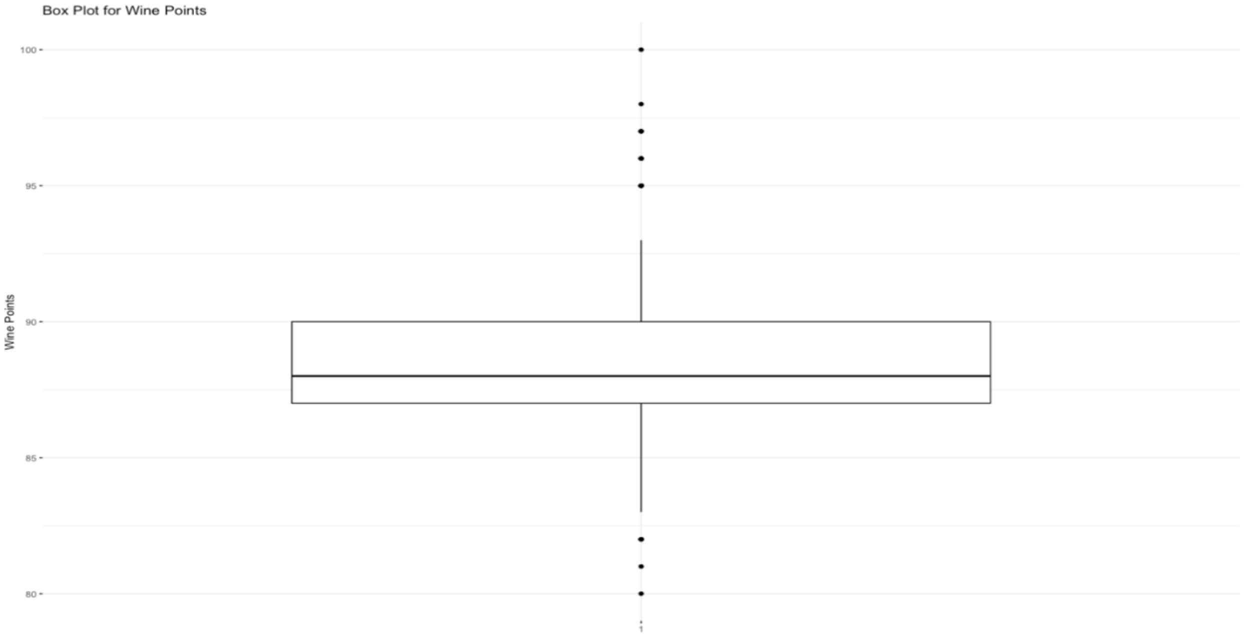
Table 8: Summary of the data on wine rating (points)



```
     points      
 Min.   : 80.00  
 1st Qu.: 87.00  
 Median : 88.00  
 Mean   : 88.58  
 3rd Qu.: 90.00  
 Max.   :100.00  
```

From this table, we observe that the wine rating of the dataset had an average mean of 88.58, minimum rating of 80, maximum rating of 100 and a median of 88.

Plot 14: Boxplot of the wine ratings

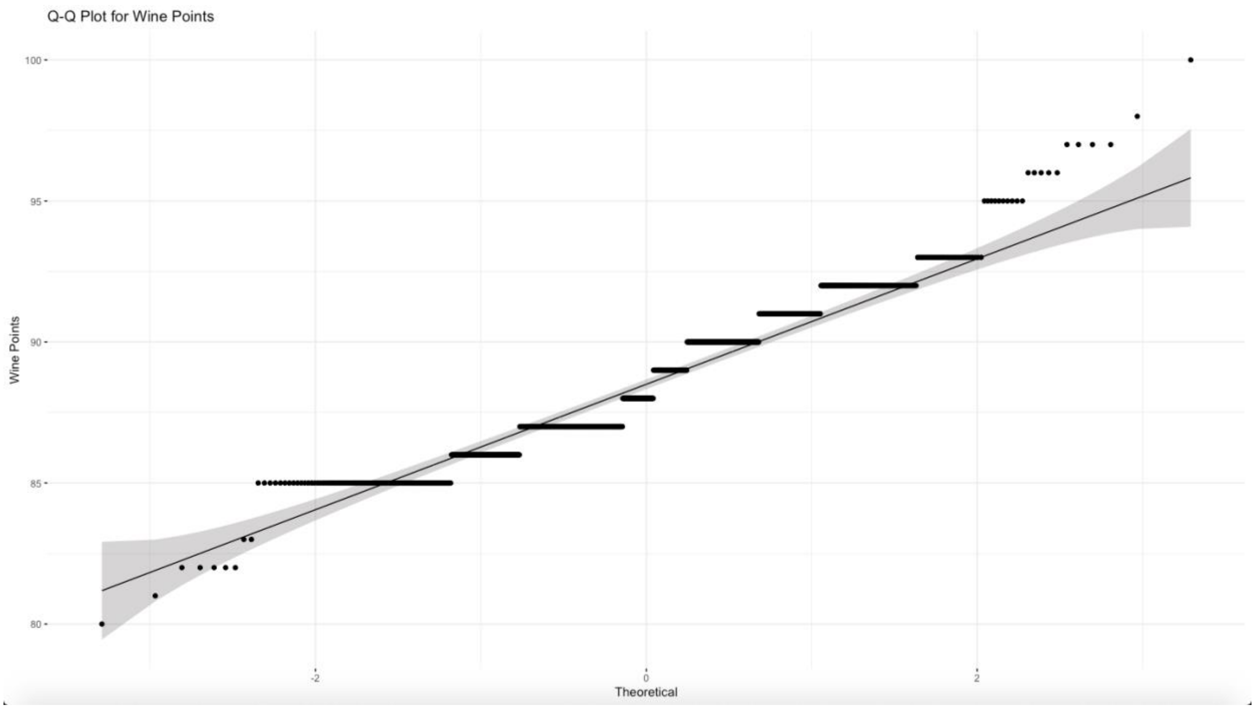


The plot above represents the wine tasting dataset’s points or wine ratings. It shows that there are outliers in the dataset.

Preliminary Test:

I conducted the preliminary test to check if the data met the conditions for test assumptions and to determine which test to conduct. To know if the data is normally distributed, I will be conducting the Shapiro Wilk Test. Also, since the data is distributed on both sides, we will use a two tailed test.

Plot 15: Q-Q Plot to check normal distribution

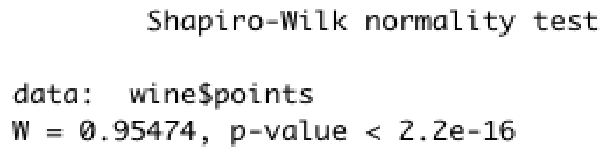


The Q-Q Plot was used to check for data normality. From this plot, it can be inferred that the data is mostly normally distributed. To be sure, we will conduct the Shapiro Wilk test.

Shapiro Wilk Test

Null Hypothesis: The data is normally distributed

Alternative Hypothesis: The data is not normally distributed



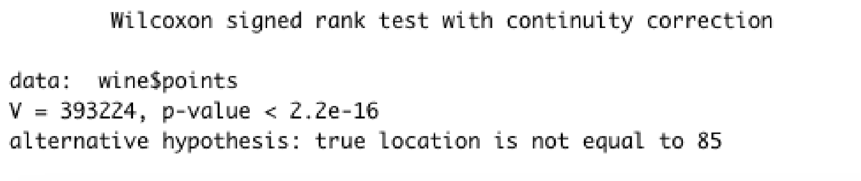
```
        Shapiro-Wilk normality test

data:  wine$points
W = 0.95474, p-value < 2.2e-16
```

From the test, we find that the p-value (2.2e-16) is lesser than significance level of 0.05 which indicates that the data is not normally distributed. As such, the Wilcoxon test can be conducted.

Computing:

A confidence interval of 95% and $\alpha = 0.05$ for the test. According to the rule, we reject the null hypothesis if the p-value is less than the 0.05.

Result:



```
        Wilcoxon signed rank test with continuity correction

data:  wine$points
V = 393224, p-value < 2.2e-16
alternative hypothesis: true location is not equal to 85
```

The Wilcoxon test results show that the p-value is 2.2e-16 which is lesser than the significance level of 0.05. Thus, we can reject the null hypothesis.

Interpretation of the results:

From this result, we can conclude that since the p-value is less than 0.05, the null hypothesis is rejected which indicates that the mean points of wine was not equal to 85.

Summary

From the tests conducted, the findings were:

- Test 1 was a two-sample test where I analyzed if there was a linear relationship between wine price and points. It had a significance level of 0.05, we rejected the null hypothesis because the p-value of 2.2e-16 is lesser than the significance level which proved that there was a significant linear correlation between the wine price and points of the wine tasting dataset.
- Test 2 was a two-sample test where I analyzed if there was a linear relationship between wine price and the direction of the wineries (east world and west world). It had a significance level of 0.05, we failed to reject the null hypothesis because the p-value of 0.5883 is greater than the significant level. Thus, the test didn't have enough evidence to show that wine price and direction had a significant linear correlation between each other.
- Test 3 was a two-sample test where I analyzed if there was a linear relationship between wine points and the direction of the wineries (east world and west world). It had a significance level of 0.05, we failed to reject the null hypothesis because the p-value of 0.4299 is greater than the significant level. Thus, the test didn't have enough evidence to show that wine points and direction had a significant linear correlation between each other.
- Test 4 was a two-sample test that had a significance level of 0.05, we failed to reject the null hypothesis because the p-value of 0.001208 is lesser than the significance level. Thus, we reject the null hypothesis and conclude that the price of wines produced in the east world and west world are significantly different.

- Test 5 was a two-sample test that had a significance level of 0.05, we failed to reject the null hypothesis in this test because the p-value of 0.6998 is greater than the significance level. Thus, the test didn’t have enough evidence to show that the mean rating was significantly different between east world and west world.
- Test 6 was a one-sample test that had a significance level of 0.05, we rejected the null hypothesis in this test since the p-value of 2.2e-16 is lesser than the significance level which proved that the mean rating or points of wine is not equal to 85.

References

1. Antoine Soetewey. (2020, May 28). *Correlation coefficient and correlation test in R.* Retrieved December 16, 2021, from https://statsandr.com/blog/correlation-coefficient-and-correlation-test-in-r/#interpretation-of-a-correlation-coefficient

2. STHDA. (n.d.). *Correlation Test Between Two Variables in R.* Retrieved December 16, 2021, from http://www.sthda.com/english/wiki/correlation-test-between-two-variables-in-r

3. Rebecca Evans. (2020, February 25). *A step-by-step guide to linear regression in R.* Retrieved December 16, 2021, from https://www.scribbr.com/statistics/linear-regression-in-r/

4. Stat Trek. (n.d.). *Hypothesis Test for Regression Slope.* Retrieved December 16, 2021, from https://stattrek.com/regression/slope-test.aspx