

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. Finally, we conclude the entire analysis, as well as discussing what we can do in the future.

Data Structure and Visualisation

The whole data set has 7 factors and 2000 observations.

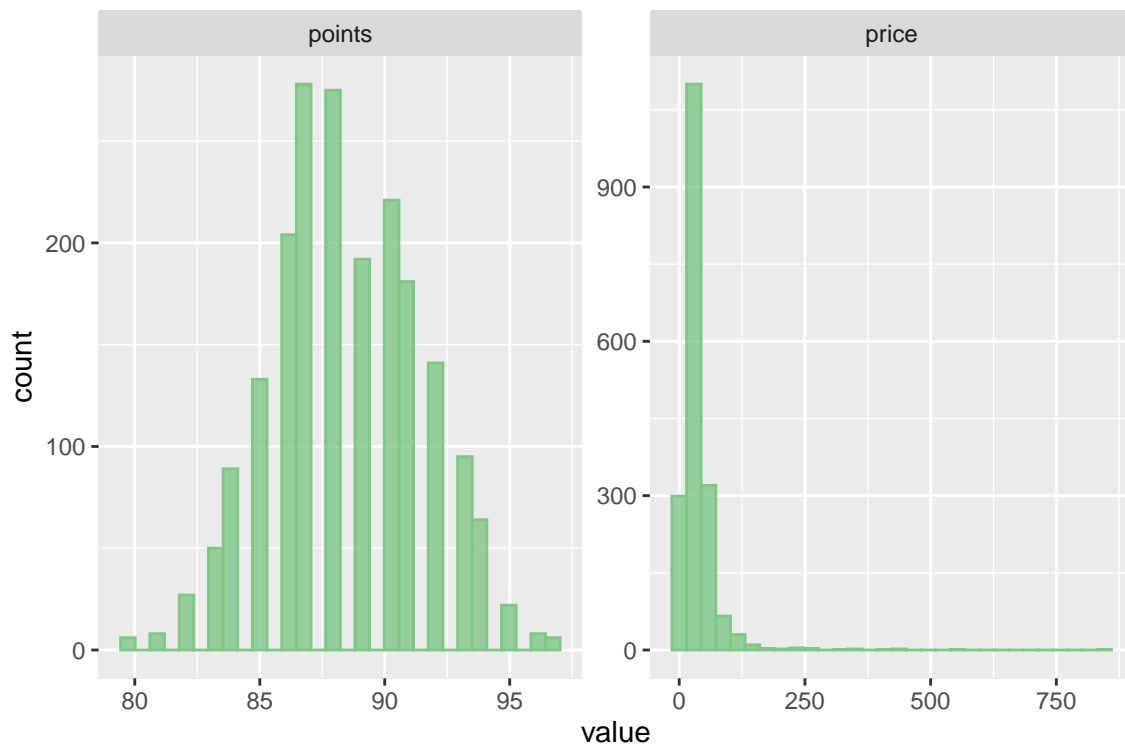


Figure 1: Histogram of Price and Points

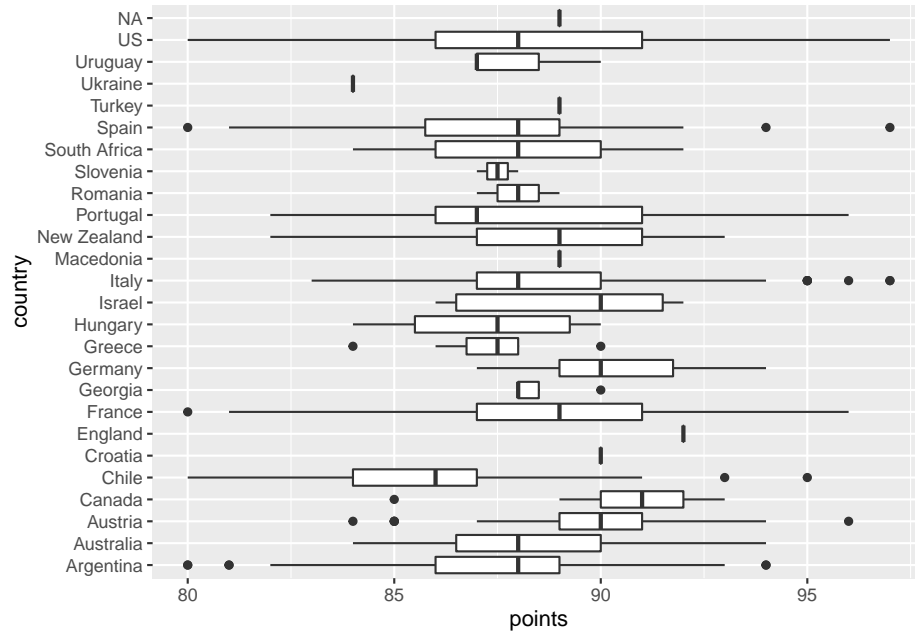


Figure 2: Boxplot of points by country

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)
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)

Pearson's Chi-squared test

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

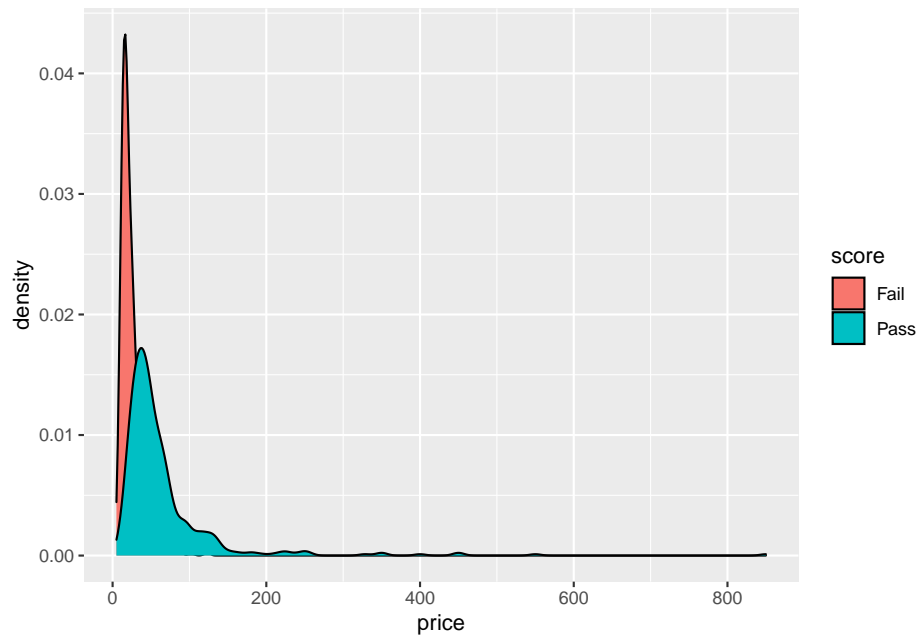


Figure 3: Density plot by score

Methodology

New Zealand and Italy are significantly different from others

GLM Model for New Zealand

GLM Model for Italy

GLM Model & Predicted Values Plot based on only Price

\$price

Table 1: Summary Statistics of GLM Model

	Estimate	Std. Error	z value	Pr(> z)
(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

Table 2: GLM Model for New Zealand

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-25.9451	1.773e+04	-0.0015	0.9988
price	0.3586	2.112e-01	1.6977	0.0896
provinceCentral Otago	59.7177	1.081e+05	0.0006	0.9996
provinceHawke's Bay	-2.8688	2.507e+04	-0.0001	0.9999
provinceKumeu	30.3742	2.507e+04	0.0012	0.9990
provinceMarlborough	17.7017	1.773e+04	0.0010	0.9992
provinceMartinborough	36.9341	1.930e+04	0.0019	0.9985
provinceWairau Valley	-1.0758	2.507e+04	0.0000	1.0000
varietyMerlot	31.4500	2.507e+04	0.0013	0.9990
varietyPinot Noir	-22.1716	1.051e+05	-0.0002	0.9998
varietyRiesling	-36.0322	1.362e+04	-0.0026	0.9979
varietyRosé	-18.4189	1.773e+04	-0.0010	0.9992

Table 3: GLM Model for Italy

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-5.4914	9.777e+03	-0.0006	0.9996
price	0.0703	1.350e-02	5.2027	0.0000
provinceItaly Other	-16.1991	2.025e+04	-0.0008	0.9994
provinceLombardy	15.5740	2.232e+04	0.0007	0.9994
provinceNortheastern Italy	-20.7067	9.798e+03	-0.0021	0.9983
provincePiedmont	1.7338	9.777e+03	0.0002	0.9999
provinceSicily & Sardinia	19.2912	5.153e+03	0.0037	0.9970
provinceSouthern Italy	0.8081	9.777e+03	0.0001	0.9999
provinceTuscany	16.5883	5.153e+03	0.0032	0.9974
provinceVeneto	17.5417	5.153e+03	0.0034	0.9973
varietyAlbana	-16.1288	2.025e+04	-0.0008	0.9994
varietyArneis	-18.5654	1.773e+04	-0.0010	0.9992
varietyBarbera	-18.5593	1.015e+04	-0.0018	0.9985
varietyCabernet	-16.7964	1.773e+04	-0.0009	0.9992
varietyCabernet Franc	-18.2811	8.309e+03	-0.0022	0.9982
varietyCabernet Sauvignon	-35.3299	1.403e+04	-0.0025	0.9980
varietyCarignano	5.7122	1.958e+04	0.0003	0.9998
varietyChardonnay	-37.2861	1.245e+04	-0.0030	0.9976
varietyCiliegiolo	-33.9118	1.958e+04	-0.0017	0.9986
varietyCorvina, Rondinella, Molinara	-17.0796	8.309e+03	-0.0021	0.9984
varietyDolcetto	-18.0611	1.021e+04	-0.0018	0.9986
varietyFalanghina	-16.9369	1.254e+04	-0.0014	0.9989
varietyFiano	-17.1478	1.773e+04	-0.0010	0.9992
varietyGlera	-33.9839	9.554e+03	-0.0036	0.9972
varietyGreco	-17.2883	1.773e+04	-0.0010	0.9992
varietyInzolia	-35.5761	1.501e+04	-0.0024	0.9981
varietyLambrusco	-16.6207	2.025e+04	-0.0008	0.9993
varietyMalvar	3.1725	2.128e+04	0.0001	0.9999
varietyMalvasia	-37.8795	1.958e+04	-0.0019	0.9985
varietyMerlot	-17.3865	8.309e+03	-0.0021	0.9983
varietyMontepulciano	-17.1486	1.232e+04	-0.0014	0.9989
varietyMoscato	-17.9329	1.773e+04	-0.0010	0.9992
varietyNegroamaro	2.9264	1.955e+00	1.4966	0.1345
varietyNerello Mascalese	2.7763	1.501e+04	0.0002	0.9999
varietyNero d'Avola	-35.4228	1.148e+04	-0.0031	0.9975
varietyPasserina	-15.9883	2.025e+04	-0.0008	0.9994
varietyPetit Manseng	2.6805	2.128e+04	0.0001	0.9999
varietyPinot Grigio	4.2208	1.290e+04	0.0003	0.9997
varietyPinot Nero	5.0630	1.039e+04	0.0005	0.9996
varietyPrimitivo	-16.7261	1.773e+04	-0.0009	0.9992
varietyProsecco	-34.4432	1.497e+04	-0.0023	0.9982
varietyPrugnolo Gentile	-34.4740	1.958e+04	-0.0018	0.9986
varietyRed Blend	-16.7683	8.309e+03	-0.0020	0.9984
varietyRibolla Gialla	3.3711	1.704e+04	0.0002	0.9998
varietyRiesling	4.7185	2.128e+04	0.0002	0.9998
varietySagrantino	-20.4858	2.025e+04	-0.0010	0.9992
varietySangiovese	-15.7580	8.309e+03	-0.0019	0.9985
varietySangiovese Grosso	-15.3647	8.309e+03	-0.0018	0.9985
varietySauvignon	3.9830	1.558e+04	0.0003	0.9998
varietySparkling Blend	-33.8110	1.504e+04	-0.0022	0.9982
varietySyrah	-16.3358	8.309e+03	-0.0020	0.9984
varietyTocai	5.49293	2.128e+04	0.0002	0.9998
varietyTrebiano	-15.7072	2.025e+04	-0.0008	0.9994
varietyTurbiana	-32.5874	2.479e+04	-0.0013	0.9990
varietyVerdicchio	-16.3699	1.256e+04	-0.0013	0.9990

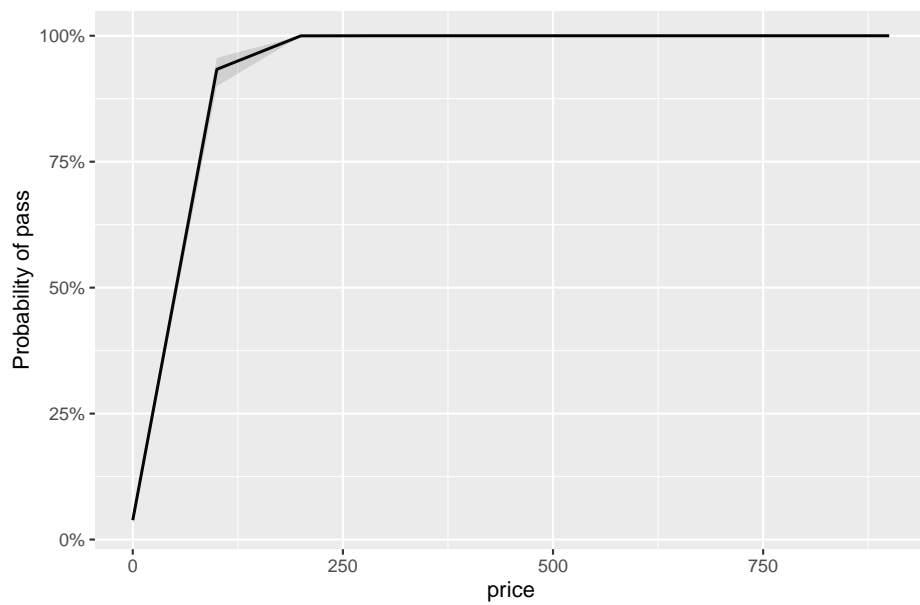


Figure 4: Pass Probability Plot by Price

Coclusion and Future Work

Reference

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