Project2

Hiba Hajali

20/03/2022

1 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:

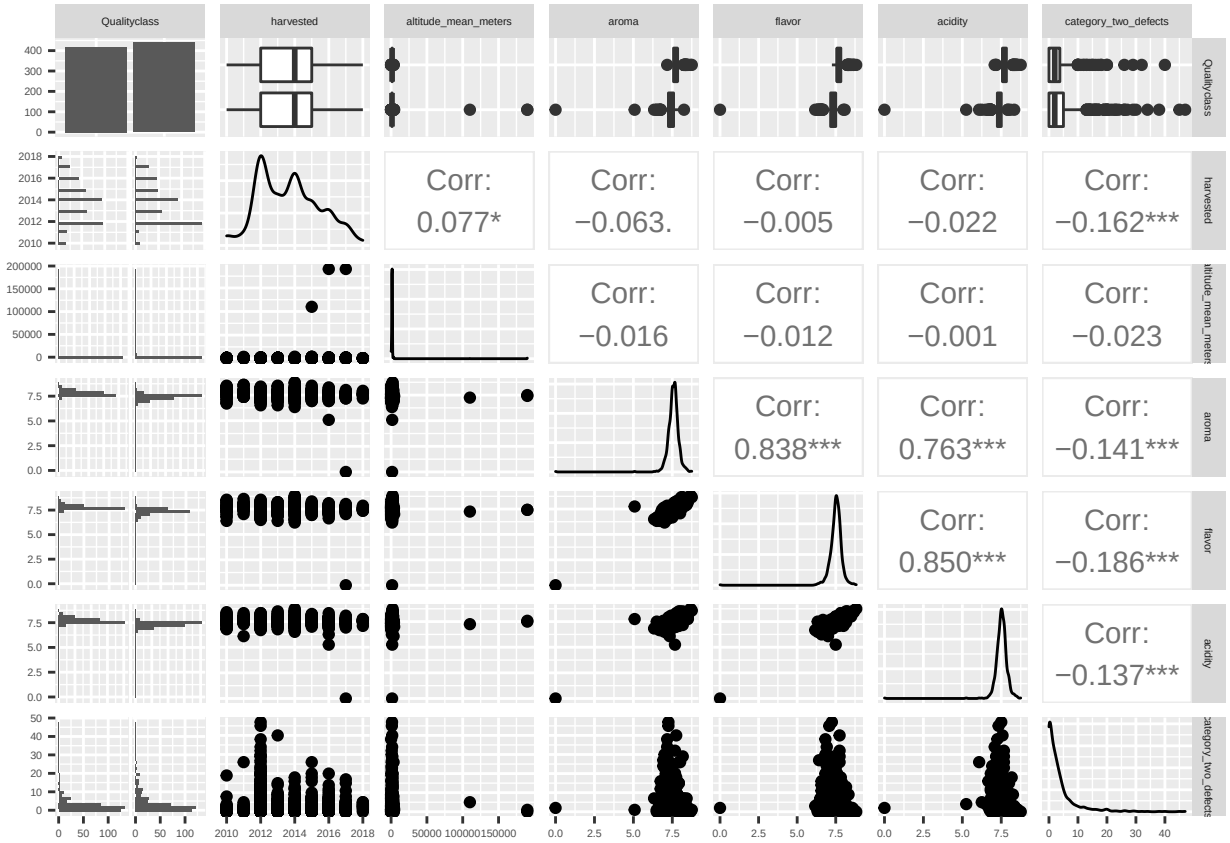
```
## [1] 34
```

The number of unique values in harvest year:

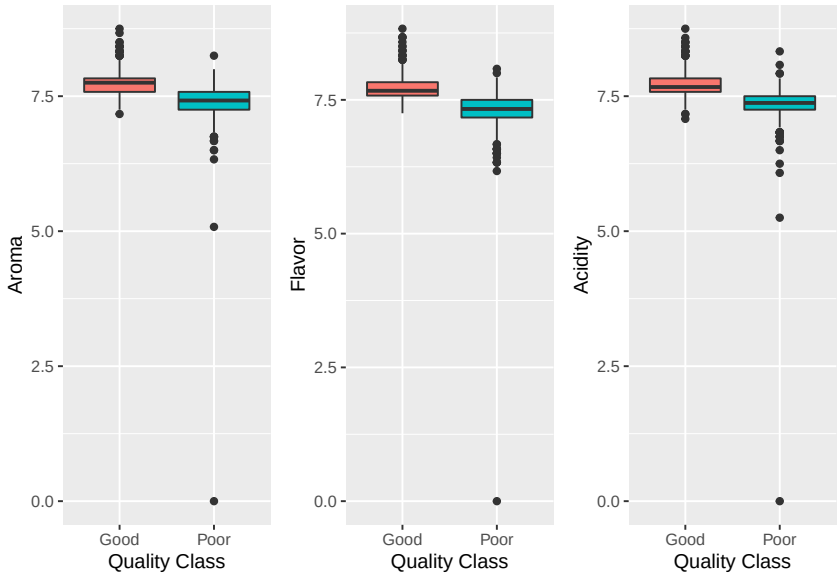
```
## [1] 9
```

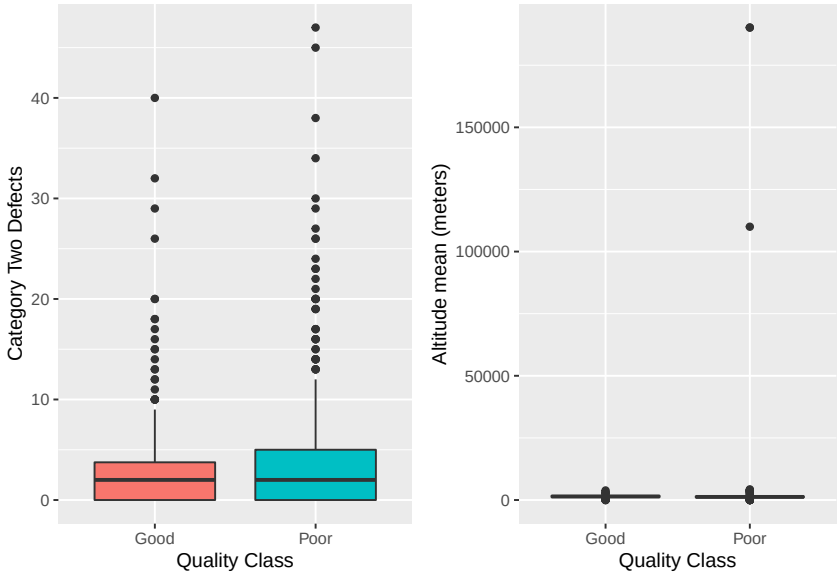
2 Explantory Analysis

The correlation between the quantitative variables:

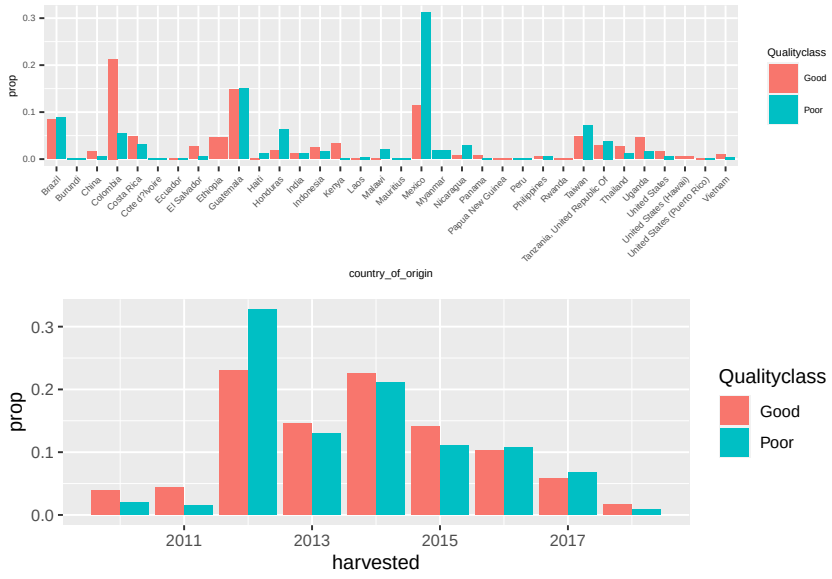


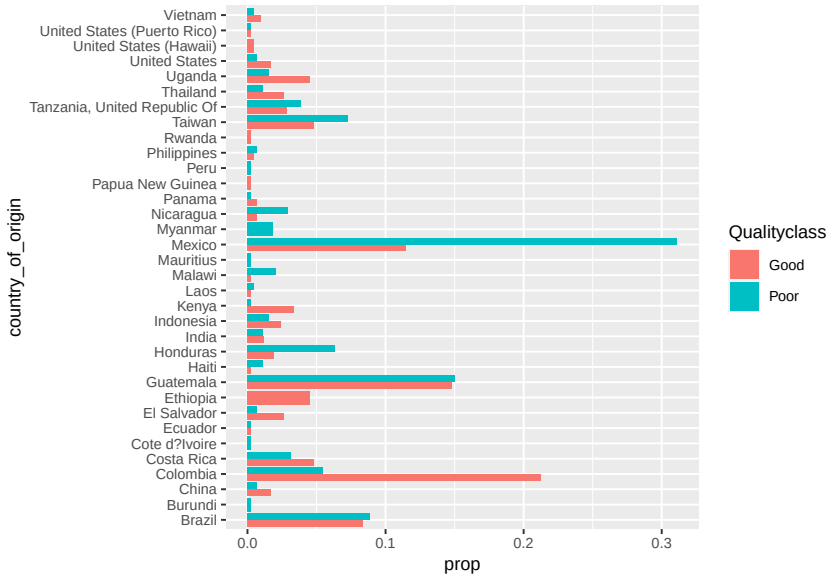
Box plots showing the distribution of the quantitative variables

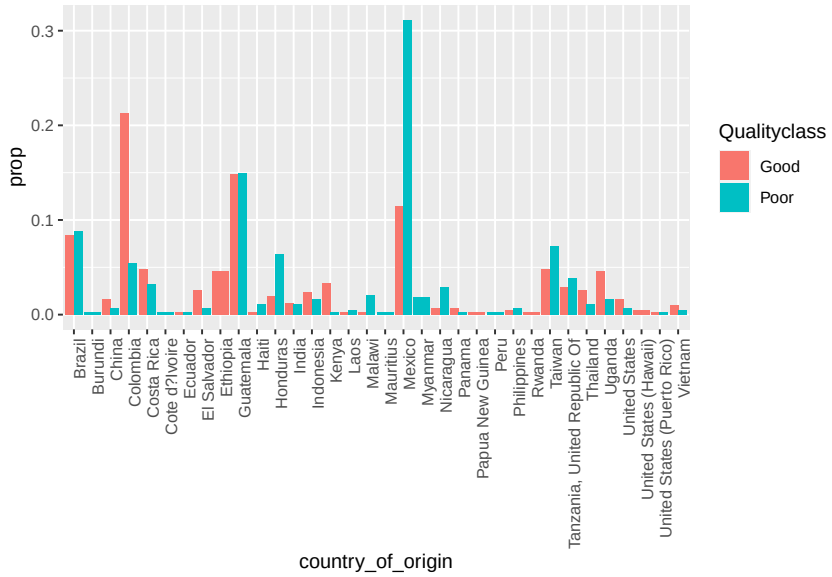




2.1 bar charts:







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)
```

3 Formal Analsis

Model 1:

$$ln\left(\frac{p_{Poor}}{1-p_{Poor}}\right) = \alpha + \beta_1 \cdot \text{Country} + \beta_2 \cdot \text{Aroma} + \beta_3 \cdot \text{Flavor} + \beta_4 \cdot \text{Acidity} + \beta_5 \cdot \text{Category Two Defects} + \beta_6 \cdot \text{Harvested} + \beta_7 \cdot \text{Altitude}$$

Model 2:

$$ln\left(\frac{p_{Poor}}{1-p_{Poor}}\right) = \alpha + \beta_1 \cdot \text{Country} + \beta_2 \cdot \text{Aroma} + \beta_3 \cdot \text{Flavor} + \beta_4 \cdot \text{Acidity} + \beta_5 \cdot \text{Category Two Defects} + \beta_6 \cdot \text{Harvested}$$

Model 3:

$$ln\left(\frac{p_{poor}}{1-p_{poor}}\right) = \alpha + \beta_1 \cdot \text{Country of origin} + \beta_2 \cdot \text{aroma} + \beta_3 \cdot \text{flavor} + \beta_4 \cdot \text{acidity} + \beta_5 \cdot \text{category two defects}$$

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 4:

$$ln\left(\frac{p_{Poor}}{1-p_{Poor}}\right) = \alpha + \beta_1 \cdot \text{Aroma} + \beta_2 \cdot \text{Flavor} + \beta_3 \cdot \text{Acidity} + \beta_4 \cdot \text{Category Two Defects}$$

Model 5:

$$ln\left(\frac{p_{Poor}}{1-p_{Poor}}\right) = \alpha + \beta_1 \cdot \text{Aroma} + \beta_2 \cdot \text{Flavor} + \beta_3 \cdot \text{Acidity}$$

3.1 Models comparison:

Table 1: The Result of Model comparison

Formula								
Qualityclass ~ country_of_origin + aroma + flavor + acidity + category_two_defects + altitude_mean_meters + harvested								
Qualityclass ~ country_of_origin + aroma + flavor + acidity + category_two_defects + harvested								
Qualityclass ~ country_of_origin + aroma + flavor + acidity + category_two_defects								
Qualityclass ~ aroma + flavor + acidity + category_two_defects								
Qualityclass ~ aroma + flavor + acidity								
Rank	Df.res	AIC	AICc	BIC	McFadden	Cox.and.Snell	Nagelkerke	p.value
40	818	508.0	512.2	702.9	0.642	0.589	0.785	0
39	819	506.4	510.4	696.6	0.641	0.589	0.785	0
38	820	506.5	510.3	691.9	0.640	0.588	0.784	0
5	853	529.5	529.6	558.0	0.565	0.543	0.724	0
4	854	527.7	527.7	551.4	0.565	0.543	0.724	0

	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

Table 2: Confidence Intervals for log odds in Model 5

	2.5 %	97.5 %
(Intercept)	101.235732	134.953218
aroma	-5.720601	-2.988415
flavor	-9.197355	-5.721335
acidity	-5.211874	-2.449520

3.2 log Odds:

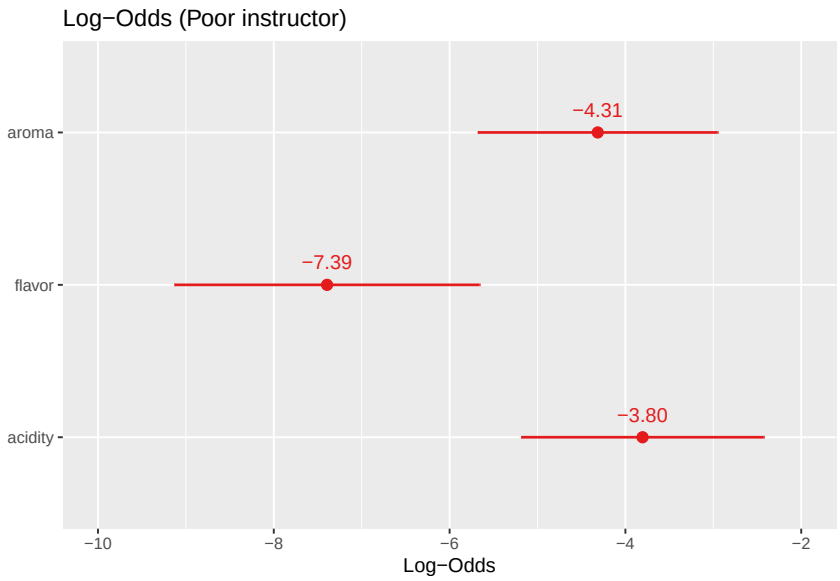


Figure 1: Log Odds r each quality class in every continents.

3.3 Confidence Intervals:

3.4 Extend Analysis-Prediction Assesment.

Confusion Matrix

```
##
## Call:
## glm(formula = Qualityclass ~ aroma + flavor + acidity, family = binomial(link = "logit"),
##     data = train_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2530  -0.4579  -0.0001   0.3704   4.0055
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 116.4801     9.5944  12.140  < 2e-16 ***
## aroma        -3.8676     0.7514  -5.147 2.64e-07 ***
```

Table 3: Accuracy of Prediction.

	Value
Accuracy	0.8779070
Kappa	0.7565382
AccuracyLower	0.8194395
AccuracyUpper	0.9228098
AccuracyNull	0.5639535
AccuracyPValue	0.0000000
McnemarPValue	0.0088288

Table 4: The Resule of Sensitivity and Specificity of Prediction.

	Value
Sensitivity	0.8247423
Specificity	0.9466667
Pos Pred Value	0.9523810
Neg Pred Value	0.8068182
Precision	0.9523810
Recall	0.8247423
F1	0.8839779
Prevalence	0.5639535
Detection Rate	0.4651163
Detection Prevalence	0.4883721
Balanced Accuracy	0.8857045

```
## flavor       -7.4868     0.9848  -7.603 2.90e-14 ***
## acidity      -4.0783     0.7916  -5.152 2.58e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 951.00  on 685  degrees of freedom
## Residual deviance: 422.28  on 682  degrees of freedom
## AIC: 430.28
##
## Number of Fisher Scoring iterations: 7
```

ROC Curve

Table 5: Confuse table.

	Actual Good	Actual Bad
0	71	17
1	4	80

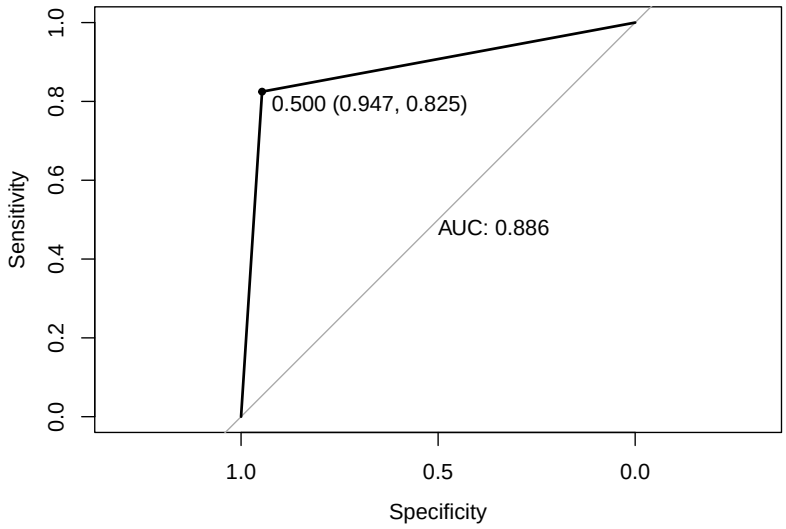


Figure 2: ROC cureve for model predicton

4 Conclusion