Final Project Submission

Vikas Ranjan

2/24/2020

Wine Reviews

Introduction

Wine, a much loved alcoholic drink has been produced and enjoyed since thousands of years. It is typically made from Sented grapes. Different varieties of grapes and strains of yeasts produce different styles of wine. This dataset consists of details of 129971 wines reviews produced across the globe by different wineries. The dataset consists of country, wine description, designation, points, price, province, region_1, region_2, taster_name, taster_twitter_handle, title, variety, winery. We would be looking at various aspects of this data to uncover some insights.

- Dataset - Wine Reviews

Problem statements

Determine which countries/region produces best wines? Determine which wineries produces best wines? Determine which countries/region produces costly wines? Determine which countries/region produces economical wines? Determine which countries/region produces economical and high quality wines? Determine which states are producing most wines? Determine which countries are producing best and worst wines? Establish corelation between a price of wines and points scored by the wine?

Summarize how you addressed this problem statement (the data used and the methodology employed).

First step of the process was to analyse the data and understand what all data fields were consistent and relevent to the questions I was looking to answer. I removed the first column which was just the sequential number and non relevent to data analysis. I retrieved the percentage of missing data per column and found that region_2 is missing for more than 50% of the observations. Therefore I'll not be using this column for analysis and removed it from dataset. For missing price, I applied mean value which is 35 in this case. To make the dataset consistent, removed 63 observations which had country and province as a NA and removed 1 observation which had variety as a NA. For missing region_1, applied "ALL". All these steps helped to get a clean, relevant and consistent dataset. With clean dataset in hand, I created various plots based on my questions. Also, did perform correlation between price and points.

Summarize the interesting insights that your analysis provided.

- US, France, Italy, Portugal, Australia, Germany, Spain and Austria are the best wine producing counties based on count of wines scoring more than 96 points with US topping the list.

- France by far produces the costliest wines in the world.
- US produces most economical wines.
- US and Protugal are top 2 countries producing highly rated economical wines.
- California, Washington, Bordeaux and Tuscany are most wine producing regions, with california topping the list.
- Austrian wines have the best mean scores and Chile has the lowest mean scores.
- Price and points have a strong positive correlation.

Summarize the implications to the consumer (target audience) of your analysis.

If you are wine lover and don't mind spending big bucks, French wines are to choose from. Otherwise, US wines are a good fit with higher point scores and affordable prices. Similar analogy would apply from business side. If a business/retailer serves upscale customers, they would probably source French wines otherwise US or Portugal wines.

Discuss the limitations of your analysis and how you, or someone else, could improve or build on it.

I would have loved to build a predictive algorithm however, I felt that it would have been more accurate and effective if I had more varibles such as age of wine, type of grapes, climate of the region, complexity, balance, etc.

R scripts to cleanup data and perform analysis

- **Structure of wineRawData Dataframe:**

```
## 'data.frame':    129971 obs. of  14 variables:
##  $ X                     : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ country               : chr  "Italy" "Portugal" "US" "US" ...
##  $ description           : chr  "Aromas include tropical fruit, broom, brimstone and dried herb. The p
##  $ designation           : chr  "VulkÃ  Bianco" "Avidagos" NA "Reserve Late Harvest" ...
##  $ points                : int  87 87 87 87 87 87 87 87 87 87 ...
##  $ price                 : int  NA 15 14 13 65 15 16 24 12 27 ...
##  $ province              : chr  "Sicily & Sardinia" "Douro" "Oregon" "Michigan" ...
##  $ region_1              : chr  "Etna" NA "Willamette Valley" "Lake Michigan Shore" ...
##  $ region_2              : chr  NA NA "Willamette Valley" NA ...
##  $ taster_name           : chr  "Kerin Oâ\200\231Keefe" "Roger Voss" "Paul Gregutt" "Alexander Peartre
##  $ taster_twitter_handle: chr  "@kerinokeefe" "@vossroger" "@paulgwineÂ " NA ...
##  $ title                 : chr  "Nicosia 2013 VulkÃ  Bianco  (Etna)" "Quinta dos Avidagos 2011 Avidag
##  $ variety               : chr  "White Blend" "Portuguese Red" "Pinot Gris" "Riesling" ...
##  $ winery                : chr  "Nicosia" "Quinta dos Avidagos" "Rainstorm" "St. Julian" ...
```

Data cleanup

1. 1st column is just the sequential number which is non relevent to data analysis, therefore removing it.

```
wineRawData1 = wineRawData %>%  select(-X)
```

2. Find the columns with missing values:

- Following are the columns with missing data:**

```
colnames(wineRawData1)[apply(wineRawData1, 02, anyNA)]
```

```
## [1] "country"               "designation"           "price"
## [4] "province"              "region_1"              "region_2"
## [7] "taster_name"           "taster_twitter_handle" "variety"
```

3. Retrieve the percentage of missing data per column:

```
missing_data = wineRawData1 %>%
    map_df(function(x) sum(is.na(x))) %>%
    gather(feature, num_nulls)%>%
    arrange(desc(num_nulls))%>%
    mutate(percent_missing = num_nulls/nrow(wineRawData1)*100)
missing_data
```

```
## # A tibble: 13 x 3
##    feature               num_nulls percent_missing
##    <chr>                     <int>           <dbl>
##  1 region_2                  79460       61.1
##  2 designation               37465       28.8
##  3 taster_twitter_handle     31213       24.0
##  4 taster_name               26244       20.2
##  5 region_1                  21247       16.3
##  6 price                      8996        6.92
##  7 country                      63        0.0485
##  8 province                     63        0.0485
##  9 variety                       1        0.000769
## 10 description                   0        0
## 11 points                        0        0
## 12 title                         0        0
## 13 winery                        0        0
```

4. Looking at the above statistics, region_2 is missing for more than 50% of the observations. Therefore we will not be using this column for analysis and can be dropped.

```
# Remove region_2
wineRawData2 = wineRawData1 %>%  select(-region_2)
```

5. For missing price, we will be applying mean value.

The mean price of wines is coming as 35.36. So we will apply 35 to observations where price is missing.

```
##   avgPrice
## 1 35.36339
```

6. We will be removing 63 observations which have country and province as a NA.

```
wineRawData3 <- subset(wineRawData2, ((!is.na(wineRawData2$province))|(!is.na(wineRawData2$country))))
```

7. We will be removing 1 observation which has variety as a NA.

```
## 'data.frame':    129907 obs. of  12 variables:
##  $ country              : chr  "Italy" "Portugal" "US" "US" ...
##  $ description          : chr  "Aromas include tropical fruit, broom, brimstone and dried herb. The p
##  $ designation          : chr  "VulkÃ  Bianco" "Avidagos" NA "Reserve Late Harvest" ...
##  $ points               : int  87 87 87 87 87 87 87 87 87 87 ...
##  $ price                : num  35 15 14 13 65 15 16 24 12 27 ...
##  $ province             : chr  "Sicily & Sardinia" "Douro" "Oregon" "Michigan" ...
##  $ region_1             : chr  "Etna" NA "Willamette Valley" "Lake Michigan Shore" ...
##  $ taster_name          : chr  "Kerin Oâ\200\231Keefe" "Roger Voss" "Paul Gregutt" "Alexander Peartr
##  $ taster_twitter_handle: chr  "@kerinokeefe" "@vossroger" "@paulgwineÃ " NA ...
##  $ title                : chr  "Nicosia 2013 VulkÃ  Bianco  (Etna)" "Quinta dos Avidagos 2011 Avidag
##  $ variety              : chr  "White Blend" "Portuguese Red" "Pinot Gris" "Riesling" ...
##  $ winery               : chr  "Nicosia" "Quinta dos Avidagos" "Rainstorm" "St. Julian" ...
```

8. For missing region_1, we will be applying "ALL".

```
str(wineDataClean)
```

Structure of the clean data frame

```
## 'data.frame':    129907 obs. of  12 variables:
##  $ country              : chr  "Italy" "Portugal" "US" "US" ...
##  $ description          : chr  "Aromas include tropical fruit, broom, brimstone and dried herb. The p
##  $ designation          : chr  "VulkÃ  Bianco" "Avidagos" NA "Reserve Late Harvest" ...
##  $ points               : int  87 87 87 87 87 87 87 87 87 87 ...
##  $ price                : num  35 15 14 13 65 15 16 24 12 27 ...
##  $ province             : chr  "Sicily & Sardinia" "Douro" "Oregon" "Michigan" ...
##  $ region_1             : chr  "Etna" NA "Willamette Valley" "Lake Michigan Shore" ...
##  $ taster_name          : chr  "Kerin Oâ\200\231Keefe" "Roger Voss" "Paul Gregutt" "Alexander Peartr
##  $ taster_twitter_handle: chr  "@kerinokeefe" "@vossroger" "@paulgwineÃ " NA ...
##  $ title                : chr  "Nicosia 2013 VulkÃ  Bianco  (Etna)" "Quinta dos Avidagos 2011 Avidag
##  $ variety              : chr  "White Blend" "Portuguese Red" "Pinot Gris" "Riesling" ...
##  $ winery               : chr  "Nicosia" "Quinta dos Avidagos" "Rainstorm" "St. Julian" ...
```

Glimpse of the clean dataset.

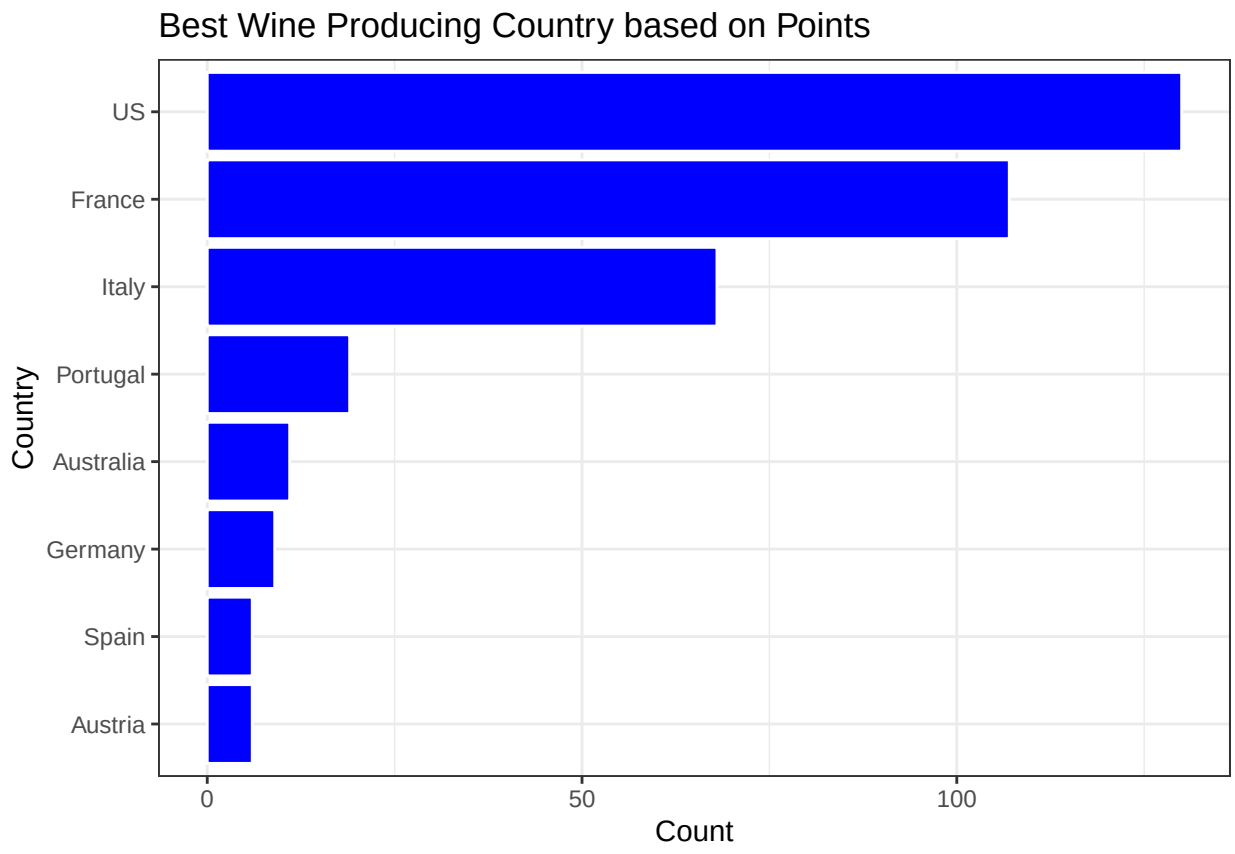
```
glimpse(wineDataClean)
```

```
## Observations: 129,907
## Variables: 12
## $ country              <chr> "Italy", "Portugal", "US", "US", "US", "Spain",...
## $ description          <chr> "Aromas include tropical fruit, broom, brimston...
## $ designation          <chr> "VulkÃ  Bianco", "Avidagos", NA, "Reserve Late ...
```

```
## $ points                <int> 87, 87, 87, 87, 87, 87, 87, 87, 87, 87, 87, 87,...
## $ price                 <dbl> 35, 15, 14, 13, 65, 15, 16, 24, 12, 27, 19, 30,...
## $ province              <chr> "Sicily & Sardinia", "Douro", "Oregon", "Michig...
## $ region_1              <chr> "Etna", NA, "Willamette Valley", "Lake Michigan...
## $ taster_name           <chr> "Kerin Oâ\200\231Keefe", "Roger Voss", "Paul Gr...
## $ taster_twitter_handle <chr> "@kerinokeefe", "@vossroger", "@paulgwineÂ ", N...
## $ title                 <chr> "Nicosia 2013 VulkÃ  Bianco  (Etna)", "Quinta d...
## $ variety               <chr> "White Blend", "Portuguese Red", "Pinot Gris", ...
## $ winery                <chr> "Nicosia", "Quinta dos Avidagos", "Rainstorm", ...
```

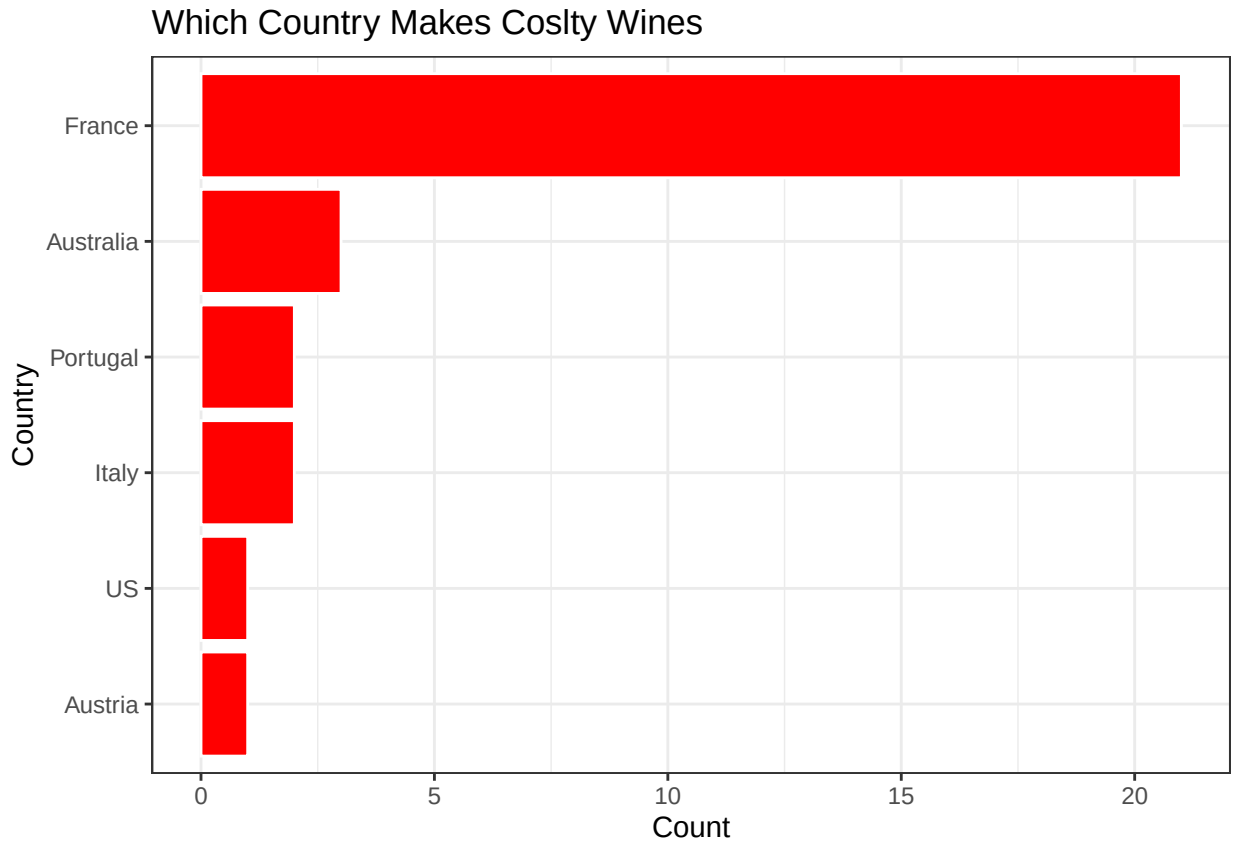
Best wine producing countries (points greater than 96):

```
CountryWiseWines <- wineDataClean %>%
    filter(points > 96) %>%
    group_by(country)%>%
    summarise(n = n()) %>%
    arrange(desc(n)) %>%
    head(n = 8) %>%
    ggplot(aes(x = reorder(country,n), y = n)) +
    geom_bar(stat='identity',colour="white", fill = c("blue")) +
    labs(x = 'Country', y = 'Count', title = 'Best Wine Producing Country based on Points') + coord_fli

CountryWiseWines
```



Countries producing costly wines:

```
Costlywines <- wineDataClean %>%
    arrange(desc(price)) %>%
    head(n = 30)%>%
    group_by(country)%>%
    summarise(n = n())%>%
    arrange(desc(n)) %>%
    ggplot(aes(x =reorder(country,n), y = n)) +
    geom_bar(stat="identity",colour="white", fill = c("red")) +
    labs(x = 'Country', y = 'Count', title = 'Which Country Makes Coslty Wines') + coord_flip() + theme_

Costlywines
```

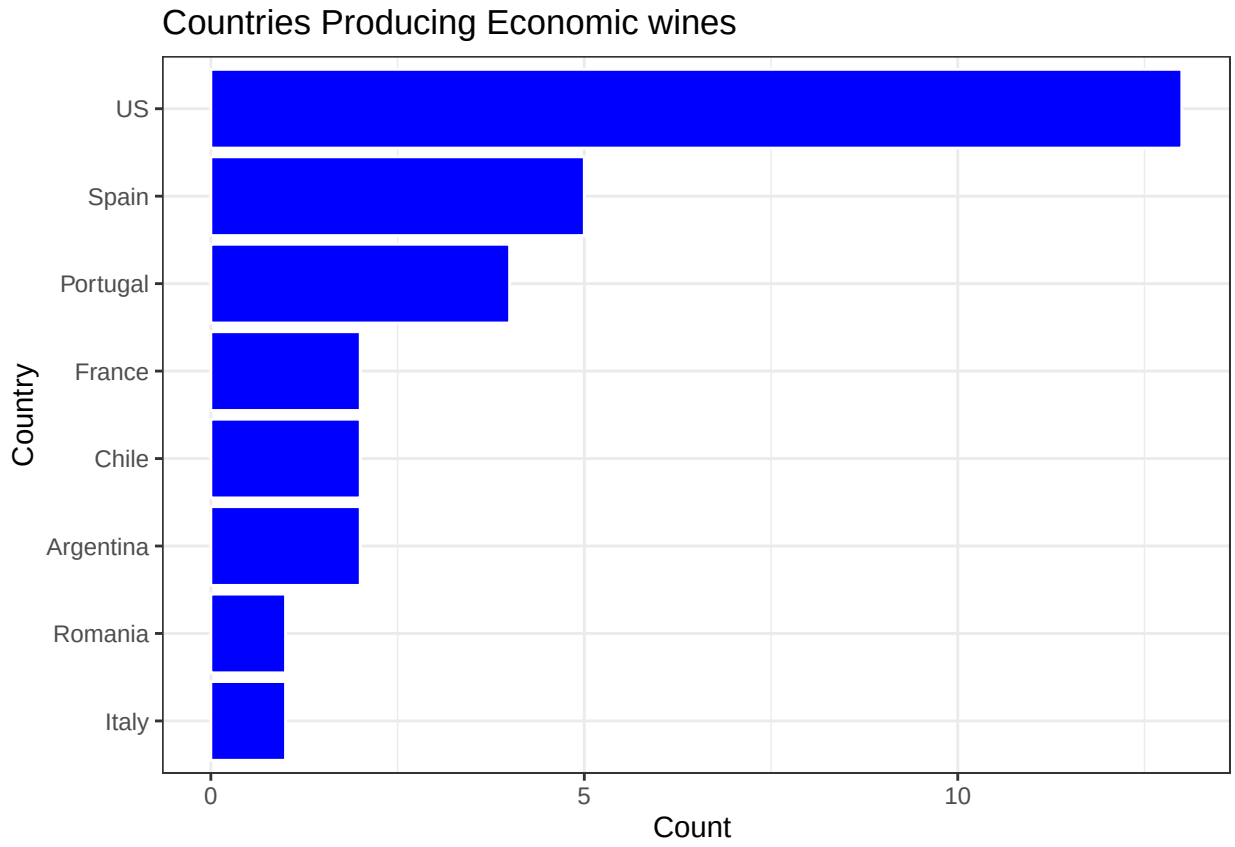


Countries producing Economical wines:

```
EcoWines <- wineDataClean %>%
    arrange(price) %>%
    head(n = 30)%>%
    group_by(country)%>%
    summarise(n = n())%>%
    arrange(desc(n))%>%
    ggplot(aes(x =reorder(country,n), y =  n )) +
```

```
    geom_bar(stat='identity',colour="white", fill = c("blue")) +
    labs(x = 'Country', y = 'Count ', title = 'Countries Producing Economic wines') + coord_flip() + th
```

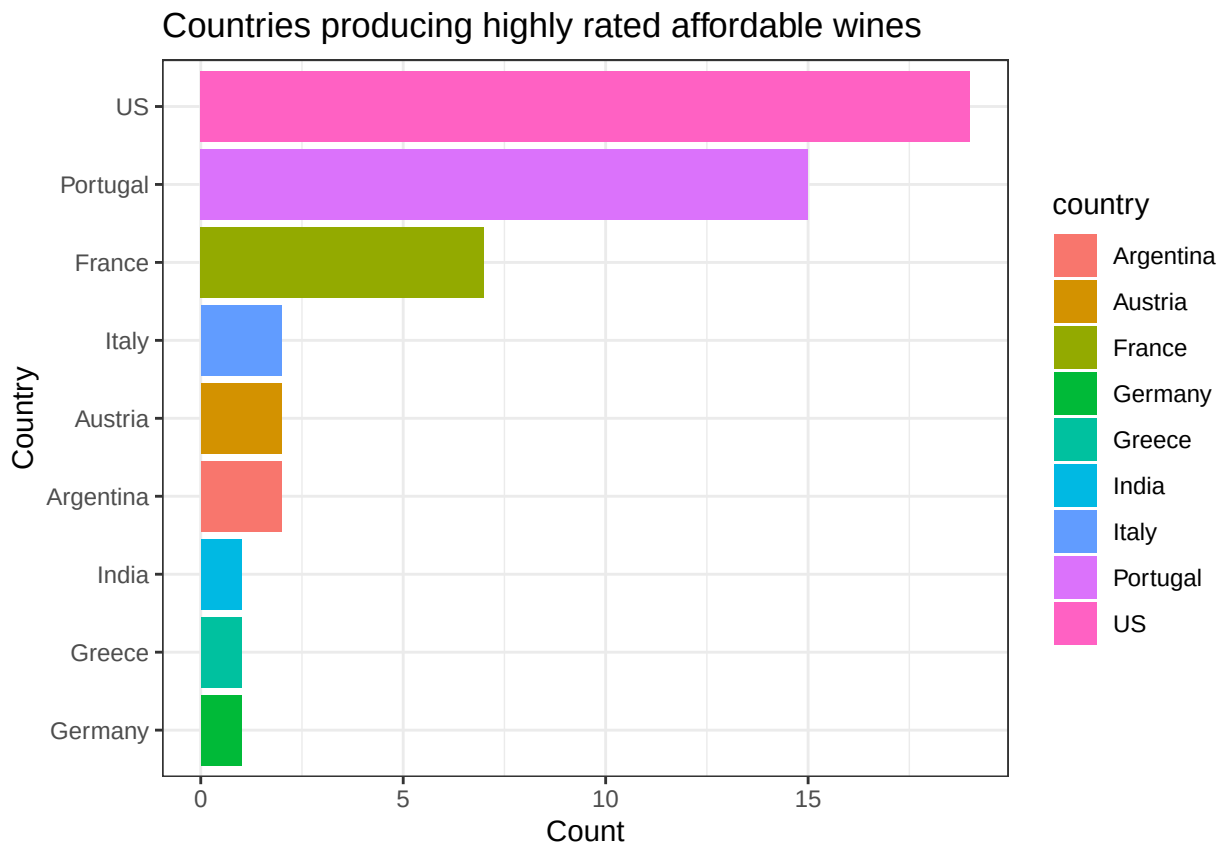
```
EcoWines
```



Countries producing Economical and high quality wines:

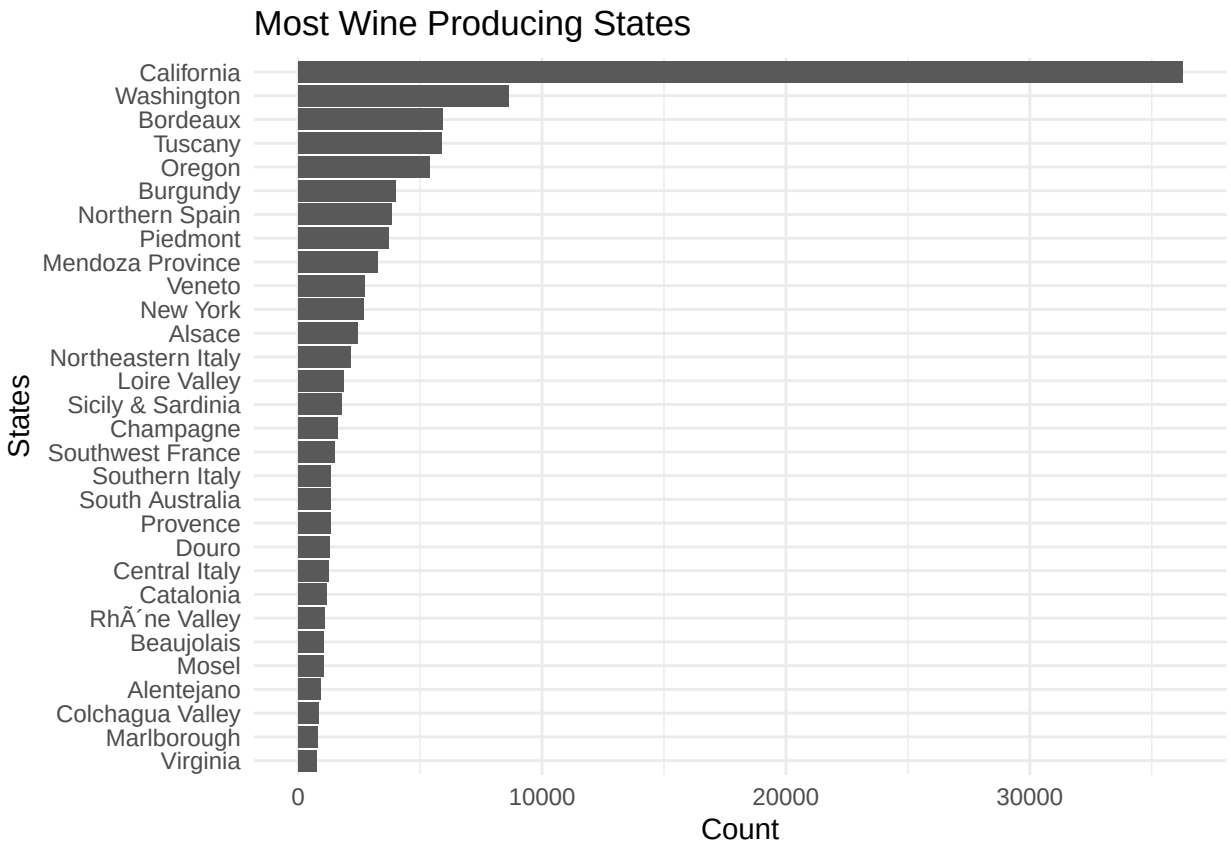
```
GoodEcoWines <- wineDataClean%>%
  filter(points > 90)%>%
  arrange(price) %>%
  head(n = 50) %>%
  group_by(country) %>%
  summarise(n = n())%>%
  arrange(desc(n))%>%
  ggplot(aes(x = reorder(country, n), y = n, fill = country))+
  geom_bar(stat = 'identity')+
  labs(x = 'Country', y ='Count', title ='Countries producing highly rated affordable wines')+
  coord_flip()+
  theme_bw()
```

```
GoodEcoWines
```



States producing most wines:

```
Most_wine_producing_states <- wineDataClean %>%
  group_by(province, country)%>%
  summarise(n = n()) %>%
  arrange(desc(n)) %>%
  head(n = 30)%>%
  ggplot(aes(x = reorder(province, n), y = n, ))+
  geom_bar(stat = 'identity')+
  labs(x = 'States', y ='Count', title ='Most Wine Producing States')+
  coord_flip()+
  theme_minimal()

Most_wine_producing_states
```



Countries with the Best and Worst Wine

```
top_countries = wineDataClean %>%
  group_by(country) %>%
  count() %>%
  filter(n>500)

Best_Wines <- wineDataClean %>%
  filter(country %in% top_countries$country) %>%
  select(country,points) %>%
  group_by(country) %>%
  summarise(Mean_Score = mean(points)) %>%
  arrange(desc(Mean_Score)) %>%
  kable()

Best_Wines
```

country	Mean_Score
Austria	90.10135
Germany	89.85173
France	88.84511
Australia	88.58051
US	88.56372

country	Mean_Score
Italy	88.56223
Israel	88.47129
New Zealand	88.30303
Portugal	88.25022
South Africa	88.05639
Spain	87.28834
Argentina	86.71026
Chile	86.49318

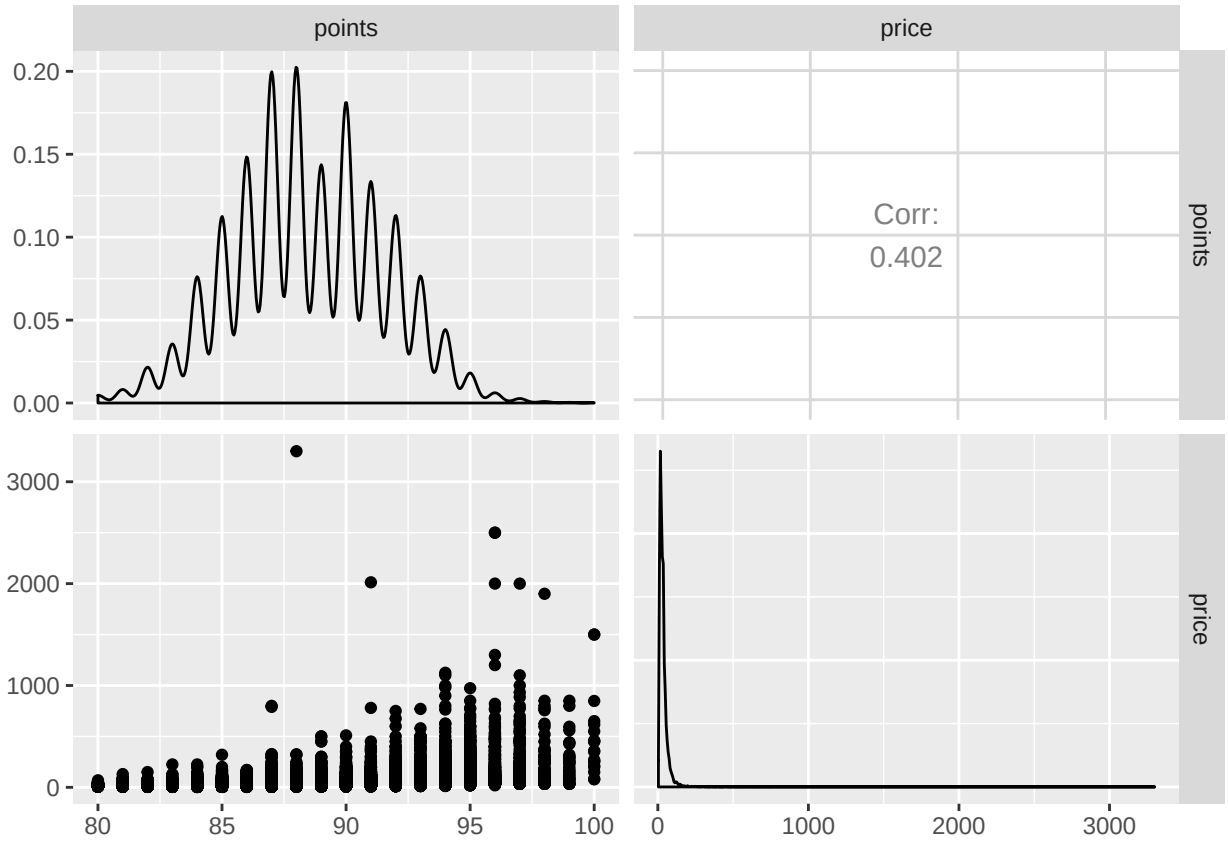
Correlation:

Based on spearman correlation, it seems like there is strong positive correlation between points and price.

```
subset_df <- wineDataClean[, 4:5]
points_df <- subset_df$points
price_df <- subset_df$price

cor_result <- cor.test(price_df, points_df,  method = "spearman", exact=FALSE)
cor_result
```

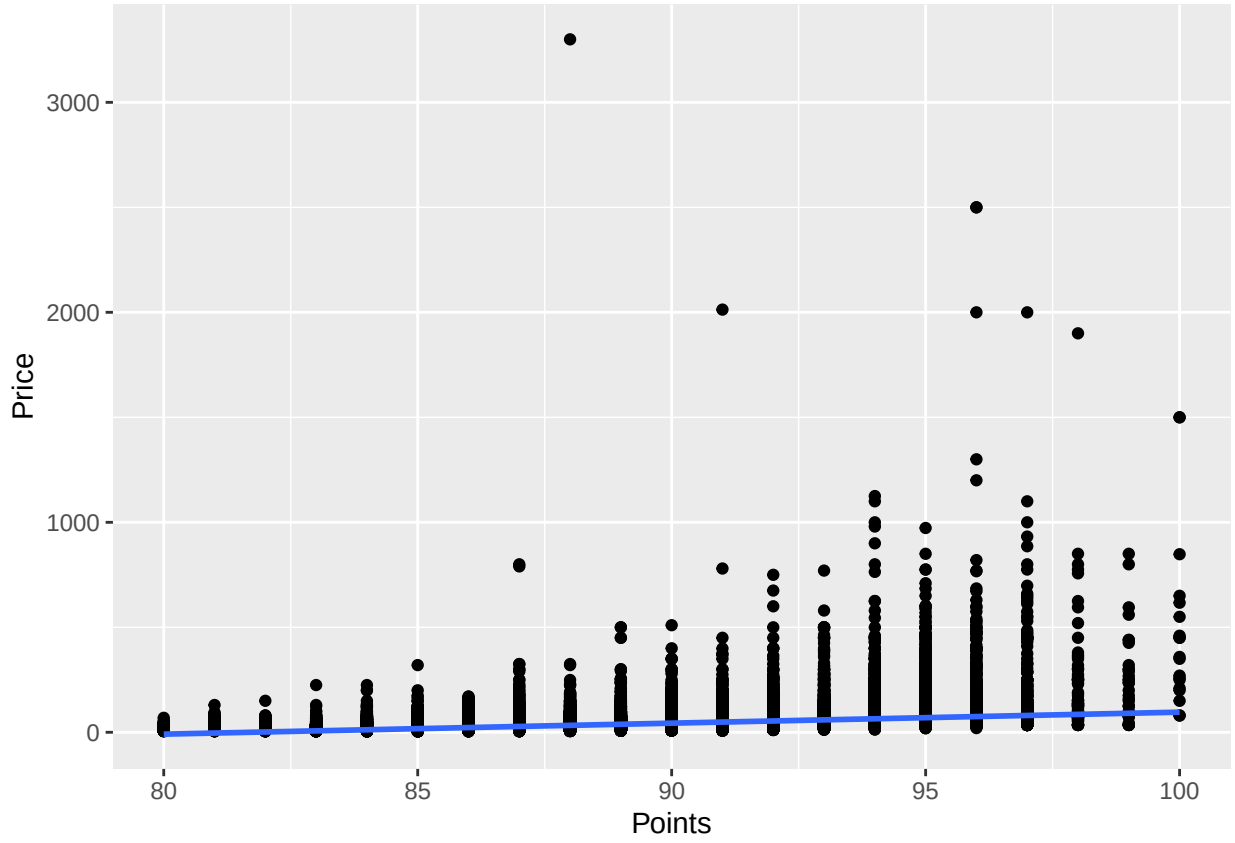
```
##
##  Spearman's rank correlation rho
##
## data:  price_df and points_df
## S = 1.5277e+14, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.5818817
```

```
ggpairs(subset_df)
```



Regression:

```
ggplot(wineDataClean, aes(points, price)) + geom_point() + geom_smooth(method=lm, se = FALSE) + labs(x=
```




```
mod1 <- lm(price ~ points, data = wineDataClean)

mod1
```

```
##
## Call:
## lm(formula = price ~ points, data = wineDataClean)
##
## Coefficients:
## (Intercept)       points
##    -427.759        5.236
```

```
summary(mod1)
```

```
##
## Call:
## lm(formula = price ~ points, data = wineDataClean)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -55.6  -14.2   -4.9    7.2 3267.0
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -427.75947    2.92741  -146.1   <2e-16 ***
```

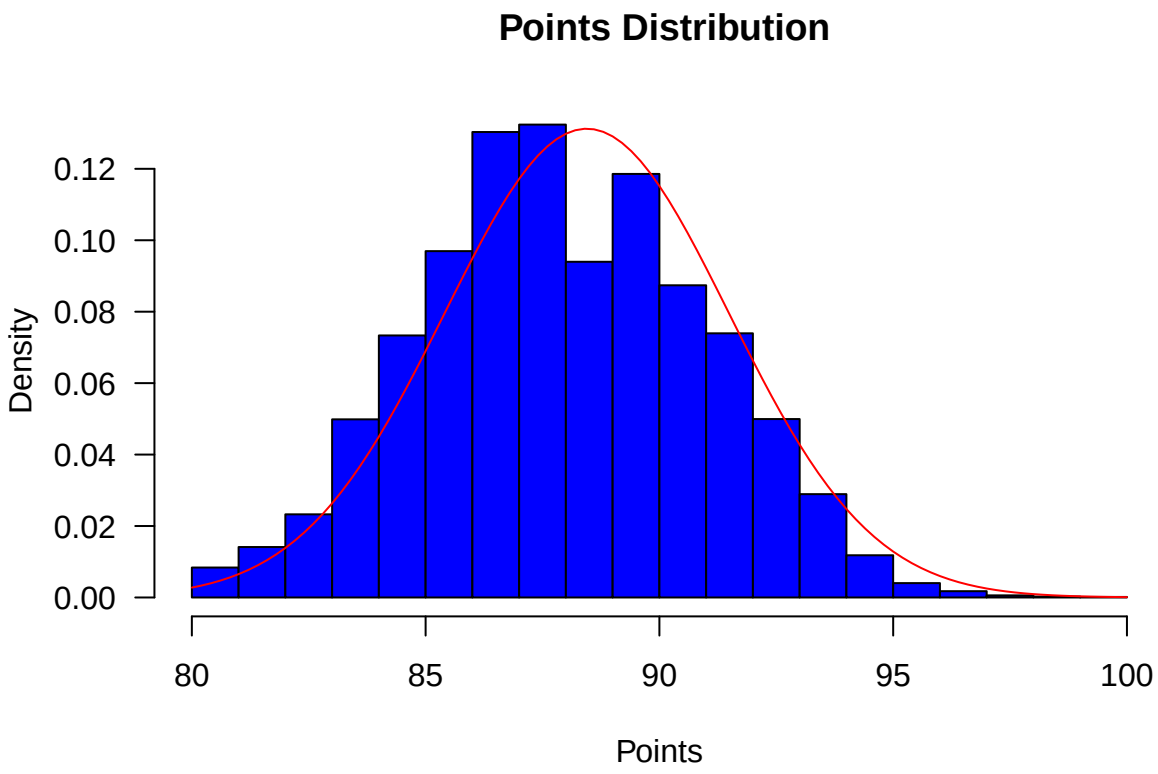
```
## points         5.23593    0.03308   158.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36.24 on 129905 degrees of freedom
## Multiple R-squared:  0.1617, Adjusted R-squared:  0.1617
## F-statistic: 2.506e+04 on 1 and 129905 DF,  p-value: < 2.2e-16
```

Points distribution:

```
summary(wineDataClean$points)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   80.00   86.00   88.00   88.45   91.00  100.00
```

```
hist(wineDataClean$points, freq=FALSE, col="blue", xlab="Points", main=" Points Distribution", las=1)
curve(dnorm(x, mean=mean(wineDataClean$points), sd=sd(wineDataClean$points)), add=TRUE, col="red")
```




```
summary(wineDataClean$price)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.00   18.00   28.00   35.34   40.00 3300.00
```

```r
hist(wineDataClean$price, freq=FALSE, col="blue", xlab="Price", main=" Price Distribution", las=1)
curve(dnorm(x, mean=mean(wineDataClean$price), sd=sd(wineDataClean$price)), add=TRUE, col="red")
```