Exploratory Data Analysis for Machine Learning

Certification Project

Ignacio Sánchez Barraza

I) Dataset

The data contained in this dataset was downloaded from sklearn toysets, which results from a chemical analysis of wines grown in a certain region by 3 different cultivators. There are 13 different columns or measurements with 178 observations (samples) each one, from the constituents of each cultivator of wine (target).
The measurements are:

1. Alcohol
2. Malic acid
3. Ash
4. Alcalinity of ash
5. Magnesium
6. Total phenols
7. Flavanoids
8. Nonflavanoid phenols
9. Proanthocyanins
10. Color intensity
11. Hue
12. OD280/OD315 of diluted wines
13. Proline

The data used will be truncated to 4 columns of my choice to reduce computation time, simplify the results and to allow for further creation of polynomial features without increasing the computation time as said before. A description of the columns chosen is given by pandas as following:

	alcohol	malic_acid	ash	alcalinity_of_ash	target
count	178.000000	178.000000	178.000000	178.000000	178.000000
mean	13.000618	2.336348	2.366517	19.494944	0.938202
std	0.811827	1.117146	0.274344	3.339564	0.775035
min	11.030000	0.740000	1.360000	10.600000	0.000000
25%	12.362500	1.602500	2.210000	17.200000	0.000000
50%	13.050000	1.865000	2.360000	19.500000	1.000000
75%	13.677500	3.082500	2.557500	21.500000	2.000000
max	14.830000	5.800000	3.230000	30.000000	2.000000

The data has not been rid of outliers or any preprocessing treatment so far and we can note the difference on scale for which we should account for in further usage of this dataset, for example, when training models for classification problems.

We may begin counting the number of values for each class or target, for later use in a classification problem, accounting for disparity on counts, where we might need to do some data augmentation procedures or to decrease the number of samples per class to match the same.

```
1    71
0    59
2    48
Name: target, dtype: int64
```

We see that for classes 0,1 and 2 (one per cultivator) differ in the number of samples as said before, but for our exploratory data analysis will not be significant so far.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 178 entries, 0 to 177
Data columns (total 5 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   alcohol           178 non-null    float64
 1   malic_acid        178 non-null    float64
 2   ash               178 non-null    float64
 3   alcalinity_of_ash 178 non-null    float64
 4   target            178 non-null    int32
dtypes: float64(4), int32(1)
memory usage: 6.4 KB
```

We may extract some information to get a notion of which type of data we are working with. All the columns selected are float-type (continuous) except for the target values that refer to each one of the classes (discrete).

II) Initial plan for data exploration

Following the Exploratory Data Analysis (EDA) we will need to preprocess the data before conducting any modeling or other analysis procedures.
First of all we will need to transform or remove columns that do not give any information at all in their current state, such as categorical data (for any model training in the future) or object-like data.
We will obtain some visualization plots to see quantiles, outliers, distributions, histograms and some statistics for any useful insight we may find.
Following the previous step, we will first need to account for outliers, removing any data value that lies too far away from the average, but preventing the data to lose too many samples in the

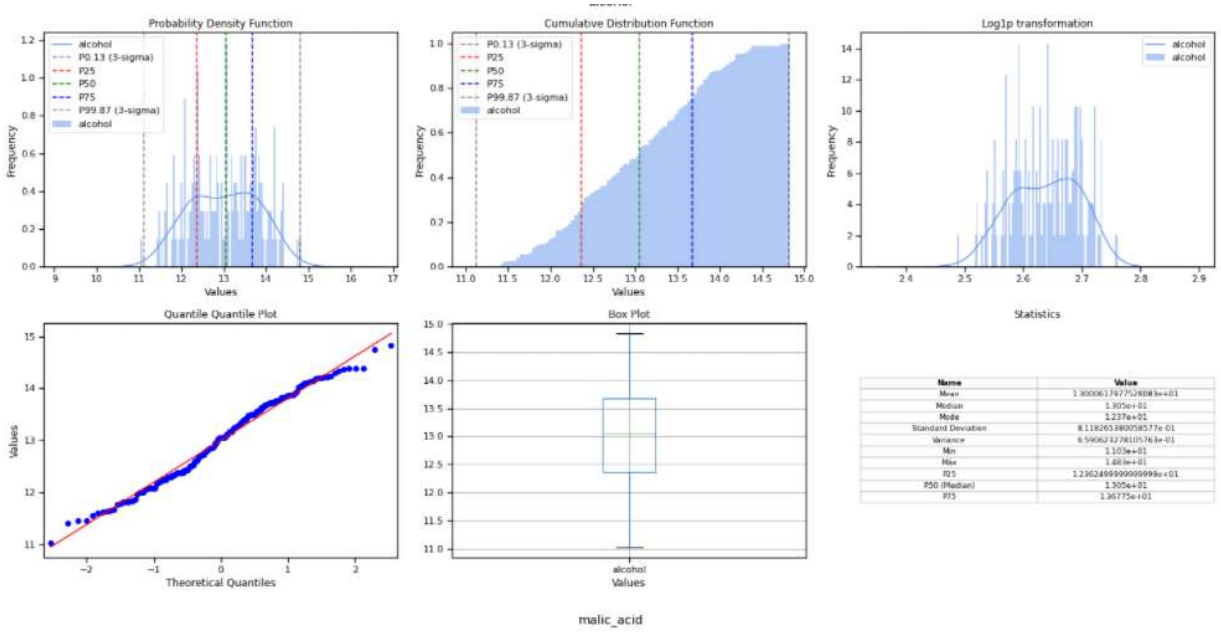
process, for which we might need to account for it interpolating the data on those outliers values, using the non-outlier data available (column-wise).
We might need to transform certain columns too, to try to achieve a normal distribution if possible in each measurement, or a known distribution that simplifies our study.
A good way to improve our analysis and future solve of the classification (or regression) problem will be creating new features on base of the existing ones, to find any non-linear relation between them or to explain several measurements on the same one.
After all, we will need to find the correlations between each column or measurement to find out any dependency, so, in addition, to extract the distribution of each one too, we will use a pair plot visualization to get handy data along these correlations.
Finally we will perform a hypothesis testing on data to investigate some premises chosen by the investigator.

III) Actions taken for data cleaning and feature engineering

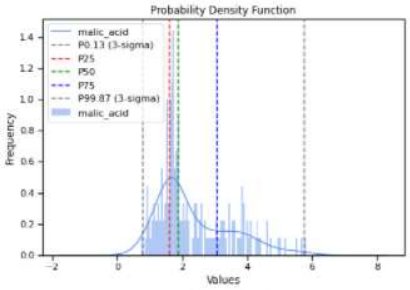
As for any data driven study, we should preprocess the data to get rid of outliers, object-type columns and transform categorical data to numerical data.

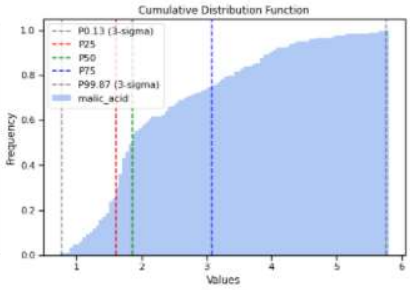
a) Data visualization:

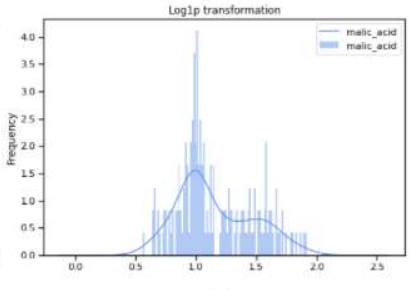
Before any preprocessing we should see the data we are working with, so we may find some easy to see and solve problems before entering any outlier removal or data transformation procedure. For this we created some custom plots shown here:

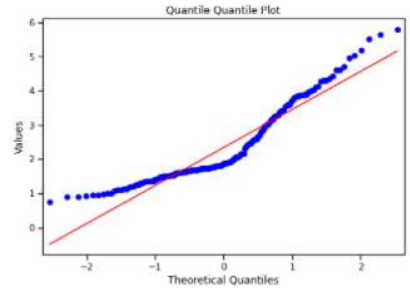


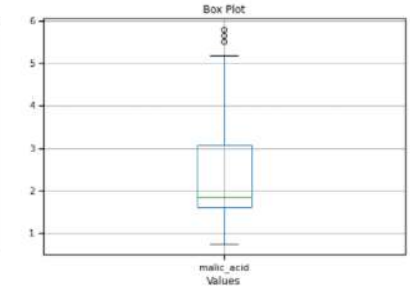
malic_acid







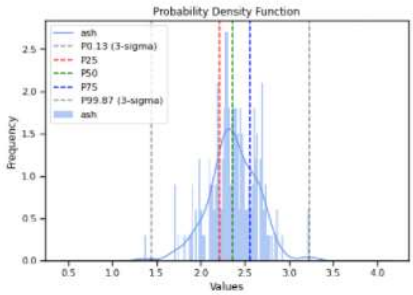


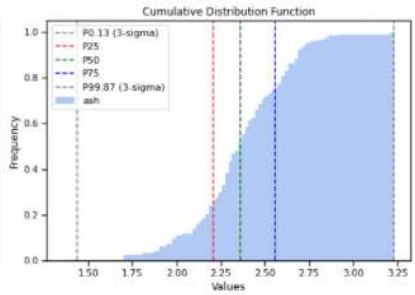


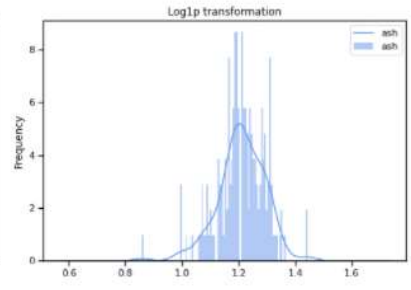
Statistics

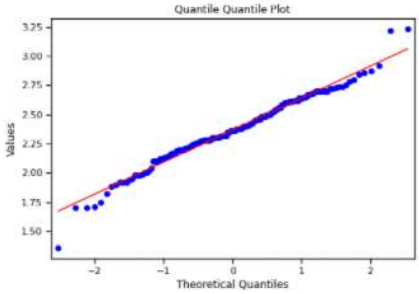
Name	Value
Mean	2.336348314606741e+00
Median	1.865000000000002e+00
Mode	1.73e+00
Standard Deviation	1.117140097814467?e+00
Variance	1.248915403415222?e+00
Min	7.4e-01
Max	5.8e+00
P25	1.602500000000003e+00
P50 (Median)	1.865000000000002e+00
P75	3.0825e+00

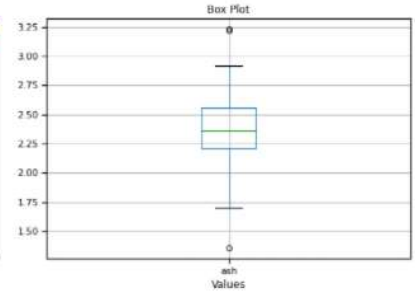
ash





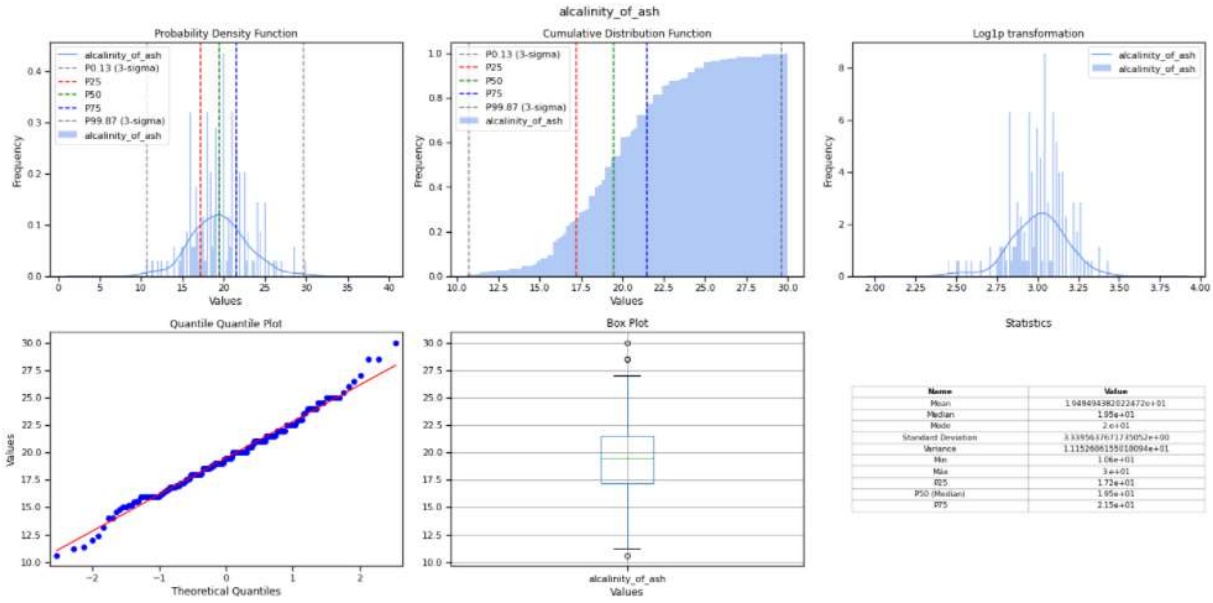






Statistics

Name	Value
Mean	2.366516853932584e+00
Median	2.36e+00
Mode	2.28e+00
Standard Deviation	2.743440090608148e-01
Variance	7.526863530756?8e-02
Min	1.36e+00
Max	3.23e+00
P25	2.21e+00
P50 (Median)	2.36e+00
P75	2.5575e+00



From these plots, we can see in the quantile-quantile plots that the data behaves like a normal distribution, which we can see on the histograms too with the adjusted probability density function. We can see too in the box plots that there is data that lies outside the 25th and 75th percentiles, showing little or no significant outliers at all, but we should analyze them separately.

b) Skew transformation:

After seeing the data, we might notice that there are some measurements (such as 'ash' and 'malic_acid') that have a little skew. We may improve them by transforming the data with a Log10(1-x), which plot is shown above together with the qq plot and others. This transformation does not account for all the skew problems, but helps a little. We could guess that there are more than one distribution probably in the data due to the different peaks we see (multimodality), so this could be affecting our measurements. This transformation was not studied in this course and will not be applied to this project.

c) Outlier removal:

For outlier removal procedures, we apply a z-score to the data, leaving behind all the data samples where any column value has a score greater than 3 (ie., any data value that lies more than 3 standard deviation values far from the mean value).
This methodology can be replaced with interpolation procedures to replace and not lose any sample (rows) in data.
We can see that the description of the old dataset and the new one with outliers removed didn't differ at all, so data was preprocessed before loading it to sklearn datasets or it had no outliers.

	alcohol	malic_acid	ash	alcalinity_of_ash	target
count	178.000000	178.000000	178.000000	178.000000	178.000000
mean	13.000618	2.336348	2.366517	19.494944	0.938202
std	0.811827	1.117146	0.274344	3.339564	0.775035
min	11.030000	0.740000	1.360000	10.600000	0.000000
25%	12.362500	1.602500	2.210000	17.200000	0.000000
50%	13.050000	1.865000	2.360000	19.500000	1.000000
75%	13.677500	3.082500	2.557500	21.500000	2.000000
max	14.830000	5.800000	3.230000	30.000000	2.000000

d) Removal/transformation of non categorical/non numeric data:

As seen in the data information obtained from pandas, there are no categorical or object-like columns we need to remove/transform in our data, so no processing is done here.

e) Feature engineering

As described before we should explore polynomial features to study any non linear correlation between the measurements or to find out if any of them can be described as the same measurement or just one (dependency). To achieve this, we use the polynomial_features on sklearn, at degree 2 (to maintain computation complexity low):

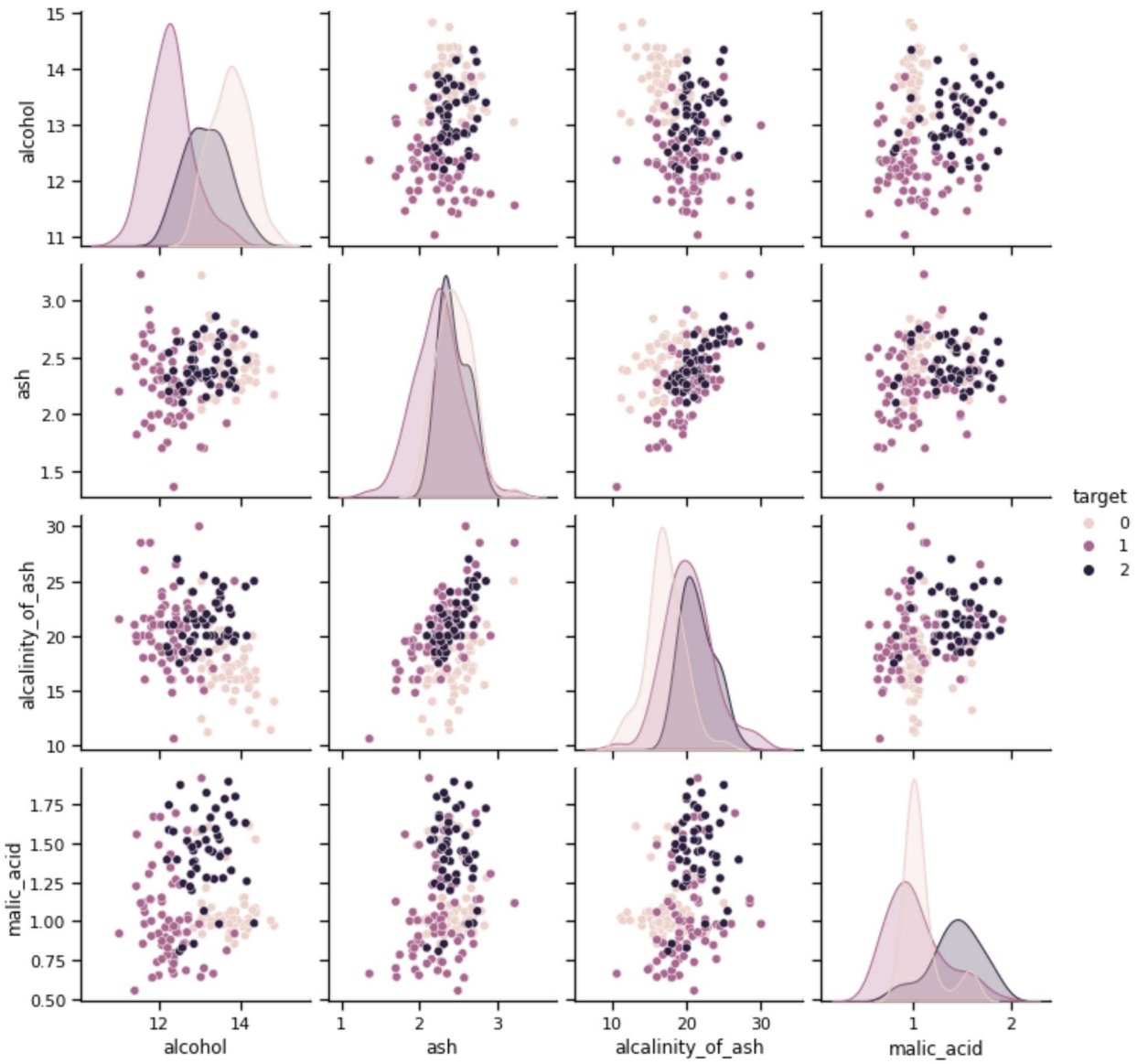
	1	alcohol	ash	alcalinity_of_ash	malic_acid	alcohol^2	alcohol ash	alcohol alcalinity_of_ash	alcohol malic_acid	ash^2	ash alcalinity_of_ash	malic_a
count	178.0	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.000000	178.000
mean	1.0	13.000618	2.366517	19.494944	1.154950	169.671428	30.813032	252.609949	15.044282	5.675244	46.539039	2.750
std	0.0	0.811827	0.274344	3.339564	0.309582	21.086573	4.340106	41.653066	4.209475	1.291991	11.571862	0.837
min	1.0	11.030000	1.360000	10.600000	0.553885	121.660900	16.823200	131.122000	6.319829	1.849600	14.416000	0.901
25%	1.0	12.362500	2.210000	17.200000	0.956471	152.831425	28.041600	226.231250	12.284415	4.884100	39.026000	2.140
50%	1.0	13.050000	2.360000	19.500000	1.052567	170.302500	30.933700	246.758000	14.006097	5.569600	45.385000	2.619
75%	1.0	13.677500	2.557500	21.500000	1.406682	187.074025	34.144950	277.112250	17.829947	6.540825	52.390000	3.379
max	1.0	14.830000	3.230000	30.000000	1.916923	219.928900	42.021000	389.700000	25.975197	10.432900	92.055000	4.945

```
Index(['1', 'alcohol', 'ash', 'alcalinity_of_ash', 'malic_acid', 'alcohol^2',
       'alcohol ash', 'alcohol alcalinity_of_ash', 'alcohol malic_acid',
       'ash^2', 'ash alcalinity_of_ash', 'ash malic_acid',
       'alcalinity_of_ash^2', 'alcalinity_of_ash malic_acid', 'malic_acid^2',
       'target'],
      dtype='object')
```

f) Correlation study

For this analysis, we use the Pair Plot learned on the course, to extract in addition the distribution of the data and separate them by target classes (cultivators of wine). The results are shown next:

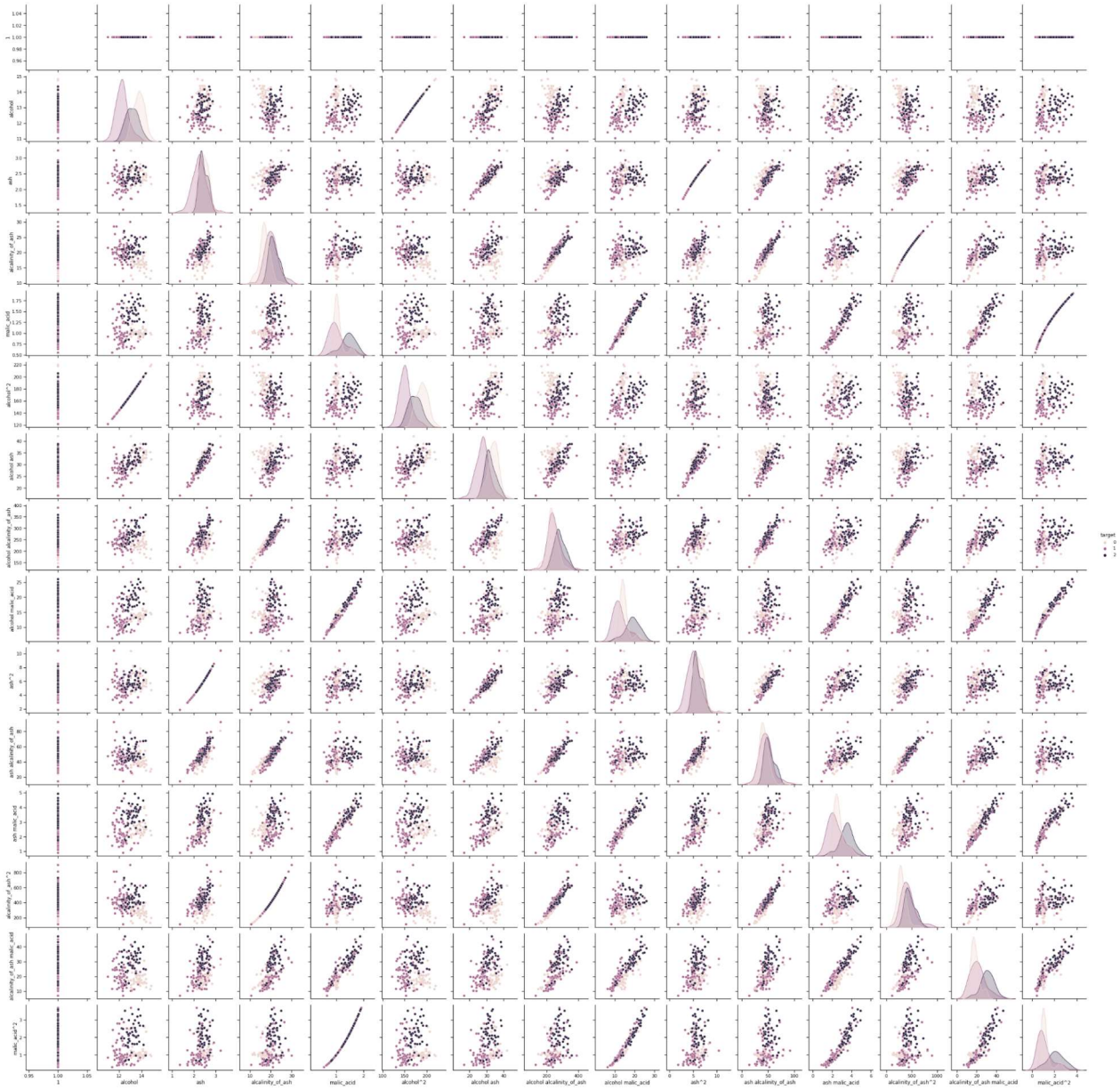
Old features:



We can see, before creating new features, a little correlation between ‘alcalinity_of_ash’ and ‘ash’ that's not entirely understandable. If there was a correlation between alkalinity of the wine and ash concentration, that would be expectable, due to an increase of ash leading to an increase of alkalinity (increase of pH). This behavior should be visible if we inspect the ash and ph relation. In any case, there is a positive correlation between ‘alcalinity_of_ash’ and ‘ash’, for which we can expect them to be dependent, depending on the value of the correlation (by the step seen in the plot, we can expect a high correlation between the two of them).

In addition we noted the guess on the previous sections, that the distribution for each measurements is different for each of the classes (wine cultivator), seeing an exception for ‘alcalinity_of_ash’ and ‘ash’ distributions where we can see a similar average value for each of the 3 classes.

New polynomial features included:



We can see that there are several correlated measurements, such as ‘alcohol^2’ and ‘alcohol’, a trivial case, and so on. There are no easily seen correlated measurements that we can guess, apart from these trivial ones.

IV) Key Findings and Insights

From the results above we can conclude that 'alcalinity_of_ash' and 'ash' are correlated in a way we can not tell about, but the procedures applied showed a strong correlation between those two measurements. In addition, we see that those same columns have distributions that differ little or nothing to the normal distribution, and separating each distribution by class of cultivator wine we see the average values are similar too, demonstrating a common property of the wine across the cultivators in the same region, that is 'ash', maybe because of the proximity with a volcano or any geological features in the area that are shared in the region.

We could guess, too, that this column (feature) is a poor one to choose for solving the classification problem, because it gives us no properties to choose between any of the classes due to the general relation of the wine with the ash in the region, being present across the cultivator classes.

V) Hypothesis testing

For the hypothesis testing section, we will present 3 premises:

1) H0: The distribution of ash in the same region follows a normal distribution
   Ha: The distribution of ash in the same region does not follow a normal distribution

2) H0 : The average quantity of ash in the same region is equal to 2.36
   Ha : The average quantity of ash in the same region is not equal to 2.36

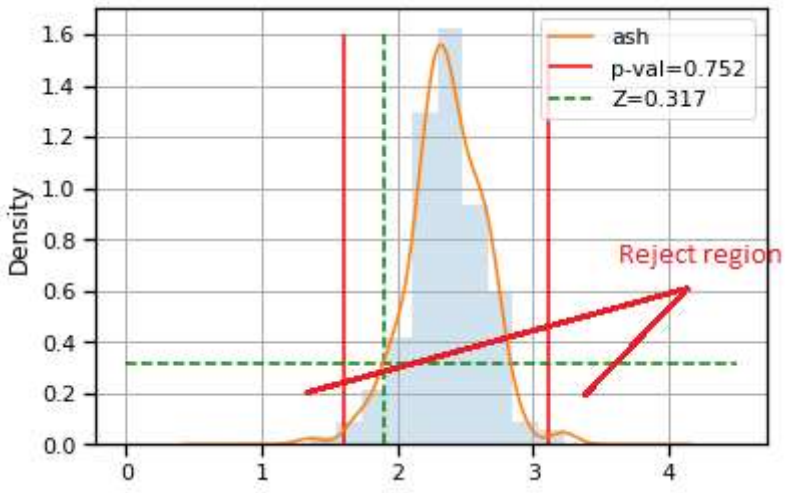
3) H0 : The average quantity of alcohol in the wine in the same region is equal to 14
   Ha : The average quantity of alcohol in the wine in the same region is not equal to 14

We will be conducting a hypothesis testing procedure for the 2) premise ***"The average quantity of ash in the same region is equal to 2.36"***

Using the scipy stats module from python, we can implement a test statistic for comparing the average value of a random sample and the average value of the data population. So, assuming a normal distribution as follows from the conclusion in the section of findings and insights above, we can estimate the p-value (probability of obtaining test results at least as extreme as the results observed) and the test statistic value at a confidence level of 95% (significance level = 1 - C = 0.05). So, if the p-value is greater than the significance level, we fail to reject the null hypothesis, and if the p-value is less or equal than the significance level, we reject the null hypothesis.

Following the steps above, we obtain a test statistic value (Z) of ***0.317*** and a p-value of ***0.752***, so we have failed to reject the null hypothesis.

	alcohol	malic_acid	ash
count	178.000000	178.000000	178.000000
mean	13.000618	2.336348	2.366517



If we compare the value that we assume the average quantity of ash would be for the same region with the actual average of 'ash' measurement, we see they are almost equal, therefore, the p-value is not as low as we need for the test statistic to lie in the reject zone.

VI) Suggestions

Some suggestions for following up with the analysis are including a more refined outlier detection and removal, together with an interpolation method, being careful with the outliers on the extremes of samples (start and end) to avoid extrapolating.

In addition, we should scale the data to compare each measurement from a common ground and to prepare the data for a future training process, to solve a supervised learning problem or even a non supervised learning problem, such as applying a Principal Component Analysis (PCA) or any other dimensionality reduction procedures, to extract the importance of each measurement on defining a new variable space, where we can, in addition, find clusters to label other properties not seen in this study (apart from the cultivator classes).

An increase of the sample size should be a first try, to obtain a better understanding of the data and to avoid getting only local insights and not general ones (the sample size of the dataset was only 178). With this in mind, we can overcome the trade off when removing outliers or any faulty sample too, so this would be a strategy we can benefit from.