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 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 different 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. For identifying good wine easier, we exploring elements of high-ranked wine.

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 data set. We, next in order, analyse 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 dimension number of the data set and build models separately for discussion. Finally, we conclude the entire analysis, as well as discussing what can be done in the future.

```
# library packages
library(tidyverse)
library(ggplot2)
library(tidymodels)
library(GGally)
library(car)
library(sjPlot)
library(skimr)
library(kableExtra)
library(janitor)
```

Exploratory Data Analysis

General information

The whole data set has 7 variables and 2000 observations.

```
# read data
Data <- read.csv("dataset28.csv")
#remove the index
Data <- Data[, -1]
# transform the character as factor
Data$country <- as.factor(Data$country)
Data$province <- as.factor(Data$province)
Data$title <- as.factor(Data$title)
Data$variety <- as.factor(Data$variety)
```

```
Data$winery <- as.factor(Data$winery)
Data$price <- as.integer(Data$price)
# generate general information about dataset
my_skim <- skim_with(base = sfl(n = length), numeric = sfl(p0 = NULL, p100 = NULL, hist = NULL))
print(my_skim(Data))
```

```
-- 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 variables.

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 variable is transformed as a dummy variable, where Pass means the point is greater than 90 while Fail is not.

```
# Group by points, those greater than 90 are pass, others are fail
Data <- Data %>% mutate(score = ifelse(points > 90, "Pass", "Fail"))
Data$score <- as.factor(Data$score)
```

Category variables

For the category variables, we should check the percentages of pass and fail. According to the data, using the tabyl function to display the proportion of score variables in different countries. First six levels are as Table 1.

```r
# cross-table of country and score
 Data %>%
  tabyl(country,score)%>%
  adorn_percentages()%>%
  adorn_pct_formatting()%>%
  adorn_ns()%>%
  head() %>%
  kbl(caption = 'Pass\\% and Fail\\% for each Country', booktabs = T)%>%
  kable_styling(latex_options = "HOLD_position")
```

Table 1: Pass% and Fail% for each Country

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)

We notice that there are several countries who have only few observations For instance, Croatia,Georgia,Turkey,Ukraine and others' score variable are 100% "Fail", and other data of some countries are selected partly.

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

```r
# chi square test
chisq.test(Data$country, Data$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 hypothesis, 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'.

```r
# summary of numbers of observations
t <- as.matrix(table(Data$variety))
# merge levels with rare observations
Data$variety <- as.vector(Data$variety)
Data$variety[which(Data$variety %in% row.names(t)[t<11])] = "other"
Data$variety <- as.factor(Data$variety)
```

Similarly, generate a cross-table of variety and score as Table 2, and test the chi-square.

```
# cross-table of variety and score
 Data %>%
  tabyl(variety,score)%>%
  adorn_percentages()%>%
  adorn_pct_formatting()%>%
  adorn_ns()%>%
  kbl(caption = 'Pass\\% and Fail\\% of each Wine Variety', booktabs = T)%>%
  kable_styling(latex_options = "HOLD_position")
```

Table 2: Pass% and Fail% of each Wine Variety

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)

```r
# chi square test
chisq.test(Data$variety, Data$score)
```

```
    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 according to Fig. 1.

```r
# density plot
ggplot(data = Data) +
  geom_density(aes(price, group = score, fill = score))
```

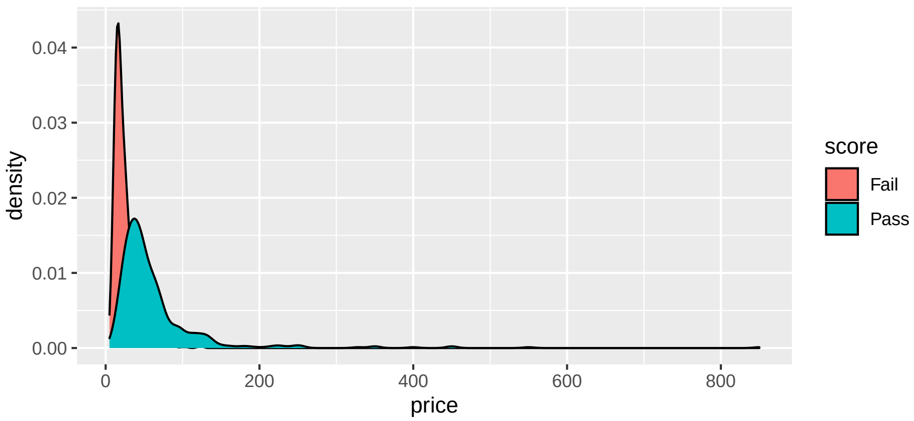

Figure 1: Density plot by score

Methodology

We conduct a generalized linear model to figure out variables having 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. $x_i, i = 0, \dots, p$ 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 generalized linear model to fit a logistic regression model with score as the response, price and variety as the explanatory variable. Summary table of the model is as Table 3.

```r
# logistic regression with price and variety
fit1 <- glm(score ~ price + variety, Data, family = binomial(link = "logit"))
fit1 %>% tidy() %>% kbl(caption = " Summary Statistics of GLM 1", booktabs = T)%>%
  kable_styling(latex_options = "HOLD_position")
```

Table 3: Summary Statistics of GLM 1

term	estimate	std.error	statistic	p.value
(Intercept)	-3.4170	1.0601	-3.2233	0.0013
price	0.0619	0.0039	15.7570	0.0000
varietyBordeaux-style Red Blend	0.1724	1.1067	0.1558	0.8762
varietyBordeaux-style White Blend	-1.4542	1.5374	-0.9459	0.3442
varietyCabernet Franc	-0.2426	1.2594	-0.1926	0.8472
varietyCabernet Sauvignon	-0.5622	1.0931	-0.5143	0.6070
varietyChampagne Blend	-1.1396	1.3062	-0.8724	0.3830
varietyChardonnay	0.2440	1.0780	0.2263	0.8209
varietyCorvina, Rondinella, Molinara	-1.1568	1.4057	-0.8229	0.4105
varietyGamay	0.7244	1.3174	0.5498	0.5824
varietyGewürztraminer	-15.4883	1082.8407	-0.0143	0.9886
varietyGlera	-15.3521	1053.7103	-0.0146	0.9884
varietyGrenache	0.9414	1.2502	0.7530	0.4514
varietyGrüner Veltliner	0.6433	1.1788	0.5457	0.5852
varietyMalbec	0.4311	1.1228	0.3840	0.7010
varietyMerlot	0.0857	1.1157	0.0768	0.9388
varietyNebbiolo	0.0821	1.1625	0.0706	0.9437
varietyother	-0.1238	1.0732	-0.1154	0.9082
varietyPetite Sirah	0.1886	1.2818	0.1471	0.8830
varietyPinot Grigio	-15.3175	979.3971	-0.0156	0.9875
varietyPinot Gris	0.5492	1.2107	0.4536	0.6501
varietyPinot Noir	0.1144	1.0754	0.1064	0.9153
varietyPort	-0.1856	1.3435	-0.1381	0.8901
varietyPortuguese Red	1.3168	1.1651	1.1302	0.2584
varietyPortuguese White	-0.1784	1.4948	-0.1194	0.9050
varietyRed Blend	-0.2232	1.0884	-0.2051	0.8375
varietyRhône-style Red Blend	0.9802	1.1816	0.8295	0.4068
varietyRiesling	1.1312	1.0942	1.0338	0.3012
varietyRosé	0.2011	1.1385	0.1766	0.8598
varietySangiovese	-0.9703	1.1931	-0.8132	0.4161
varietySangiovese Grosso	-0.4122	1.2942	-0.3185	0.7501
varietySauvignon Blanc	0.4099	1.1030	0.3716	0.7102
varietySparkling Blend	0.4929	1.1699	0.4214	0.6735
varietySyrah	0.6363	1.0981	0.5795	0.5623
varietyTempranillo	-15.4991	766.8771	-0.0202	0.9839
varietyViognier	0.6929	1.3241	0.5233	0.6008
varietyWhite Blend	-0.1325	1.2984	-0.1021	0.9187
varietyZinfandel	-0.4681	1.1940	-0.3921	0.6950

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 as Table 4.

```r
# logistic regression with price and country
fit2 <- glm(score ~ price + country,
            data = Data,
            family = binomial(link = "logit"))
fit2 %>% tidy() %>% kbl(caption = " Summary Statistics of GLM 2", booktabs = T)%>%
  kable_styling(latex_options = "HOLD_position")
```

Table 4: Summary Statistics of GLM 2

term	estimate	std.error	statistic	p.value
(Intercept)	-3.2480	0.3928	-8.2679	0.0000
price	0.0618	0.0037	16.7567	0.0000
countryAustralia	-0.3414	0.7322	-0.4662	0.6410
countryAustria	0.8276	0.5284	1.5662	0.1173
countryCanada	1.1124	1.0559	1.0535	0.2921
countryChile	-0.9035	0.6626	-1.3635	0.1727
countryCroatia	-14.3687	2399.5447	-0.0060	0.9952
countryEngland	17.4038	2399.5447	0.0073	0.9942
countryFrance	0.1712	0.4120	0.4155	0.6778
countryGeorgia	-14.4954	1180.6531	-0.0123	0.9902
countryGermany	0.3629	0.6423	0.5651	0.5720
countryGreece	-14.8608	819.5126	-0.0181	0.9855
countryHungary	-16.4099	1034.4621	-0.0159	0.9873
countryIsrael	0.5013	0.9311	0.5384	0.5903
countryItaly	-1.0064	0.4326	-2.3262	0.0200
countryMacedonia	-14.5541	2399.5447	-0.0061	0.9952
countryNew Zealand	1.1326	0.6097	1.8578	0.0632
countryPortugal	0.5513	0.5131	1.0745	0.2826
countryRomania	-14.2089	1678.1159	-0.0085	0.9932
countrySlovenia	-14.8497	1680.7738	-0.0088	0.9930
countrySouth Africa	0.3572	0.6493	0.5502	0.5822
countrySpain	0.2386	0.4926	0.4844	0.6281
countryTurkey	-14.2451	2399.5447	-0.0059	0.9953
countryUkraine	-13.8743	2399.5448	-0.0058	0.9954
countryUruguay	-14.7202	1385.0393	-0.0106	0.9915
countryUS	-0.0669	0.3887	-0.1721	0.8634

We choose variables with p-values less than 0.1. From the summary Table 5, '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.

```r
# transform the country variable
Data <- Data %>% mutate(
  country = case_when(
    country == "New Zealand" ~ "New Zealand",
    country == "Italy" ~ "Italy",
    !country %in% c( 'New Zealand', 'Italy') ~ "others"
  )
)
```

```
# logistic with new country variable
fit3 <- glm(score ~ price + country,
            data = Data,
            family = binomial(link = "logit"))
fit3 %>% tidy() %>% kbl(caption = " Summary Statistics of GLM 3", booktabs = T)%>%
  kable_styling(latex_options = "HOLD_position")
```

Table 5: Summary Statistics of GLM 3

term	estimate	std.error	statistic	p.value
(Intercept)	-4.2458	0.2636	-16.110	0e+00
price	0.0616	0.0036	17.326	0e+00
countryNew Zealand	2.1344	0.5266	4.053	1e-04
countryothers	1.0255	0.2170	4.727	0e+00

```
# AIC
AIC(fit2)
```

```
[1] 1570
```

```
AIC(fit3)
```

```
[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.

```
# divide dataset by country
New_Zealand <- Data %>% filter(country == "New Zealand")
# logistic of New Zealand
glm(score ~ price + province,
    data = New_Zealand,
    family = binomial(link = "logit"),
    ) %>%
  tidy() %>% kbl(caption = " Summary Statistics of GLM with Country = New Zealand", booktabs = T)%>%
  kable_styling(latex_options = "HOLD_position")
```

Table 6: Summary Statistics of GLM with Country = New Zealand

term	estimate	std.error	statistic	p.value
(Intercept)	-26.6868	1.075e+04	-0.0025	0.9980
price	0.4747	2.315e-01	2.0509	0.0403
provinceCentral Otago	34.3850	1.521e+04	0.0023	0.9982
provinceHawke's Bay	9.3597	1.075e+04	0.0009	0.9993
provinceKumeu	24.8907	1.521e+04	0.0016	0.9987
provinceMarlborough	15.7035	1.075e+04	0.0015	0.9988
provinceMartinborough	17.9046	1.075e+04	0.0017	0.9987
provinceWairau Valley	-1.4241	1.521e+04	-0.0001	0.9999

```r
# logistic of Italy
Italy <- Data %>% filter(country == "Italy")
glm(score ~ price + province,
    data = Italy,
    family = binomial(link = "logit"),
    ) %>%
  tidy() %>% kbl(caption = " Summary Statistics of GLM  wiht country = Italy", booktabs = T)%>%
  kable_styling(latex_options = "HOLD_position")
```

Table 7: Summary Statistics of GLM wiht country = Italy

term	estimate	std.error	statistic	p.value
(Intercept)	-22.0106	1.954e+03	-0.0113	0.9910
price	0.0701	1.110e-02	6.3142	0.0000
provinceItaly Other	1.3227	1.093e+04	0.0001	0.9999
provinceLombardy	0.1738	5.005e+03	0.0000	1.0000
provinceNortheastern Italy	0.5979	2.845e+03	0.0002	0.9998
provincePiedmont	18.1047	1.954e+03	0.0093	0.9926
provinceSicily & Sardinia	18.4812	1.954e+03	0.0095	0.9925
provinceSouthern Italy	17.7883	1.954e+03	0.0091	0.9927
provinceTuscany	17.0473	1.954e+03	0.0087	0.9930
provinceVeneto	16.9776	1.954e+03	0.0087	0.9931

From the summary tables (Table 6, 7) 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.

```r
# check overdispersion
deviance(fit3)/df.residual(fit3)
```

```
[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 *<0.0001*, *<0.0001* and *<0.0001*), thus we qualify the effect of them.

```
# odds plot
plot_model(fit3, show.values = TRUE,
           title = "Odds (price)", show.p = TRUE)
```

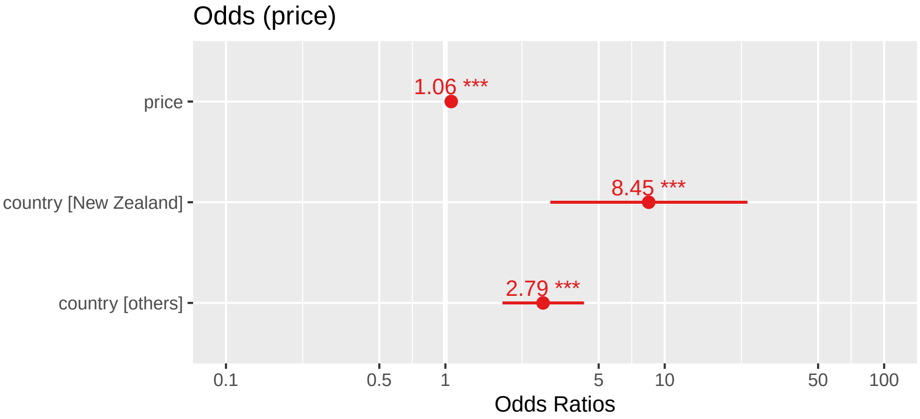

Figure 2: Odds Ratio Graph Based on GLM Model

```
# prediction plot
plot_model(fit3,
           type = "pred", title = "",
           axis.title = c("price", "Probability of pass"))
```

```
$price
```

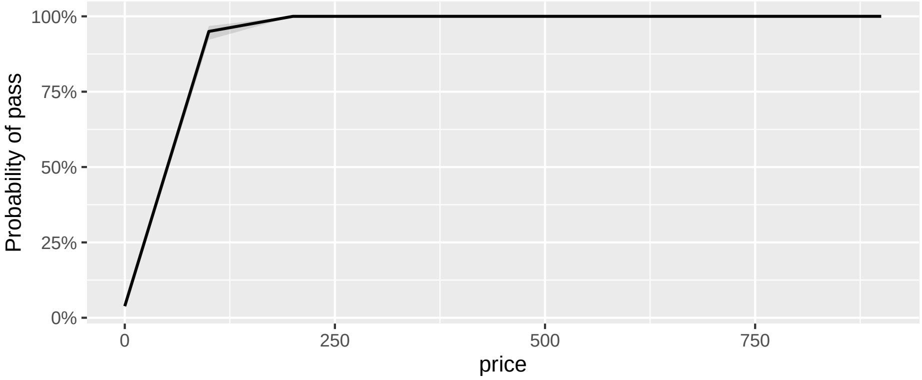

Figure 3: Probability of Pass for Italy & New Zealand

```
$country
```

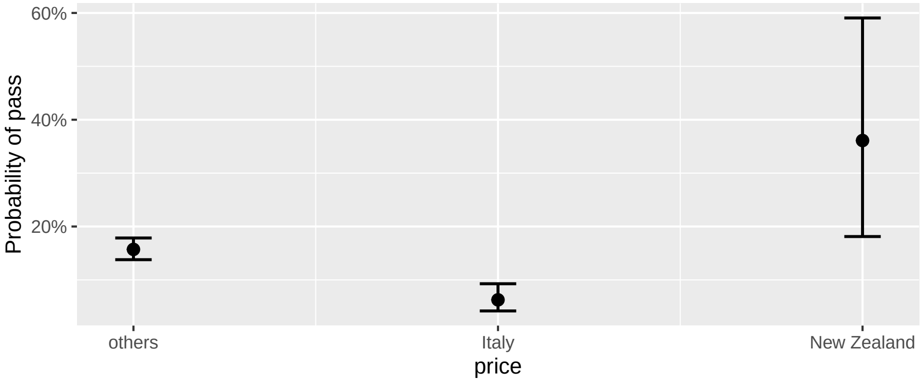

Figure 4: Probability of Pass for Italy & New Zealand

We interpret the odds ratios (Fig. 2, 3) as follows: New Zealand produced wine's odds of passing are 8.54 times those of Italy produced wine, the other countries produced wine's odds of passing are 2.79 times those of Italy produced wine. And for every unit increase in price, the expensive wine's odds of passing are 1.06 times those of the cheaper one.

In the probability figure as Fig. 4, We can have a more intuitive awareness of the above tendency. For example, Wine with a price of more than 80 pounds has a high probability of getting a score of 90 or more. And wine from Italy only has a 5% chance of getting more than 90 points, while New Zealand produced wine has a more than 35% chance of passing.

Conclusions

Discussing reasons and background information, we detected the properties related to high-rated wine. Checking the data set, we discovered that the correlation between variety and score is obvious. There is, also, potential relation between price of wine and score. Exploring the structure and hidden relation inside the data, we applied generalized linear models for data analysis. After establishing several generalised linear models and having comparisons, we found that the price is the most significant factor on wine scores. We, additionally, found that Riesling wine or Israel wine are easier to reach 90 points than other varieties of wine after completing the analysis.

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 through a picture. This application would be useful for assigning price and prevent purchasing unqualified wines.

Reference

Barret Schloerke, Di Cook, Joseph Larmarange, Francois Briatte, Moritz Marbach, Edwin Thoen, Amos Elberg and Jason Crowley, 2021, GGally: Extension to 'ggplot2'. R package version 2.1.2., Ref: https://CRAN.R-project.org/package=GGally

Christina, 2019, What is Wine?, Ref: https://cellar.asia/wine/what-is-wine/‘

Elin Waring, Michael Quinn, Amelia McNamara, Eduardo Arino de la Rubia, Hao Zhu and Shannon Ellis, 2021, skimr: Compact and Flexible Summaries of Data, R package version 2.1.3, Ref: https://CRAN.R-project.org/package=skimr

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Hao Zhu, 2021, kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.3.4, Ref: https://CRAN.R-project.org/package=kableExtra

John Fox and Sanford Weisberg, 2019, An {R} Companion to Applied Regression, Third Edition, Thousand Oaks CA: Sage. Ref: https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Kuhn et al., 2020, Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles., Ref: https://www.tidymodels.org

Lüdecke D, 2021, *sjPlot: Data Visualization for Statistics in Social Science*, R package version 2.8.8, <Ref: https://CRAN.R-project.org/package=sjPlot>.

Sam Firke, 2021, janitor: Simple Tools for Examining and Cleaning Dirty Data. R package version 2.1.0, Ref: https://CRAN.R-project.org/package=janitor

Wickham et al., 2. Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, Ref: https://doi.org/10.21105/joss.01686