Properties that influence wine score

Group 28: Aishwin Tikku, Mengran Li, Steven Kwok, Shaoquan Li, Shuning Li

Introduction

Wine is an alcoholic drink, produced by various kinds of fermented fruits like grapes, apple or blueberry. There are four kinds of wines, involving white wine, red wine, rose wine and sparkling wine. The difference of wines depends on various kinds of factors, including type of grapes, soil status, and province state. We analysis a data set from the Wine Enthusiast, a famous American wine provider, in this project. Thousands of wines were rated in this data set, where wine with points lower than 80 were filtered.

The aim of our project is discovering properties leading the occurrence of high rated wine, the wine with points larger than 90. The first session visualize the structure, properties, as well as correlations inside the dataset. We, next in order, analysis factors of wine, leading high ranking, thorough the best generalized linear model. Due to the excessive classification of nations, provinces,varieties and wineries, we need to reduce the dimensionality of the data set and build models separately for discussion. Finally, we conclude the entire analysis, as well as discussing what we can do in the future.

Variables of study and data

General information

The whole data set has 7 variables and 2000 observations.

```
-- Data Summary ------------------------
                           Values
Name                       Data
Number of rows             2000
Number of columns          7
_______________________
Column type frequency:
  factor                   5
  numeric                  2
________________________
Group variables            None

-- Variable type: factor -------------------------------------------------------
# A tibble: 5 x 5
  skim_variable     n ordered n_unique top_counts
* <chr>         <int> <lgl>      <int> <chr>
1 country        2000 FALSE         25 US: 855, Fra: 359, Ita: 298, Spa: 84
2 province       2000 FALSE        140 Cal: 576, Was: 138, Bor: 100, Tus: 98
3 title          2000 FALSE       1997 Dom: 2, Gim: 2, Wil: 2, :No: 1
4 variety        2000 FALSE        178 Pin: 204, Cha: 187, Cab: 152, Red: 138
5 winery         2000 FALSE       1712 Geo: 6, Lou: 6, Hen: 5, Bra: 4

-- Variable type: numeric ------------------------------------------------------
```

```
# A tibble: 2 x 7
  skim_variable     n  mean    sd   p25   p50   p75
* <chr>         <int> <dbl> <dbl> <dbl> <dbl> <dbl>
1 points         2000  88.5  3.00    86    88    91
2 price          2000  35.5 40.8     17    25    42
```

The variables of points and price are continuous while country, province, title, variety and winery are category varibales.

The levels of title and winery are beyond one thousand, which is meaningless to predict, thus we ignore the two variables.

To explore the factors points of wines over 90, the points varaible is transformed as a dummy variable, where Pass means the point is greater than 90 while Fail is not.

Category variables

For the category varibales, we should check the percentages of pass and fail.

```
      country         Fail          Pass
    Argentina  78.3%  (54)  21.7%  (15)
    Australia  78.3%  (18)  21.7%   (5)
      Austria  60.0%  (27)  40.0%  (18)
       Canada  28.6%   (2)  71.4%   (5)
        Chile  93.2%  (68)   6.8%   (5)
      Croatia 100.0%   (1)   0.0%   (0)
      England   0.0%   (0) 100.0%   (1)
       France  67.7% (243)  32.3% (116)
      Georgia 100.0%   (4)   0.0%   (0)
      Germany  61.5%  (16)  38.5%  (10)
       Greece 100.0%   (8)   0.0%   (0)
      Hungary 100.0%   (4)   0.0%   (0)
       Israel  57.1%   (4)  42.9%   (3)
        Italy  80.2% (239)  19.8%  (59)
    Macedonia 100.0%   (1)   0.0%   (0)
  New Zealand  65.2%  (15)  34.8%   (8)
     Portugal  72.4%  (55)  27.6%  (21)
      Romania 100.0%   (2)   0.0%   (0)
     Slovenia 100.0%   (2)   0.0%   (0)
 South Africa  80.8%  (21)  19.2%   (5)
        Spain  78.6%  (66)  21.4%  (18)
       Turkey 100.0%   (1)   0.0%   (0)
      Ukraine 100.0%   (1)   0.0%   (0)
      Uruguay 100.0%   (3)   0.0%   (0)
           US  73.3% (627)  26.7% (228)
         <NA> 100.0%   (1)   0.0%   (0)
```

We notice that there are several countries who have only few observations.

Chi-squared test is applied to exmain the dependence of country and score.

```
    Pearson's Chi-squared test

data:  Data$country and Data$score
X-squared = 59, df = 24, p-value = 1e-04
```

At the level of 0.05, refuse the null hypotheis, which means there is dependence between country and response variable score.

We conduct statistics on the number of samples according to the type of wine. There are so many levels with rare observations. To deduce dimensions, We classify the types with a sample number of less than 10 as 'others'.

```
       V1
 Min.   :  1.0
 1st Qu.:  1.0
 Median :  2.0
 Mean   : 11.2
 3rd Qu.:  5.0
 Max.   :204.0
```

Simairly, generate a cross-table of variety and score, and test the chi-square.

```
                      variety         Fail        Pass
                     Albariño  81.8%   (9) 18.2%  (2)
      Bordeaux-style Red Blend  62.7%  (64) 37.3% (38)
    Bordeaux-style White Blend  80.0%  (20) 20.0%  (5)
               Cabernet Franc  77.3%  (17) 22.7%  (5)
           Cabernet Sauvignon  75.7% (115) 24.3% (37)
              Champagne Blend  66.7%   (8) 33.3%  (4)
                   Chardonnay  72.7% (136) 27.3% (51)
 Corvina, Rondinella, Molinara  85.7%  (12) 14.3%  (2)
                        Gamay  81.2%  (13) 18.8%  (3)
               Gewürztraminer 100.0%  (13)  0.0%  (0)
                        Glera 100.0%  (15)  0.0%  (0)
                     Grenache  45.5%   (5) 54.5%  (6)
             Grüner Veltliner  72.0%  (18) 28.0%  (7)
                       Malbec  63.0%  (34) 37.0% (20)
                       Merlot  79.0%  (49) 21.0% (13)
                     Nebbiolo  45.7%  (16) 54.3% (19)
                        other  81.9% (249) 18.1% (55)
                 Petite Sirah  75.0%   (9) 25.0%  (3)
                 Pinot Grigio 100.0%  (17)  0.0%  (0)
                   Pinot Gris  81.8%  (18) 18.2%  (4)
                   Pinot Noir  60.3% (123) 39.7% (81)
                         Port  63.6%   (7) 36.4%  (4)
               Portuguese Red  56.7%  (17) 43.3% (13)
             Portuguese White  95.0%  (19)  5.0%  (1)
                    Red Blend  76.8% (106) 23.2% (32)
        Rhône-style Red Blend  63.2%  (12) 36.8%  (7)
                     Riesling  62.7%  (47) 37.3% (28)
                         Rosé  88.3%  (53) 11.7%  (7)
                   Sangiovese  80.0%  (36) 20.0%  (9)
            Sangiovese Grosso  57.1%   (8) 42.9%  (6)
              Sauvignon Blanc  83.7%  (72) 16.3% (14)
              Sparkling Blend  70.4%  (19) 29.6%  (8)
                        Syrah  64.5%  (40) 35.5% (22)
                  Tempranillo 100.0%  (26)  0.0%  (0)
                     Viognier  72.7%   (8) 27.3%  (3)
                  White Blend  85.7%  (24) 14.3%  (4)
                    Zinfandel  87.9%  (29) 12.1%  (4)
```

```
    Pearson's Chi-squared test

data:  Data$variety and Data$score
```

```
X-squared = 128, df = 36, p-value = 3e-12
```

The dependence between variety and score is significant at $\alpha = 0.05$.

Continuous variable

Finally, we compare the distributions of price in different score group. The price in Pass group has a obvious higher mean value than that in Fail group. There is a potential relationship between price and score.

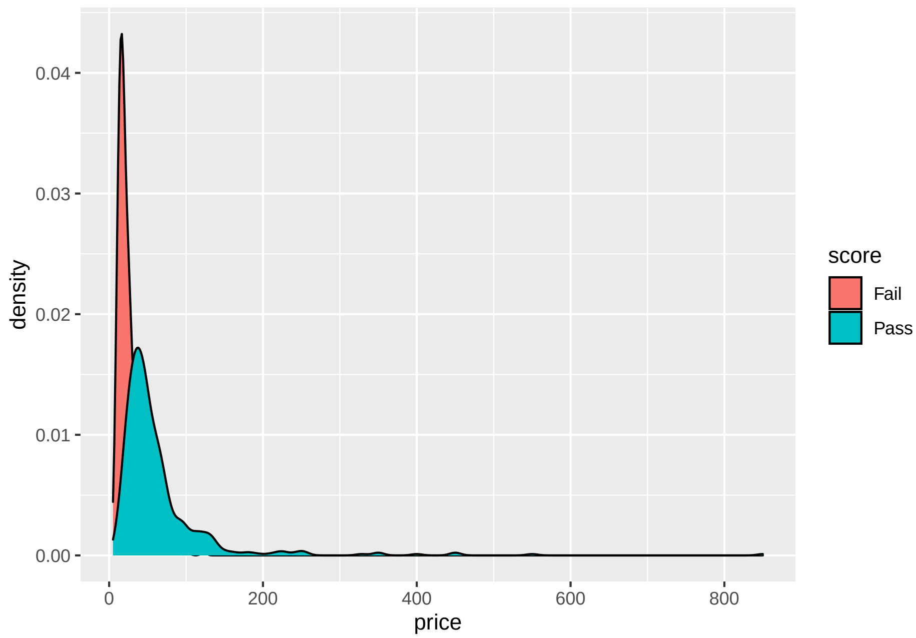


Figure 1: Density plot by score

Methodology

We conduct a generalized linear model to figure out variables have an influence on whether the point of wine can lie above 90. Our main challenge is that the category variables have too many levels which makes situation tricky.

Generalized linear model

$$g(\mu) = \sum \beta_i x_i$$

Where μ is the mean of Y. Y is the response variable. $x_i, i = 0, \ldots, p$ are the explanatory variables. g if the link function. Our response variable is binary, thus the link function takes the form as $log(\frac{\mu}{1-\mu})$. This model is so-called logistic regression model.

Framework

We aim to develop a reasonable model which contains rare variables. Price is a continuous variable and entry the model directly. We test the variety variable first to examine the significance. Then add the country variable and point out the countries who have better wine. We subset the selected countries and explore the influence of province. After checking the overdispersion, we obtain the best model to explain and predict if the point of a wine is greater than 90, which we call Pass here.

Result

Price and variety

We will use the generalised linear model to fit a logistic regression model with score as the response, price and variety as the explanatory variable.

```
Call:
glm(formula = score ~ price + variety, family = binomial(link = "logit"),
    data = Data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9035  -0.6225  -0.4212   0.0003   2.4788

Coefficients:
                                      Estimate Std. Error z value Pr(>|z|)
(Intercept)                          -3.42e+00   1.06e+00   -3.22   0.0013 **
price                                 6.19e-02   3.93e-03   15.76   <2e-16 ***
varietyBordeaux-style Red Blend       1.72e-01   1.11e+00    0.16   0.8762
varietyBordeaux-style White Blend    -1.45e+00   1.54e+00   -0.95   0.3442
varietyCabernet Franc                -2.43e-01   1.26e+00   -0.19   0.8472
varietyCabernet Sauvignon            -5.62e-01   1.09e+00   -0.51   0.6070
varietyChampagne Blend               -1.14e+00   1.31e+00   -0.87   0.3830
varietyChardonnay                     2.44e-01   1.08e+00    0.23   0.8209
varietyCorvina, Rondinella, Molinara -1.16e+00   1.41e+00   -0.82   0.4105
varietyGamay                          7.24e-01   1.32e+00    0.55   0.5824
varietyGewürztraminer                -1.55e+01   1.08e+03   -0.01   0.9886
varietyGlera                         -1.54e+01   1.05e+03   -0.01   0.9884
varietyGrenache                       9.41e-01   1.25e+00    0.75   0.4514
varietyGrüner Veltliner               6.43e-01   1.18e+00    0.55   0.5852
varietyMalbec                         4.31e-01   1.12e+00    0.38   0.7010
varietyMerlot                         8.57e-02   1.12e+00    0.08   0.9388
varietyNebbiolo                       8.21e-02   1.16e+00    0.07   0.9437
varietyother                         -1.24e-01   1.07e+00   -0.12   0.9082
varietyPetite Sirah                   1.89e-01   1.28e+00    0.15   0.8830
varietyPinot Grigio                  -1.53e+01   9.79e+02   -0.02   0.9875
varietyPinot Gris                     5.49e-01   1.21e+00    0.45   0.6501
varietyPinot Noir                     1.14e-01   1.08e+00    0.11   0.9153
varietyPort                          -1.86e-01   1.34e+00   -0.14   0.8901
varietyPortuguese Red                 1.32e+00   1.17e+00    1.13   0.2584
varietyPortuguese White              -1.78e-01   1.49e+00   -0.12   0.9050
varietyRed Blend                     -2.23e-01   1.09e+00   -0.21   0.8375
varietyRhône-style Red Blend          9.80e-01   1.18e+00    0.83   0.4068
varietyRiesling                       1.13e+00   1.09e+00    1.03   0.3012
varietyRosé                           2.01e-01   1.14e+00    0.18   0.8598
varietySangiovese                    -9.70e-01   1.19e+00   -0.81   0.4161
varietySangiovese Grosso             -4.12e-01   1.29e+00   -0.32   0.7501
varietySauvignon Blanc                4.10e-01   1.10e+00    0.37   0.7102
varietySparkling Blend                4.93e-01   1.17e+00    0.42   0.6735
varietySyrah                          6.36e-01   1.10e+00    0.58   0.5623
varietyTempranillo                   -1.55e+01   7.67e+02   -0.02   0.9839
varietyViognier                       6.93e-01   1.32e+00    0.52   0.6008
varietyWhite Blend                   -1.33e-01   1.30e+00   -0.10   0.9187
varietyZinfandel                     -4.68e-01   1.19e+00   -0.39   0.6950
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2091.9  on 1844  degrees of freedom
Residual deviance: 1506.8  on 1807  degrees of freedom
  (155 observations deleted due to missingness)
AIC: 1583

Number of Fisher Scoring iterations: 16
```

Notice that no variety of wine is significant at the 5% significance level.

Country and province

Similarly, We can use the same method to eliminate interference from too many categories. In order to find this standard, we fit a logistic regression model and check its summary table.

```
Call:
glm(formula = score ~ price + country, family = binomial(link = "logit"),
    data = Data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0741  -0.5938  -0.4427   0.0003   2.5933

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         -3.25e+00   3.93e-01   -8.27   <2e-16 ***
price                6.18e-02   3.69e-03   16.76   <2e-16 ***
countryAustralia    -3.41e-01   7.32e-01   -0.47    0.641
countryAustria       8.28e-01   5.28e-01    1.57    0.117
countryCanada        1.11e+00   1.06e+00    1.05    0.292
countryChile        -9.03e-01   6.63e-01   -1.36    0.173
countryCroatia      -1.44e+01   2.40e+03   -0.01    0.995
countryEngland       1.74e+01   2.40e+03    0.01    0.994
countryFrance        1.71e-01   4.12e-01    0.42    0.678
countryGeorgia      -1.45e+01   1.18e+03   -0.01    0.990
countryGermany       3.63e-01   6.42e-01    0.57    0.572
countryGreece       -1.49e+01   8.20e+02   -0.02    0.986
countryHungary      -1.64e+01   1.03e+03   -0.02    0.987
countryIsrael        5.01e-01   9.31e-01    0.54    0.590
countryItaly        -1.01e+00   4.33e-01   -2.33    0.020 *
countryMacedonia    -1.46e+01   2.40e+03   -0.01    0.995
countryNew Zealand   1.13e+00   6.10e-01    1.86    0.063 .
countryPortugal      5.51e-01   5.13e-01    1.07    0.283
countryRomania      -1.42e+01   1.68e+03   -0.01    0.993
countrySlovenia     -1.48e+01   1.68e+03   -0.01    0.993
countrySouth Africa  3.57e-01   6.49e-01    0.55    0.582
countrySpain         2.39e-01   4.93e-01    0.48    0.628
countryTurkey       -1.42e+01   2.40e+03   -0.01    0.995
countryUkraine      -1.39e+01   2.40e+03   -0.01    0.995
countryUruguay      -1.47e+01   1.39e+03   -0.01    0.992
countryUS           -6.69e-02   3.89e-01   -0.17    0.863
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2091.3  on 1843  degrees of freedom
Residual deviance: 1517.9  on 1818  degrees of freedom
  (156 observations deleted due to missingness)
AIC: 1570

Number of Fisher Scoring iterations: 15
```

We choose variables with p-values less than 0.1. From the summary table, 'New Zealand' and 'Italy' in country and price have a significant influence on the score.

Therefore, we set all countries except 'New Zealand' and 'Italy' as 'others' to make this variable a categorical variable with three levels.

```
Call:
glm(formula = score ~ price + country, family = binomial(link = "logit"),
    data = Data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9980  -0.5847  -0.4380   0.0003   2.5911

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        -4.24575    0.26356  -16.11  < 2e-16 ***
price               0.06162    0.00356   17.33  < 2e-16 ***
countryNew Zealand  2.13442    0.52657    4.05  5.0e-05 ***
countryothers       1.02554    0.21695    4.73  2.3e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2091.9  on 1844  degrees of freedom
Residual deviance: 1546.9  on 1841  degrees of freedom
  (155 observations deleted due to missingness)
AIC: 1555

Number of Fisher Scoring iterations: 6
```

```
[1] 1570
```

```
[1] 1555
```

Compared with the model with all countries, the model with merged counties has a smaller AIC and all variables are significant.

Next, We select 'New Zealand' and 'Italy' to check 'province' variable. Filter samples whose country is either 'New Zealand' or 'Italy' and fit a logistic regression model with score as response, price and province as explanatory variables.

```
Call:
glm(formula = score ~ price + province, family = binomial(link = "logit"),
    data = New_Zealand)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5686  -0.2578  -0.0574   0.0001   1.8398

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)             -26.687  10754.014    0.00     1.00
price                     0.475      0.231    2.05     0.04 *
provinceCentral Otago    34.385  15208.471    0.00     1.00
provinceHawke's Bay       9.360  10754.028    0.00     1.00
provinceKumeu            24.891  15208.473    0.00     1.00
provinceMarlborough      15.703  10754.013    0.00     1.00
provinceMartinborough    17.905  10754.013    0.00     1.00
provinceWairau Valley    -1.424  15208.471    0.00     1.00
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 29.7202  on 22  degrees of freedom
Residual deviance:  7.6594  on 15  degrees of freedom
AIC: 23.66

Number of Fisher Scoring iterations: 18

Call:
glm(formula = score ~ price + province, family = binomial(link = "logit"),
    data = Italy)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-2.053  -0.403  -0.221   0.000   2.628

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                -2.20e+01   1.95e+03   -0.01     0.99
price                       7.01e-02   1.11e-02    6.31  2.7e-10 ***
provinceItaly Other         1.32e+00   1.09e+04    0.00     1.00
provinceLombardy            1.74e-01   5.01e+03    0.00     1.00
provinceNortheastern Italy  5.98e-01   2.84e+03    0.00     1.00
provincePiedmont            1.81e+01   1.95e+03    0.01     0.99
provinceSicily & Sardinia   1.85e+01   1.95e+03    0.01     0.99
provinceSouthern Italy      1.78e+01   1.95e+03    0.01     0.99
provinceTuscany             1.70e+01   1.95e+03    0.01     0.99
provinceVeneto              1.70e+01   1.95e+03    0.01     0.99
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 252.09  on 260  degrees of freedom
Residual deviance: 137.49  on 251  degrees of freedom
  (37 observations deleted due to missingness)
```

```
AIC: 157.5

Number of Fisher Scoring iterations: 18
```

From the summary table of the above model, it can be seen that the province has no significant effect on the score.

Overdispersion

To avoid the overdispersion, we need to compare the value of deviance divided by residual deviance with 1.

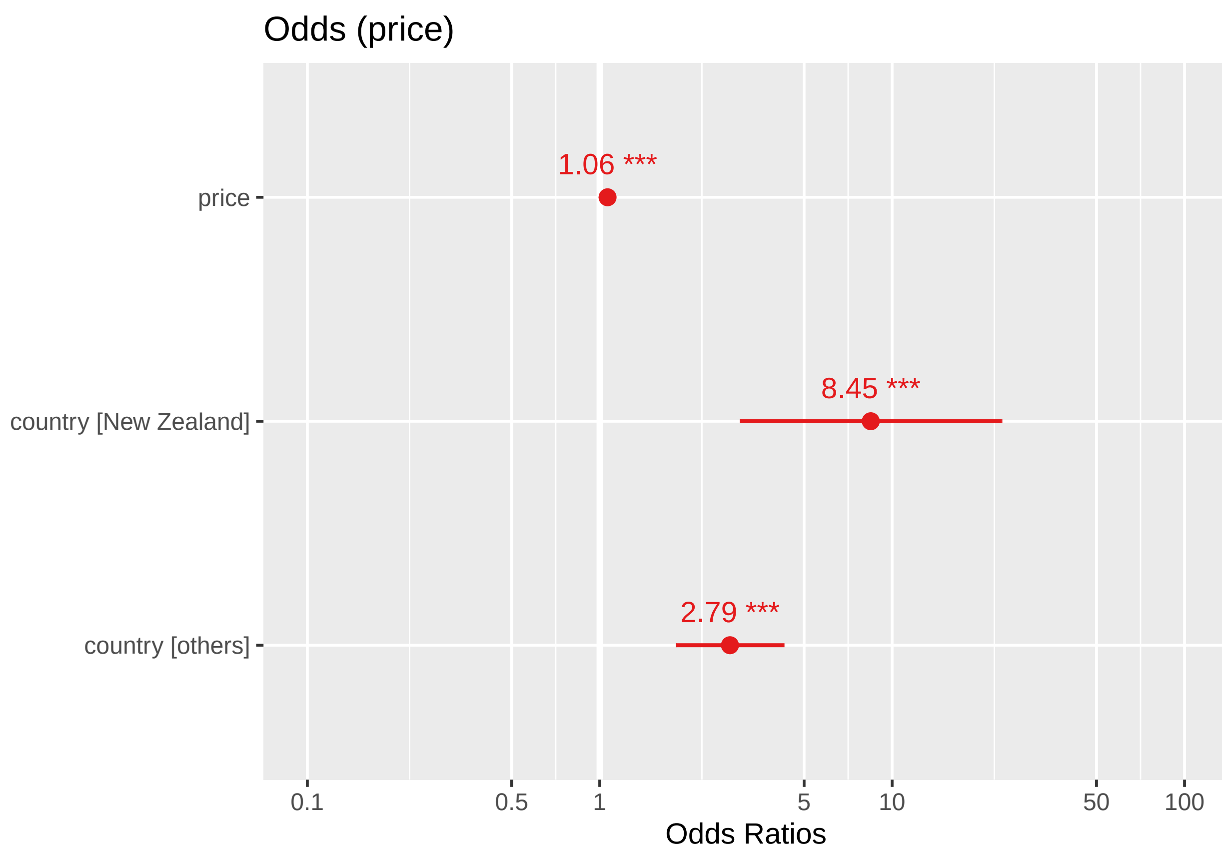
```
[1] 0.8402
```

The result is less than 1, there is no overdispersion.

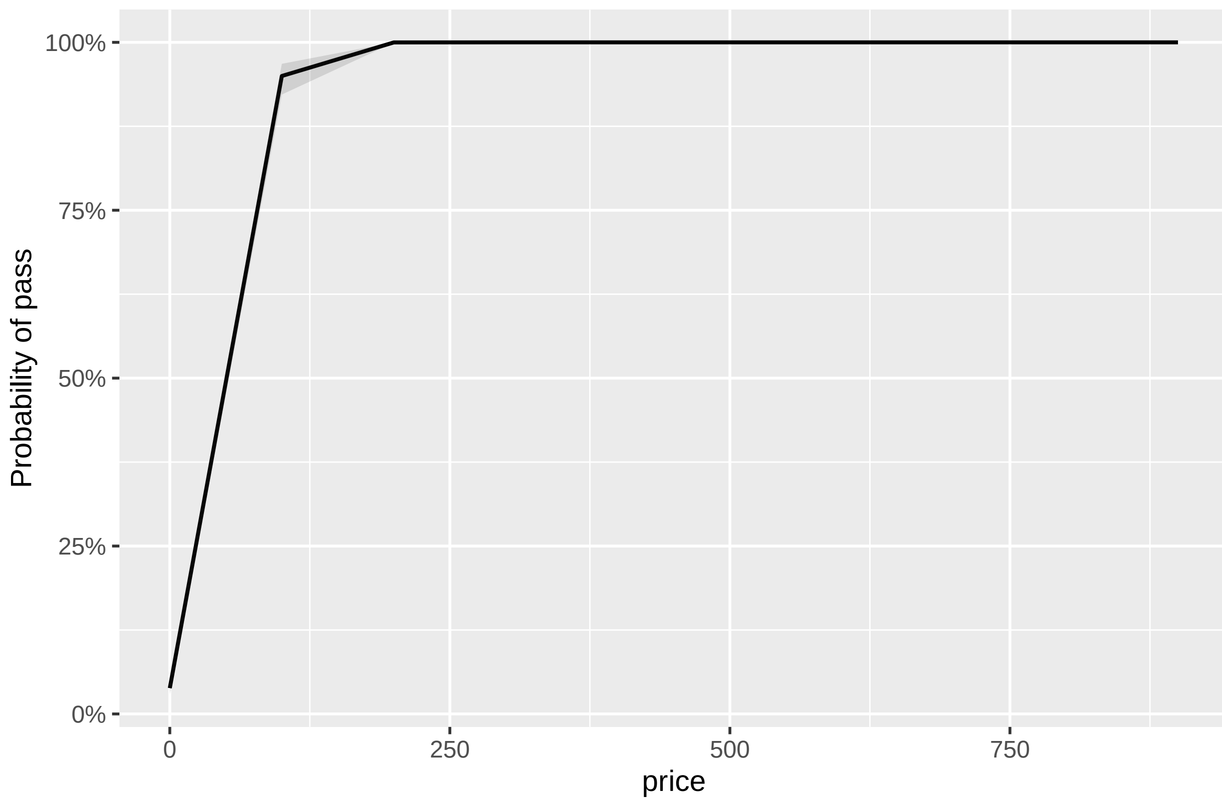
Odds and prediction

Notice that the coefficients of price, countryNew Zealand, other countries are positive, which means that expensive wine are more likely to pass(points is greater than 90). And the all coefficients are significant (p-value of *0.0616*, *2.1344* and *1.0255*), so we qualify the effect of them.

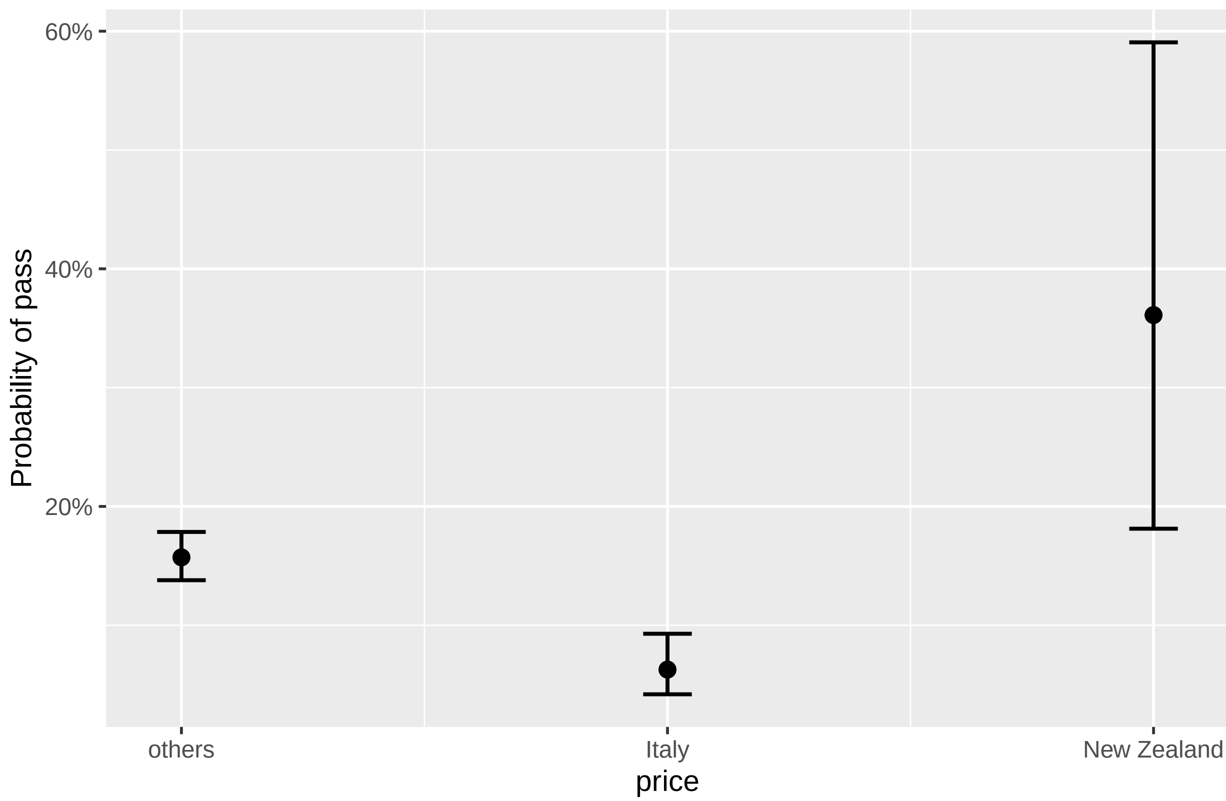
And the all coefficients are significant (p-value of *3.00757432715095e-67*, *5.04786285700546e-05* and *2.27837947903262e-06*), so we qualify the effect of them.




```
$price
```



$country



Coclusion and Future Work

After establishing the generalised linear model and comparing them, we found that the price is the most significant factor on wine scores. In addition, after analysis, we found that Riesling wine or Israel wine are easier to reach 90 points than other varieties of wine.

For discovering factors affecting the quality of wine more accurately, we can combine multiple data sets with details of Wineries. Different model, additionally, will be trained and tested for selecting the best model. The targeted model will be neural network, a self-learning model which can handle large amount of new data. Lastly, we will construct a system, which can help business searching and identifying good model. This application would be useful for assigning price and prevent purchasing unqualified wines.

Reference

Christina, https://cellar.asia/wine/what-is-wine/‘