Data summarization

Yuqi Pan

1 Data summarization

The summary statistics are tabled below:

Table 1: Summary statistics on all numerical variables

Variables	Missing	Mean	SD	Min	Median	Max
aroma	0	7.57	0.39	0	7.58	8.75
flavor	0	7.52	0.40	0	7.58	8.67
acidity	0	7.54	0.39	0	7.50	8.58
category_two_defects	0	3.67	5.41	0	2.00	55.00
altitude_mean_meters	201	1850.69	9392.09	1	1310.64	190164.00
harvested	60	2013.67	1.81	2010	2014.00	2018.00

We noticed that there are 201 missing observations of `altitude_mean_meters` and 60 missing observations of `harvested`. We will delete these missing observations and have a new summarize statistics tabled below:

Table 2: Summary statistics on all numerical variables after deleting the missing value

Variables	Missing	Mean	SD	Min	Median	Max
aroma	0	7.57	0.39	0	7.58	8.75
flavor	0	7.52	0.41	0	7.50	8.67
acidity	0	7.53	0.40	0	7.50	8.58
category_two_defects	0	3.64	5.36	0	2.00	47.00
altitude_mean_meters	0	1856.24	9436.94	1	1310.64	190164.00
harvested	0	2013.70	1.81	2010	2014.00	2018.00

1.1 Cleaning data

From table 2, we found the minimum value of `aroma`, `flavor` and `acidity` are 0 but it is not consistent with normal situation. So we find this country and delete this observation.

Also for the variable of `altitude_mean_meters`, the maximum value is 190164 meters and it is out of question. Even the highest mountain in the world is less than 9000 meters, we decide to remove the observations of `altitude_mean_meters` greater than 9000 meters.

The final version of out data set will be like this:

```
## Rows: 930
## Columns: 7
```

```
## $ aroma                <dbl> 7.25, 8.33, 8.42, 7.17, 7.75, 7.92, 7.83, 7.00, 7~
## $ flavor               <dbl> 7.42, 7.92, 8.00, 7.08, 7.67, 7.75, 7.67, 6.83, 7~
## $ acidity              <dbl> 7.50, 7.92, 8.00, 7.25, 7.50, 7.75, 7.58, 7.17, 7~
## $ category_two_defects <int> 4, 1, 7, 3, 5, 0, 2, 2, 1, 0, 8, 0, 2, 0, 0, 2, 0~
## $ altitude_mean_meters <dbl> 1219.20, 1600.00, 1700.00, 1300.00, 1880.00, 1400~
## $ harvested            <int> 2015, 2013, 2014, 2012, 2012, 2014, 2015, 2013, 2~
## $ Qualityclass         <chr> "Poor", "Good", "Good", "Poor", "Good", "Good", "~
```

Table 3: Summary statistics on all numerical variables after cleaning the outliers.

Variables	Missing	Mean	SD	Min	Median	Max
aroma	0	7.58	0.31	5.08	7.58	8.75
flavor	0	7.53	0.32	6.17	7.58	8.67
acidity	0	7.53	0.31	5.25	7.50	8.58
category_two_defects	0	3.64	5.35	0.00	2.00	47.00
altitude_mean_meters	0	1325.65	484.31	1.00	1310.64	4287.00
harvested	0	2013.69	1.81	2010.00	2014.00	2018.00

1.2 Visualize the data

After deleting the missing observations and the outliers, we will plot boxplots of `Qualityclass` against the other variables.

From the boxplots, we can see that the difference between quality class of good and poor against aroma, acidity and flavor is obvious since the the boxplots do not overlap with each other. But the difference of quality class of good and poor against with other variables are not significant since they have some overlap between each other. Then we would expect that aroma, acidity and flavor will have a strong influence to the quality class of a certain batch of coffee.

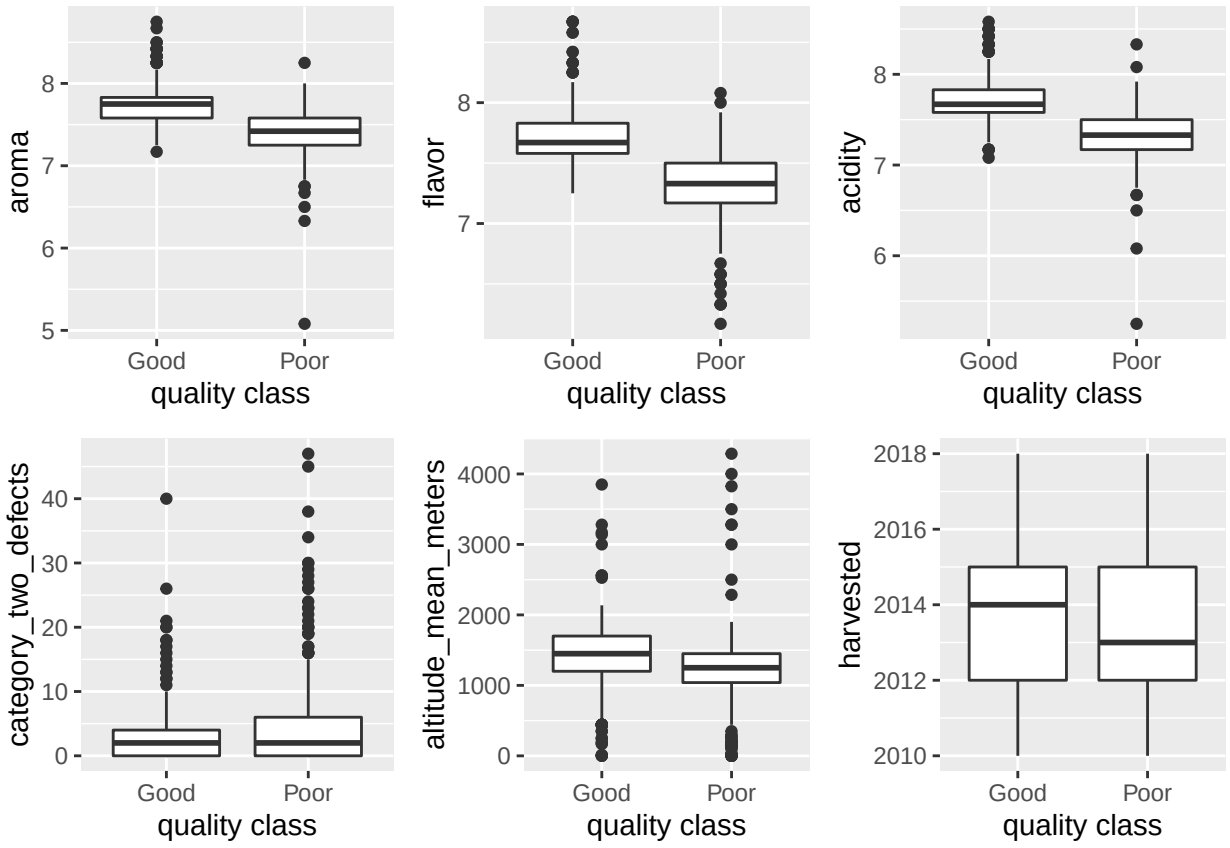

Figure 1: Boxplot of the Qualityclass against the other feature variables.

The correlation plot is plotted below:

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

From correlation plot, we notice that the correlation coefficient between acidity and flavor is 0.744 and the correlation coefficient between acidity and aroma is 0.591. We think there might be some problem with multicolinearity and we will discuss this issue in the next section.

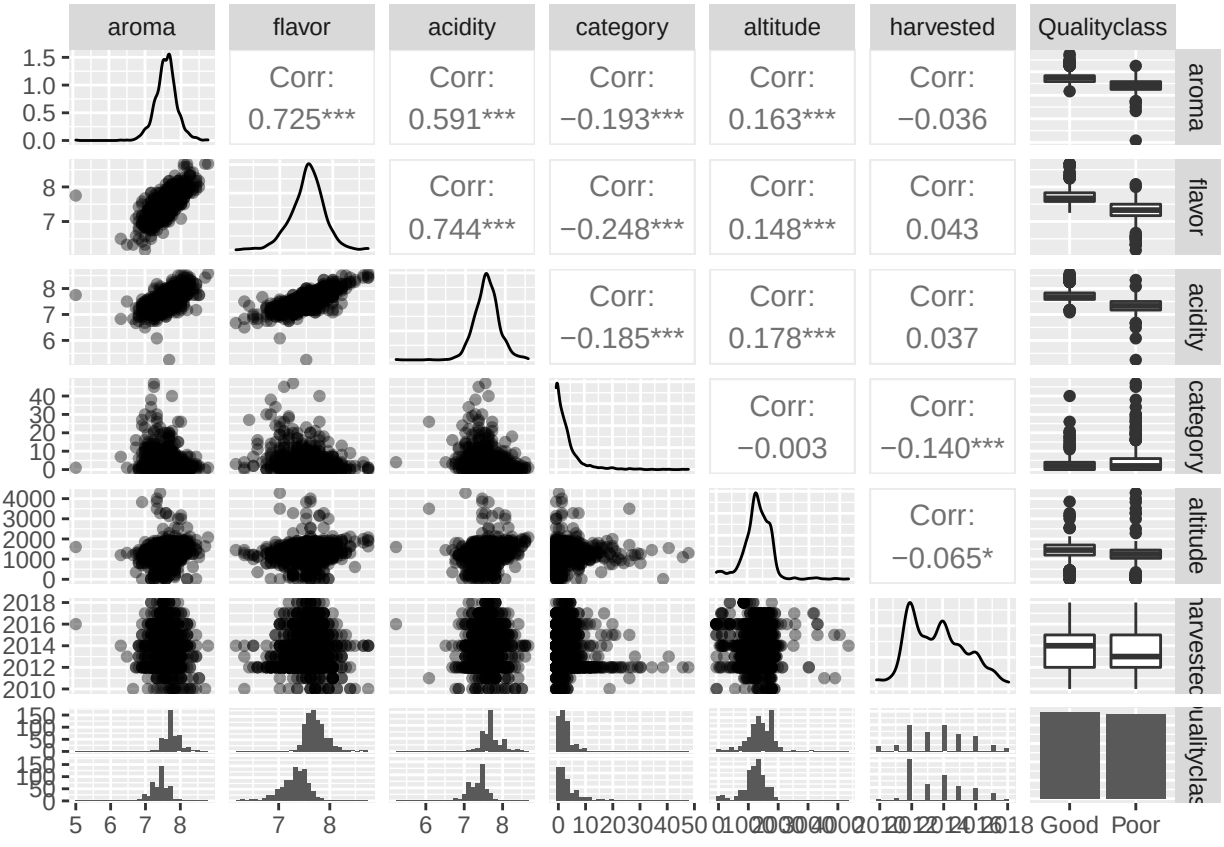

Figure 2: Correlation plot of the variables.