Project: Wine It!

1.Problem Statement and Scope:

1.1 Background and main issue:
Using a dataset of some 130k wine reviews, this project aims to come up with an effective pricing strategy for wine. With data about the wine variety, origin winery, description and ratings by Wine Enthusiasts, we aim to predict what factors will impress a wine sommelier.

Sommeliers are key influencers for the decision of choosing the right wine. They don't just educate people on what is the best wine but are also great content marketers. This can enable winemakers to market well by targeted descriptions of the wine that makes the reader want to buy it. This analysis will be useful to bucket consumers according to taste/price/wine choice to be able to compose wine descriptions that will resonate with that category of wine consumers and narrow down on the factors to be highlighted in the product pitch for a wine. As a next step, we can successfully target these consumer segments via channels (like restaurants, retail stores, social media handles of sommeliers, wine tastings, and vineyard tours) that have more likelihood to reach those potential customers.

1.2 Structuring the problem:
To convert the business problem into a statistical challenge, we have framed some analytical questions that the data can potentially help us take better marketing decisions. Here are some of the key questions that will help us analyze the data:

- Exploring the data for categorical data, missing data and inconsistencies.
- What is the relationship between ratings and known factors like variety, type of vineyard, and origin of the wine?
- What are the pricing trends across factors like country, region, variety, etc.?
- Are there any patterns or personal bias of the taster that impacts rating? Can we identify patterns in factors that help increase ratings?

1.3 Data-set chosen:
Data-set for this project has been taken from https://www.kaggle.com/zynicide/wine-reviews. The data set has 13-15 columns and 130k rows. We tried to ensure that the data has the least continuous text (string) variables - 2 fields: description and the name of the wine. The data has a mix of continuous and categorical variables. The rows of data set are different wines and the columns are the characteristics like the origin, price, etc.

2.Exploration and Data-cleaning:

#check missing data: nearly 9000 prices rows missing

```
print(colSums(is.na(wine.df)))
```

```
> print(colSums(is.na(wine.df)))
                    X               country           description           designation
                    0                     0                     0                     0
               points                 price              province              region_1
                    0                  8996                     0                     0
             region_2           taster_name taster_twitter_handle                 title
                    0                     0                     0                     0
              variety                winery
                    0                     0
```

#removing rows with NA data

```
wine.df <- wine.df[!is.na(wine.df$price), ]
```

```
> wine.df <- wine.df[!is.na(wine.df$price), ]
> print(colSums(is.na(wine.df)))
                    X               country           description           designation
                    0                     0                     0                     0
               points                 price              province              region_1
                    0                     0                     0                     0
             region_2           taster_name taster_twitter_handle                 title
                    0                     0                     0                     0
              variety                winery
                    0                     0
```

#checking number of unique values in each column

```
unique_wine.df <- lapply(wine.df, unique)
sapply(unique_wine.df, length)
```

```
unique_wine.df <- lapply(wine.df, unique)
sapply(unique_wine.df, length)
                    X               country           description           designation
               120975                    43                111567                 35777
               points                 price              province              region_1
                   21                   390                   423                  1205
             region_2           taster_name taster_twitter_handle                 title
                   18                    20                    16                110638
              variety                winery
                  698                 15855
```

Observations:

- 20 unique tasters provided the reviews to wines from 43 countries.
- There were 698 unique varieties of wines reviewed, both red and white.

#removing NAs and factoring categorical variables

```
wine_structured.df <- wine.df
wine_structured.df$province <- as.factor(wine_structured.df$province)
wine_structured.df$variety <- as.factor(wine_structured.df$variety)
wine_structured.df$country <- as.factor(wine_structured.df$country)
wine_structured.df$winery <- as.factor(wine_structured.df$winery)
wine_structured.df$designation <- as.factor(wine_structured.df$designation)
wine_structured.df$description <- NULL
wine_structured.df$region_1 <- NULL
wine_structured.df$region_2 <- NULL
wine_structured.df$taster_twitter_handle <- NULL
wine_structured.df$title <- NULL
wine_structured.df<-na.omit(wine_structured.df)
```

```
summary(wine_structured.df)
str(wine_structured.df)
```

```
> summary(wine_structured.df)
       X                country              designation          points           price                    province    
 Min.   :     1   US       :54265                    :34779   Min.   : 80.00   Min.   :   4.00   California     :36104  
 1st Qu.: 32575   France   :17776   Reserve          : 1980   1st Qu.: 86.00   1st Qu.:  17.00   Washington     : 8583  
 Median : 65144   Italy    :16914   Estate           : 1318   Median : 88.00   Median :  25.00   Oregon         : 5359  
 Mean   : 65046   Spain    : 6573   Reserva          : 1219   Mean   : 88.42   Mean   :  35.36   Tuscany        : 5128  
 3rd Qu.: 97507   Portugal: 4875    Estate Grown:      618    3rd Qu.: 91.00   3rd Qu.:  42.00   Bordeaux       : 4002  
 Max.   :129970   Chile    : 4416   Riserva          :  607   Max.   :100.00   Max.   :3300.00   Northern Spain: 3797  
                  (Other) :16156    (Other)          :80454                                      (Other)        :58002  
             taster_name                      variety                        winery      
                   :24496   Pinot Noir              :12787   Testarossa            :   217  
 Roger Voss        :20172   Chardonnay              :11080   Williams Selyem       :   211  
 Michael Schachner:14951    Cabernet Sauvignon      : 9386   DFJ Vinhos            :   209  
 Kerin Oâ€™Keefe   : 9874    Red Blend               : 8476   Wines & Winemakers    :   209  
 Virginie Boone    : 9507   Bordeaux-style Red Blend: 5340   Chateau Ste. Michelle:    193  
 Paul Gregutt      : 9498   Riesling                : 4972   Louis Latour          :   173  
 (Other)           :32477   (Other)                 :68934   (Other)               :119763  
```

From the summary, we can see the top 6 most reviewed wine varieties.

```
> str(wine_structured.df)
'data.frame':   120975 obs. of  9 variables:
 $ X          : int  1 2 3 4 5 6 7 8 9 10 ...
 $ country    : Factor w/ 44 levels "","Argentina",..: 33 44 44 44 39 24 17 19 17 44 ...
 $ designation: Factor w/ 37980 levels "","'61 RosÃ©",..: 2404 1 28205 36717 2049 3107 1 30993 20073 23124 ...
 $ points     : int  87 87 87 87 87 87 87 87 87 87 ...
 $ price      : num  15 14 13 65 15 16 24 12 27 19 ...
 $ province   : Factor w/ 426 levels "","Ã–sterreichischer Perlwein",..: 115 274 225 274 268 340 17 314 17 57 ...
 $ taster_name: Factor w/ 20 levels "","Alexander Peartree",..: 17 16 2 16 14 11 17 3 17 20 ...
 $ variety    : Factor w/ 708 levels "","Ã‡alkarasÄ±",..: 454 441 483 445 593 191 213 213 441 88 ...
 $ winery     : Factor w/ 16757 levels ":Nota Bene","1+1=3",..: 12989 13063 14433 14665 14742 15048 15436 8443 9000 9342 ...
 - attr(*, "na.action")= 'omit' Named int  1 14 31 32 33 51 55 80 138 160 ...
  ..- attr(*, "names")= chr  "1" "14" "31" "32" ...
```

There are 20 Tasters, 16.8k wineries, 44 countries, and 708 varieties.

Some analysis on the points variable:

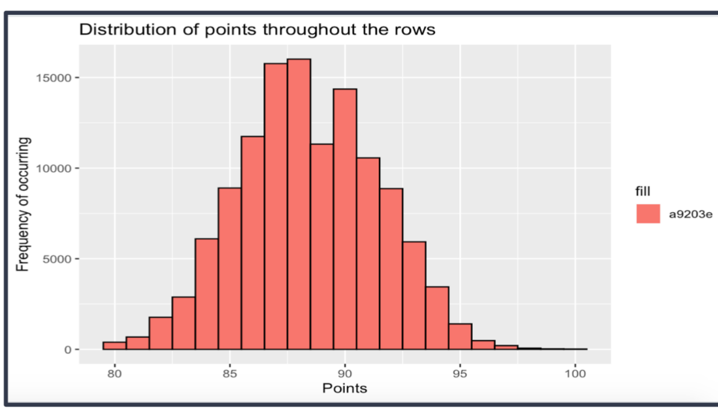
```
summary(wine.df$points)
```

```
summary(wine.df$points)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  80.00   86.00   88.00   88.42   91.00  100.00 
```

Min points =80
Max points =100

#analysing distribution: Normally distributed

```
ggplot(data = wine.df, aes(x= points, colour = I("black"), fill = "a9203e"))+
geom_histogram(binwidth=1)+ labs(x = "Points", y= "Frequency of occurring", title =
"Distribution of points throughout the rows")
```



Normally Distributed

Some analysis on the price variable:

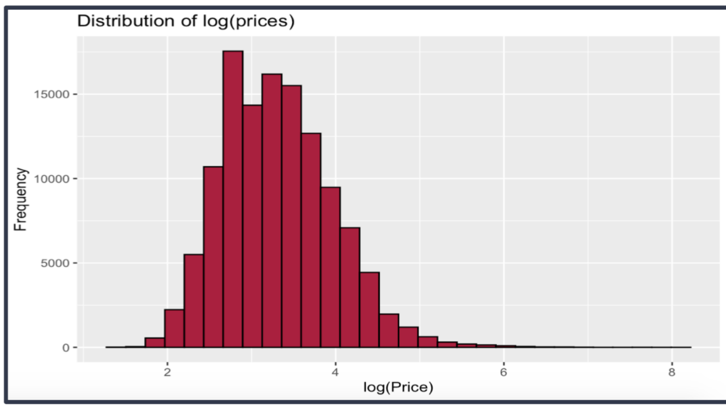
```
summary(wine.df$price)
```

```
summary(wine.df$price)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 4.00   17.00   25.00   35.36   42.00 3300.00
```

Min points =4.00
Max points =3300.00

#log data to respond to left-skewness of price data

```
ggplot(data = wine.df, aes(x= log(price), colour = I('black'), fill =
I('#a9203e')))+geom_histogram()+ labs(x = "log(Price)", y= "Frequency",title =
"Distribution of log(prices)")
```



Not normally distributed

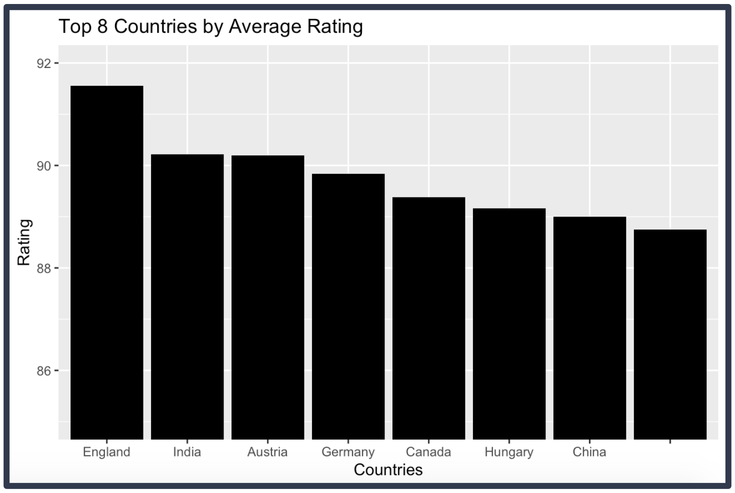
Analysing if highest rated wines come from France

France is intuitively the country known for best quality wines, hence it is important to analyze if it is actually true. Extracting top 15 mean data for country-wise ratings:

```
wineRating = wine.df %>%
group_by(country) %>%
summarise_at(vars(points), funs(points = mean(., na.rm=T))) %>%
arrange(desc(points)) %>%
head(8)
```

After plotting wineRating against the countries, it can be concluded that the highest rated wines actually come from England and not France, as shown below:

```
ggplot(data=wineRating, aes(x=reorder(country,-points), y= points)) +
geom_bar(stat="identity", fill = "purple") +
coord_cartesian(ylim=c(85,92)) +
labs(x="Countries", y="Rating",title="Top 15 Countries by Average Rating")
```



Analysing where the costliest wines come from:

To corroborate the price analysis further, an analysis to determine the origin of costliest wines has been performed.

```
install.packages("kableExtra")
library(kableExtra)

costly_wines<-wine.df %>%
          select(country,points,price,province,winery) %>%
          arrange(desc(price)) %>%
          head(n = 25)
costly_wines_number<-wine.df %>%
          arrange(desc(price)) %>%
          head(n = 25)%>%
          group_by(country)%>%
          summarise(n = n())%>%
          arrange(desc(n))
```

After extracting data for top 25 costliest wines by sorting them in descending price order, it has been found that the country France occurs the greatest number of times.

```
kable(costly_wines_number,"html") %>% kable_styling("striped",full_width=T,position =
"float_right")%>% row_spec(1,bold=F,color="white",background="#ffb6c1")
```

country	n
France	18
Australia	2
Portugal	2
Austria	1
Italy	1
US	1

```
kable(costly_wines,"html") %>% kable_styling("striped",full_width=T) %>%
column_spec(1:3,bold=T,background="white") %>%
row_spec(c(1,2,3,5:13,16,18,19,21,24,25),bold=F,color="white",background="#ffb6c1")
```

country	points	price	province	winery
France	88	3300	Bordeaux	Château les Ormes Sorbet
France	96	2500	Bordeaux	Château Pétrus
France	96	2500	Burgundy	Domaine du Comte Liger-Belair
US	91	2013	California	Blair
France	97	2000	Bordeaux	Château Pétrus
France	96	2000	Burgundy	Domaine du Comte Liger-Belair
France	98	1900	Bordeaux	Château Margaux
France	100	1500	Bordeaux	Château Lafite Rothschild
France	100	1500	Bordeaux	Château Cheval Blanc
France	96	1300	Bordeaux	Château Mouton Rothschild
France	96	1200	Bordeaux	Château Haut-Brion
France	94	1125	Burgundy	Domaine du Comte Liger-Belair
France	97	1100	Bordeaux	Château La Mission Haut-Brion

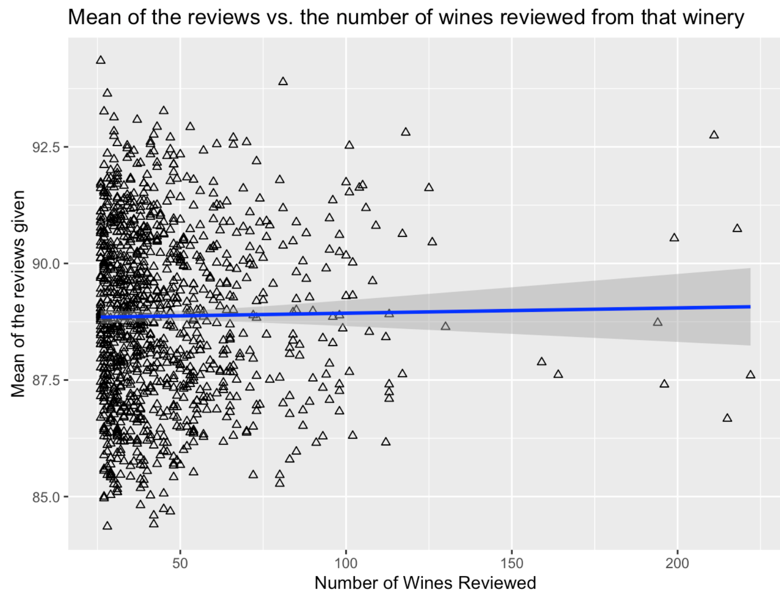
Therefore, it can be established that France is the origin of costliest wines reviewed in our data set.

Analyzing if the best-reviewed wines come from most reviewed wineries:

After finding the top reviewed wines, it is important to find if the best-rated wines were actually from the popular, the most rated wineries because it may be the case that the better a winery, the more it's wines are reviewed. Or the opposite can be true that the number of reviews is a function of the volume of wine produced at a winery or smaller the winery, the better the wine.
To answer this question, after grouping all the wineries based on a number of reviews of different wines from there, the mean of all the ratings given is calculated and scatter plotted against the total number of wines reviewed.

```
wineries = wine.df%>%
  group_by(winery) %>%
  count()

winery_mean<-wine.df %>%
  select(winery,points) %>%
  group_by(winery) %>%
  summarise(Mean_points = mean(points))

winery_reviews_mean<-wineries %>%
  inner_join(winery_mean, by = "winery") %>%
   filter(n>25)
```

```
ggplot(winery_reviews_mean, aes(x=n, y=Mean_points)) +  geom_smooth(method=lm ,
color="blue", se=TRUE) + geom_point(shape=2) +
   labs(title ="Mean of the reviews vs. the number of wines reviewed from that
winery", x = "Number of Wines Reviewed", y = "Mean of the reviews given")
```



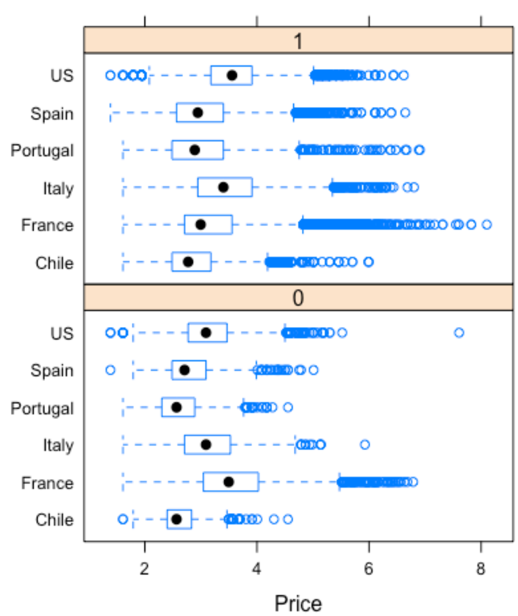
As can be seen from the plot, there is barely any correlation. Therefore, it is not necessary that the most reviewed wineries are producing the best-rated wines.

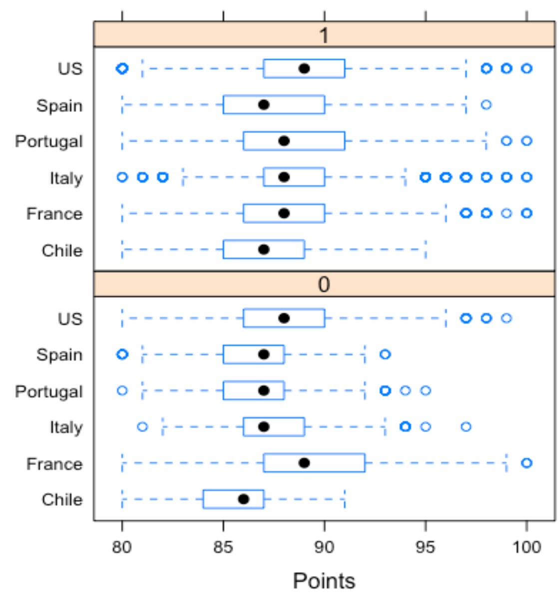
Creating box plots for Price and Points:

Since the data came from so many different countries and produced in different varieties, it is great to have an aggregated look. This can help in checking on some key assumptions like best wines are produced in France or American wines are cheap. To narrow down and work on significant data, the Pareto rule was followed to choose the data set. Six countries were picked which produced a maximum number of wines. Following that, the wines were broadly classified into white wines and red wines.

```
wine.boxplot.1.df<- wine_structured.df[wine_clear$country == "US"
                                        | wine_clear$country == "France"
                                       | wine_clear$country == "Italy"
                                       |wine_clear$country == "Spain"
                                       | wine_clear$country == "Portugal"
                                       | wine_clear$country == "Chile" , ]
   wine.boxplot.1.df$color<- sapply(wine.boxplot.1.df$variety, function (x)
{          ifelse(x == "Chardonnay" |
              x == "Riesling" |
              x == "Sauvignon Blanc" |
              x == "White Blend" |
              x == "Sparkling Blend" |
              x == "Pinot Gris" |
              x == "Champagne Blend" |
              x == "GrÃ ¼ner Veltliner" |
```

```
                x == "Pinot Grigio" |
                x== "Portuguese White" |
                x == "Viognier" |
                x == "GewÃ ¼rztraminer" |
                x == "GewÃ ¼rztraminer",
              0 , 1)})

    wine.boxplot.1.df$color <- as.factor(wine.boxplot.1.df$color)
    wine.boxplot.2.df <- wine.boxplot.1.df[ wine.boxplot.1.df$price < 4000, ]

    library(lattice)
    bwplot(country ~ log(price)| color, data=wine.boxplot.2.df, xlab="Price")
    bwplot(country ~ points | color, data=wine.boxplot.1.df, xlab="Points")
```





Inference:
From the box, the plot we can infer that: against popular opinion about wines,
1. Red wines from the USA, followed by Portugal are of the best quality.
2. Italy, followed by the USA produces expensive red wine.
3. White wines from France, followed by the USA are of the best quality.
4. French white wines are priced costlier than the rest of white wine

Testing for wine data set:

French wines are always hyped up in terms of both quality and price. It was essentially, to qualify these assumptions. The wines were grouped in different data sets based on the french or non-french origin. The points and price were compared using the t-test.

```
wine_france.df <- wine_structured.df[wine_structured.df$country=='France',]

wine_no_france.df <- wine_structured.df[wine_structured.df$country!='France',]
summary(wine_no_france.df)
```

```
t.test(wine_france.df$points,
       wine_no_france.df$points,
       alternative = "two.sided",
       var.equal = FALSE)

t.test(wine_france.df$price,
       wine_no_france.df$price,
       alternative = "two.sided",
       var.equal = FALSE)
```

```
data:  wine_france.df$points and wine_no_france.df$points
t = 14.971, df = 24457, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.3188617 0.4149354
sample estimates:
mean of x mean of y
 88.73487  88.36797
```

```
data:  wine_france.df$price and wine_no_france.df$price
t = 12.042, df = 18948, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 5.668572 7.872627
sample estimates:
mean of x mean of y
 41.13912  34.36852
```

Based on the results, it is inferred that when compared with the wines from across the globe, French wines get higher points. The difference which to the naked eye does not look significant, is statistically significant, thereby rejecting the null hypothesis (wines get about the same points). The mean rating obtained for French wines is 88.73 as compared to 88.37 for others.

Additionally, it can be said that French wines are more expensive as compared to wine produced globally. The average price for French wines is 41.14 which makes it expensive as compared to the other.

Predicting the price of wine based on origin, color and points:

Based on the data available, predicting the price of the wine seemed achievable. Decision tree classifiers like Random Forest can be used to predict the class of an observation. Since there were too many levels within the categorical variable of country and variety, it was necessary to keep only the important levels and group the variety variable, so that the model would be easier for the model to give the prediction.

The top three wine producing countries - the USA, France and Italy were selected. The wines were categorized into two types - red and white.

```
    wine_clear.df$color<- sapply(wine_clear.df$variety, function (x) {

                            ifelse(x == "Chardonnay" |
                           x == "Riesling" |
                           x == "Sauvignon Blanc" |
                           x == "White Blend" |
                           x == "Sparkling Blend" |
                           x == "Pinot Gris" |
                           x == "Champagne Blend" |
                           x == "GrÃ ¼ner Veltliner" |
                           x == "Pinot Grigio" |
                           x== "Portuguese White" |
                           x == "Viognier" |
                           x == "GewÃ ¼rztraminer" |
                           x == "GewÃ ¼rztraminer",
                           0 , 1)})

    wine_clear.df$color <- as.factor(wine_clear.df$color)
    summary(wine_clear.df)


    wine_clear_2.df<- wine_clear.df[wine_clear$country == "US" |  wine_clear$country
== "France" | wine_clear$country == "Italy",]

    wine_clear_2.df$color <- as.factor(wine_clear_2.df$color)
    wine_clear_2.df$country <- as.factor(wine_clear_2.df$country)

    wine_clear_2.df$country<- sapply(wine_clear_2.df$country, function (x) {

      if(x == "US")
        x = 1
      else if (x == "France")
      x = 2
      else if ( x == "Italy")
        x = 3
        })

    summary(wine_clear_2.df)
    wine_clear_2.df$variety <- NULL

    library(randomForest)
    model.for.wine <- randomForest(price~country+color+points,
                                data = wine_clear_2.df, maxnodes = 500)
    summary(model.for.wine)
    sample <- sample.int(n = nrow(wine_clear_2.df), size =
floor(.75*nrow(wine_clear_2.df)), replace = F)
    wine.train <- wine_clear_2.df[sample, ]
    wine.test  <- wine_clear_2.df[-sample, ]
    exam <- data.frame("country" = c(1,2,3), "points" = c(95, 95, 95),
                    "color" = c(0, 1 ,1))
```

```
exam$price <- predict(model.for.wine, exam)
exam

wine.train.2 <- predict(model.for.wine, wine.train)
wine.test.2 <- predict(model.for.wine, wine.test)
```

	country	points	color	price
1	1	95	0	51.74332
2	2	95	1	110.03076
3	3	95	1	87.08203

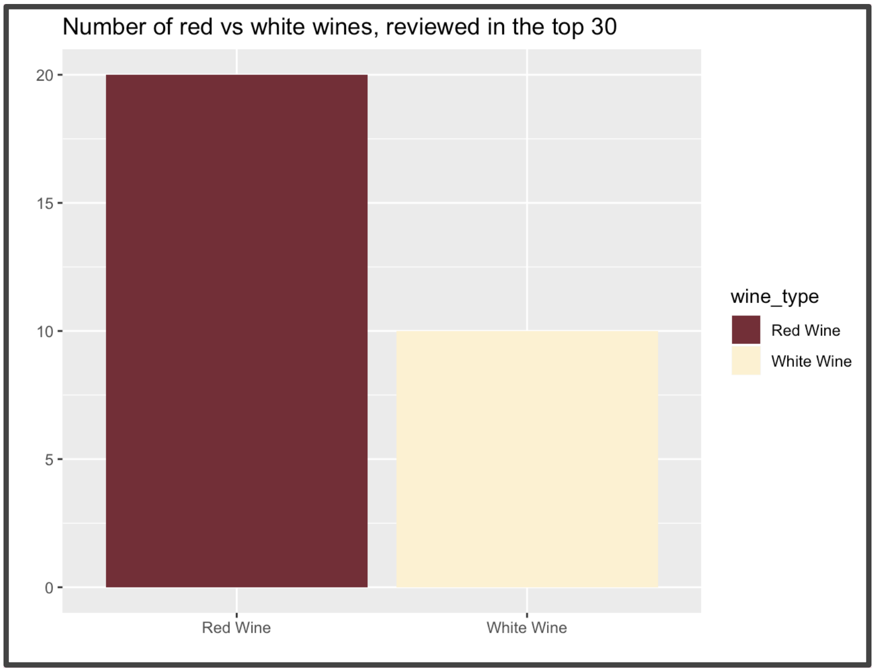
Analyzing if red wines are reviewed more than white wines:

In an attempt to find the most popular wines amongst wine testers, it is imperative to know if red wine is more popular among wine enthusiasts or white wine is.
After extracting the top 30 most reviewed wines, the wines have been categorized into red or white from subject matter expertise about various wines and their respective categories, as shown below:

Extract top 30 reviewed wines	`most_reviewed <- wine.df %>% group_by(variety) %>% summarise(count = n())%>% arrange(desc(count)) %>% head(n=30)`
Assigning red or white attribute to the wines from subject-matter expertise	`red_or_white <- most_reviewed %>% mutate(wine_type = ifelse(variety == "Chardonnay" | variety == "Riesling" | variety == "Sauvignon Blanc" | variety == "White Blend" | variety == "Sparkling Blend" | variety == "Pinot Gris" | variety == "Champagne Blend" | variety == "GrÃÂ¼ner Veltliner" | variety == "Pinot Grigio" | variety == "Portuguese White" | variety == "Viognier" | variety == "GewÃÂ¼rztraminer" | variety == "GewÃÂ¼rztraminer", "White Wine", "Red Wine"))`

After assigning red or white types, the number of each wine color that have been reviewed in the top 30 varieties are determined as below:

How many of each wine color have been reviewed in the top 30 varieties?	`red_or_white %>% group_by(wine_type) %>% summarise(n = n()) %>% ggplot(aes(x=wine_type, y=n, fill = wine_type))+ geom_bar(stat = "identity") + scale_fill_manual(values = c("#722f37", "#fcf1d2")) + labs(title = "Number of red vs white wines, reviewed in the top 30", x= "", y= "")`



As can be seen, 20 out of 30 top-reviewed wines are Red

Independence of Factors

We have two important categorical variables - country and variety. Also, we have two important continuous variables - Price and Points. We used Chi-Square tests to check association between categorical variables and one-way anova tests to check association between a categorical and a continuous variable for e.g. points and variety, points and country, price and variety, price and country.

```
wine.aov3<-aov(points~variety, data = wine_structured.df)
summary(wine.aov3)
wine.aov4<-aov(price~variety, data = wine_structured.df)
summary(wine.aov4)
wine.aov5<-aov(points~country, data = wine_structured.df)
summary(wine.aov5)
wine.aov6<-aov(price~country, data = wine_structured.df)
summary(wine.aov6)
chisq.test(wine_structured.df$variety, wine_structured.df$province)
chisq.test(wine_structured.df$variety, wine_structured.df$country)
chisq.test(wine_structured.df$variety, wine_structured.df$taster_name)
chisq.test(wine_structured.df$points, wine_structured.df$taster_name)
```

The Chi-square tests and Anova tests showed dependence showed association between all the combinations of variables tested. Hence, the next step was to dive deeper into what these associations are and if we can come up with some insights based on these associations among the various variables.

```
> wine.aov3<-aov(points~variety, data = wine_structured.df)
> summary(wine.aov3)
                Df  Sum Sq Mean Sq F value Pr(>F)    
variety        697  108712  155.97   18.53 <2e-16 ***
Residuals   120277 1012600    8.42                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> wine.aov4<-aov(price~variety, data = wine_structured.df)
> summary(wine.aov4)
                Df    Sum Sq Mean Sq F value Pr(>F)    
variety        697  17142617   24595   15.87 <2e-16 ***
Residuals   120277 186435133    1550                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> wine.aov5<-aov(points~country, data = wine_structured.df)
> summary(wine.aov5)
                Df  Sum Sq Mean Sq F value Pr(>F)    
country         42   57717  1374.2   156.3 <2e-16 ***
Residuals   120932 1063594     8.8                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> wine.aov6<-aov(price~country, data = wine_structured.df)
> summary(wine.aov6)
                Df    Sum Sq Mean Sq F value Pr(>F)    
country         42   3885651   92516   56.03 <2e-16 ***
Residuals   120932 199692099    1651                   
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> chisq.test(wine_structured.df$variety, wine_structured.df$province)
Chi-squared approximation may be incorrect
        Pearson's Chi-squared test

data:  wine_structured.df$variety and wine_structured.df$province
X-squared = 6305572, df = 294134, p-value < 2.2e-16

> chisq.test(wine_structured.df$variety, wine_structured.df$country)
Chi-squared approximation may be incorrect
        Pearson's Chi-squared test

data:  wine_structured.df$variety and wine_structured.df$country
X-squared = 1294772, df = 29274, p-value < 2.2e-16

> chisq.test(wine_structured.df$variety, wine_structured.df$taster_name)
Chi-squared approximation may be incorrect
        Pearson's Chi-squared test

data:  wine_structured.df$variety and wine_structured.df$taster_name
X-squared = 368074, df = 13243, p-value < 2.2e-16

> chisq.test(wine_structured.df$points, wine_structured.df$taster_name)
Chi-squared approximation may be incorrect
        Pearson's Chi-squared test

data:  wine_structured.df$points and wine_structured.df$taster_name
X-squared = 18371, df = 380, p-value < 2.2e-16
```

Top 6 Varieties and Countries

Currently in our data we have too many levels of categorical variables. It will make sense to do some analysis on the most popular wine varieties and the most popular countries.

We first analyzed the 6 most popular varieties:

```
wine_structured.df.most.reviewed<- wine_structured.df[wine_structured.df$variety=="Red Blend" | wine_structured.df$variety=="Chardonnay"
wine_structured.df$variety == "Pinot Noir" | wine_structured.df$variety=="Cabernet Sauvignon" | wine_structured.df$variety== "Riesling"
|wine_structured.df$variety== "Bordeaux-style Red Blend", ]
summary(wine_structured.df.most.reviewed)
wine.most.avg<-aggregate(points~variety, data=wine_structured.df.most.reviewed,mean)
wine.most.avg
wine.most.sd<-wine.most.avg<-aggregate(points~variety, data=wine_structured.df.most.reviewed,sd)
wine.most.sd
wine.price.avg<-aggregate(price~variety, data=wine_structured.df.most.reviewed,mean)
wine.price.avg
wine.price.iqr<-aggregate(price~variety, data=wine_structured.df.most.reviewed,IQR)
wine.price.iqr
```

Mean Points by variety

variety <fctr>	points <dbl>
Bordeaux-style Red Blend	88.79213
Cabernet Sauvignon	88.61059
Chardonnay	88.30298
Pinot Noir	89.40885
Red Blend	88.37978
Riesling	89.43805

Std Dev Points by variety

variety <fctr>	points <dbl>
Bordeaux-style Red Blend	3.075173
Cabernet Sauvignon	3.321676
Chardonnay	3.234521
Pinot Noir	3.131741
Red Blend	2.800811
Riesling	2.856380

The mean and std dev aggregates for top 6 varieties of wine show that the average points do not have much difference across varieties. Which means that highly rated wines are not specific to a certain variety of wine. Riesling and Pinot Noir have the highest avg. points among the most popular varieties.

For price, we could not use std. Dev. as the measure of spread as the distribution for price is not normal and so we decided to use interquartile range.

Means

variety <fctr>	price <dbl>
Bordeaux-style Red Blend	47.21086
Cabernet Sauvignon	47.94002
Chardonnay	34.52202
Pinot Noir	47.52890
Red Blend	35.88119
Riesling	32.00040

IQR

variety <fctr>	price <dbl>
Bordeaux-style Red Blend	30.25
Cabernet Sauvignon	45.00
Chardonnay	22.00
Pinot Noir	27.00
Red Blend	27.00
Riesling	18.00

To dig further into the analysis of top 6 varieties, we used linear regression models taking points as the dependent variable and variety and price as the independent variables.

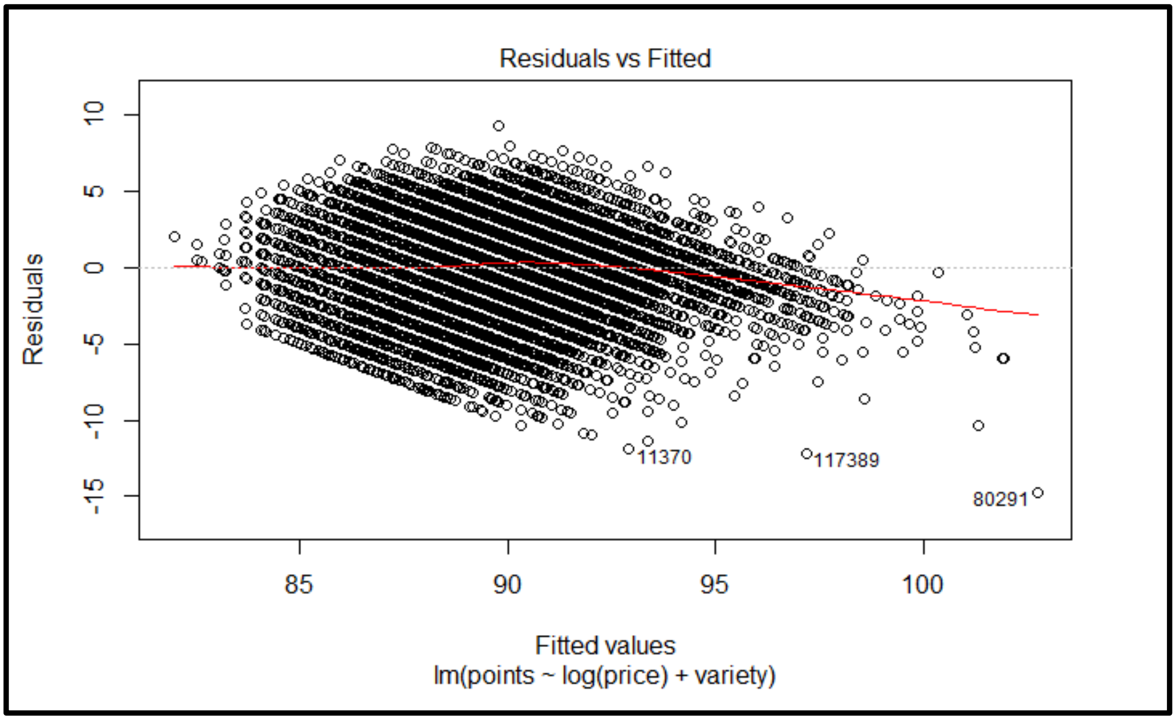
```
linear.0<-lm(points~price+variety, data=wine_structured.df.most.reviewed)
summary(linear.0)
plot(linear.0)

linear.1<-lm(points~variety, data=wine_structured.df.most.reviewed)
summary(linear.1)
linear.2<-lm(log(price)~variety, data=wine_structured.df.most.reviewed)
summary(linear.2)
linear.3<-lm(points~log(price)+variety, data=wine_structured.df.most.reviewed)
summary(linear.3)
linear.4<-lm(points~log(price), data=wine_structured.df.most.reviewed)
summary(linear.4)
linear.5<-lm(points~log(price), data=wine_structured.df)
summary(linear.5)

plot(linear.3)
```

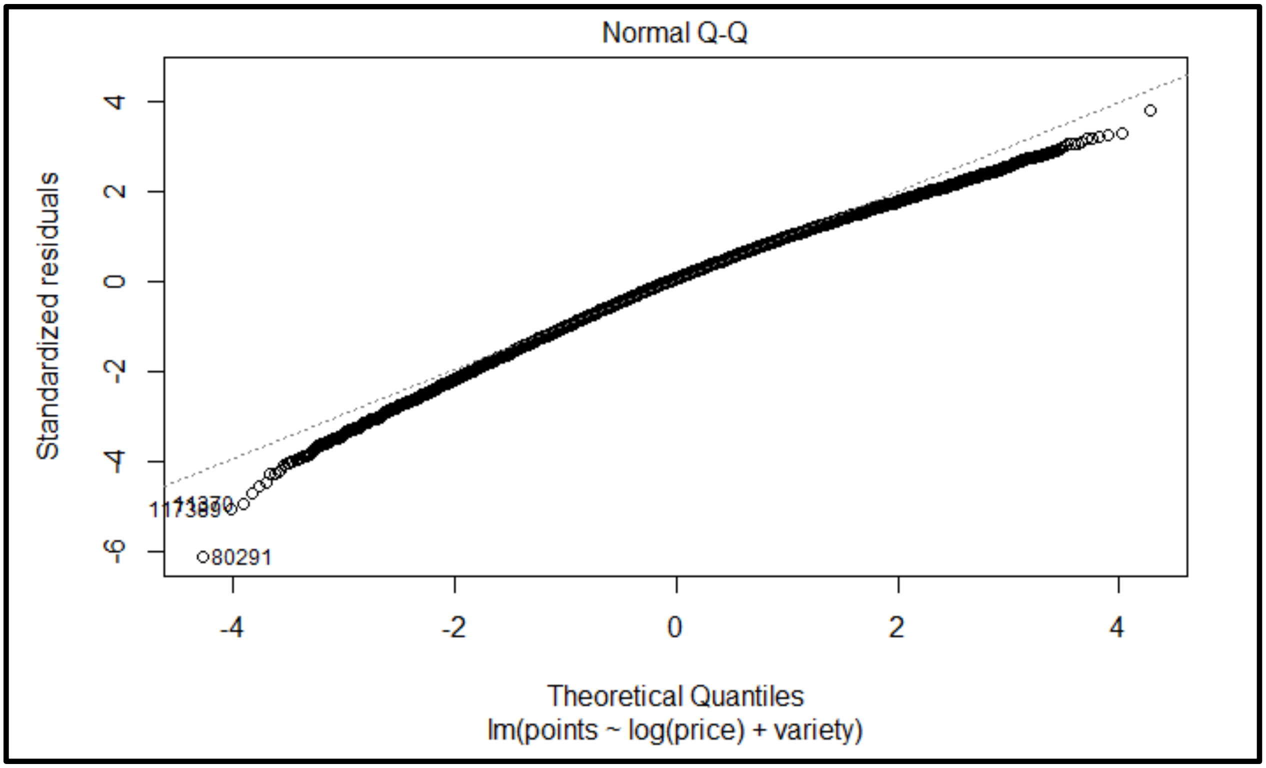
The Residual curves of the first linear model that we created showed us that we will need to take log of price variable. Hence, for all the models, price was substituted as log(price).

```
Call:
lm(formula = points ~ log(price) + variety, data = wine_structured.df.most.reviewed)

Residuals:
     Min       1Q   Median       3Q      Max 
-14.7454  -1.5222   0.1591   1.7052   9.2361 

Coefficients:
                          Estimate Std. Error  t value Pr(>|t|)    
(Intercept)               78.22399    0.06583 1188.295   <2e-16 ***
log(price)                 3.02671    0.01632  185.516   <2e-16 ***
varietyCabernet Sauvignon -0.45028    0.04135  -10.891   <2e-16 ***
varietyChardonnay          0.08631    0.04028    2.143   0.0321 *  
varietyPinot Noir          0.03511    0.03940    0.891   0.3728    
varietyRed Blend          -0.03980    0.04217   -0.944   0.3452    
varietyRiesling            1.50170    0.04773   31.461   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.411 on 52034 degrees of freedom
Multiple R-squared:  0.4114,    Adjusted R-squared:  0.4113 
F-statistic:  6061 on 6 and 52034 DF,  p-value: < 2.2e-16
```

- 41% of the variation in points can be explained by variation in price and variety.
- Coefficients for Chardonnay, Pinot Noir and Red blend are not statistically significant.
- Going from Bordeaux Red Blend to Cabernet Sauvignon will lead to decrease in points by .45 units and going to Riesling will increase points by 1.5 units.
- 1% change in price will change points by 3 units

The residual analytics graphs show that this a good linear model:





Similar analysis was done for Top 6 countries.

Aggregates:

```
wine_structured.df.most.reviewed.country<- wine_structured.df[wine_structured.df$country=="US" | wine_structured.df$country=="France" |
wine_structured.df$country == "Italy" | wine_structured.df$country=="Spain" | wine_structured.df$country== "Portugal"
|wine_structured.df$country== "Chile", ]
summary(wine_structured.df.most.reviewed.country)
wine.most.avg.country<-aggregate(points~country, data=wine_structured.df.most.reviewed.country,mean)
wine.most.avg.country
wine.most.sd.country<-wine.most.avg<-aggregate(points~country, data=wine_structured.df.most.reviewed.country,sd)
wine.most.sd.country
wine.price.avg.country<-aggregate(price~country, data=wine_structured.df.most.reviewed.country,mean)
wine.price.avg.country
wine.price.iqr.country<-aggregate(price~country, data=wine_structured.df.most.reviewed.country,IQR)
wine.price.iqr.country
```

Mean Points by country

country <fctr>	points <dbl>
Chile	86.49547
France	88.73487
Italy	88.61819
Portugal	88.31672
Spain	87.29073
US	88.56639

Std Dev Points by variety

country <fctr>	points <dbl>
Chile	2.700443
France	3.012972
Italy	2.660785
Portugal	3.016879
Spain	3.070916
US	3.116825

Means

country <fctr>	price <dbl>
Chile	8
France	27
Italy	32
Portugal	16
Spain	17
US	25

IQR

country <fctr>	price <dbl>
Chile	20.78646
France	41.13912
Italy	39.66377
Portugal	26.21826
Spain	28.21527
US	36.57346

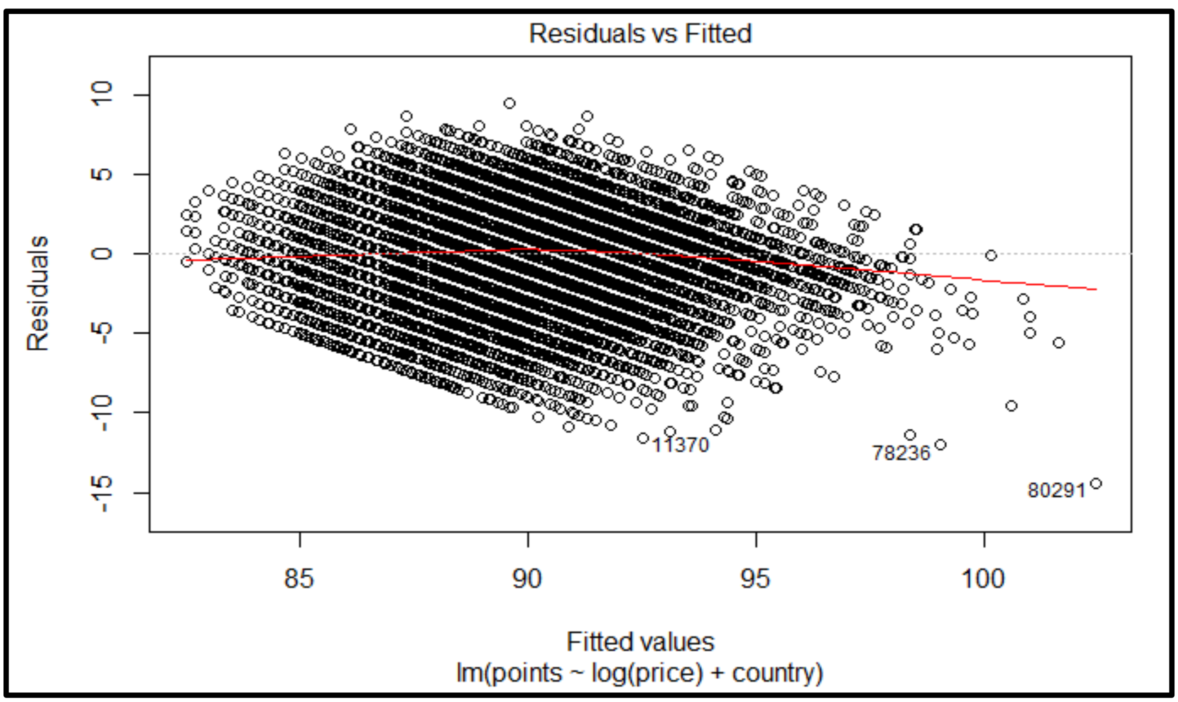
Linear regression:

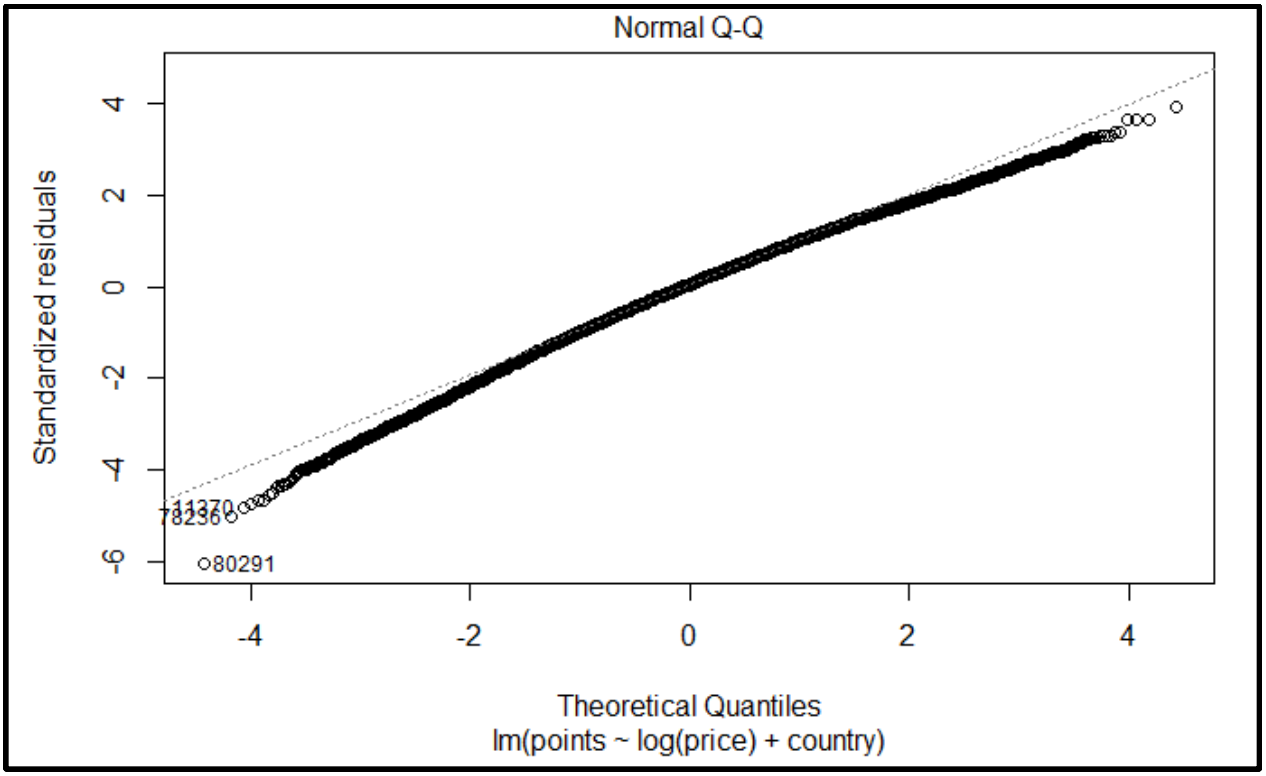
```
Call:
lm(formula = points ~ log(price) + country, data = wine_structured.df.most.reviewed.country)

Residuals:
     Min       1Q   Median       3Q      Max 
-14.4160  -1.5056   0.0926   1.6756   9.4017 

Coefficients:
                Estimate Std. Error  t value Pr(>|t|)    
(Intercept)     78.38369    0.04890 1603.061  < 2e-16 ***
log(price)       2.87388    0.01171  245.474  < 2e-16 ***
countryFrance    0.74905    0.04072   18.394  < 2e-16 ***
countryItaly     0.38844    0.04108    9.455  < 2e-16 ***
countryPortugal  1.46313    0.04977   29.395  < 2e-16 ***
countrySpain     0.16118    0.04667    3.454 0.000554 ***
countryUS        0.33926    0.03814    8.896  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.395 on 104812 degrees of freedom
Multiple R-squared:  0.3835,	Adjusted R-squared:  0.3835 
F-statistic: 1.087e+04 on 6 and 104812 DF,  p-value: < 2.2e-16
```

The model shows that all countries have higher points than Chile.

Also, 1% change in price will change points by 2.87 units.

Residual Analytics:





We also did anova to confirm that the linear models with variety and country are better than the model with only price :

```
> anova(linear.3,linear.4)
Analysis of Variance Table

Model 1: points ~ log(price) + variety
Model 2: points ~ log(price)
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1  52034 302388
2  52039 315003 -5    -12616 434.17 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> anova(linear.8, linear.9)
Analysis of Variance Table

Model 1: points ~ log(price) + country
Model 2: points ~ log(price)
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1 104812 601171
2 104817 609912 -5   -8741.4 304.81 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Clustering

Based on points and price there should be some segments or clusters of wines that we can uncover.

Knowing about these clusters can help in positioning new wines for a winery.

We first use the elbow method to see the ideal number of segments which gave us 5. Taking k as 5 we identified the clusters of wine and tried to interpret these clusters.

We interpreted the clusters as:

1. Connoisseur's Choice Wine
2. Premium Wine
3. Regular Wine
4. Luxury Wine
5. Budget Wine

Finally, we used cl_predict to come up with a way for a winery to predict which cluster a new wine is a member of.

K-means is unsupervised learning. So, if there is a large number of new data, reclustering would be required.

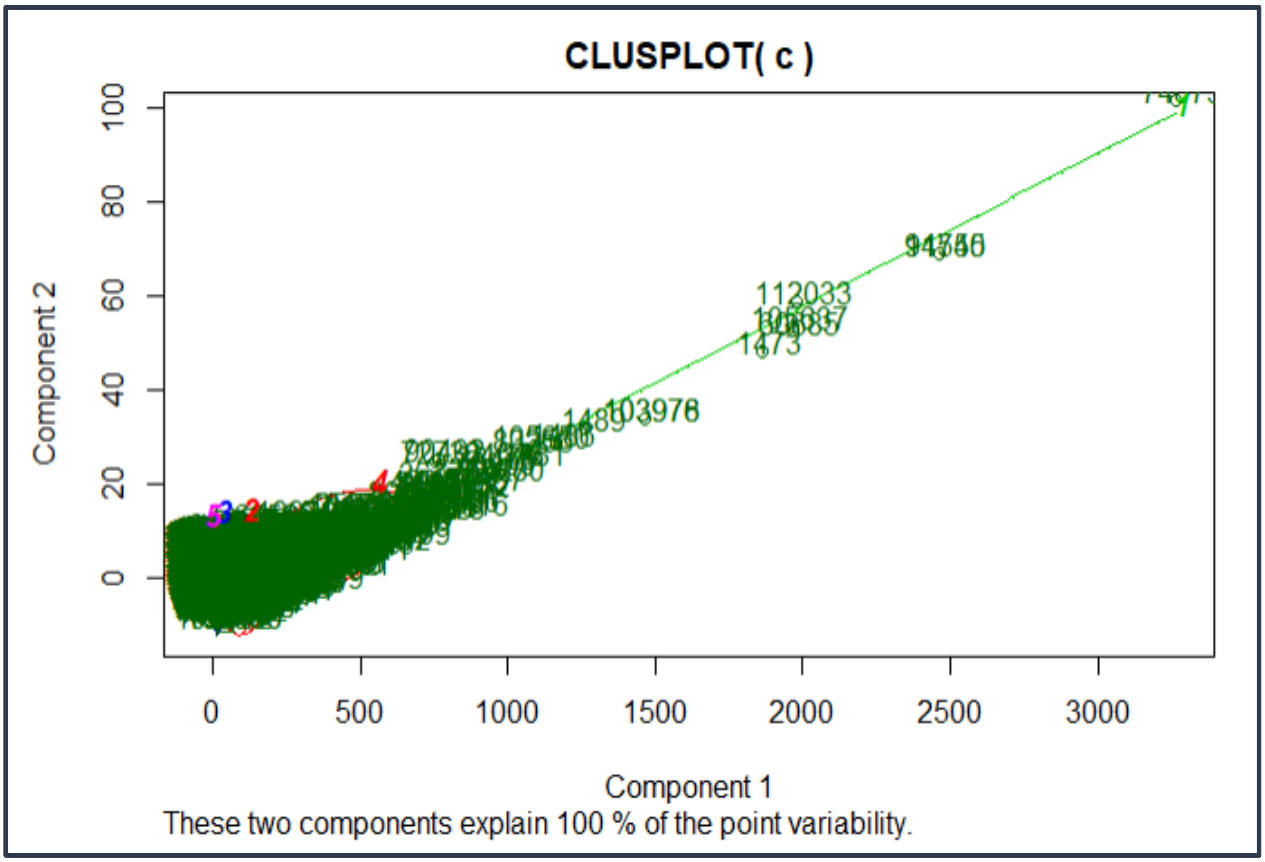
```
wss<-sapply(1:20, function(k){kmeans(c, k)$tot.withinss})
plot(1:20, wss,
     type="b", pch = 19, frame = FALSE,
     xlab="Number of clusters K",
     ylab="Total within-clusters sum of squares")
bet<-sapply(1:20, function(k){kmeans(c, k)$betweens/kmeans(c, k)$totss})
plot(1:20, bet,
     type="b", pch = 19, frame = FALSE,
     xlab="Number of clusters K",
     ylab="percentage of variance explained")
k.wine<-kmeans(c, 5)
k.wine$centers
clusplot(c, k.wine$cluster, color=TRUE, shade=TRUE,
         labels=2, lines=0)
k.wine$size
library(cluster)
library(fpc)
plotcluster(c, k.wine$cluster)
predict.cluster<-cl_predict(k.wine, newdata=c[200:210,], type = "memberships")
predict.cluster
c[200:210,]
```

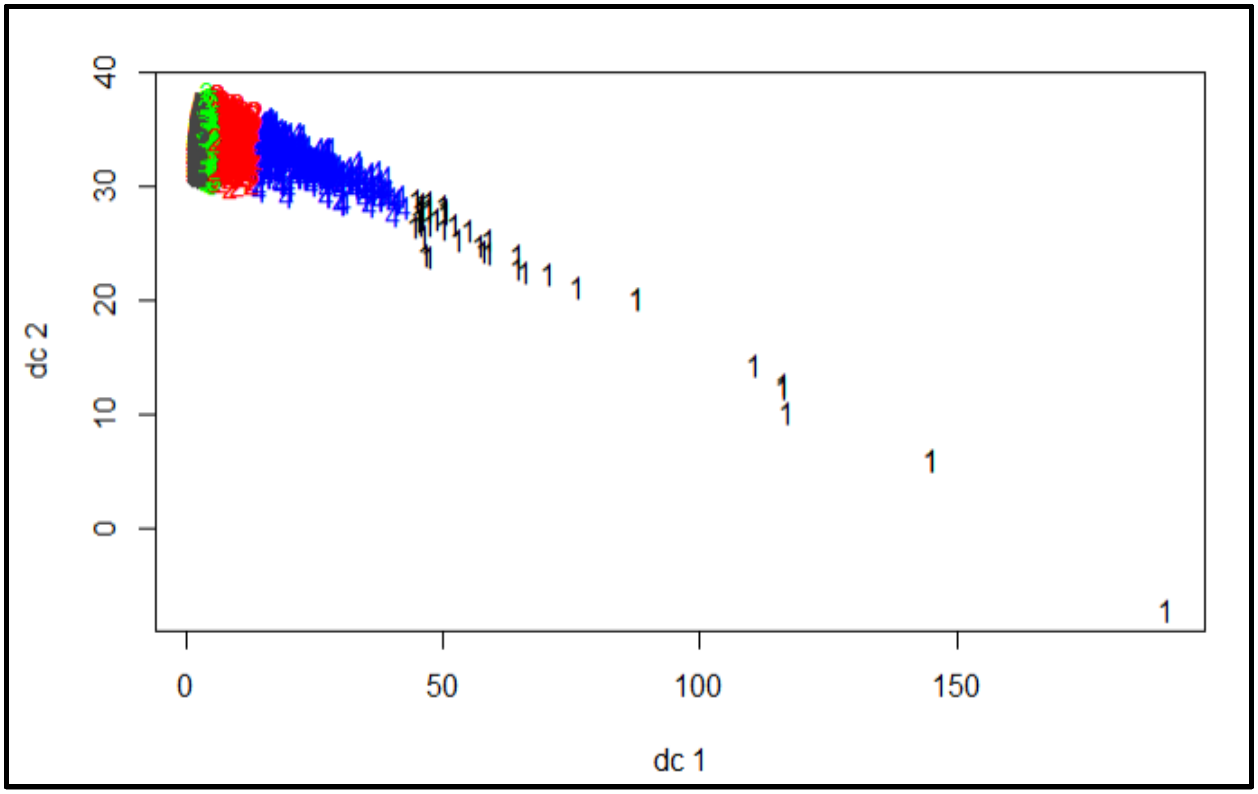
```
> k.wine$size
[1]    43  8349 35607   596 76380
```

```
k.wine<-kmeans(c, 5)
k.wine$centers
wine_structured.df$price wine_structured.df$points
              1143.60465                  95.39535
               101.82034                  91.81710
                47.37462                  90.01604
               334.98490                  94.31040
                19.53776                  87.25771
```

```
> predict.cluster<-cl_predict(k.wine, newdata=c[200:210,], type = "memberships")
> predict.cluster
Memberships:
     1 2 5
 [1,] 0 1 0
 [2,] 0 1 0
 [3,] 0 0 1
 [4,] 0 1 0
 [5,] 1 0 0
 [6,] 1 0 0
 [7,] 0 1 0
```

Cluster Visualizations:





Conclusion and Recommendations:

1. England has the greatest number of highest rated wines out of the top 8 countries; therefore, the wineries can look up to England style of winemaking.
2. Apart from one out of the top 10 costliest wines, all wines originated from France.
3. The number of red wines being reviewed more is an indication for the wineries of the popularity of red against white wines.
4. For a winemaker: the ideal country to open up a new winery is USA as it produces great white and red wine. The wines produced in the USA are well priced thereby guaranteed high revenues.
5. Country of origin will have a stronger influence on ratings than the variety.
6. Depending on the rating, the wineries can have a high price range going from budget to luxury wines in each type or variety.
7. Out of all the varieties Reisling is likely to get higher ratings from sommeliers.
8. The rain forest model created can be used by both winemakers and buyers to predict the price (pricing strategy) of the wine depending upon the country of origin, type and points.
9. The K-means model can be to segment the wines they produce based on the price and points and use this understanding to position the new wine to right market.