Group_9_Analysis

Brent Strong, Enyu Li, Haotian Wang, Honjin Ren, Mu He

3/7/2022

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

##
## Attaching package: 'kableExtra'

## The following object is masked from 'package:dplyr':
##
##     group_rows

##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine

##
## Attaching package: 'olsrr'

## The following object is masked from 'package:datasets':
##
##     rivers

## Install package "strengejacke" from GitHub (`devtools::install_github("strengejacke/strengejacke")`)

##
## Attaching package: 'MASS'

## The following object is masked from 'package:olsrr':
##
##     cement
```

```
## The following object is masked from 'package:dplyr':
##
##     select

##
## Attaching package: 'janitor'

## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test

## Loading required package: lattice

##
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':
##
##     lift

## Rows: 1145 Columns: 8

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): country_of_origin, Qualityclass
## dbl (6): aroma, flavor, acidity, category_two_defects, altitude_mean_meters,...

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.

## Rows: 1,145
## Columns: 8
## $ country_of_origin    <chr> "Myanmar", "Uganda", "Ethiopia", "Mexico", "Burun~
## $ aroma                <dbl> 7.25, 8.33, 8.42, 7.17, 7.75, 7.92, 7.92, 7.83, 7~
## $ flavor               <dbl> 7.42, 7.92, 8.00, 7.08, 7.67, 7.75, 7.83, 7.67, 6~
## $ acidity              <dbl> 7.50, 7.92, 8.00, 7.25, 7.50, 7.75, 7.67, 7.58, 7~
## $ category_two_defects <dbl> 4, 1, 7, 3, 5, 0, 1, 2, 2, 1, 0, 8, 0, 2, 0, 0, 2~
## $ altitude_mean_meters <dbl> 1219.20, 1600.00, 1700.00, 1300.00, 1880.00, 1400~
## $ harvested            <dbl> 2015, 2013, 2014, 2012, 2012, 2014, NA, 2015, 201~
## $ Qualityclass         <chr> "Poor", "Good", "Good", "Poor", "Good", "Good", "~

## [1] 0.5135371
```

Table 1: Summary statistics of altitude mean meters and harvested.

Variable	Mean	SD	Min.	1st Q.	Median	3rd Q.	Max.
aroma	7.57	0.39	0	7.42	7.58	7.75	8.75
flavor	7.52	0.40	0	7.33	7.58	7.75	8.67
acidity	7.54	0.39	0	7.33	7.50	7.75	8.58
category_two_defects	3.67	5.41	0	0.00	2.00	5.00	55.00
altitude_mean_meters	1850.69	9392.09	1	1100.00	1310.64	1600.00	190164.00
harvested	2013.67	1.81	2010	2012.00	2014.00	2015.00	2018.00

```
my_skim <- skim_with(base = sfl(n = length))
my.analysis <- analysis %>%
  mutate(Qualityclassindicator = as.numeric(Qualityclass=="Good"))
my.analysis %>%
  group_by(country_of_origin) %>%
  dplyr::select(country_of_origin, Qualityclassindicator) %>%
  my_skim() %>%
  dplyr::select(country_of_origin, n, numeric.mean) %>%
  transmute(country_of_origin=country_of_origin,
            number_of_batch=n,
            Proportion_of_good_quality=numeric.mean) %>%
  kable(caption = '\\label{tab:countryskim} Summary statistics of the sepal length by species of irises
  kable_styling(font_size = 10, latex_options = "HOLD_position")
```

Table 2: Summary statistics of the sepal length by species of irises

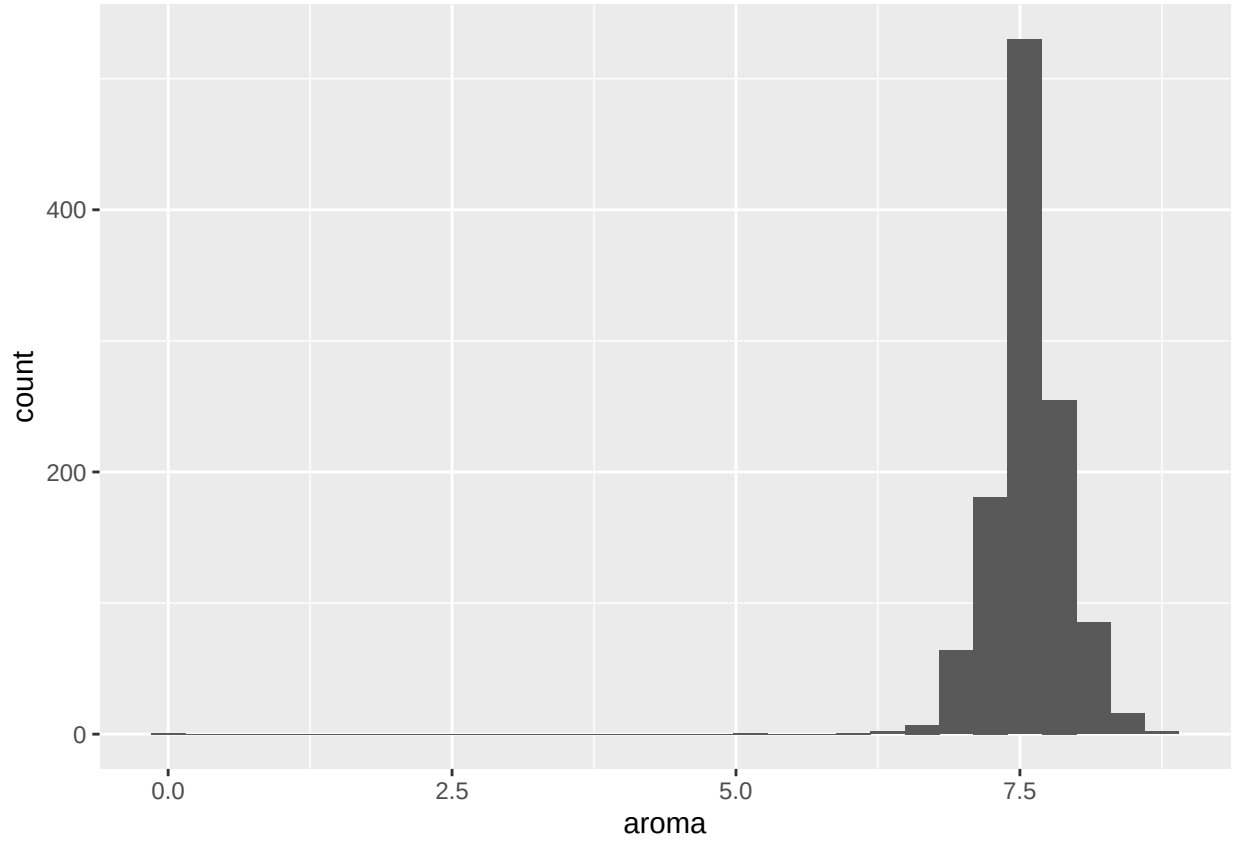
country_of_origin	number_of_batch	Proportion_of_good_quality
Brazil	116	0.47
Burundi	2	0.50
China	14	0.64
Colombia	158	0.80
Costa Rica	41	0.56
Cote d?Ivoire	1	0.00
Ecuador	3	0.33
El Salvador	20	0.70
Ethiopia	38	0.92
Guatemala	152	0.50
Haiti	5	0.20
Hawaii	62	0.55
Honduras	48	0.25
India	10	0.50
Indonesia	16	0.56
Japan	1	1.00
Kenya	24	0.92
Laos	2	0.00
Malawi	11	0.09
Mauritius	1	0.00
Mexico	203	0.27
Myanmar	6	0.00
Nicaragua	23	0.22
Panama	4	0.75
Peru	9	0.56
Philippines	5	0.40
Puerto Rico	3	0.33
Taiwan	62	0.42
Tanzania	32	0.50
Thailand	23	0.70
Uganda	32	0.78
United States	9	0.67
Vietnam	8	0.50
Zambia	1	0.00

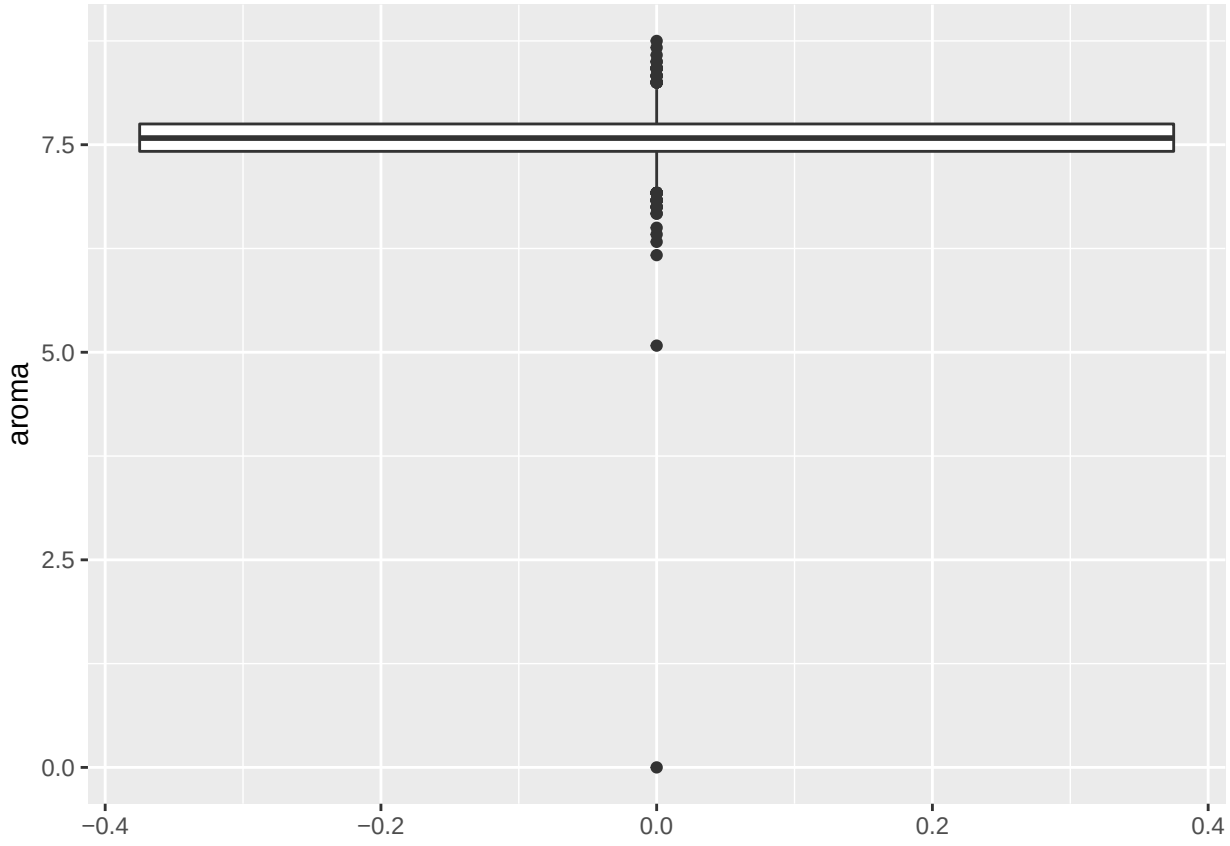
```
analysis %>%
  group_by(Qualityclass) %>%
  dplyr::select(Qualityclass, aroma, flavor, acidity, category_two_defects) %>%
  my_skim() %>%
  transmute(Variable=skim_variable, Qualityclass=Qualityclass, n=n, Mean=numeric.mean, SD=numeric.sd,
            Min=numeric.p0, Median=numeric.p50,  Max=numeric.p100,
            IQR = numeric.p75-numeric.p50) %>%
  kable(caption = '\\label{tab:catskim} Summary statistics of the sepal length by species of irises', bo
  kable_styling(font_size = 10, latex_options = "HOLD_position")
```

Table 3: Summary statistics of the sepal length by species of irises

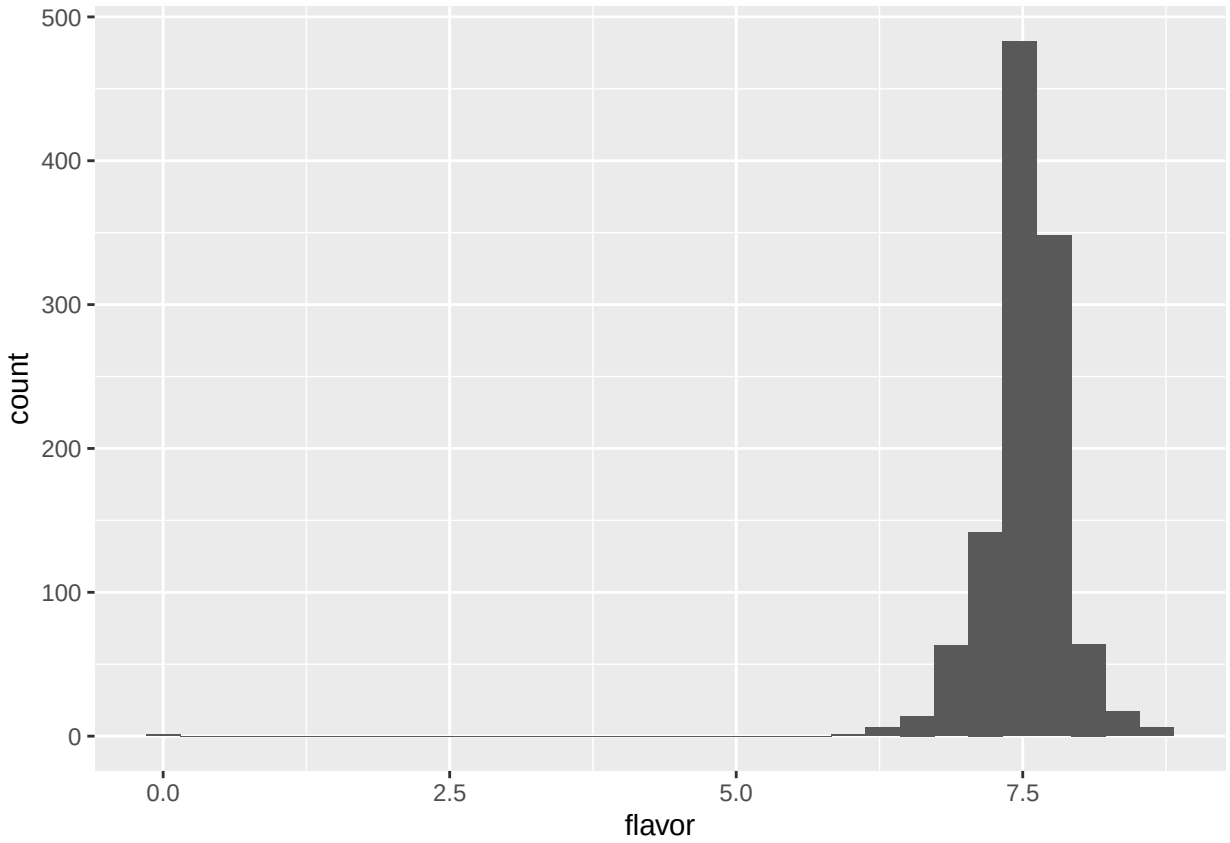
Variable	Qualityclass	n	Mean	SD	Min	Median	Max	IQR
aroma	Good	588	7.76	0.23	7.08	7.75	8.75	0.08
aroma	Poor	557	7.37	0.41	0.00	7.42	8.25	0.16
flavor	Good	588	7.74	0.23	7.00	7.67	8.67	0.16
flavor	Poor	557	7.29	0.42	0.00	7.33	8.08	0.17
acidity	Good	588	7.72	0.25	6.75	7.67	8.58	0.16
acidity	Poor	557	7.34	0.40	0.00	7.33	8.33	0.17
category_two_defects	Good	588	2.87	3.82	0.00	2.00	40.00	2.00
category_two_defects	Poor	557	4.52	6.60	0.00	2.00	55.00	4.00

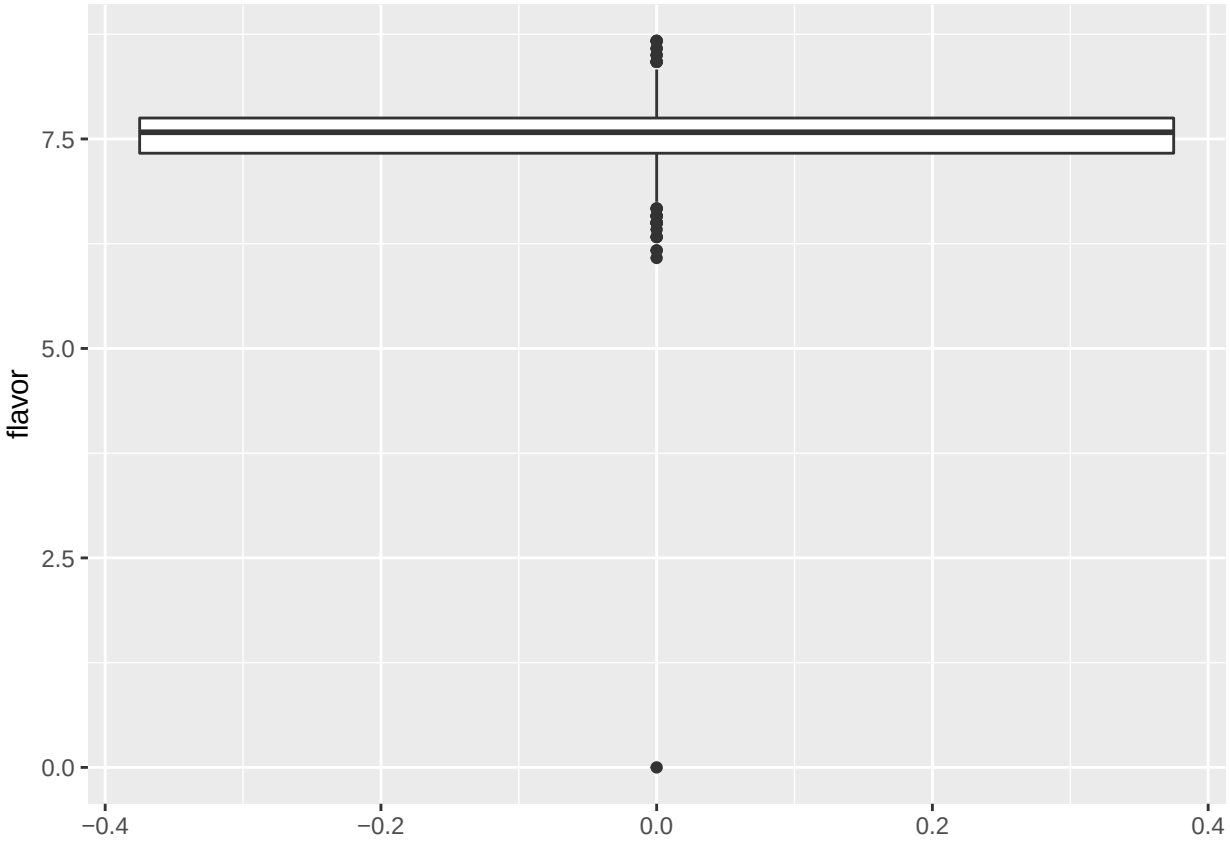
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



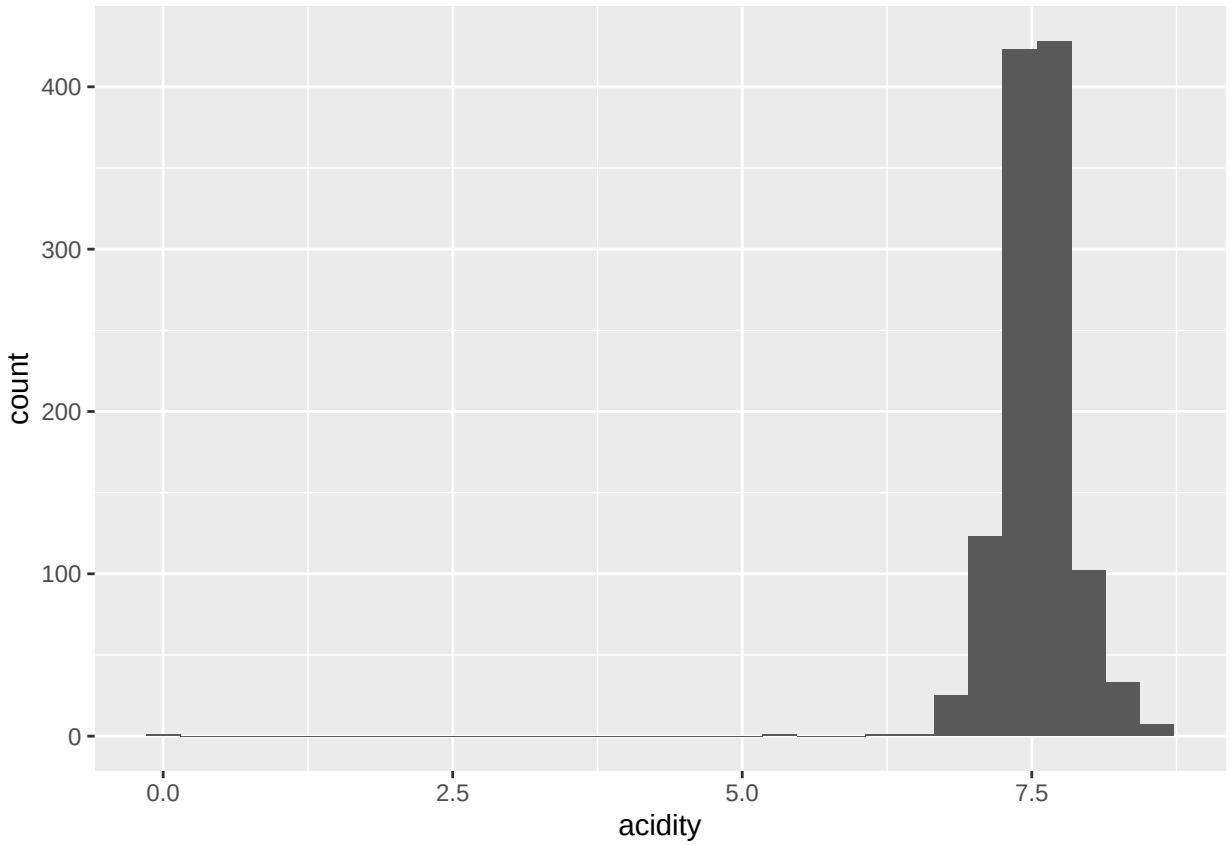


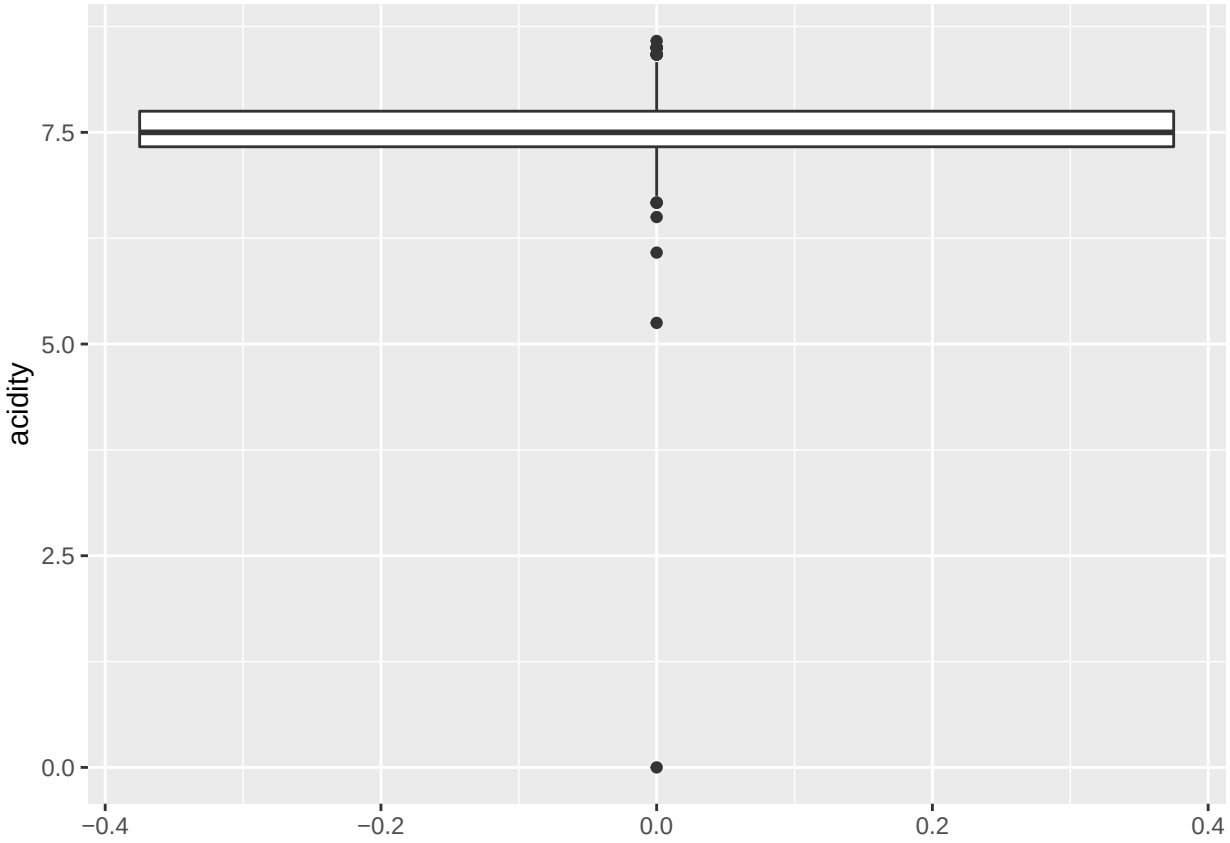

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```


500
400
300
200
100
0
count
0.0
2.5
5.0
7.5
flavor

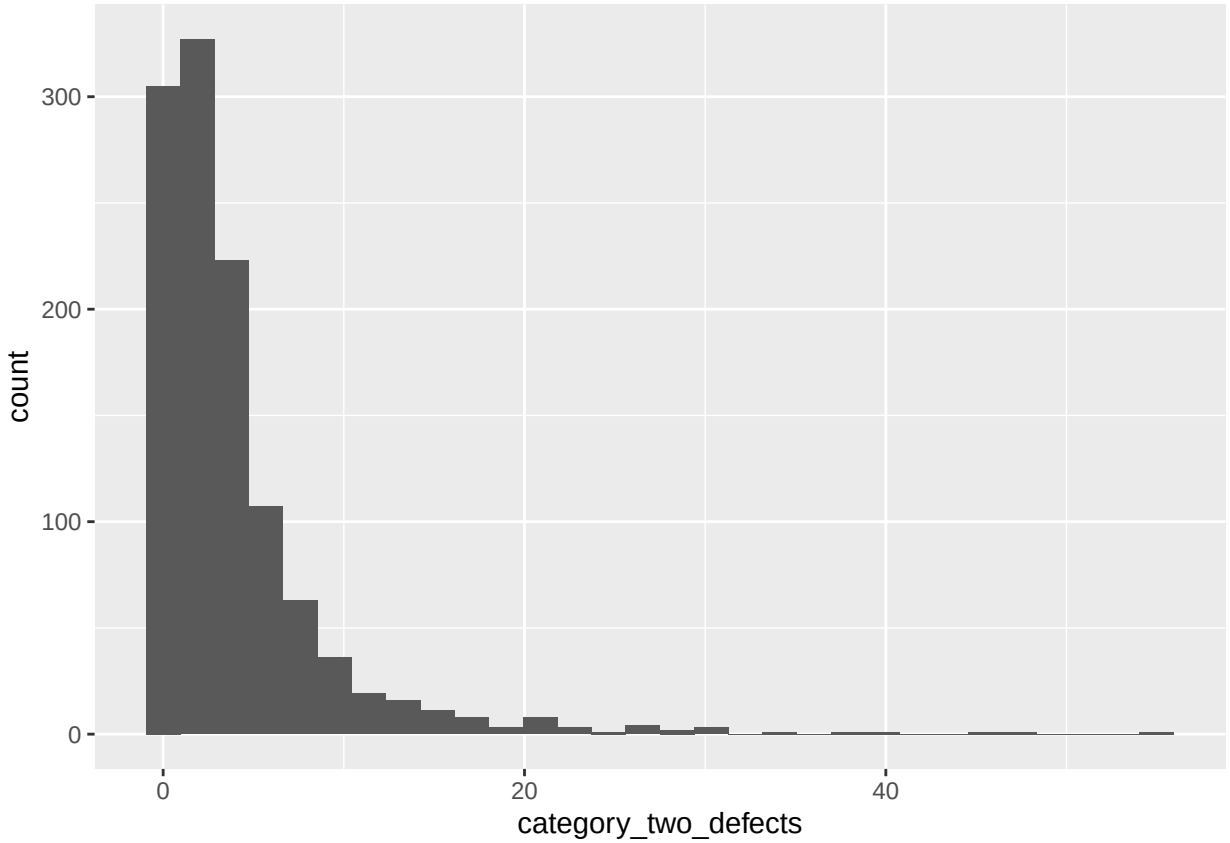


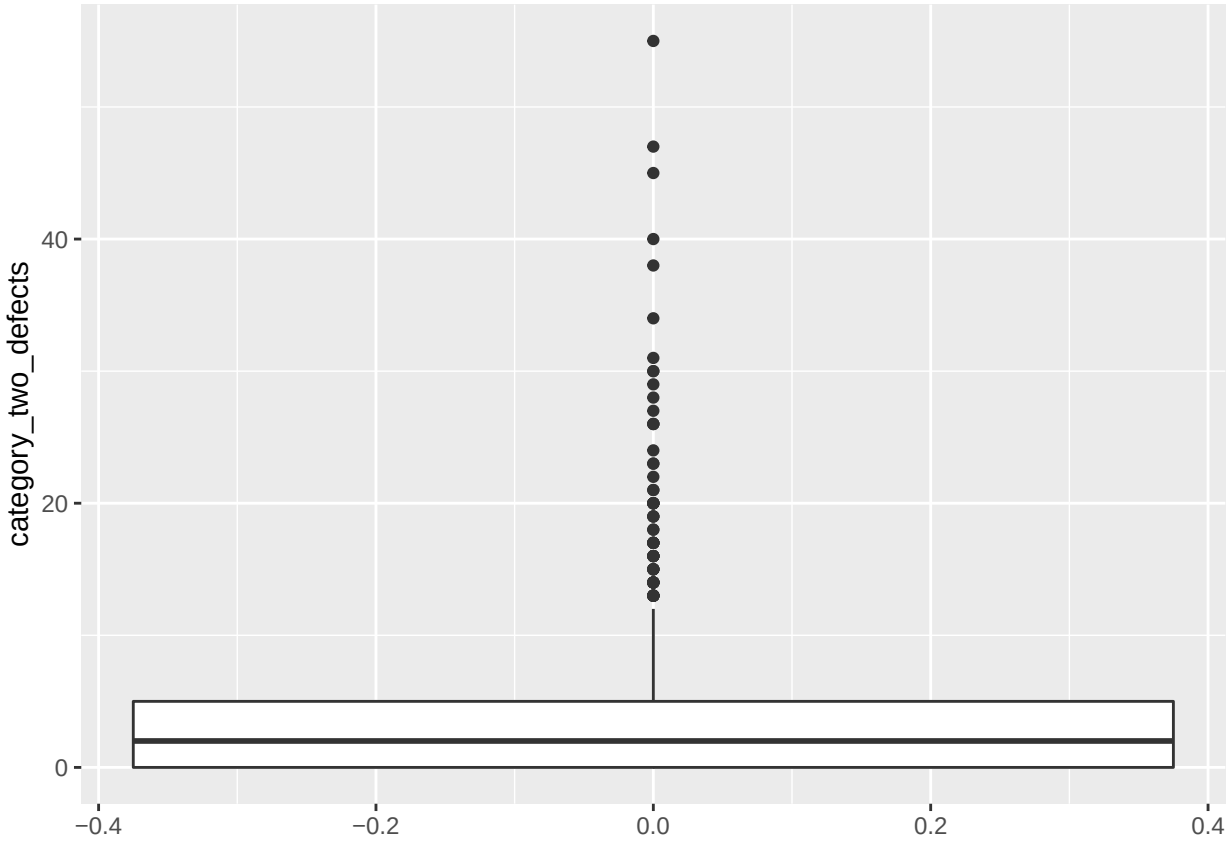

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```


400
300
200
100
0
count
0.0
2.5
5.0
7.5
acidity

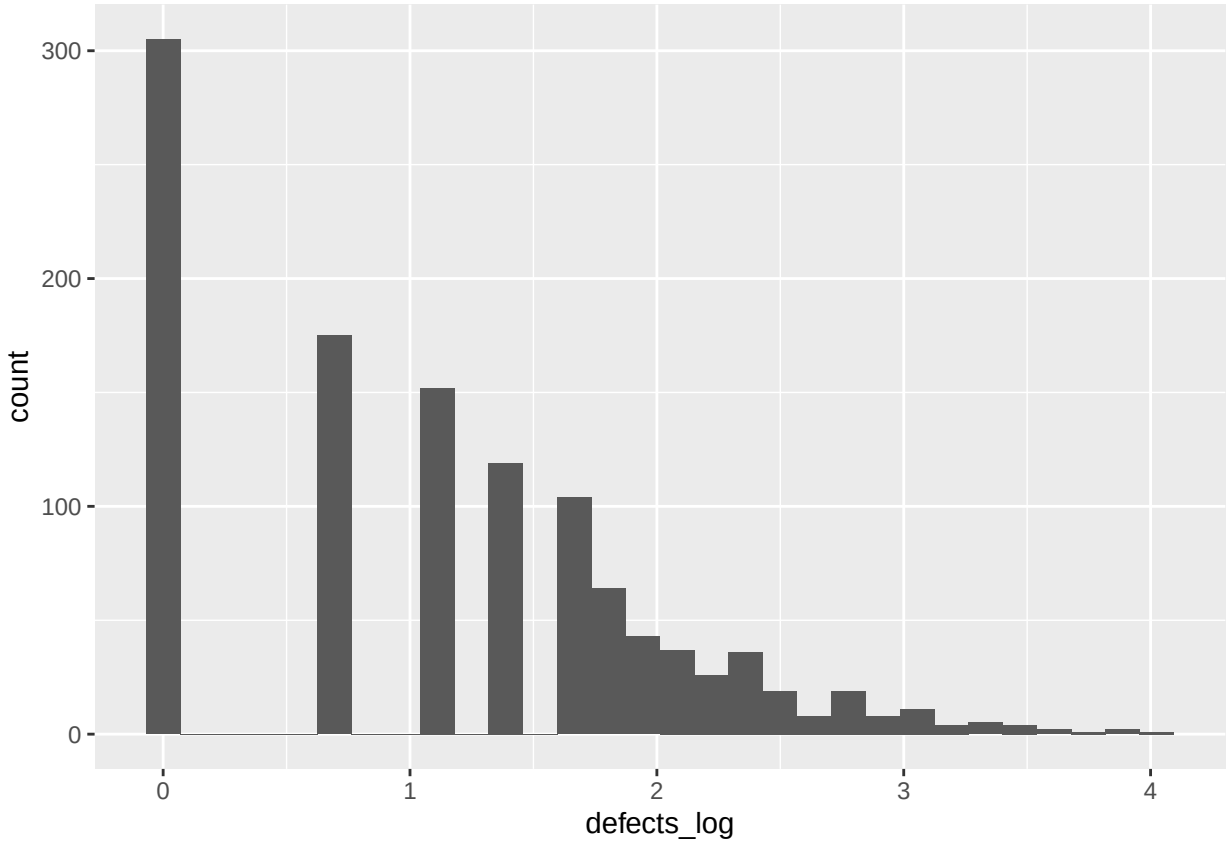


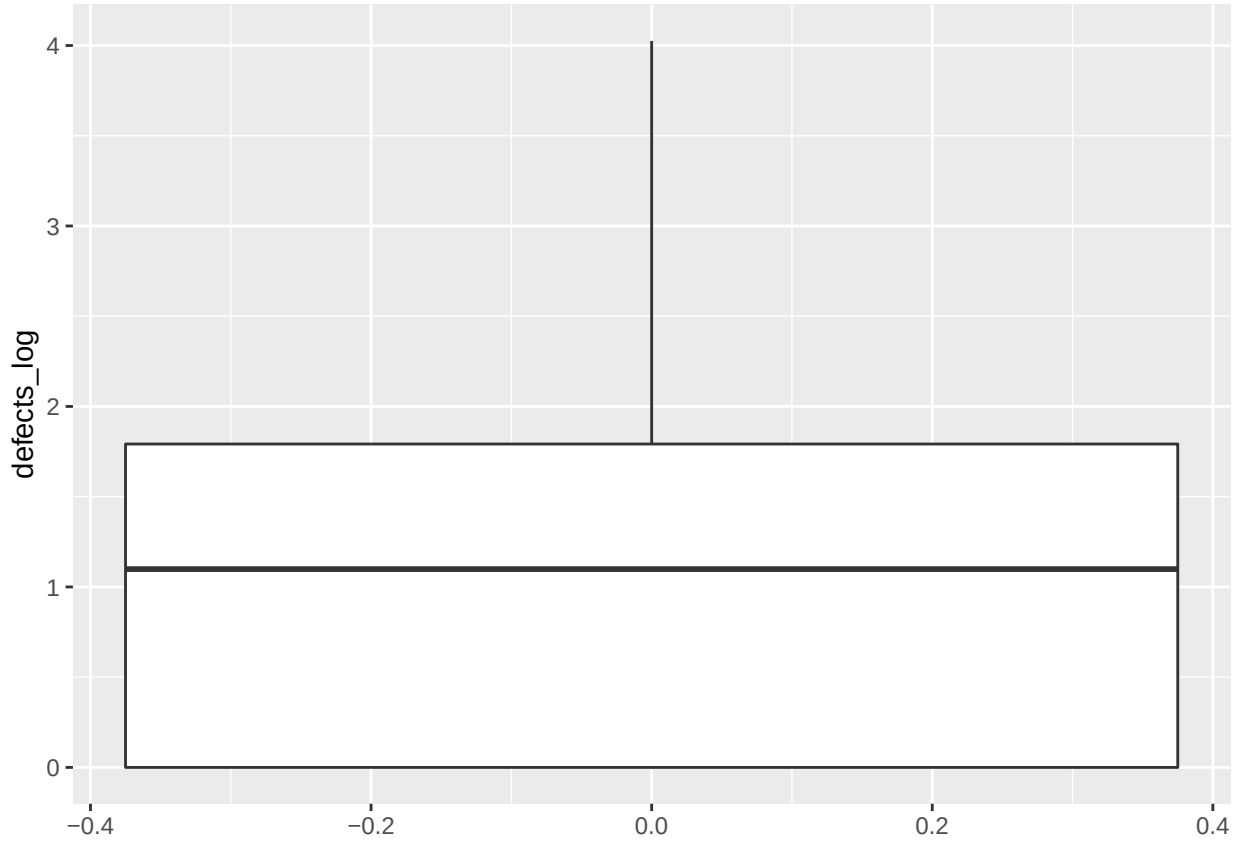

```
## # A tibble: 1 x 8
##   country_of_origin aroma flavor acidity category_two_defects altitude_mean_met~
##   <chr>             <dbl>  <dbl>   <dbl>                <dbl>              <dbl>
## 1 Honduras              0      0       0                    2               1400
## # ... with 2 more variables: harvested <dbl>, Qualityclass <chr>

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```


300
200
100
0
count
0
20
40
category_two_defects



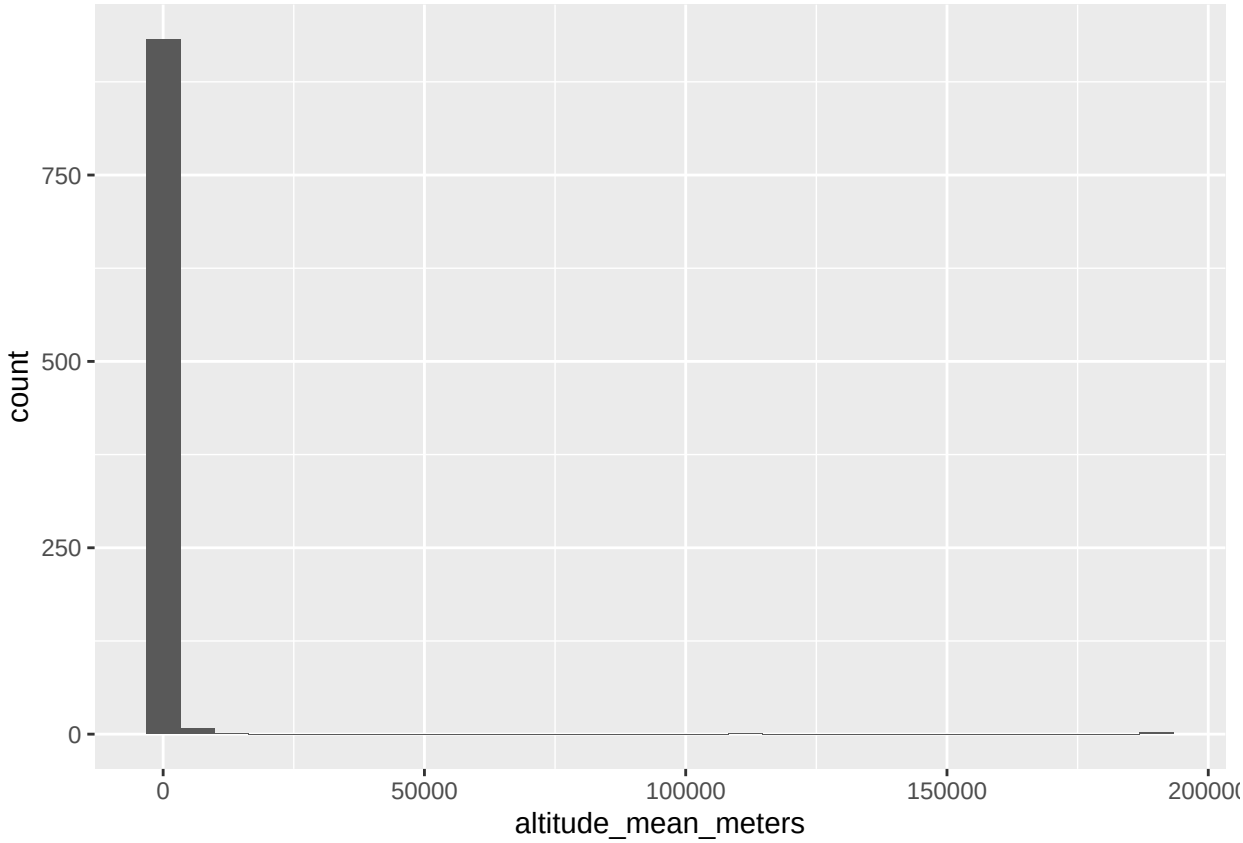

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```


300
200
100
0
count
0
1
2
3
4
defects_log

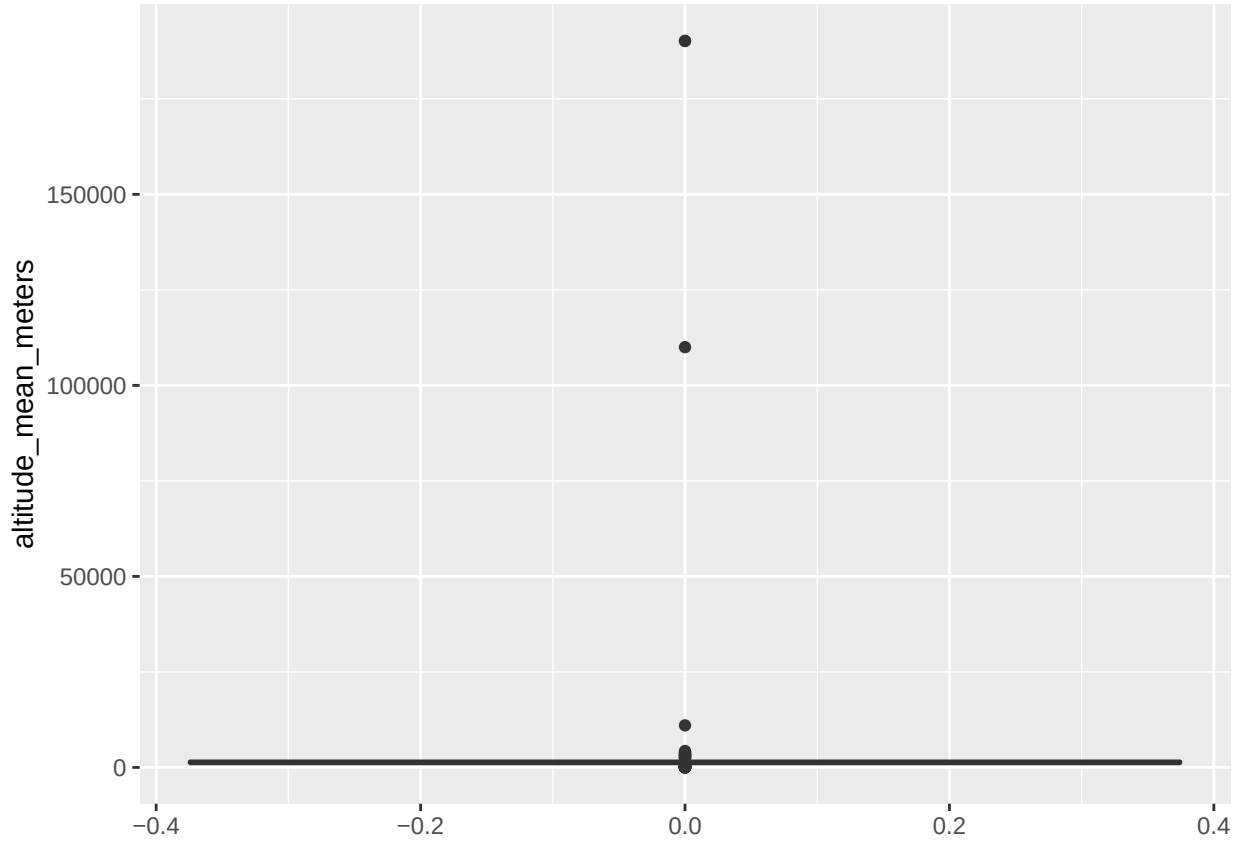



```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

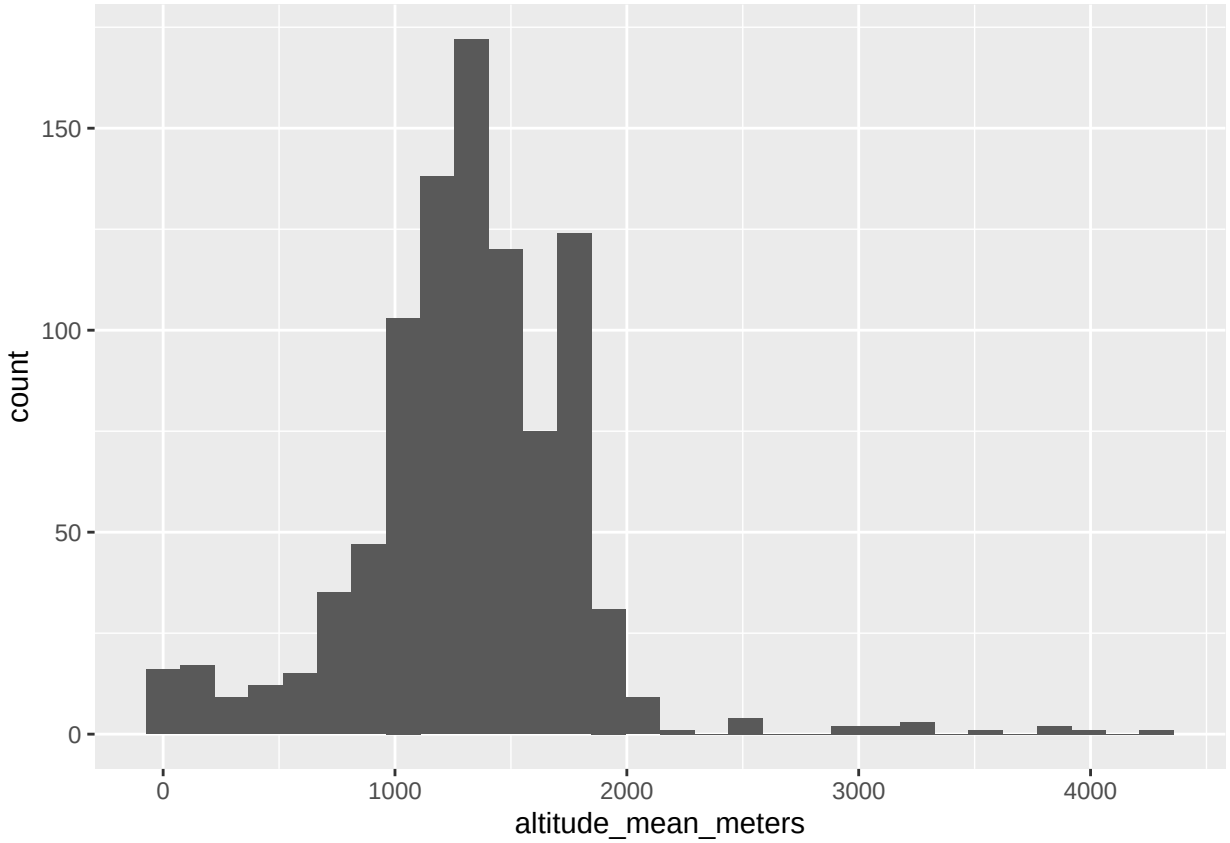
```
## Warning: Removed 201 rows containing non-finite values (stat_bin).
```

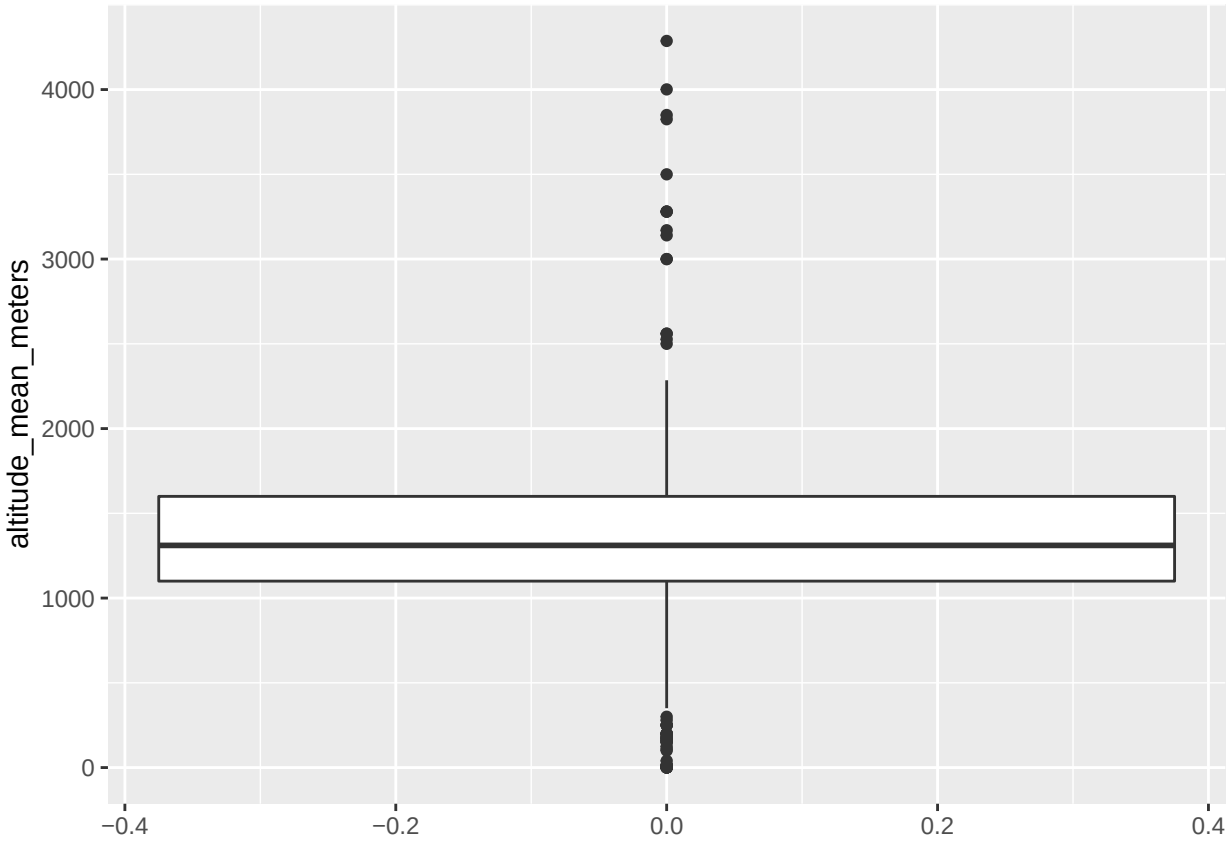



```
## Warning: Removed 201 rows containing non-finite values (stat_boxplot).
```



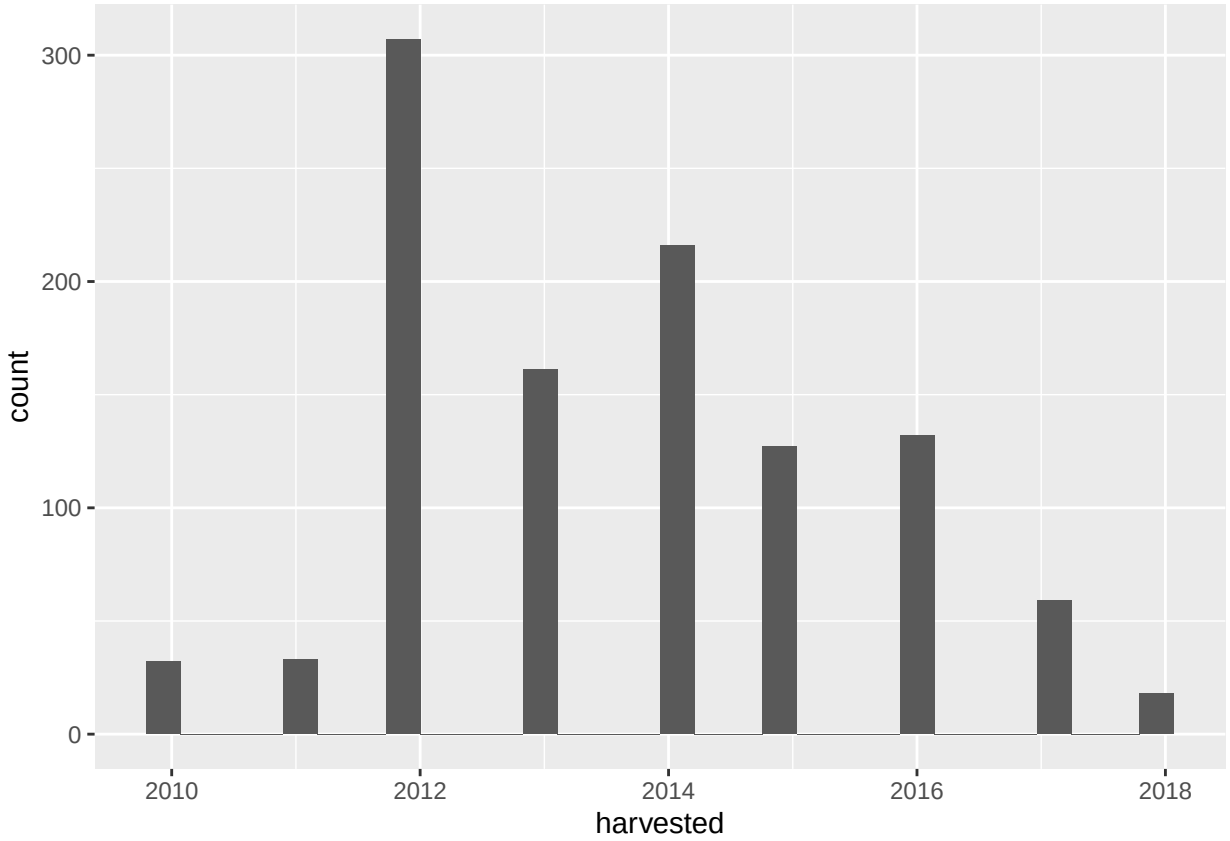

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```


150
100
50
0
count
0
1000
2000
3000
4000
altitude_mean_meters

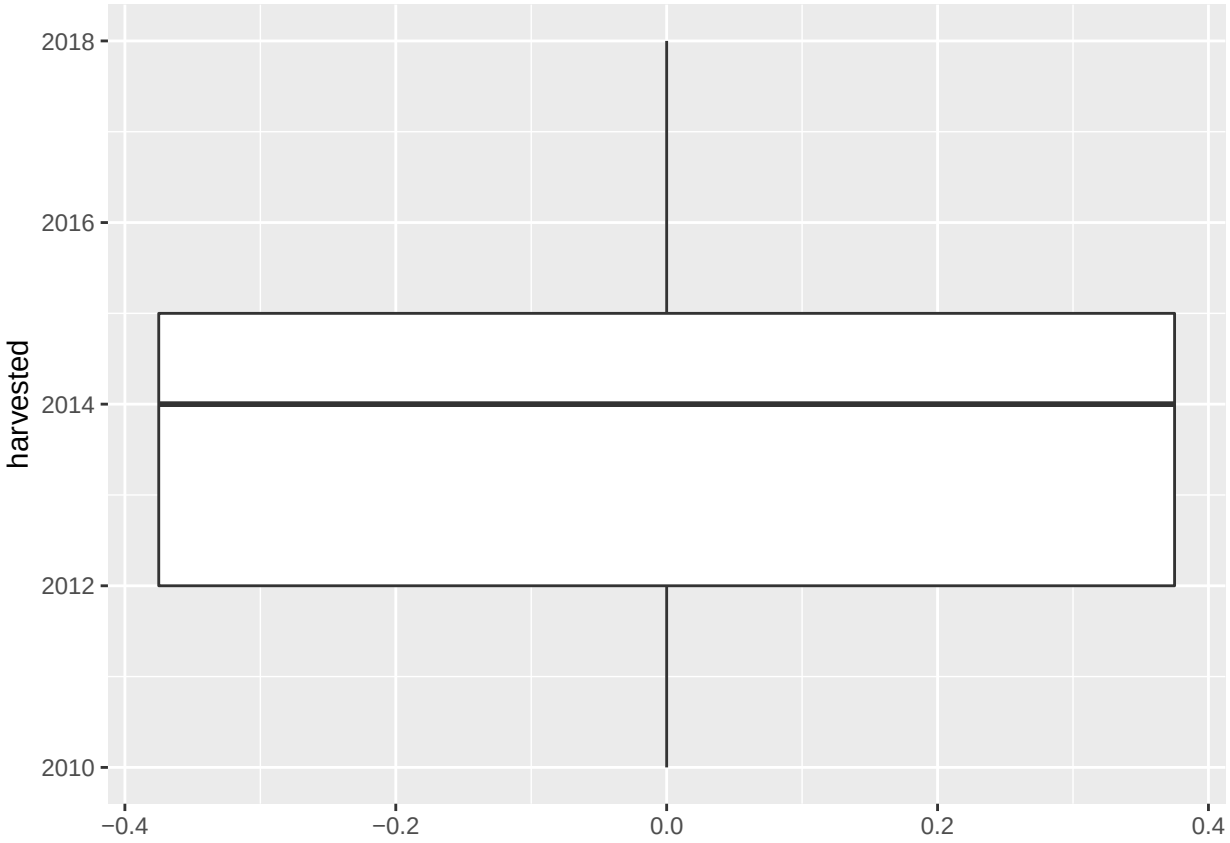


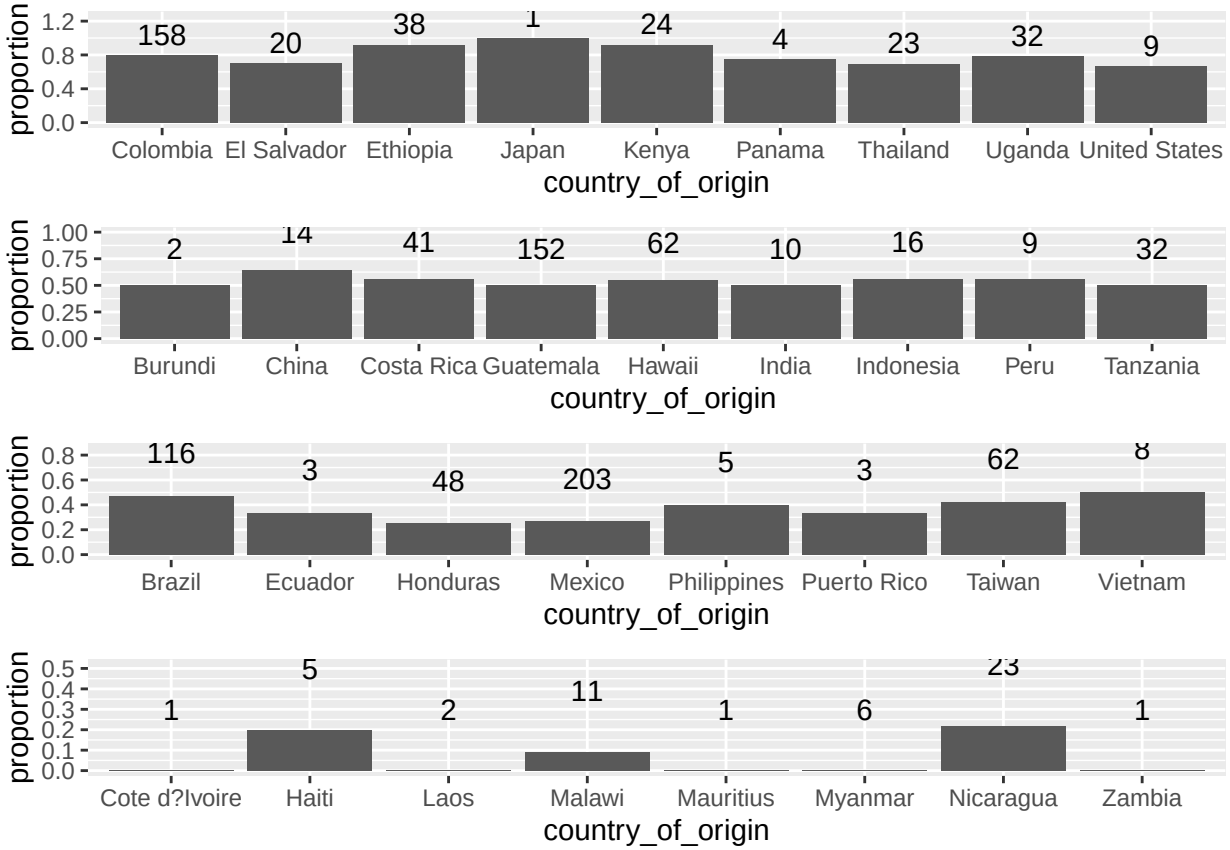

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 60 rows containing non-finite values (stat_bin).
```




```
## Warning: Removed 60 rows containing non-finite values (stat_boxplot).
```


harvested
2018
2016
2014
2012
2010
−0.4
−0.2
0.0
0.2
0.4




```
## Rows: 930
## Columns: 8
## $ country_of_origin <chr> "Myanmar", "Uganda", "Ethiopia", "Mexico", "Burundi"~
## $ aroma             <dbl> -1.084245709, 2.451292211, 2.745920371, -1.346137407~
## $ flavor            <dbl> -0.32474610, 1.21692758, 1.46359537, -1.37308420, 0.~
## $ acidity           <dbl> -0.1080856, 1.2368728, 1.4930554, -0.9086561, -0.108~
## $ defects_log       <dbl> 1.6094379, 0.6931472, 2.0794415, 1.3862944, 1.791759~
## $ year              <fct> 2015, 2013, 2014, 2012, 2012, 2014, 2015, 2013, 2013~
## $ level             <chr> "3", "3", "3", "3", "3", "3", "3", "3", "3", "3", "3~
## $ Qualityclass      <dbl> 0, 1, 1, 0, 1, 1, 1, 0, 0, 1, 0, 1, 0, 1, 1, 0, 1, 1~
```

```
## # A tibble: 9 x 2
##   year      n
##   <fct> <int>
## 1 2010     26
## 2 2011     30
## 3 2012    255
## 4 2013    134
## 5 2014    194
## 6 2015    118
## 7 2016    103
## 8 2017     52
## 9 2018     18
```

```
# We generate a table to have a look at the proportion and counts for both Good and Poor quality based
```

```
coffee_final %>%
  tabyl(Qualityclass, country_of_origin) %>%
  adorn_percentages() %>%
  adorn_pct_formatting() %>%
  adorn_ns()
```

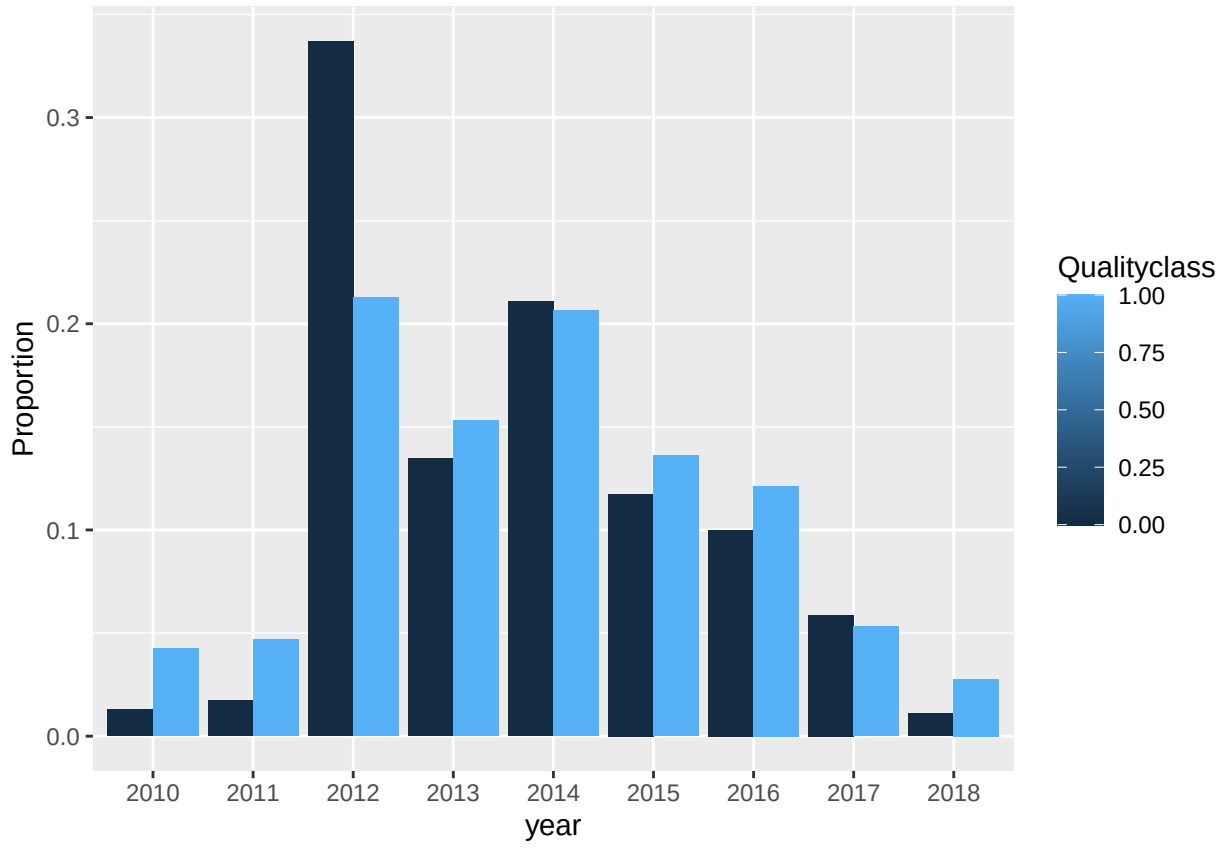
```
##  Qualityclass     Brazil  Burundi    China    Colombia Costa Rica Cote d?Ivoire
##             0  9.6% (44) 0.2% (1) 1.1% (5)  4.8%  (22)  3.5% (16)      0.2% (1)
##             1 10.0% (47) 0.2% (1) 1.9% (9) 22.3% (105)  4.3% (20)      0.0% (0)
##   Ecuador El Salvador  Ethiopia  Guatemala    Haiti   Hawaii  Honduras    India
##  0.2% (1)   1.1%  (5) 0.0%  (0) 13.3% (61) 0.9% (4) 0.0% (0) 7.4% (34) 1.1% (5)
##  0.2% (1)   2.8% (13) 4.9% (23) 14.0% (66) 0.2% (1) 0.2% (1) 2.6% (12) 1.1% (5)
##  Indonesia     Kenya     Laos    Malawi Mauritius      Mexico  Myanmar
##   1.3% (6) 0.4%  (2) 0.4% (2) 2.2% (10)  0.2% (1) 32.2% (148) 1.3% (6)
##   1.7% (8) 3.8% (18) 0.0% (0) 0.2%  (1)  0.0% (0) 11.1%  (52) 0.0% (0)
##  Nicaragua   Panama     Peru Philippines Puerto Rico    Taiwan  Tanzania
##  2.2% (10) 0.2% (1) 0.2% (1)    0.7% (3)    0.4% (2) 7.4% (34) 3.3% (15)
##  0.6%  (3) 0.6% (3) 0.0% (0)    0.4% (2)    0.2% (1) 4.9% (23) 3.0% (14)
##  Thailand    Uganda United States  Vietnam   Zambia
##  1.3% (6) 1.5%  (7)      0.7% (3) 0.7% (3) 0.2% (1)
##  1.7% (8) 4.9% (23)      1.3% (6) 0.9% (4) 0.0% (0)
```

```
coffee_final %>%
  tabyl(Qualityclass, year) %>%
  adorn_percentages() %>%
  adorn_pct_formatting() %>%
  adorn_ns()
```

```
##  Qualityclass      2010      2011        2012       2013       2014       2015
##             0 1.3%  (6) 1.7%  (8) 33.7% (155) 13.5% (62) 21.1% (97) 11.7% (54)
##             1 4.3% (20) 4.7% (22) 21.3% (100) 15.3% (72) 20.6% (97) 13.6% (64)
##        2016      2017      2018
##  10.0% (46) 5.9% (27) 1.1%  (5)
##  12.1% (57) 5.3% (25) 2.8% (13)
```

```
# Plot a bar chart to get intuition of whether the quality of coffee is influenced by the year of harve
```

```
ggplot(coffee_final, aes(x= year,  y = ..prop.., group=Qualityclass, fill=Qualityclass)) +
    geom_bar(position="dodge", stat="count") +
    labs(y = "Proportion")
```



Formal Analysis

Build the models

Only base on the altitude

```
model_level <- glm(Qualityclass ~ level - 1, data = coffee_final,family = binomial(link = "logit"))
summary(model_level)
```

```
##
## Call:
## glm(formula = Qualityclass ~ level - 1, family = binomial(link = "logit"),
##     data = coffee_final)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.287  -1.287   1.071   1.071   1.369
##
## Coefficients:
##        Estimate Std. Error z value Pr(>|z|)
## level1 -0.43891    0.18321  -2.396  0.01659 *
## level2 -0.40968    0.14513  -2.823  0.00476 **
## level3  0.25508    0.08184   3.117  0.00183 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1289.3  on 930  degrees of freedom
## Residual deviance: 1265.4  on 927  degrees of freedom
## AIC: 1271.4
##
## Number of Fisher Scoring iterations: 4
```

Base on the year of harvest

```
model_year <- glm(Qualityclass ~ year, data = coffee_final,family = binomial(link = "logit"))
summary(model_year)
```

```
##
## Call:
## glm(formula = Qualityclass ~ year, family = binomial(link = "logit"),
##     data = coffee_final)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7125  -1.1774   0.7244   1.1146   1.3683
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.2040     0.4655   2.587 0.009694 **
## year2011     -0.1924     0.6222  -0.309 0.757181
## year2012     -1.6422     0.4828  -3.401 0.000671 ***
## year2013     -1.0544     0.4967  -2.123 0.033753 *
## year2014     -1.2040     0.4871  -2.472 0.013450 *
## year2015     -1.0341     0.5008  -2.065 0.038941 *
## year2016     -0.9896     0.5059  -1.956 0.050466 .
## year2017     -1.2809     0.5419  -2.364 0.018099 *
## year2018     -0.2485     0.7026  -0.354 0.723600
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1289.1  on 929  degrees of freedom
## Residual deviance: 1256.0  on 921  degrees of freedom
## AIC: 1274
##
## Number of Fisher Scoring iterations: 4
```

Base on the country

```
model_country <- glm(Qualityclass ~ country_of_origin, data = coffee_final,family = binomial(link = "log
summary(model_country)
```

```
##
```

```
## Call:
## glm(formula = Qualityclass ~ country_of_origin, family = binomial(link = "logit"),
##     data = coffee_final)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.14597  -1.01655   0.00036   1.08424   2.18993
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      0.06596    0.20977   0.314  0.75320
## country_of_originBurundi        -0.06596    1.42969  -0.046  0.96320
## country_of_originChina           0.52183    0.59592   0.876  0.38121
## country_of_originColombia        1.49696    0.31461   4.758 1.95e-06 ***
## country_of_originCosta Rica      0.15719    0.39561   0.397  0.69112
## country_of_originCote d?Ivoire -16.63203 2399.54473  -0.007  0.99447
## country_of_originEcuador        -0.06596    1.42969  -0.046  0.96320
## country_of_originEl Salvador     0.88955    0.56650   1.570  0.11636
## country_of_originEthiopia       16.50011  500.33971   0.033  0.97369
## country_of_originGuatemala       0.01282    0.27486   0.047  0.96279
## country_of_originHaiti          -1.45225    1.13754  -1.277  0.20172
## country_of_originHawaii         16.50011 2399.54473   0.007  0.99451
## country_of_originHonduras       -1.10741    0.39592  -2.797  0.00516 **
## country_of_originIndia          -0.06596    0.66634  -0.099  0.92115
## country_of_originIndonesia       0.22172    0.57937   0.383  0.70194
## country_of_originKenya           2.13127    0.77431   2.752  0.00591 **
## country_of_originLaos          -16.63203 1696.73436  -0.010  0.99218
## country_of_originMalawi         -2.36854    1.06958  -2.214  0.02680 *
## country_of_originMauritius     -16.63203 2399.54473  -0.007  0.99447
## country_of_originMexico         -1.11193    0.26456  -4.203 2.63e-05 ***
## country_of_originMyanmar       -16.63203  979.61005  -0.017  0.98645
## country_of_originNicaragua      -1.26993    0.69090  -1.838  0.06605 .
## country_of_originPanama          1.03265    1.17360   0.880  0.37891
## country_of_originPeru          -16.63203 2399.54473  -0.007  0.99447
## country_of_originPhilippines    -0.47142    0.93666  -0.503  0.61475
## country_of_originPuerto Rico    -0.75911    1.24258  -0.611  0.54126
## country_of_originTaiwan         -0.45682    0.34190  -1.336  0.18150
## country_of_originTanzania       -0.13495    0.42673  -0.316  0.75182
## country_of_originThailand        0.22172    0.57937   0.383  0.70194
## country_of_originUganda          1.12363    0.47994   2.341  0.01922 *
## country_of_originUnited States   0.62719    0.73757   0.850  0.39513
## country_of_originVietnam         0.22172    0.79205   0.280  0.77952
## country_of_originZambia        -16.63203 2399.54473  -0.007  0.99447
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1289.1  on 929  degrees of freedom
## Residual deviance: 1072.1  on 897  degrees of freedom
## AIC: 1138.1
##
## Number of Fisher Scoring iterations: 15
```

According the result before, we choose some significant country as a class variable.

```
coffee_final$Colombia <- ifelse(coffee_final$country_of_origin == 'Colombia',1,0)
coffee_final$Mexico <- ifelse(coffee_final$country_of_origin == 'Mexico',1,0)
coffee_final$Honduras <- ifelse(coffee_final$country_of_origin == 'Honduras',1,0)
coffee_final$Kenya <- ifelse(coffee_final$country_of_origin == 'Kenya',1,0)
```

```
model_co_4 <- glm(Qualityclass ~ Colombia + Mexico + Honduras + Kenya-1, data = coffee_final,family = bi
summary(model_co_4)
```

```
##
## Call:
## glm(formula = Qualityclass ~ Colombia + Mexico + Honduras + Kenya -
##     1, family = binomial(link = "logit"), data = coffee_final)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.146  -1.177   0.459   1.177   1.641
##
## Coefficients:
##          Estimate Std. Error z value Pr(>|z|)
## Colombia   1.5629     0.2345   6.666 2.64e-11 ***
## Mexico    -1.0460     0.1612  -6.488 8.68e-11 ***
## Honduras  -1.0415     0.3358  -3.102  0.00192 **
## Kenya      2.1972     0.7453   2.948  0.00320 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1289.3  on 930  degrees of freedom
## Residual deviance: 1156.6  on 926  degrees of freedom
## AIC: 1164.6
##
## Number of Fisher Scoring iterations: 4
```

Base on the year and country

```
model_cn_ye <- glm(Qualityclass ~ country_of_origin + year, data = coffee_final,family = binomial(link =
summary(model_cn_ye)
```

```
##
## Call:
## glm(formula = Qualityclass ~ country_of_origin + year, family = binomial(link = "logit"),
##     data = coffee_final)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.17448  -0.97437   0.00032   1.00309   2.18993
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)                      -0.03468    0.59460  -0.058  0.95349
## country_of_originBurundi          0.15624    1.44037   0.108  0.91362
## country_of_originChina            0.77187    0.63053   1.224  0.22089
## country_of_originColombia         1.84054    0.34098   5.398 6.75e-08 ***
## country_of_originCosta Rica       0.49081    0.42149   1.164  0.24423
## country_of_originCote d?Ivoire  -16.57149 2399.54474  -0.007  0.99449
## country_of_originEcuador          0.32438    1.45807   0.222  0.82395
## country_of_originEl Salvador      1.15147    0.58475   1.969  0.04893 *
## country_of_originEthiopia        16.82884  497.80415   0.034  0.97303
## country_of_originGuatemala        0.41515    0.30673   1.353  0.17591
## country_of_originHaiti           -0.94200    1.15374  -0.816  0.41423
## country_of_originHawaii          16.60075 2399.54479   0.007  0.99448
## country_of_originHonduras        -0.88360    0.41906  -2.109  0.03499 *
## country_of_originIndia            0.27588    0.69004   0.400  0.68930
## country_of_originIndonesia        0.42228    0.60048   0.703  0.48191
## country_of_originKenya            2.57473    0.79133   3.254  0.00114 **
## country_of_originLaos           -16.57882 1696.72545  -0.010  0.99220
## country_of_originMalawi          -1.91280    1.09277  -1.750  0.08005 .
## country_of_originMauritius      -16.57149 2399.54474  -0.007  0.99449
## country_of_originMexico          -0.71989    0.31440  -2.290  0.02204 *
## country_of_originMyanmar        -16.52163  976.27716  -0.017  0.98650
## country_of_originNicaragua       -1.11252    0.70712  -1.573  0.11565
## country_of_originPanama           1.47628    1.18697   1.244  0.21360
## country_of_originPeru           -16.24816 2399.54474  -0.007  0.99460
## country_of_originPhilippines     -0.11413    0.95334  -0.120  0.90471
## country_of_originPuerto Rico     -0.03897    1.26851  -0.031  0.97549
## country_of_originTaiwan          -0.10856    0.37283  -0.291  0.77092
## country_of_originTanzania         0.21667    0.45820   0.473  0.63630
## country_of_originThailand         0.61483    0.60175   1.022  0.30691
## country_of_originUganda           1.68023    0.51792   3.244  0.00118 **
## country_of_originUnited States    1.07504    0.76568   1.404  0.16031
## country_of_originVietnam          0.66336    0.81159   0.817  0.41372
## country_of_originZambia         -16.17628 2399.54474  -0.007  0.99462
## year2011                          0.45958    0.71314   0.644  0.51928
## year2012                         -0.28323    0.58365  -0.485  0.62749
## year2013                         -0.61950    0.59127  -1.048  0.29476
## year2014                         -0.35511    0.58834  -0.604  0.54613
## year2015                          0.05473    0.59555   0.092  0.92678
## year2016                          0.04011    0.59731   0.067  0.94646
## year2017                         -0.37725    0.63786  -0.591  0.55423
## year2018                          0.81484    0.79630   1.023  0.30617
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1289.1  on 929  degrees of freedom
## Residual deviance: 1060.4  on 889  degrees of freedom
## AIC: 1142.4
##
## Number of Fisher Scoring iterations: 15
```

Base on the altitude and country

```r
model_al_co <- glm(Qualityclass ~ level + Colombia + Mexico + Honduras + Kenya, data = coffee_final,fami
summary(model_al_co)
```

```
##
## Call:
## glm(formula = Qualityclass ~ level + Colombia + Mexico + Honduras +
##     Kenya, family = binomial(link = "logit"), data = coffee_final)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2327  -0.9800   0.4155   1.0322   1.9251
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.4838     0.1948  -2.484   0.0130 *
## level2        0.4061     0.2476   1.640   0.1010
## level3        0.8352     0.2144   3.896 9.79e-05 ***
## Colombia      1.3248     0.2544   5.207 1.92e-07 ***
## Mexico       -1.1988     0.1859  -6.448 1.13e-10 ***
## Honduras     -1.3846     0.3520  -3.933 8.39e-05 ***
## Kenya         2.0547     0.7551   2.721   0.0065 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1289.1  on 929  degrees of freedom
## Residual deviance: 1137.0  on 923  degrees of freedom
## AIC: 1151
##
## Number of Fisher Scoring iterations: 4
```

Base on the 3

```r
model_al_co <- glm(Qualityclass ~ level + country_of_origin + year - 1, data = coffee_final,family = bi
summary(model_al_co)
```

```
##
## Call:
## glm(formula = Qualityclass ~ level + country_of_origin + year -
##     1, family = binomial(link = "logit"), data = coffee_final)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.24878  -0.87349   0.00031   0.98739   2.12047
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## level1                        -6.057e-01  6.460e-01  -0.938  0.34842
## level2                         6.539e-02  6.237e-01   0.105  0.91650
## level3                         4.660e-01  6.327e-01   0.736  0.46144
## country_of_originBurundi      -3.436e-01  1.454e+00  -0.236  0.81321
```

```
## country_of_originChina          4.626e-01  6.421e-01   0.721  0.47119
## country_of_originColombia       1.525e+00  3.612e-01   4.223 2.42e-05 ***
## country_of_originCosta Rica     2.346e-01  4.355e-01   0.539  0.59011
## country_of_originCote d?Ivoire -1.606e+01  2.400e+03  -0.007  0.99466
## country_of_originEcuador        9.024e-01  1.481e+00   0.609  0.54239
## country_of_originEl Salvador    7.797e-01  5.933e-01   1.314  0.18874
## country_of_originEthiopia       1.639e+01  4.970e+02   0.033  0.97369
## country_of_originGuatemala     -2.782e-03  3.394e-01  -0.008  0.99346
## country_of_originHaiti         -7.446e-01  1.185e+00  -0.628  0.52988
## country_of_originHawaii         1.717e+01  2.400e+03   0.007  0.99429
## country_of_originHonduras      -1.345e+00  4.471e-01  -3.008  0.00263 **
## country_of_originIndia          3.083e-01  7.105e-01   0.434  0.66438
## country_of_originIndonesia      4.157e-03  6.167e-01   0.007  0.99462
## country_of_originKenya          2.400e+00  8.037e-01   2.986  0.00283 **
## country_of_originLaos          -1.707e+01  1.696e+03  -0.010  0.99197
## country_of_originMalawi        -2.146e+00  1.097e+00  -1.957  0.05035 .
## country_of_originMauritius     -1.606e+01  2.400e+03  -0.007  0.99466
## country_of_originMexico        -9.546e-01  3.255e-01  -2.933  0.00336 **
## country_of_originMyanmar       -1.682e+01  9.735e+02  -0.017  0.98622
## country_of_originNicaragua     -1.245e+00  7.120e-01  -1.748  0.08044 .
## country_of_originPanama         1.155e+00  1.192e+00   0.969  0.33234
## country_of_originPeru          -1.669e+01  2.400e+03  -0.007  0.99445
## country_of_originPhilippines   -2.897e-01  9.557e-01  -0.303  0.76183
## country_of_originPuerto Rico    6.030e-01  1.289e+00   0.468  0.63987
## country_of_originTaiwan         2.949e-01  3.980e-01   0.741  0.45863
## country_of_originTanzania      -1.655e-01  4.788e-01  -0.346  0.72962
## country_of_originThailand       7.166e-01  6.329e-01   1.132  0.25752
## country_of_originUganda         1.292e+00  5.338e-01   2.420  0.01554 *
## country_of_originUnited States  8.490e-01  7.818e-01   1.086  0.27750
## country_of_originVietnam        4.664e-01  8.165e-01   0.571  0.56787
## country_of_originZambia        -1.658e+01  2.400e+03  -0.007  0.99449
## year2011                        2.581e-01  7.296e-01   0.354  0.72350
## year2012                       -3.418e-01  6.032e-01  -0.567  0.57092
## year2013                       -6.905e-01  6.106e-01  -1.131  0.25814
## year2014                       -4.561e-01  6.076e-01  -0.751  0.45289
## year2015                       -1.422e-02  6.148e-01  -0.023  0.98154
## year2016                        9.710e-02  6.157e-01   0.158  0.87469
## year2017                       -4.201e-01  6.555e-01  -0.641  0.52156
## year2018                        1.284e+00  8.242e-01   1.557  0.11939
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1289.3  on 930  degrees of freedom
## Residual deviance: 1047.4  on 887  degrees of freedom
## AIC: 1133.4
##
## Number of Fisher Scoring iterations: 15
```

Colombia + Mexico + Honduras + Kenya

Consider everything

```
model_all <- glm(Qualityclass ~ aroma + flavor + acidity + country_of_origin + defects_log + level + yea
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(model_all)
```

```
##
## Call:
## glm(formula = Qualityclass ~ aroma + flavor + acidity + country_of_origin +
##     defects_log + level + year, family = binomial(link = "logit"),
##     data = coffee_final)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.5914  -0.2397   0.0000   0.2843   3.5781
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -1.57595    1.09508  -1.439  0.15012
## aroma                            1.58526    0.25890   6.123 9.18e-10 ***
## flavor                           2.78982    0.34848   8.006 1.19e-15 ***
## acidity                          1.64741    0.25864   6.369 1.90e-10 ***
## country_of_originBurundi         1.88240    5.12830   0.367  0.71357
## country_of_originChina           0.49916    1.08844   0.459  0.64652
## country_of_originColombia        1.84638    0.57358   3.219  0.00129 **
## country_of_originCosta Rica      0.26961    0.76612   0.352  0.72491
## country_of_originCote d?Ivoire -12.11826 6522.63865  -0.002  0.99852
## country_of_originEcuador        -1.02265    1.52999  -0.668  0.50388
## country_of_originEl Salvador     0.32640    0.96977   0.337  0.73644
## country_of_originEthiopia       13.49329  894.76317   0.015  0.98797
## country_of_originGuatemala      -0.75268    0.57572  -1.307  0.19108
## country_of_originHaiti           2.27451    2.16150   1.052  0.29267
## country_of_originHawaii          4.41740 6522.63879   0.001  0.99946
## country_of_originHonduras       -0.72501    0.71286  -1.017  0.30913
## country_of_originIndia          -2.55120    1.07559  -2.372  0.01770 *
## country_of_originIndonesia      -0.38258    1.01141  -0.378  0.70524
## country_of_originKenya           0.52684    1.54516   0.341  0.73313
## country_of_originLaos          -15.24675 4515.00054  -0.003  0.99731
## country_of_originMalawi         -0.65398    1.30094  -0.503  0.61518
## country_of_originMauritius     -11.76872 6522.63865  -0.002  0.99856
## country_of_originMexico         -0.80196    0.52029  -1.541  0.12323
## country_of_originMyanmar       -15.49786 2401.00369  -0.006  0.99485
## country_of_originNicaragua       0.53829    1.98308   0.271  0.78605
## country_of_originPanama          3.27141    1.79738   1.820  0.06874 .
## country_of_originPeru          -14.50164 6522.63864  -0.002  0.99823
## country_of_originPhilippines     2.89981    2.57307   1.127  0.25975
## country_of_originPuerto Rico    -2.65794    1.78541  -1.489  0.13657
## country_of_originTaiwan          1.18951    0.70762   1.681  0.09276 .
## country_of_originTanzania        0.91717    0.75964   1.207  0.22729
## country_of_originThailand        2.87480    0.99592   2.887  0.00389 **
## country_of_originUganda         -1.53625    0.79415  -1.934  0.05306 .
## country_of_originUnited States   0.19578    1.52935   0.128  0.89814
```

```
## country_of_originVietnam          2.24627    1.15874   1.939  0.05256 .
## country_of_originZambia         -13.96552 6522.63865  -0.002  0.99829
## defects_log                       0.33145    0.17162   1.931  0.05345 .
## level2                            0.52403    0.48450   1.082  0.27943
## level3                            1.03968    0.48225   2.156  0.03109 *
## year2011                         -0.22625    1.12956  -0.200  0.84125
## year2012                          0.03098    0.98109   0.032  0.97481
## year2013                          0.48471    0.98717   0.491  0.62342
## year2014                         -0.07904    0.99385  -0.080  0.93661
## year2015                         -0.14258    0.98571  -0.145  0.88499
## year2016                          0.78470    1.03677   0.757  0.44913
## year2017                          0.46753    1.03839   0.450  0.65254
## year2018                          2.35570    1.32235   1.781  0.07484 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1289.15  on 929  degrees of freedom
## Residual deviance:  448.42  on 883  degrees of freedom
## AIC: 542.42
##
## Number of Fisher Scoring iterations: 17
```

```
model_all_2 <- glm(Qualityclass ~ aroma + flavor + acidity + Colombia + Mexico + Honduras + Kenya + defe
summary(model_all_2)
```

```
##
## Call:
## glm(formula = Qualityclass ~ aroma + flavor + acidity + Colombia +
##     Mexico + Honduras + Kenya + defects_log + level + year, family = binomial(link = "logit"),
##     data = coffee_final)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.2576  -0.2933   0.0010   0.3296   3.6482
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.3371     0.9359  -1.429   0.1531
## aroma         1.3517     0.2267   5.964 2.47e-09 ***
## flavor        2.3658     0.2899   8.162 3.30e-16 ***
## acidity       1.4612     0.2311   6.324 2.55e-10 ***
## Colombia      1.9282     0.4095   4.708 2.50e-06 ***
## Mexico       -0.7003     0.3512  -1.994   0.0461 *
## Honduras     -0.5767     0.5473  -1.054   0.2920
## Kenya         0.8497     1.3961   0.609   0.5427
## defects_log   0.3119     0.1509   2.067   0.0387 *
## level2        0.4545     0.4188   1.085   0.2779
## level3        0.6754     0.3769   1.792   0.0731 .
## year2011     -0.1599     1.0447  -0.153   0.8783
## year2012      0.1019     0.8983   0.113   0.9097
## year2013      0.1358     0.8937   0.152   0.8793
## year2014      0.4155     0.8995   0.462   0.6441
```

```
## year2015     -0.1081     0.9090  -0.119   0.9054
## year2016      0.8173     0.9402   0.869   0.3847
## year2017      0.2811     0.9682   0.290   0.7716
## year2018      2.0529     1.1977   1.714   0.0865 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1289.15  on 929  degrees of freedom
## Residual deviance:  493.82  on 911  degrees of freedom
## AIC: 531.82
##
## Number of Fisher Scoring iterations: 7
```

```
stepAIC(model_all_2, direction = 'both')
```

```
## Start:  AIC=531.82
## Qualityclass ~ aroma + flavor + acidity + Colombia + Mexico +
##     Honduras + Kenya + defects_log + level + year
##
##               Df Deviance    AIC
## - year         8   503.11 525.11
## - Kenya        1   494.24 530.24
## - Honduras     1   494.97 530.97
## - level        2   497.14 531.14
## <none>             493.82 531.82
## - Mexico       1   497.85 533.85
## - defects_log  1   498.16 534.16
## - Colombia     1   520.03 556.03
## - acidity      1   543.95 579.95
## - aroma        1   544.44 580.44
## - flavor       1   584.68 620.68
##
## Step:  AIC=525.11
## Qualityclass ~ aroma + flavor + acidity + Colombia + Mexico +
##     Honduras + Kenya + defects_log + level
##
##               Df Deviance    AIC
## - level        2   503.99 521.99
## - Kenya        1   503.79 523.79
## - Honduras     1   504.15 524.15
## <none>             503.11 525.11
## - defects_log  1   508.43 528.43
## + year         8   493.82 531.82
## - Mexico       1   511.95 531.95
## - Colombia     1   530.49 550.49
## - aroma        1   551.81 571.81
## - acidity      1   557.40 577.40
## - flavor       1   593.89 613.89
##
## Step:  AIC=521.99
## Qualityclass ~ aroma + flavor + acidity + Colombia + Mexico +
##     Honduras + Kenya + defects_log
```

```
##
##               Df Deviance    AIC
## - Honduras     1   504.71 520.71
## - Kenya        1   504.78 520.78
## <none>             503.99 521.99
## + level        2   503.11 525.11
## - defects_log  1   509.68 525.68
## - Mexico       1   512.91 528.91
## + year         8   497.14 531.14
## - Colombia     1   535.47 551.47
## - aroma        1   554.81 570.81
## - acidity      1   560.36 576.36
## - flavor       1   593.96 609.96
##
## Step:  AIC=520.71
## Qualityclass ~ aroma + flavor + acidity + Colombia + Mexico +
##     Kenya + defects_log
##
##               Df Deviance    AIC
## - Kenya        1   505.57 519.57
## <none>             504.71 520.71
## + Honduras     1   503.99 521.99
## - defects_log  1   509.97 523.97
## + level        2   504.15 524.15
## - Mexico       1   513.00 527.00
## + year         8   497.78 529.78
## - Colombia     1   538.07 552.07
## - aroma        1   556.11 570.11
## - acidity      1   561.95 575.95
## - flavor       1   594.88 608.88
##
## Step:  AIC=519.57
## Qualityclass ~ aroma + flavor + acidity + Colombia + Mexico +
##     defects_log
##
##               Df Deviance    AIC
## <none>             505.57 519.57
## + Kenya        1   504.71 520.71
## + Honduras     1   504.78 520.78
## + level        2   504.92 522.92
## - defects_log  1   510.96 522.96
## - Mexico       1   514.26 526.26
## + year         8   498.34 528.34
## - Colombia     1   538.58 550.58
## - aroma        1   556.46 568.46
## - acidity      1   563.79 575.79
## - flavor       1   597.26 609.26


##
## Call:  glm(formula = Qualityclass ~ aroma + flavor + acidity + Colombia +
##     Mexico + defects_log, family = binomial(link = "logit"),
##     data = coffee_final)
##
## Coefficients:
```

```
## (Intercept)        aroma       flavor      acidity     Colombia       Mexico
##     -0.5381       1.2882       2.3012       1.5419       1.8858      -0.8483
## defects_log
##      0.3276
##
## Degrees of Freedom: 929 Total (i.e. Null);  923 Residual
## Null Deviance:       1289
## Residual Deviance: 505.6     AIC: 519.6
```

```
model_best <- glm(Qualityclass ~ aroma + flavor + acidity + Colombia + Mexico + defects_log, data = cof
summary(model_best)
```

```
##
## Call:
## glm(formula = Qualityclass ~ aroma + flavor + acidity + Colombia +
##     Mexico + defects_log, family = binomial(link = "logit"),
##     data = coffee_final)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.0930  -0.3190   0.0012   0.3521   3.5017
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.5381     0.1981  -2.716  0.00661 **
## aroma         1.2882     0.2134   6.036 1.58e-09 ***
## flavor        2.3012     0.2781   8.276  < 2e-16 ***
## acidity       1.5419     0.2238   6.889 5.61e-12 ***
## Colombia      1.8858     0.3584   5.261 1.43e-07 ***
## Mexico       -0.8483     0.2919  -2.906  0.00366 **
## defects_log   0.3276     0.1424   2.300  0.02142 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1289.15  on 929  degrees of freedom
## Residual deviance:  505.57  on 923  degrees of freedom
## AIC: 519.57
##
## Number of Fisher Scoring iterations: 7
```

```
set.seed(9)
folds <- createFolds(y=coffee_final$Qualityclass, k=10)
accuracy <- as.numeric()
sensitivity <- as.numeric()
specificity <- as.numeric()
for(i in 1:10){
  fold_test <- coffee_final[folds[[i]],]
  fold_train <- coffee_final[-folds[[i]],]
  fold_pre <- glm(Qualityclass ~ aroma + flavor + acidity + Colombia + Mexico + defects_log,family = bi
  fold_predict <- predict(fold_pre,type='response',newdata=fold_test)
  fold_predict <- ifelse(fold_predict >= 0.5, 1, 0)
```

```
  accuracy[i] <- mean(fold_predict == fold_test[,8])
  sensitivity[i] <- sum(fold_predict + fold_test[,8] == 2) / sum(fold_test[,8] == 1)
  specificity[i] <- sum(fold_predict + fold_test[,8] == 0) / sum(fold_test[,8] == 0)
}
mean(accuracy)
```

```
## [1] 0.8924731
```

```
mean(sensitivity)
```

```
## [1] 0.9032471
```

```
mean(specificity)
```

```
## [1] 0.8811642
```