Exploratory Analysis

YZ Analytics

```r
library(tidyverse)
library(GGally)
```

Countries

It will be important to understand the countries that are represented in our dataset in order to be able to know what types of mapping capabilities we have to have to create a good experience.

```r
path <- "../data/winemag-data-130k-v2.csv"
Wine <- read_csv(path,
                 col_types = cols(
                     X1 = col_double(),
                     country = col_character(),
                     description = col_character(),
                     designation = col_character(),
                     points = col_double(),
                     price = col_double(),
                     province = col_character(),
                     region_1 = col_character(),
                     region_2 = col_character(),
                     taster_name = col_character(),
                     taster_twitter_handle = col_character(),
                     title = col_character(),
                     variety = col_character(),
                     winery = col_character()),
                 progress = FALSE
                 ) %>%
    rename(id = X1)
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

```r
Wine %>% glimpse()
```

```
## Observations: 129,971
## Variables: 14
## $ id                    <dbl> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12...
## $ country               <chr> "Italy", "Portugal", "US", "US", "US", "...
## $ description           <chr> "Aromas include tropical fruit, broom, b...
## $ designation           <chr> "Vulkà Bianco", "Avidagos", NA, "Reserve...
## $ points                <dbl> 87, 87, 87, 87, 87, 87, 87, 87, 87, 87, ...
## $ price                 <dbl> NA, 15, 14, 13, 65, 15, 16, 24, 12, 27, ...
## $ province              <chr> "Sicily & Sardinia", "Douro", "Oregon", ...
## $ region_1              <chr> "Etna", NA, "Willamette Valley", "Lake M...
## $ region_2              <chr> NA, NA, "Willamette Valley", NA, "Willam...
## $ taster_name           <chr> "Kerin O'Keefe", "Roger Voss", "Paul Gre...
## $ taster_twitter_handle <chr> "@kerinokeefe", "@vossroger", "@paulgwin...
## $ title                 <chr> "Nicosia 2013 Vulkà Bianco  (Etna)", "Qu...
## $ variety               <chr> "White Blend", "Portuguese Red", "Pinot ...
## $ winery                <chr> "Nicosia", "Quinta dos Avidagos", "Rains...
```

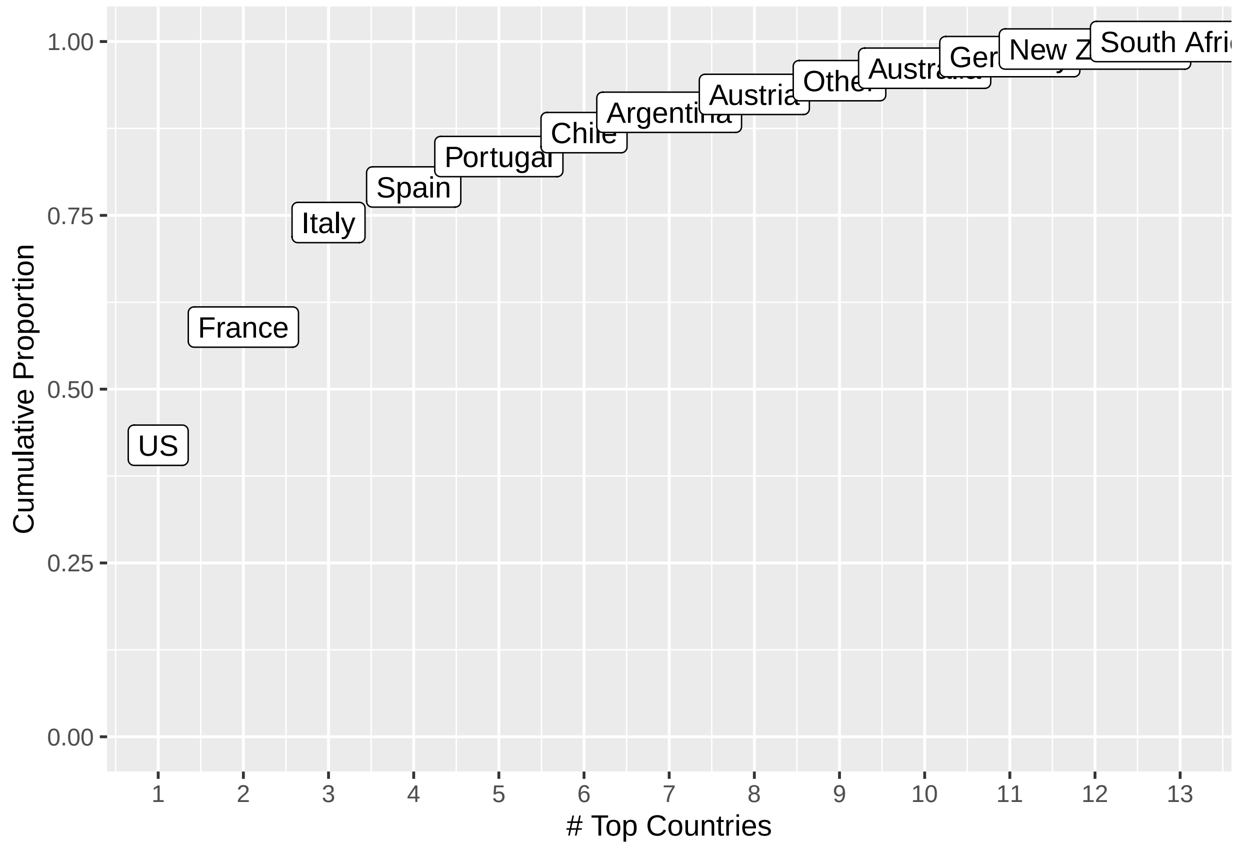
```r
top_countries_tbl <- Wine %>%
    mutate(country = fct_explicit_na(country)) %>%
    mutate(country = fct_lump(country, 12)) %>%
    count(country, sort = TRUE) %>%
    mutate(prop = n / sum(n))

top_countries_tbl
```

```
## # A tibble: 13 x 3
##    country          n   prop
##    <fct>        <int>  <dbl>
##  1 US           54504 0.419
##  2 France       22093 0.170
##  3 Italy        19540 0.150
##  4 Spain         6645 0.0511
##  5 Portugal      5691 0.0438
##  6 Chile         4472 0.0344
##  7 Argentina     3800 0.0292
##  8 Austria       3345 0.0257
##  9 Other         2567 0.0198
## 10 Australia     2329 0.0179
## 11 Germany       2165 0.0167
## 12 New Zealand   1419 0.0109
## 13 South Africa  1401 0.0108
```

The top 13 categories, including the lumped-together category of "Other" consist of those categories which have a count consisting of more than 1% of the observations in the dataset.

```r
top_countries_tbl %>%
    mutate(prop_cumulative = cumsum(prop)) %>%
    ggplot(aes(x = seq_along(country), y = prop_cumulative)) +
    geom_point() +
    geom_label(aes(label = country)) +
    scale_y_continuous(limits = c(0, 1)) +
    scale_x_continuous(breaks = seq(0, 13)) +
    labs(x = "# Top Countries" , y = "Cumulative Proportion")
```



Note that most of the observations, in fact, more than 90% of the observations are contained in the 8 most represented countries and 80% on the top 4, and 60% on the top 2 (USA and France).

It looks like it will be possible to create an interactive map. Now we need to geolocate the wineries. Worst case scenario we have the countries and their representation in the dataset.

Another interesting fact is that since 40% of the observations come from the USA, then perhaps it will be possible to get historical information to add quantitative predictors to our dataset, but it is not crucial since our focus is in the presentation of the data.

Wineries

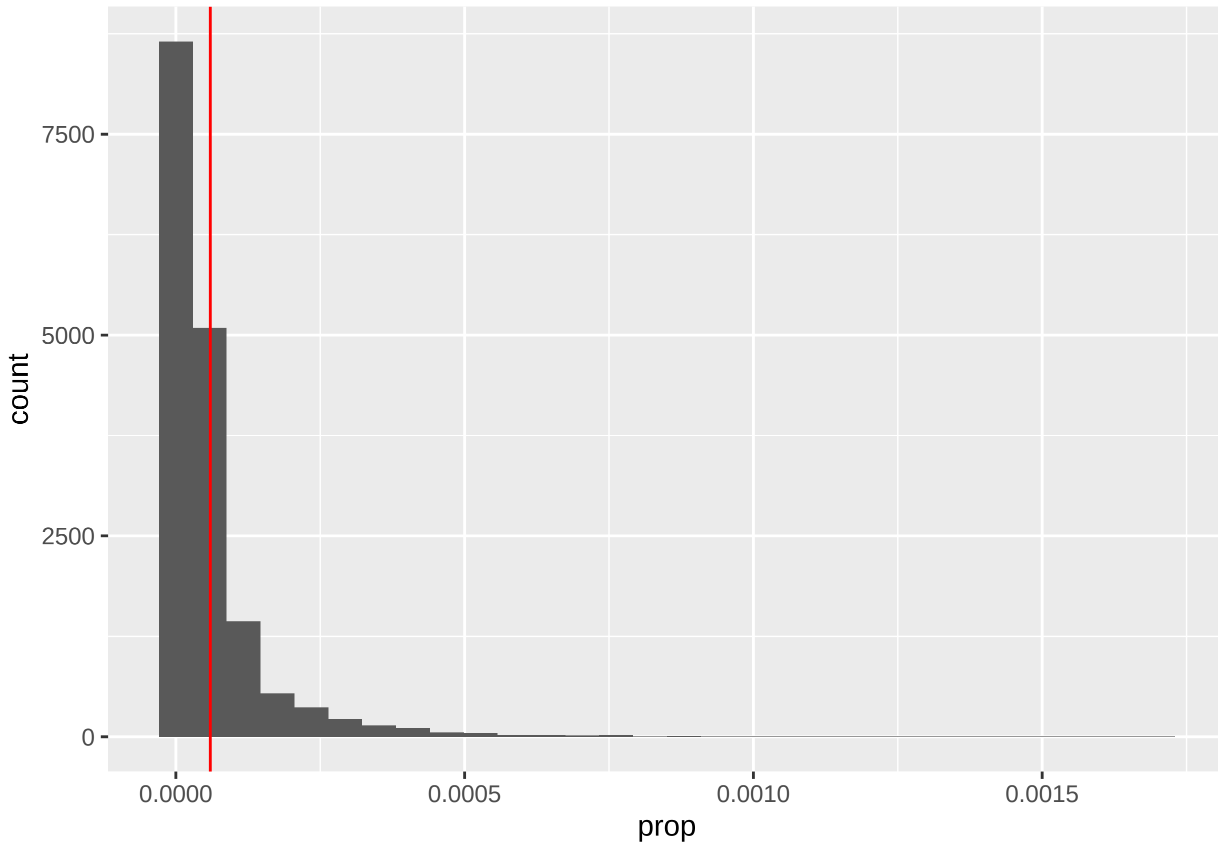
Is there a similar concentration for the wineries? Turns out no. Lumping won't work because there are just so many wineries and there aren't any ones that particularly dominate.

```
top_wineries_tbl <- Wine %>%
    count(winery, sort = TRUE) %>%
    mutate(prop = n / sum(n)) %>%
    mutate(prop_cumulative = cumsum(prop))

mean_prop <- mean(top_wineries_tbl$prop)
mean_prop
```

```
## [1] 5.967655e-05
```

```
top_wineries_tbl %>%
    ggplot(aes(x = prop)) +
    geom_histogram(bins = 30) +
    geom_vline(xintercept = mean_prop, color = "red")
```



The table below summarizes the information for the proportions of each winery.

```r
summary(top_wineries_tbl)
```

```
##     winery                n                prop          
##  Length:16757       Min.   :  1.000   Min.   :7.694e-06  
##  Class :character   1st Qu.:  1.000   1st Qu.:7.694e-06  
##  Mode  :character   Median :  3.000   Median :2.308e-05  
##                     Mean   :  7.756   Mean   :5.968e-05  
##                     3rd Qu.:  8.000   3rd Qu.:6.155e-05  
##                     Max.   :222.000   Max.   :1.708e-03  
##  prop_cumulative  
##  Min.   :0.001708  
##  1st Qu.:0.721630  
##  Median :0.892214  
##  Mean   :0.807783  
##  3rd Qu.:0.967770  
##  Max.   :1.000000
```

From the summary we note that half of the wineries contain more than 90% of the observations, and 25% of the wineries contain more than 70% of the observations. However there are 16757 wineries which means that 25% of the observations is around 4000 wineries.

If we need to geolocate the wineries and we run into trouble, then perhaps only doing half will suffice for our visualization.

Variety

Variety of a wine refers to the type of grape that is used - for example, among white grape wines, there are varieties such as Sauvignon Blanc, Chardonnay, and Riesling. Among red grape wines, some varieties include Merlot, Cabernet Sauvignon, and Pinot Noir.

```
top_varieties_tbl <- Wine %>%
    count(variety, sort = TRUE) %>%
    mutate(prop = n / sum(n)) %>%
    mutate(prop_cumulative = cumsum(prop))

summary(top_varieties_tbl)
```

```
##    variety                  n                 prop
##  Length:708         Min.   :    1.00   Min.   :7.690e-06
##  Class :character   1st Qu.:    2.00   1st Qu.:1.539e-05
##  Mode  :character   Median :    6.00   Median :4.616e-05
##                     Mean   :  183.57   Mean   :1.412e-03
##                     3rd Qu.:   28.25   3rd Qu.:2.174e-04
##                     Max.   :13272.00   Max.   :1.021e-01
##  prop_cumulative
##  Min.   :0.1021
##  1st Qu.:0.9749
##  Median :0.9937
##  Mean   :0.9654
##  3rd Qu.:0.9984
##  Max.   :1.0000
```

```
top_varieties_tbl %>%
  arrange(desc(prop)) %>%
  head()
```

```
## # A tibble: 6 x 4
##   variety                      n   prop prop_cumulative
##   <chr>                    <int>  <dbl>           <dbl>
## 1 Pinot Noir               13272 0.102            0.102
## 2 Chardonnay               11753 0.0904           0.193
## 3 Cabernet Sauvignon        9472 0.0729           0.265
## 4 Red Blend                 8946 0.0688           0.334
## 5 Bordeaux-style Red Blend  6915 0.0532           0.387
## 6 Riesling                  5189 0.0399           0.427
```

We can see from the summary that with among wine variety, 25% of the varieties include more than 97% of of the wines and half of the varieties account for more than 99% of the observed wines. Additionally, Pinot Noir accounts for more than 10% of the wines, followed by Chardonnay with 9.0%, Cabernet Sauvignon with 7.3%, and Red Blend with 6.9%.

Designation

Designation is a tricky variable to work with. It refers to a label placed on the wine by the winemaker in regulation with rules of the country, although not every country has the same rules. For example, the designation of "Reserve" wine generally means the wine has been set aside to age for a longer time than other wines generally would, and it often implies a higher quality. While "Reserva" refers to reserve wines in Spain, and "Riserva" to those in Italy, the two countries have different rules about how long the wine must be aged for in order to receive their respective designations. Other countries, like the U.S., don't have any rules in general. Given this general lack of universality of designation, this variable likely will not mean much in our project, but we can still look at its characteristics.

```r
top_designation_tbl <- Wine %>%
    count(designation, sort = TRUE) %>%
    mutate(prop = n / sum(n)) %>%
    mutate(prop_cumulative = cumsum(prop))

summary(top_designation_tbl)
```

```
##  designation              n                   prop
##  Length:37980       Min.   :    1.00   Min.   :7.690e-06
##  Class :character   1st Qu.:    1.00   1st Qu.:7.690e-06
##  Mode  :character   Median :    1.00   Median :7.690e-06
##                     Mean   :    3.42   Mean   :2.633e-05
##                     3rd Qu.:    2.00   3rd Qu.:1.539e-05
##                     Max.   :37465.00   Max.   :2.883e-01
##  prop_cumulative
##  Min.   :0.2883
##  1st Qu.:0.7374
##  Median :0.8539
##  Mean   :0.8205
##  3rd Qu.:0.9269
##  Max.   :1.0000
```

```r
top_designation_tbl %>%
  arrange(desc(prop)) %>%
  head()
```

```
## # A tibble: 6 x 4
##   designation      n    prop prop_cumulative
##   <chr>        <int>   <dbl>           <dbl>
## 1 <NA>         37465 0.288             0.288
## 2 Reserve       2009 0.0155            0.304
## 3 Estate        1322 0.0102            0.314
## 4 Reserva       1259 0.00969           0.324
## 5 Riserva        698 0.00537           0.329
## 6 Estate Grown   621 0.00478           0.334
```

While 28.8% of the wines do not have a designation, 25% of the designations contain more than 73% of the wines. We see that of the most common 5 designations, three of them are related to reserve wines but in different languages, while the other two refer to estate wines - wines in which the grapes are grown and the wine is made in the same location.

Taster

The tasters are Wine Enthusiast Magazine wine reviewers.

```r
top_taster_tbl <- Wine %>%
    mutate(taster_name = fct_explicit_na(taster_name)) %>%
    mutate(taster_name = fct_lump(taster_name, 15)) %>%
    count(taster_name, sort = TRUE) %>%
    mutate(prop = n / sum(n))

top_taster_tbl
```

```
## # A tibble: 16 x 3
##    taster_name            n    prop
##    <fct>              <int>   <dbl>
```

```
##   1 (Missing)          26244 0.202
##   2 Roger Voss         25514 0.196
##   3 Michael Schachner  15134 0.116
##   4 Kerin O’Keefe      10776 0.0829
##   5 Virginie Boone      9537 0.0734
##   6 Paul Gregutt        9532 0.0733
##   7 Matt Kettmann       6332 0.0487
##   8 Joe Czerwinski      5147 0.0396
##   9 Sean P. Sullivan    4966 0.0382
##  10 Anna Lee C. Iijima  4415 0.0340
##  11 Jim Gordon          4177 0.0321
##  12 Anne Krebiehl MW    3685 0.0284
##  13 Lauren Buzzeo       1835 0.0141
##  14 Susan Kostrzewa     1085 0.00835
##  15 Other               1078 0.00829
##  16 Mike DeSimone        514 0.00395
```

While 20% of the wines do not have tasters listed, 19.6% of the wines were tasted by Roger Voss, followed by 11.6% which were tasted by Michael Schachner. A potentially interesting side project could be to try and differentiate the wine descriptions between tasters, or to search for patterns in each taster’s preferred wines.

We can speculate if any of the tasters are biased for more positive or negative reviews by looking at mean points per taster:

```
Wine %>%
  group_by(taster_name) %>%
  summarize(meanpoints = mean(points)) %>%
  arrange(desc(meanpoints))
```

```
## # A tibble: 20 x 2
##    taster_name        meanpoints
##    <chr>                   <dbl>
##  1 Anne Krebiehl MW         90.6
##  2 Matt Kettmann            90.0
##  3 Virginie Boone           89.2
##  4 Mike DeSimone            89.1
##  5 Paul Gregutt             89.1
##  6 Kerin O’Keefe            88.9
##  7 Sean P. Sullivan         88.8
##  8 Roger Voss               88.7
##  9 Jim Gordon               88.6
## 10 Joe Czerwinski           88.5
## 11 Anna Lee C. Iijima       88.4
## 12 Jeff Jenssen             88.3
## 13 Christina Pickard        87.8
## 14 <NA>                     87.8
## 15 Lauren Buzzeo            87.7
## 16 Michael Schachner        86.9
## 17 Fiona Adams              86.9
## 18 Susan Kostrzewa          86.6
## 19 Carrie Dykes             86.4
## 20 Alexander Peartree       85.9
```

The mean points per taster range between 85.9 and 90.6. Although there are likely many factors underlying these differences in points between reviewers, if I were a wine maker, I would want Anne Krebiehl MW or Matt Kettmann reviewing my wine, not Alexander Peartree.

Points

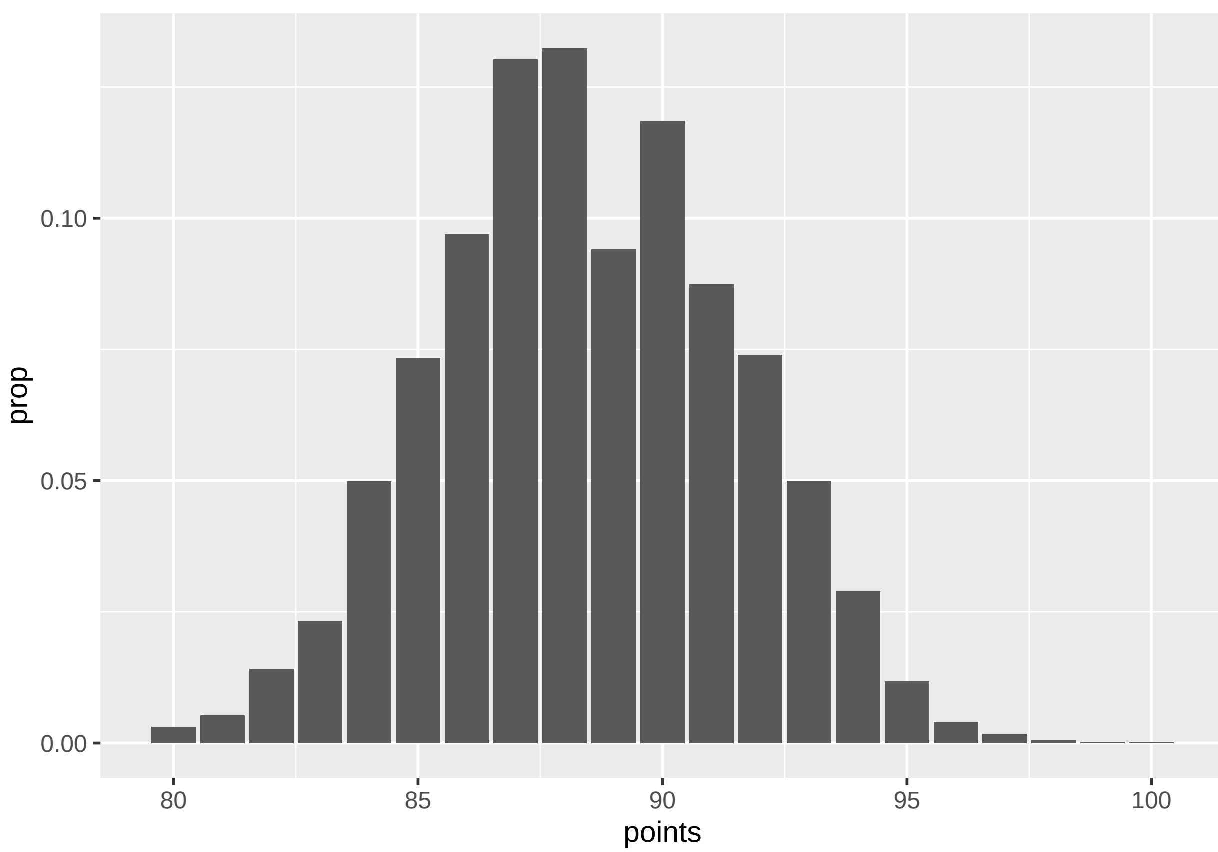
Points is the variable we will be trying to predict.

```
points_tbl <- Wine %>%
    count(points, sort = TRUE) %>%
    mutate(prop = n / sum(n)) %>%
    mutate(prop_cumulative = cumsum(prop))

summary(points_tbl)
```

```
##      points          n               prop          prop_cumulative
##  Min.   : 80   Min.   :   19   Min.   :0.0001462   Min.   :0.1324
##  1st Qu.: 85   1st Qu.:  523   1st Qu.:0.0040240   1st Qu.:0.6596
##  Median : 90   Median : 3758   Median :0.0289141   Median :0.9356
##  Mean   : 90   Mean   : 6189   Mean   :0.0476191   Mean   :0.7916
##  3rd Qu.: 95   3rd Qu.:11359   3rd Qu.:0.0873964   3rd Qu.:0.9942
##  Max.   :100   Max.   :17207   Max.   :0.1323911   Max.   :1.0000
```

```
points_tbl %>%
  ggplot(aes(x = points, y = prop)) +
  geom_bar(stat = "identity")
```



Price

```
price_tbl <- Wine %>%
    count(price, sort = TRUE) %>%
    mutate(prop = n / sum(n)) %>%
```

```r
    mutate(prop_cumulative = cumsum(prop))

summary(price_tbl)
```

```
##      price              n                 prop           prop_cumulative
##  Min.   :   4.0   Min.   :   1.0   Min.   :7.690e-06   Min.   :0.06922
##  1st Qu.: 101.2   1st Qu.:   1.0   1st Qu.:7.690e-06   1st Qu.:0.98640
##  Median : 203.5   Median :   4.0   Median :3.078e-05   Median :0.99761
##  Mean   : 293.9   Mean   : 332.4   Mean   :2.558e-03   Mean   :0.95011
##  3rd Qu.: 369.8   3rd Qu.:  47.0   3rd Qu.:3.616e-04   3rd Qu.:0.99925
##  Max.   :3300.0   Max.   :8996.0   Max.   :6.922e-02   Max.   :1.00000
##  NA's   :1
```

We can see from the table that the price for wine ranges between 4 and 3,300 USD. More than 98% of the wines are under 101.20 USD, and more than 99.7% of the wines are less than 203.5 USD.

Description

Here is an example of the description.

```r
Wine %>% pull(description) %>% pluck(1)
```

```
## [1] "Aromas include tropical fruit, broom, brimstone and dried herb. The palate isn't overly express
```

This is one example. We will want to extract features from the description in order to incorporate this information into any model we do.