Project2

Hiba Hajali

20/03/2022

Introduction

The numbers of the missing values in each column:

```
##    country_of_origin                aroma               flavor
##                    0                    0                    0
##              acidity category_two_defects altitude_mean_meters
##                    0                    0                  162
##            harvested         Qualityclass
##                   55                    0
```

The data after we remove the missing values:

```
## Rows: 858
## Columns: 8
## $ country_of_origin    <chr> "Guatemala", "China", "Colombia", "Guatemala", "C~
## $ aroma                <dbl> 7.92, 7.67, 7.75, 7.83, 7.67, 8.17, 7.83, 7.67, 7~
## $ flavor               <dbl> 7.67, 7.67, 7.50, 7.67, 7.42, 8.00, 7.50, 7.75, 7~
## $ acidity              <dbl> 7.75, 7.67, 7.50, 7.33, 7.33, 7.17, 7.42, 7.67, 7~
## $ category_two_defects <int> 3, 3, 0, 1, 5, 0, 2, 1, 4, 0, 10, 0, 4, 4, 2, 4, ~
## $ altitude_mean_meters <dbl> 1650.00, 1600.00, 1750.00, 1310.64, 1600.00, 1750~
## $ harvested            <int> 2015, 2015, 2013, 2013, 2011, 2014, 2013, 2015, 2~
## $ Qualityclass         <chr> "Good", "Good", "Good", "Poor", "Poor", "Good", "~
```

The number of unique values in country of origin:

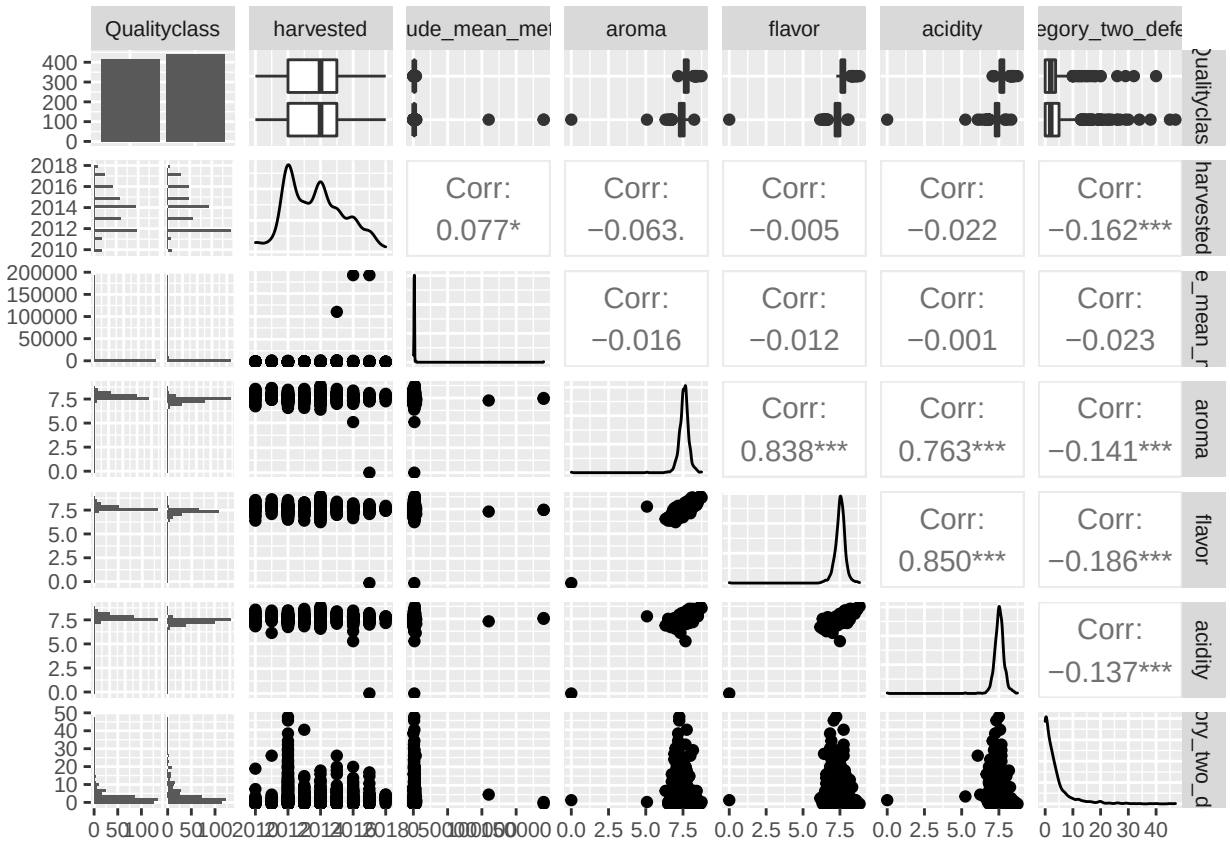
```
## [1] 34
```

The number of unique values in harvest year:

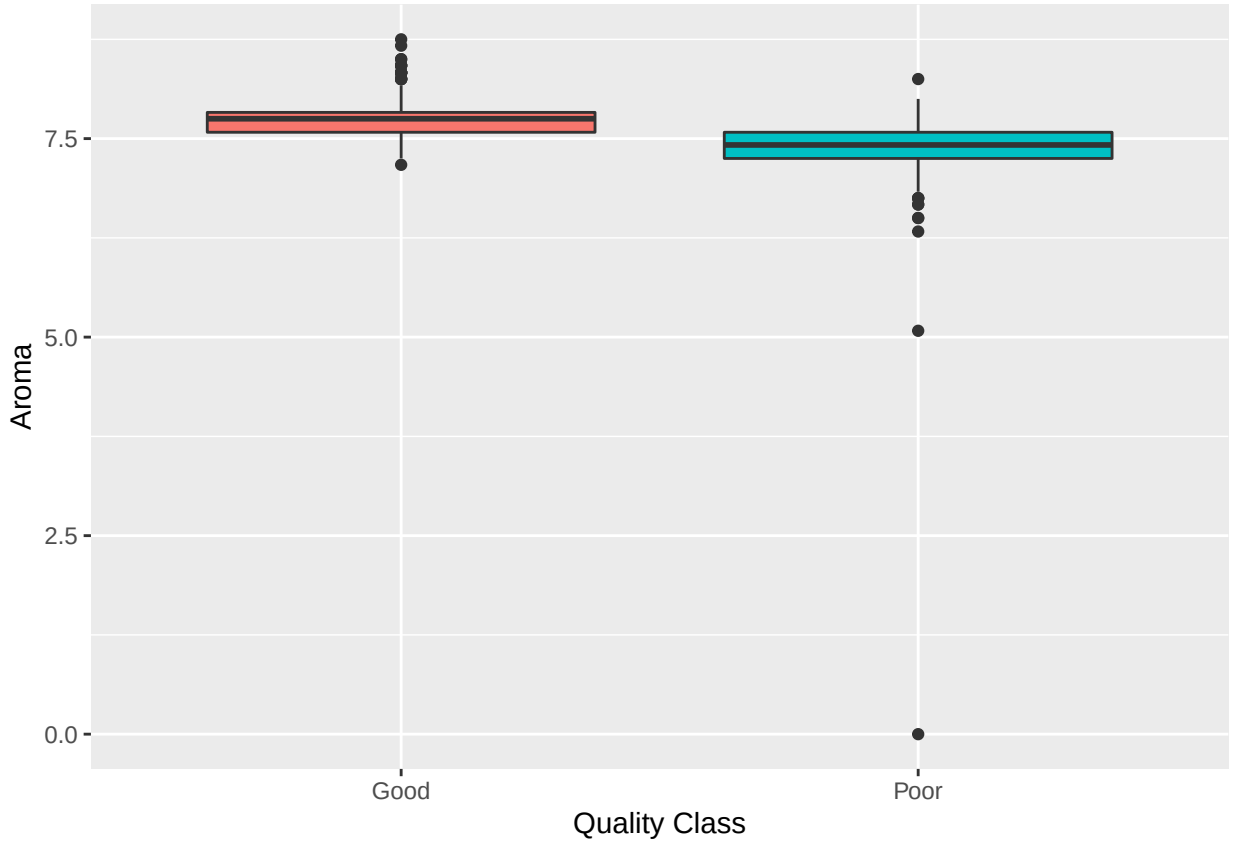
```
## [1] 9
```

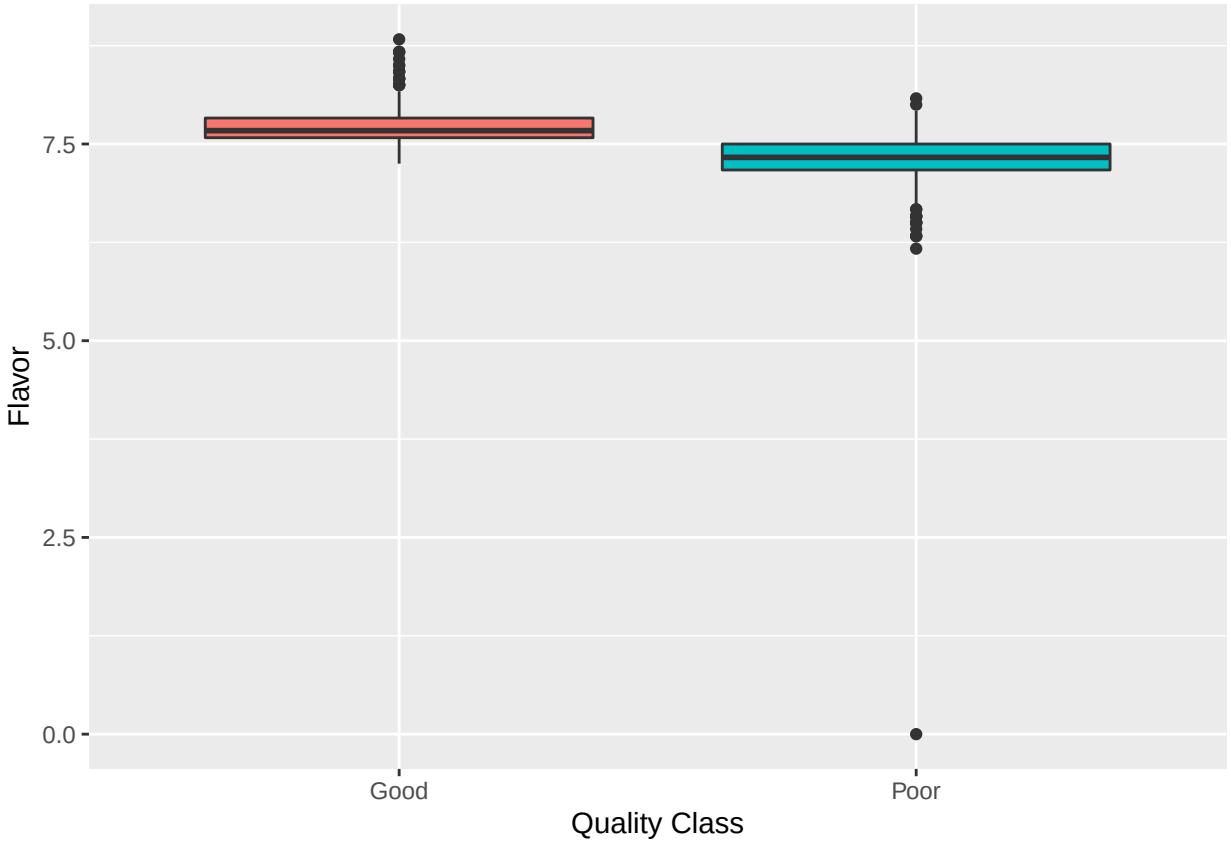
Explantory Analysis

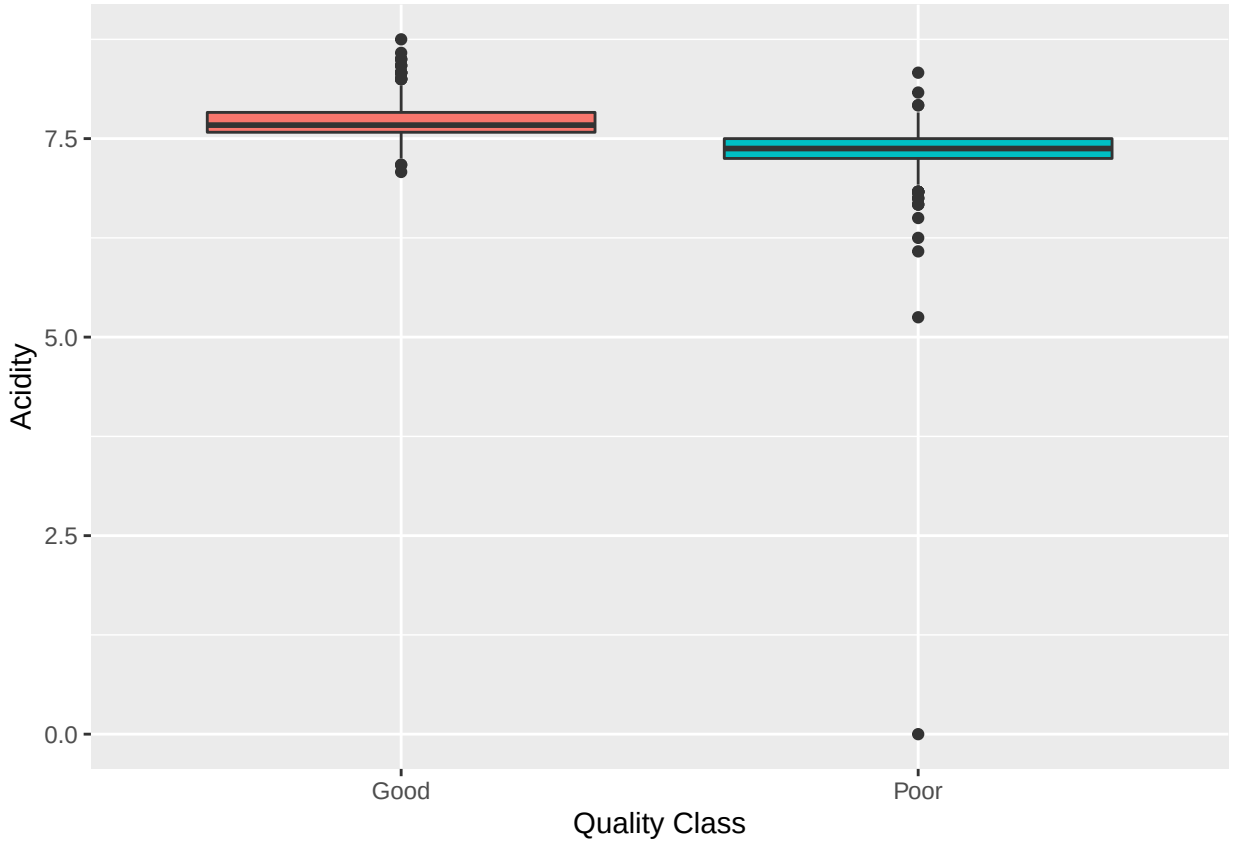
The correlation between the quantitative variables:

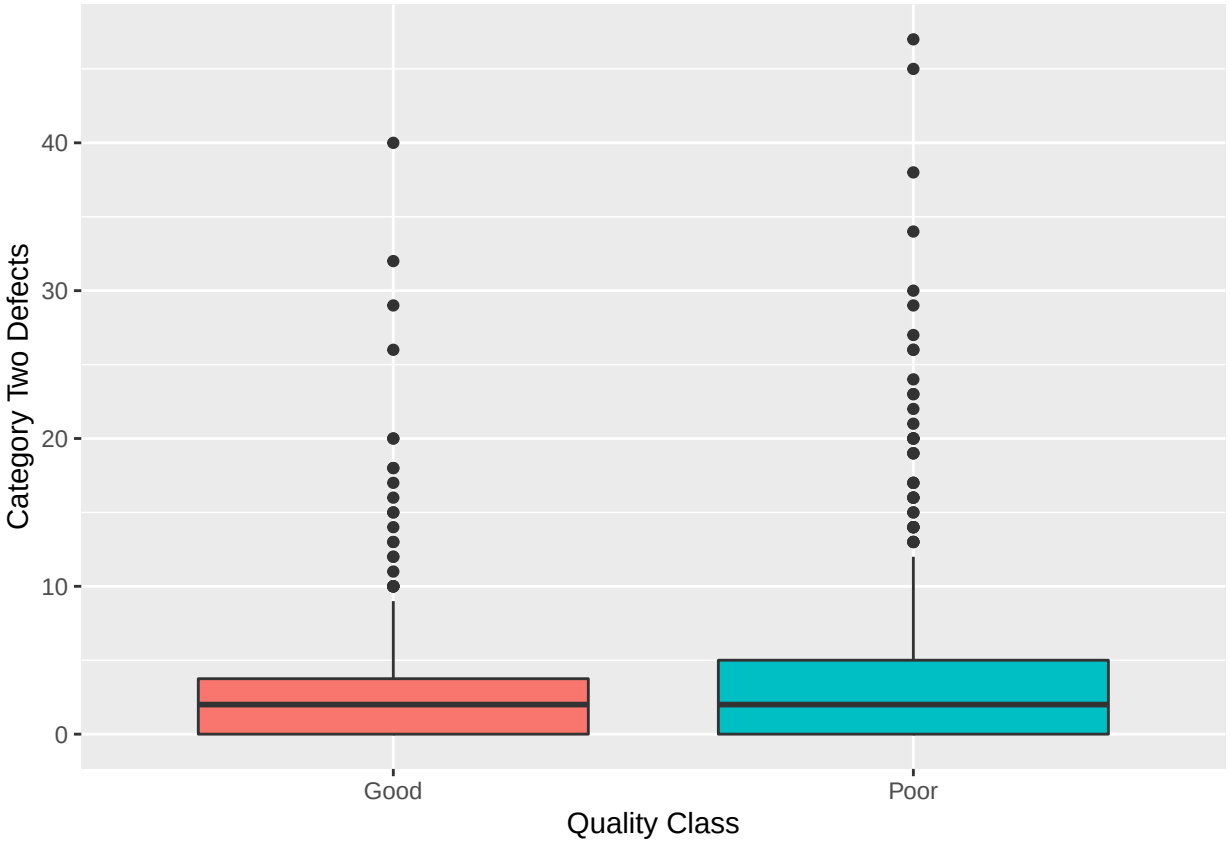


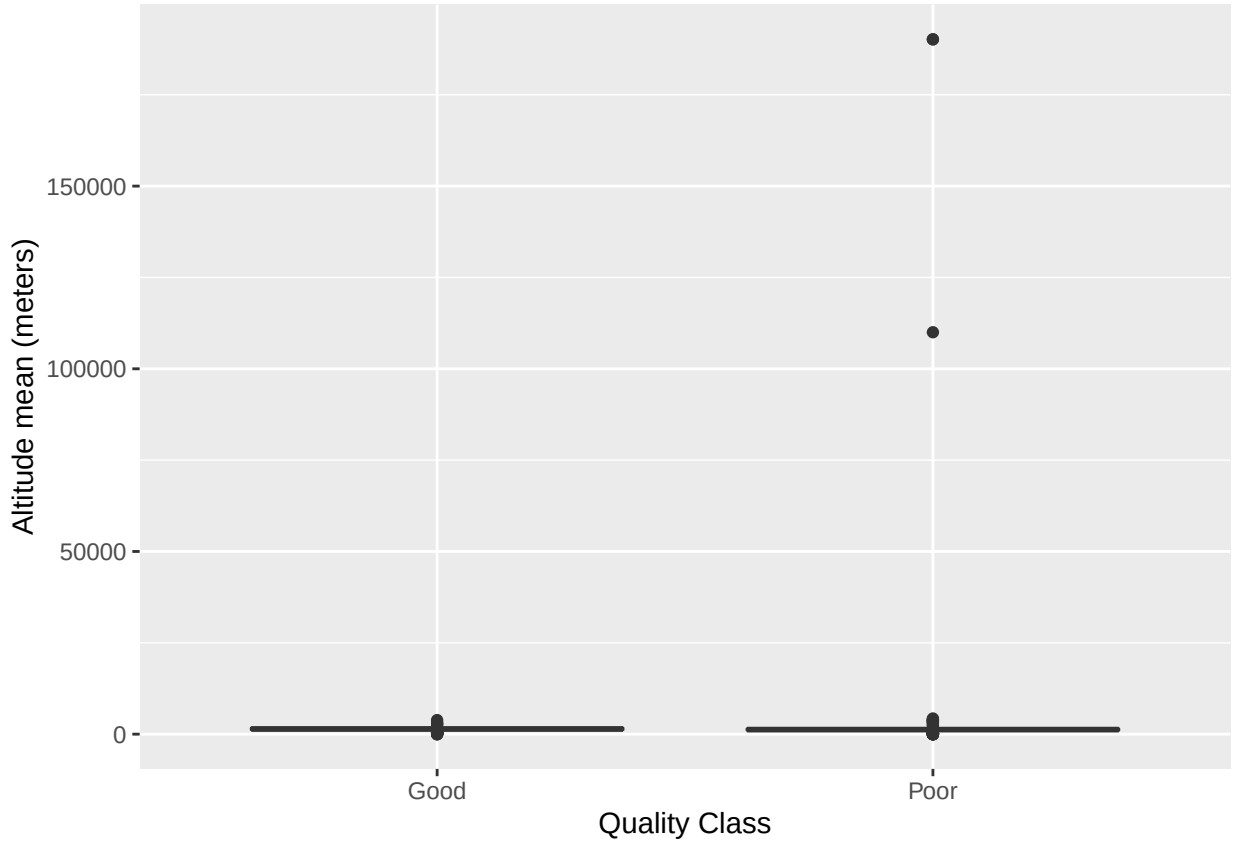
Box plots showing the distribution of the quantitative variables


Aroma
7.5
5.0
2.5
0.0
Good
Poor
Quality Class

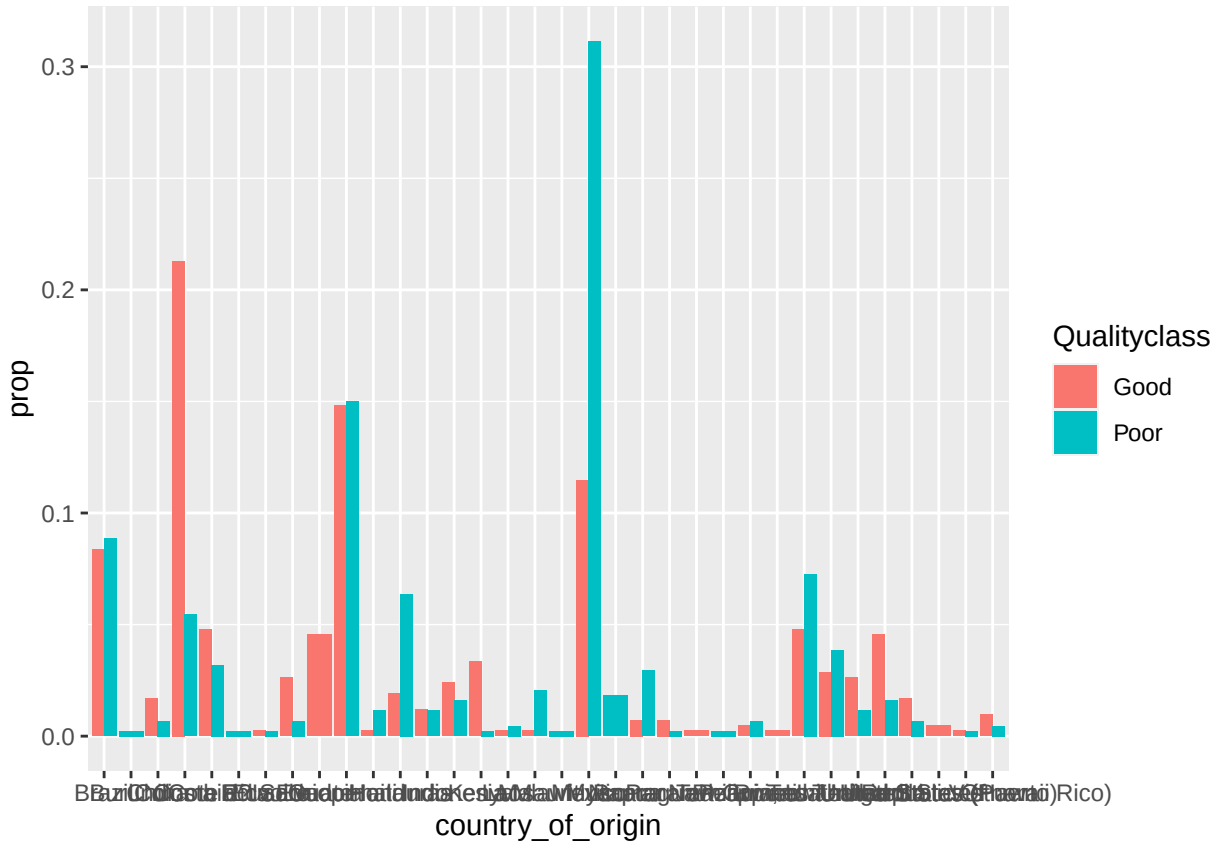

Flavor
7.5
5.0
2.5
0.0
Good
Poor
Quality Class

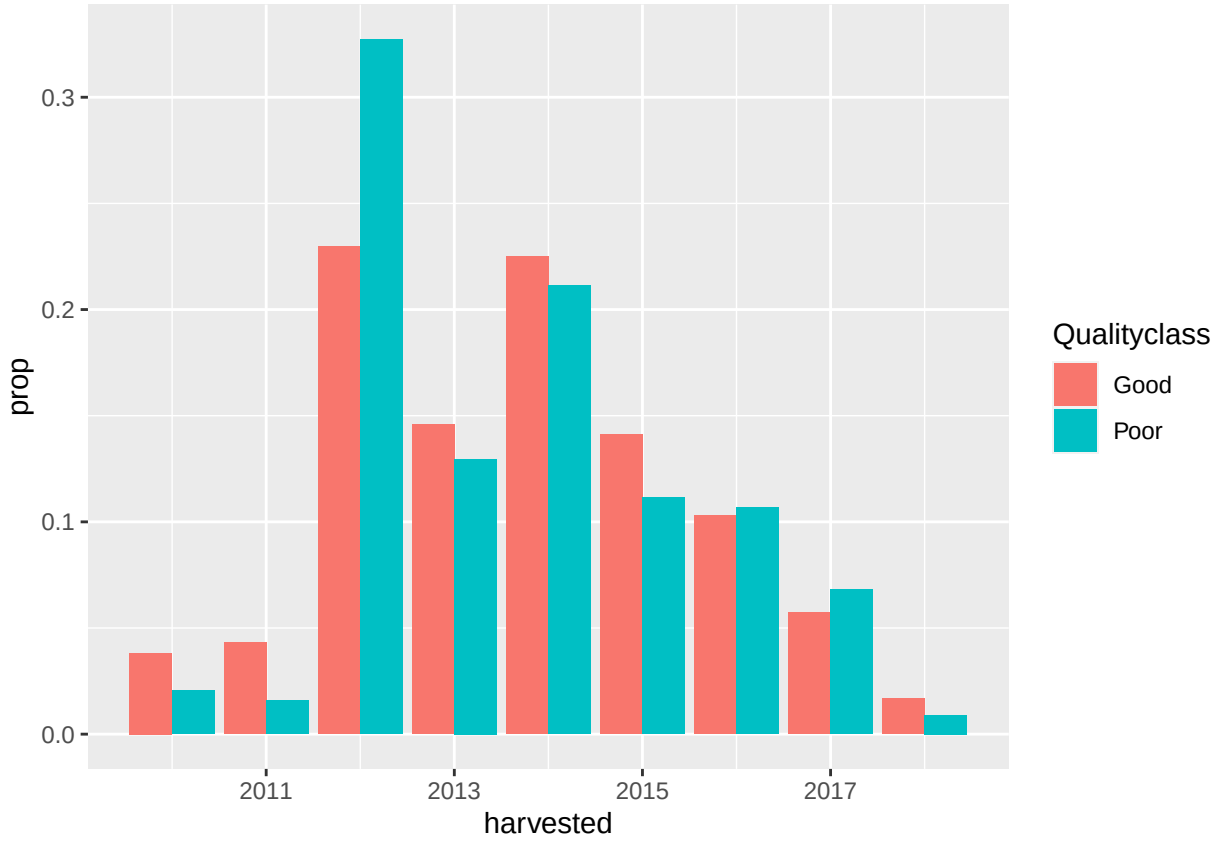

7.5
5.0
2.5
0.0
Acidity
Good
Poor
Quality Class


Category Two Defects
40
30
20
10
0
Good
Poor
Quality Class


150000
100000
50000
0
Altitude mean (meters)
Good
Poor
Quality Class

bar charts:


0.3
0.2
0.1
0.0
prop
Qualityclass
Good
Poor
country_of_origin



The percentages:

Table showing the percentage of the quality classes for each country

country_of_origin	Good	Poor
Brazil	47.3% (35)	52.7% (39)
Burundi	0.0% (0)	100.0% (1)
China	70.0% (7)	30.0% (3)
Colombia	78.8% (89)	21.2% (24)
Costa Rica	58.8% (20)	41.2% (14)
Cote d?Ivoire	0.0% (0)	100.0% (1)
Ecuador	50.0% (1)	50.0% (1)
El Salvador	78.6% (11)	21.4% (3)
Ethiopia	100.0% (19)	0.0% (0)
Guatemala	48.4% (62)	51.6% (66)
Haiti	16.7% (1)	83.3% (5)
Honduras	22.2% (8)	77.8% (28)
India	50.0% (5)	50.0% (5)
Indonesia	58.8% (10)	41.2% (7)
Kenya	93.3% (14)	6.7% (1)
Laos	33.3% (1)	66.7% (2)
Malawi	10.0% (1)	90.0% (9)
Mauritius	0.0% (0)	100.0% (1)
Mexico	25.9% (48)	74.1% (137)
Myanmar	0.0% (0)	100.0% (8)
Nicaragua	18.8% (3)	81.2% (13)
Panama	75.0% (3)	25.0% (1)

```
##              Papua New Guinea 100.0%  (1)   0.0%   (0)
##                          Peru   0.0%  (0) 100.0%   (1)
##                   Philippines  40.0%  (2)  60.0%   (3)
##                        Rwanda 100.0%  (1)   0.0%   (0)
##                        Taiwan  38.5% (20)  61.5%  (32)
##  Tanzania, United Republic Of  41.4% (12)  58.6%  (17)
##                       Thailand  68.8% (11)  31.2%   (5)
##                        Uganda  73.1% (19)  26.9%   (7)
##                 United States  70.0%  (7)  30.0%   (3)
##        United States (Hawaii) 100.0%  (2)   0.0%   (0)
##   United States (Puerto Rico)  50.0%  (1)  50.0%   (1)
##                       Vietnam  66.7%  (4)  33.3%   (2)
```

Table showing the percentage of the quality classes for each harvest year:

```
##  harvested       Good        Poor
##       2010 64.0% (16) 36.0%   (9)
##       2011 72.0% (18) 28.0%   (7)
##       2012 40.0% (96) 60.0% (144)
##       2013 51.7% (61) 48.3%  (57)
##       2014 50.3% (94) 49.7%  (93)
##       2015 54.6% (59) 45.4%  (49)
##       2016 47.8% (43) 52.2%  (47)
##       2017 44.4% (24) 55.6%  (30)
##       2018 63.6%  (7) 36.4%   (4)
```

Formal Analsis

Model 1:

$$Qualityclass \sim country\_of\_origin + aroma + flavor + acidity + category\_two\_defects + altitude\_mean\_meters$$

Observations	858
Dependent variable	Qualityclass
Type	Generalized linear model
Family	binomial
Link	logit

$\chi^2(39)$	762.87
Pseudo-R^2 (Cragg-Uhler)	0.79
Pseudo-R^2 (McFadden)	0.64
AIC	506.01
BIC	696.19

Model 2:

$$Qualityclass \sim country\_of\_origin + aroma + flavor + acidity + category\_two\_defects + harvested$$

	Est.	S.E.	z val.	p
(Intercept)	374.59	163.12	2.30	0.02
country_of_originBurundi	12.54	6522.64	0.00	1.00
country_of_originChina	-0.81	1.07	-0.75	0.45
country_of_originColombia	-2.22	0.53	-4.19	0.00
country_of_originCosta Rica	-0.88	0.79	-1.12	0.26
country_of_originCote d?Ivoire	12.55	6522.64	0.00	1.00
country_of_originEcuador	1.37	1.48	0.93	0.35
country_of_originEl Salvador	-1.66	1.17	-1.42	0.16
country_of_originEthiopia	-14.53	1069.95	-0.01	0.99
country_of_originGuatemala	0.37	0.48	0.77	0.44
country_of_originHaiti	-2.40	1.79	-1.34	0.18
country_of_originHonduras	0.45	0.71	0.63	0.53
country_of_originIndia	2.58	0.93	2.76	0.01
country_of_originIndonesia	-0.23	0.86	-0.27	0.78
country_of_originKenya	-0.51	1.60	-0.32	0.75
country_of_originLaos	-0.88	1.81	-0.49	0.63
country_of_originMalawi	0.59	1.22	0.48	0.63
country_of_originMauritius	12.52	6522.64	0.00	1.00
country_of_originMexico	0.56	0.50	1.11	0.27
country_of_originMyanmar	15.57	2066.24	0.01	0.99
country_of_originNicaragua	-0.28	1.65	-0.17	0.87
country_of_originPanama	-3.33	1.77	-1.89	0.06
country_of_originPapua New Guinea	-4.44	6522.64	-0.00	1.00
country_of_originPeru	13.75	6522.64	0.00	1.00
country_of_originPhilippines	-2.69	2.51	-1.07	0.28
country_of_originRwanda	-13.14	6522.64	-0.00	1.00
country_of_originTaiwan	0.03	0.68	0.04	0.96
country_of_originTanzania, United Republic Of	-1.39	0.71	-1.96	0.05
country_of_originThailand	-2.12	0.86	-2.46	0.01
country_of_originUganda	0.99	0.74	1.33	0.18
country_of_originUnited States	-0.31	1.42	-0.22	0.83
country_of_originUnited States (Hawaii)	-7.69	4217.60	-0.00	1.00
country_of_originUnited States (Puerto Rico)	-1.36	8.89	-0.15	0.88
country_of_originVietnam	-2.60	1.29	-2.02	0.04
aroma	-4.30	0.82	-5.24	0.00
flavor	-8.83	1.10	-8.01	0.00
acidity	-4.85	0.84	-5.76	0.00
category_two_defects	-0.06	0.03	-1.77	0.08
altitude_mean_meters	0.00	0.00	0.33	0.74
harvested	-0.12	0.08	-1.48	0.14

Standard errors: MLE

Observations	858
Dependent variable	Qualityclass
Type	Generalized linear model
Family	binomial
Link	logit

$\chi^2(38)$	762.49
Pseudo-R^2 (Cragg-Uhler)	0.79
Pseudo-R^2 (McFadden)	0.64
AIC	504.39
BIC	689.81

	Est.	S.E.	z val.	p
(Intercept)	371.50	163.09	2.28	0.02
country_of_originBurundi	12.53	6522.64	0.00	1.00
country_of_originChina	-0.80	1.07	-0.75	0.45
country_of_originColombia	-2.21	0.53	-4.17	0.00
country_of_originCosta Rica	-0.87	0.79	-1.11	0.27
country_of_originCote d?Ivoire	12.51	6522.64	0.00	1.00
country_of_originEcuador	1.37	1.48	0.92	0.36
country_of_originEl Salvador	-1.66	1.17	-1.42	0.16
country_of_originEthiopia	-14.51	1069.16	-0.01	0.99
country_of_originGuatemala	0.40	0.48	0.83	0.41
country_of_originHaiti	-2.40	1.79	-1.34	0.18
country_of_originHonduras	0.45	0.71	0.64	0.52
country_of_originIndia	2.59	0.94	2.77	0.01
country_of_originIndonesia	-0.22	0.86	-0.26	0.79
country_of_originKenya	-0.50	1.60	-0.31	0.76
country_of_originLaos	-0.88	1.82	-0.48	0.63
country_of_originMalawi	0.59	1.22	0.49	0.63
country_of_originMauritius	12.49	6522.64	0.00	1.00
country_of_originMexico	0.57	0.50	1.14	0.26
country_of_originMyanmar	15.59	2066.08	0.01	0.99
country_of_originNicaragua	-0.26	1.65	-0.16	0.87
country_of_originPanama	-3.33	1.77	-1.88	0.06
country_of_originPapua New Guinea	-4.38	6522.64	-0.00	1.00
country_of_originPeru	13.75	6522.64	0.00	1.00
country_of_originPhilippines	-2.69	2.52	-1.07	0.29
country_of_originRwanda	-13.11	6522.64	-0.00	1.00
country_of_originTaiwan	0.03	0.69	0.04	0.97
country_of_originTanzania, United Republic Of	-1.38	0.71	-1.95	0.05
country_of_originThailand	-2.12	0.86	-2.46	0.01
country_of_originUganda	1.01	0.74	1.35	0.18
country_of_originUnited States	-0.28	1.42	-0.20	0.84
country_of_originUnited States (Hawaii)	-7.65	4215.57	-0.00	1.00
country_of_originUnited States (Puerto Rico)	-1.36	8.97	-0.15	0.88
country_of_originVietnam	-2.60	1.29	-2.01	0.04
aroma	-4.31	0.82	-5.25	0.00
flavor	-8.87	1.10	-8.04	0.00
acidity	-4.86	0.84	-5.76	0.00
category_two_defects	-0.06	0.03	-1.79	0.07
harvested	-0.12	0.08	-1.46	0.15

Standard errors: MLE

Model 3:

$$Qualityclass \sim country\_of\_origin + aroma + flavor + acidity + category\_two\_defects$$

Observations	858
Dependent variable	Qualityclass
Type	Generalized linear model
Family	binomial
Link	logit

$\chi^2(37)$	760.37
Pseudo-R² (Cragg-Uhler)	0.78
Pseudo-R² (McFadden)	0.64
AIC	504.51
BIC	685.18

Model 4:

$$Qualityclass \sim aroma + flavor + acidity + category\_two\_defects$$

Model 5:

$$Qualityclass \sim aroma + flavor + acidity$$

Models comparison:

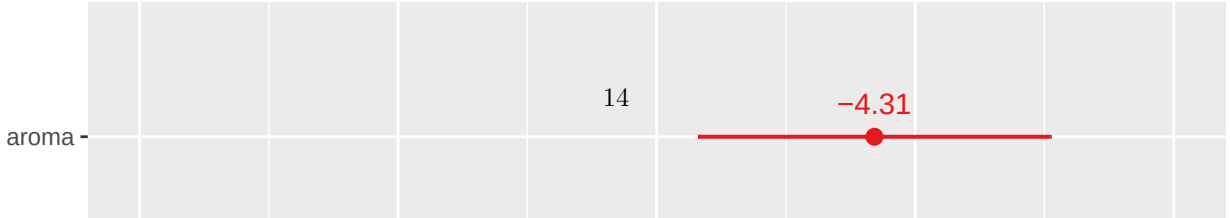
```
## $Models
##   Formula
## 1 "Qualityclass ~ country_of_origin + aroma + flavor + acidity + category_two_defects + altitude_mea
## 2 "Qualityclass ~ country_of_origin + aroma + flavor + acidity + category_two_defects + harvested"
## 3 "Qualityclass ~ country_of_origin + aroma + flavor + acidity + category_two_defects"
## 4 "Qualityclass ~ aroma + flavor + acidity + category_two_defects"
## 5 "Qualityclass ~ aroma + flavor + acidity"
##
## $Fit.criteria
##   Rank Df.res   AIC  AICc   BIC McFadden Cox.and.Snell Nagelkerke    p.value
## 1   40    818 508.0 512.2 702.9   0.6417        0.5890     0.7855 2.281e-135
## 2   39    819 506.4 510.4 696.6   0.6414        0.5888     0.7852 6.046e-136
## 3   38    820 506.5 510.3 691.9   0.6396        0.5878     0.7839 3.644e-136
## 4    5    853 529.5 529.6 558.0   0.5648        0.5428     0.7238 2.673e-144
## 5    4    854 527.7 527.7 551.4   0.5646        0.5426     0.7237 1.835e-145
```

Odds Plot:

	Est.	S.E.	z val.	p
(Intercept)	135.47	10.98	12.34	0.00
country_of_originBurundi	12.42	6522.64	0.00	1.00
country_of_originChina	-0.87	1.07	-0.81	0.42
country_of_originColombia	-1.99	0.51	-3.94	0.00
country_of_originCosta Rica	-0.88	0.78	-1.12	0.26
country_of_originCote d?Ivoire	12.36	6522.64	0.00	1.00
country_of_originEcuador	1.36	1.48	0.92	0.36
country_of_originEl Salvador	-1.82	1.17	-1.55	0.12
country_of_originEthiopia	-14.55	1059.52	-0.01	0.99
country_of_originGuatemala	0.39	0.48	0.80	0.42
country_of_originHaiti	-2.16	1.81	-1.19	0.23
country_of_originHonduras	0.30	0.70	0.43	0.67
country_of_originIndia	2.51	0.93	2.71	0.01
country_of_originIndonesia	-0.21	0.86	-0.24	0.81
country_of_originKenya	-0.68	1.65	-0.41	0.68
country_of_originLaos	-0.95	1.81	-0.53	0.60
country_of_originMalawi	0.65	1.22	0.53	0.59
country_of_originMauritius	12.35	6522.64	0.00	1.00
country_of_originMexico	0.80	0.48	1.67	0.10
country_of_originMyanmar	15.56	2074.66	0.01	0.99
country_of_originNicaragua	-0.30	1.74	-0.17	0.86
country_of_originPanama	-3.12	1.81	-1.72	0.08
country_of_originPapua New Guinea	-4.22	6522.64	-0.00	1.00
country_of_originPeru	14.05	6522.64	0.00	1.00
country_of_originPhilippines	-2.65	2.44	-1.09	0.28
country_of_originRwanda	-13.22	6522.64	-0.00	1.00
country_of_originTaiwan	0.03	0.67	0.05	0.96
country_of_originTanzania, United Republic Of	-1.34	0.71	-1.89	0.06
country_of_originThailand	-2.01	0.85	-2.36	0.02
country_of_originUganda	1.08	0.74	1.46	0.15
country_of_originUnited States	-0.13	1.43	-0.09	0.93
country_of_originUnited States (Hawaii)	-7.25	4214.84	-0.00	1.00
country_of_originUnited States (Puerto Rico)	-1.19	8.77	-0.14	0.89
country_of_originVietnam	-2.62	1.32	-1.99	0.05
aroma	-4.21	0.82	-5.15	0.00
flavor	-8.79	1.10	-8.02	0.00
acidity	-4.88	0.84	-5.81	0.00
category_two_defects	-0.06	0.03	-1.79	0.07

Standard errors: MLE

Observations	858
Dependent variable	Qualityclass
Type	Generalized linear model
Family	binomial
Link	logit



$\chi^2(4)$	671.42
Pseudo-R² (Cragg-Uhler)	0.72
Pseudo-R² (McFadden)	0.56
AIC	527.45
BIC	551.23

	Est.	S.E.	z val.	p
(Intercept)	117.46	8.62	13.63	0.00
aroma	-4.33	0.70	-6.21	0.00
flavor	-7.41	0.88	-8.38	0.00
acidity	-3.81	0.70	-5.42	0.00
category_two_defects	-0.01	0.03	-0.47	0.64

Standard errors: MLE

Observations	858
Dependent variable	Qualityclass
Type	Generalized linear model
Family	binomial
Link	logit

$\chi^2(3)$	671.21
Pseudo-R² (Cragg-Uhler)	0.72
Pseudo-R² (McFadden)	0.56
AIC	525.67
BIC	544.69

	Est.	S.E.	z val.	p
(Intercept)	117.18	8.59	13.65	0.00
aroma	-4.31	0.70	-6.20	0.00
flavor	-7.39	0.89	-8.35	0.00
acidity	-3.80	0.70	-5.41	0.00

Standard errors: MLE

Confidence intervals:

```
##                  2.5 %     97.5 %
## (Intercept) 101.235732 134.953218
## aroma        -5.720601  -2.988415
## flavor       -9.197355  -5.721335
## acidity      -5.211874  -2.449519
```

Conclusion